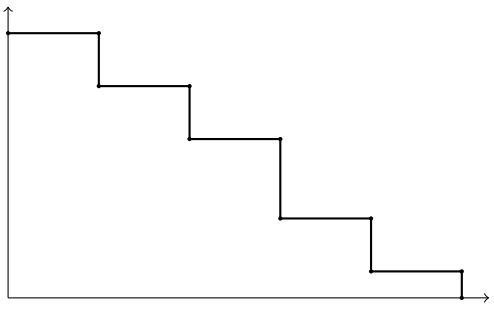
AARHUS UNIVERSITY

PHD THESIS

Survival Prediction for Precision Medicine

by Christian Marius Lillelund



PHD THESIS

Survival Prediction for Precision Medicine

Author:
Christian Marius Lillelund

Supervisors:
Assoc. Prof. Christian Fischer Pedersen
Prof. Russell Greiner

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Electrical and Computer Engineering
Faculty of Technical Sciences
Aarhus University
Denmark

September 28, 2026

Cover artwork. The step-shaped curve on the cover is an artistic illustration of the famous Kaplan–Meier curve in survival analysis.

Abstract

Survival prediction is a branch of survival analysis concerned with estimating when an event will occur, often from data in which some event times are censored. This thesis addresses four challenges in developing and applying such models: quantifying predictive uncertainty, predicting multiple events, evaluating predictions under censoring, and extending survival prediction beyond mortality. To quantify uncertainty in individual time-to-event outcomes, we develop BaySurv, a probabilistic survival modeling framework, and evaluate variants trained using variational inference, Monte Carlo dropout, and spectral-normalized neural Gaussian processes. To predict multiple related outcomes, we introduce MENSA, a single model that can predict the time to multiple competing or co-occurring events. Its trajectory-based likelihood term uses information about the temporal ordering of events. MENSA is competitive with strong deep learning and tree-based baselines across four multi-event benchmark datasets. We also apply multi-event survival analysis to functional decline in patients with amyotrophic lateral sclerosis (ALS). For model evaluation, we present a position paper showing how commonly used metrics can be misaligned with the intended prediction objective, particularly when ranking measures are used to assess time-to-event or probability predictions. We then introduce a copula-based Brier score that reduces evaluation bias when event and censoring times are dependent. Finally, we apply survival methods to predict fall risk in home care clients, and to predict the onset of degradation in ball bearings. Together, these contributions advance survival prediction for precision medicine by making individual predictions more informative, extending them to multiple outcomes, and strengthening how they are evaluated. The applications show that the same principles can also address time-to-event problems beyond healthcare.

Resumé

Overlevelsesprædiktion er en gren af overlevelsesanalyse, der beskæftiger sig med at estimere, hvornår en hændelse vil indtræffe, ofte ud fra data, hvor nogle hændelsestider er censurerede. Denne afhandling adresserer fire udfordringer i udviklingen og anvendelsen af sådanne modeller: kvantificering af prædiktiv usikkerhed, prædiktion af flere hændelser, evaluering af prædiktioner under censurering samt udvidelse af overlevelsesprædiktion ud over dødelighed. For at kvantificere usikkerheden i individuelle tid-til-hændelse-udfald udvikler vi BaySurv, et probabilistisk framework til overlevelsesmodellering, og evaluerer varianter trænet med variationel inferens, Monte Carlo-dropout og spektralnormaliserede neurale Gaussiske processer. For at prædiktere flere relaterede udfald introducerer vi MENSA, en enkelt model, der kan prædiktere tiden til flere konkurrerende eller samtidigt forekommende hændelser. Modellens trajektoriebaserede likelihood-led udnytter information om hændelsernes tidsmæssige rækkefølge. MENSA er konkurrencedygtig med stærke deep learning- og træbaserede referencemodeller på tværs af fire benchmarkdatasæt med flere hændelser. Vi anvender desuden overlevelsesanalyse med flere hændelser på funktionstab hos patienter med amyotrofisk lateral sklerose (ALS). Inden for modelevaluering præsenterer vi en positionsartikel, der viser, hvordan almindeligt anvendte evalueringsmål kan være uforenelige med det tilsigtede prædiktionsformål, især når rangeringsmål bruges til at vurdere prædiktioner af tid til hændelse eller sandsynligheder. Dernæst introducerer vi en copula-baseret Brier-score, der reducerer evalueringsbias, når hændelses- og censureringstider er afhængige. Endelig anvender vi overlevelsesmetoder til at prædiktere faldrisiko hos modtagere af hjemmepleje og til at prædiktere begyndende nedbrydning i kuglelejer. Samlet set styrker disse bidrag overlevelsesprædiktion inden for præcisionsmedicin ved at gøre individuelle prædiktioner mere informative, udvide dem til flere udfald og styrke måden, de evalueres på. Anvendelserne viser, at de samme principper også kan løse tid-til-hændelse-problemer uden for sundhedsområdet.

Acknowledgements

I would like to thank the many people who helped me through this journey with their ideas, comments, discussions, and support. In alphabetical order, I thank Ali Hossein Gharari Foomani, Christian Fischer Pedersen, Dikte Straadt, Fernando Pannullo, Lena Todnem Bach Christensen, Martin Magris, Michael Harbo, Morten Opprud Jakobsen, Peter Bangsgaard Stoustrup, Ruben Pauwels, Russell Greiner, Sanjay Kalra, Shi-ang Qi, Stratos Vassis, Thomas Klit Pedersen, and Weijie Sun. I would especially like to thank my supervisors, Christian Fischer Pedersen and Russell Greiner.

This PhD project was funded by the PRECISE project under Grant Agreement No. AAL-2021-8-90-CP from the European AAL Association, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), the Alberta Machine Intelligence Institute (Amii), the Siemens Foundation, and DigiRehab A/S. I am very grateful for this support and would like to especially thank Michael Harbo and Niels Heuer from DigiRehab A/S.

Lastly, I would like to give special thanks to Manuel Morante for taking the time to read my work and provide valuable feedback.

List of Publications

This thesis is based on the following publications:

- **Christian Marius Lillelund**, Martin Magris and Christian Fischer Pedersen "Uncertainty Estimation in Deep Bayesian Survival Models." *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2023.
- **Christian Marius Lillelund**, Fernando Pannullo, Morten Opprud Jakobsen, and Christian Fischer Pedersen. "Predicting Survival Time of Ball Bearings in the Presence of Censoring." *AAAI Fall Symposium 2023: Survival Prediction Algorithms, Challenges & Applications (SPACA)*, 2023.
- **Christian Marius Lillelund**, Michael Harbo, and Christian Fischer Pedersen. "Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis." *Journal of Public Health*, 34, 2026, pp. 905–917.
- Lena Todnem Bach Christensen, Dikte Straadt, Stratos Vassis, **Christian Marius Lillelund**, Peter Bangsgaard Stoustrup, Ruben Pauwels, Thomas Klit Pedersen, and Christian Fischer Pedersen. "An Explainable and Conformal AI Model to Detect Temporomandibular Joint Involvement in Children Suffering from Juvenile Idiopathic Arthritis." *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, 2024.
- **Christian Marius Lillelund**, Martin Magris, and Christian Fischer Pedersen. "Efficient Training of Probabilistic Neural Networks for Survival Analysis." *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6157–6166, 2025.
- **Christian Marius Lillelund**, Sanjay Kalra, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). "A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis." *PLOS ONE*, vol. 20, no. 11, pp. 1–16, 2025.
- **Christian Marius Lillelund**, Ali Hossein Gharari Foomani, Weijie Sun, Shi-ang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). "MENSA: A Multi-Event Network for Survival Analysis with Trajectory-Based Likelihood Estimation." *ML4H 2025*, Proceedings of Machine Learning Research, vol. 297, pp. 543–571, 2026.
- **Christian Marius Lillelund**, Shi-ang Qi, and Russell Greiner. "Overcoming Dependent Censoring in the Evaluation of Survival Models." *UAI 2026*, Proceedings of Machine Learning Research, vol. 337, pp. 3841–3866, 2026.
- **Christian Marius Lillelund**, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. "Position: Stop Chasing the C-index when Evaluating Survival Analysis Models." *Accepted to ICML 2026 Position Paper Track*. Spotlight. Preprint: arXiv:2506.02075, 2026.

Contents

Abstract iii

Resumé v

Acknowledgements vii

List of Publications ix

1 Introduction 1
- 1.1 Precision Medicine and Applications . . . 1
- 1.2 Survival Analysis and Time-to-Event Outcomes . . . 3
- 1.3 Machine Learning for Survival Analysis . . . 4
- 1.4 Challenges in Survival Analysis . . . 4
  - 1.4.1 Uncertainty in Survival Prediction Tasks . . . 5
  - 1.4.2 Competing Risks and Multiple Events . . . 5
  - 1.4.3 Evaluation under Censoring . . . 6
  - 1.4.4 Applications Beyond Mortality Prediction . . . 7
- 1.5 Thesis Overview . . . 7
- 1.6 Preliminaries . . . 8
  - 1.6.1 Prerequisites . . . 9
  - 1.6.2 Notation . . . 10

2 Background 11
- 2.1 Survival Data . . . 11
- 2.2 Censoring Assumptions . . . 15
- 2.3 Likelihood-based Approaches . . . 18
- 2.4 Evaluation Metrics . . . 24
  - 2.4.1 Discrimination . . . 26
  - 2.4.2 Absolute and Squared Errors . . . 29
  - 2.4.3 Calibration . . . 30
- 2.5 Competing Risks and Multiple Events . . . 33
  - 2.5.1 Single Event . . . 34
  - 2.5.2 Competing Risks . . . 35
  - 2.5.3 Multiple Events . . . 37
- 2.6 Uncertainty Estimation . . . 39
  - 2.6.1 Bayesian Learning . . . 41
  - 2.6.2 Conformal Prediction . . . 46

3 Uncertainty in Machine Learning and Survival Analysis 49
- 3.1 Uncertainty Estimation in Survival Models . . . 49
  - 3.1.1 Introduction . . . 49
  - 3.1.2 Materials and Method . . . 49
  - 3.1.3 Experiments and Results . . . 51

3.1.4 Discussion . . . 54
3.2 Prediction of TMJ Involvement in Children With JIA . . . 55
3.2.1 Introduction . . . 55
3.2.2 Materials and Method . . . 55
3.2.3 Experiments and Results . . . 56
3.2.4 Discussion . . . 57

4 Competing Risks and Multiple Events **59**
4.1 MENSA: A Multi-Event Network for Survival Analysis . . . 59
4.1.1 Introduction . . . 59
4.1.2 Materials and Method . . . 60
4.1.3 Experiments and Results . . . 63
4.1.4 Discussion . . . 67
4.2 Prediction of Functional Decline in ALS Patients . . . 67
4.2.1 Introduction . . . 67
4.2.2 Materials and Method . . . 68
4.2.3 Experiments and Results . . . 70
4.2.4 Discussion . . . 73

5 Evaluation under Censoring **75**
5.1 Common Pitfalls in Model Evaluation . . . 75
5.1.1 Introduction . . . 75
5.1.2 Objective-Metric Alignment . . . 77
5.1.3 Aspects of Evaluation and Common Metrics . . . 77
5.1.4 Model-Metric Consistency . . . 79
5.1.5 Practical Recommendations . . . 81
5.2 Overcoming Dependent Censoring in Model Evaluation . . . 82
5.2.1 Introduction . . . 82
5.2.2 Materials and Method . . . 82
5.2.3 Experiments and Results . . . 85
5.2.4 Discussion . . . 90

6 Applications Beyond Mortality Prediction **91**
6.1 Prediction of Fall Risk in Home Care Clients . . . 91
6.1.1 Introduction . . . 91
6.1.2 Materials and Method . . . 92
6.1.3 Experiments and Results . . . 93
6.1.4 Discussion . . . 95
6.2 Prediction of Survival Time in Ball Bearings . . . 96
6.2.1 Introduction . . . 96
6.2.2 Materials and Method . . . 97
6.2.3 Experiments and Results . . . 99
6.2.4 Discussion . . . 99

7 Conclusion **101**
7.1 Concluding Remarks . . . 101
7.2 Future Work . . . 102
7.2.1 Interpretability . . . 102
7.2.2 Empirical Evaluation . . . 104
7.2.3 Foundation Models . . . 105

A Uncertainty Estimation in Deep Bayesian Survival Models **109**

B Predicting Survival Time of Ball Bearings in the Presence of Censoring **115**

C Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis **123**

D An Explainable and Conformal AI Model to Detect Temporomandibular Joint Involvement in Children Suffering from Juvenile Idiopathic Arthritis **137**

E Efficient Training of Probabilistic Neural Networks for Survival Analysis **143**

F A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis **155**

G MENSA: A Multi-Event Network for Survival Analysis with Trajectory-based Likelihood Estimation **173**

H Overcoming Dependent Censoring in the Evaluation of Survival Models **203**

I Position: Stop Chasing the C-index when Evaluating Survival Analysis Models **231**

Bibliography **275**

Chapter 1

Introduction

1.1 Precision Medicine and Applications

On January 20, 2015, President Obama announced the "Precision Medicine Initiative" in his State of the Union Address. It was a 215 million dollar program designed to kick start a new approach to modern medicine by creating a more targeted or personalized approach to disease prevention and treatment [1]. Instead of relying on standard diagnostic procedures, population estimates and one-size-fits-all treatments, doctors were now beginning to consider each patient's unique biomedical profile, along with their lifestyle and environment, when making medical decisions. This personalized approach – known as *precision medicine* – has gained traction recently due to both scientific advances and political interest [2].

Precision medicine is a healthcare model that incorporates a wide range of individual data, including clinical, lifestyle, genetic, and other biomarkers to provide personalized treatment solutions. Artificial intelligence (AI) and machine learning (ML) techniques can carve through large amounts of data and identify complex patterns and interactions in patients' biomedical information. For example, when the COVID-19 pandemic hit in early 2020, such a model could be used to predict who would be at risk and potentially provide additional risk information to clinicians caring for such individuals [3]. What is driving this new era of human prediction is the explosion of massive datasets and the rise of powerful ML techniques. The rapid growth of clinical data – including imaging, text-based records, and real-time monitoring – along with advances in molecular technologies like genomics, and the widespread use of devices and wearables, is making it possible to gather detailed information about people's biology and environment at a scale we have never seen before. Potentially, ML could result in new taxonomies of disease through genomic, phonemic, and environmental predictors [4].

Figure 1.1 illustrates a general pipeline for developing and using ML models in precision medicine. Patient data are collected and prepared, then used to train a model to predict a defined clinical outcome. For a new patient, the trained model uses that patient's characteristics to produce an individualized prediction, such as the probability of an event within a specified period. A clinician may use this prediction alongside other information when assessing the patient or planning care. The stages are summarized below; further details are provided by König et al. [2] and Johnson et al. [5].

- In the "Phenotyping" stage (left column of Figure 1.1), data are collected within a suitable group of patients. This could be from cohort studies if the focus is on the development of diagnostic or prognostic models. For example, a medical history provides information about a patient's past health conditions, treatments, and responses to therapies, whereas immunology data involve the

study of the patient's immune system, including immune responses, biomarkers, and the presence of autoimmune conditions or infections.

- In "Track 1" (right column of Figure 1.1), researchers perform data preprocessing. This includes exploring the collected datasets to uncover new patterns, correlations, and insights that can be used in the subsequent modeling phase. This step also involves actually cleaning and organizing the raw data, like handling missing data, removing outliers, normalizing the data, etc. The goal is to make sure the data are clean, reliable, and good to go for modeling.
- In "Track 2" (right column of Figure 1.1), a learning algorithm uses the preprocessed data to train a diagnostic or prognostic model. For example, a prognostic model could combine multiple features to estimate the probability of a particular clinical outcome or event (*e.g.*, death or cancer relapse) either at the current time (diagnostic model) or in the future (prognostic model). The learning algorithm estimates a prediction rule for the clinical outcome of interest.
- In "Track 3" (right column of Figure 1.1), the trained model is evaluated on patients who were not used for training. If its predictions are sufficiently accurate and suitable for the intended setting, it can be applied to a new patient's data to estimate an individual outcome. For example, a prognostic model might estimate the patient's probability of cancer relapse within five years. Such a prediction can inform clinical decisions, but does not by itself determine which treatment the patient should receive.

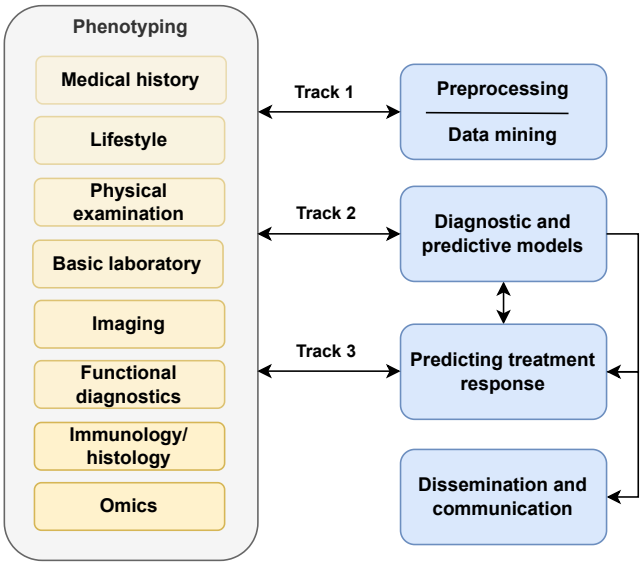


FIGURE 1.1: Modeling pipeline for precision medicine. In "Track 1", the data are preprocessed, including feature selection, and prepared for model building. In "Track 2", an algorithm uses the features to train a diagnostic or prognostic ML model. In "Track 3", the model is inspected, validated, and deployed. Finally, the results from Tracks 1–3 are fed back into the phenotyping stage to guide what data should be collected going forward. Figure recreated from [2].

1.2 Survival Analysis and Time-to-Event Outcomes

Among the different types of predictive tasks considered in Track 2, many clinically relevant outcomes are inherently time-dependent. Rather than simply predicting whether an event occurs, the goal is often to model *when* it occurs. This naturally leads to the framework of survival analysis and time-to-event prediction. Time-to-event outcomes have an extensive history in research, where the focus was initially to predict the time until death. Early analyses introduced the use of "life tables", which summarize counts such as births and deaths over time [6]. Back in 1662, John Graunt created the first life table, where he examined the chance of survival for different population groups in London based on mortality data [7]. The particular dataset he used was challenging, as the survival times (age at the time of death) were not actually recorded, which corresponds to a *censoring problem*. Rather than relying on exact data, Graunt estimated survival times using recorded causes of death, although these records were not always accurate. Later, in 1693, the well-known English astronomer Edmond Halley analyzed birth and death records from the city of Breslau [8]. He estimated the probability of dying within the next year and used these probabilities to determine how to price an annuity (an expected payout over a person's remaining lifetime). For a historical account of life tables and, more generally, reasoning about survival times, see the books by Bacaër [9] and Namboodiri and Suchindran [10].

As we have discussed, the early research on time-to-event outcomes focused primarily on the time until death, which is why the field became widely known as *survival analysis*. Many survival analysis models aim to estimate the probability that an event of interest has not yet occurred by a given time point. However, learning these models can be challenging compared to standard regression tasks because the training data often contain many censored observations – instances where the exact event time is not fully observed. This includes, for example, people who are still alive at the end of a study, as well as others who left the study and were lost to follow-up. We will return to the challenge of censored data and examine how survival techniques handle this major limitation. Since these early developments, the statistical, medical, and reliability engineering communities have written numerous textbooks on survival analysis [11–16].

From the beginning, we clarify that the term "survival prediction" (or "survival analysis") does not necessarily refer to predicting when someone dies. Traditionally, survival analysis focuses on understanding how covariates influence survival outcomes at the population level, whereas survival prediction emphasizes estimating when an event will occur for a specific individual. Although death is a common outcome, these methods can also be used to predict disease onset, the remaining useful life of mechanical components, or events in the economic and social sciences. Moreover, the "time" variable does not always represent literal time. For example, in a clinical study it may correspond to the number of chemotherapy rounds a cancer patient undergoes before treatment becomes ineffective. In this case, the goal is to predict the number of rounds until treatment failure or improvement based on patient characteristics and treatment factors. In essence, survival prediction models aim to predict the time (or count) until a specific event occurs. Next, we will look at the three fundamental types of survival prediction models.

1.3 Machine Learning for Survival Analysis

Researchers in the ML community have recently made significant methodological advances in survival analysis by leveraging the representation learning power of deep neural networks [17]. Generally, depending on the assumptions made and the way parameters are used in the model, we can divide models for survival analysis into three categories: nonparametric models, semiparametric models, and parametric models. For a broader list of models beyond what we cover here, see the excellent surveys by Wang et al. [18], Wiegrebe et al. [19] and Chen [6].

- *Nonparametric* models estimate the survival function directly from data without assuming a specific form for the underlying distribution. This makes them flexible and broadly applicable when the true distribution is unknown or difficult to specify. A wide range of models have been proposed for this setting, including the Kaplan-Meier [20] and Nelson-Aalen [21, 22] estimators. However, these approaches are limited to population-level descriptions and do not produce individual predictions.

- *Semiparametric* models make estimates by assuming a specific functional form for certain parts of the model (usually the so-called hazard function) but do not fully specify the entire distribution of the data. The most famous method is the Cox proportional hazard (CoxPH) model [23, 24]. It assumes that the hazard function at any time is a product of a baseline hazard function (which is unspecified and nonparametric) and a parametric component that relates the hazard to individual covariates (*e.g.*, treatment group, age, etc.). Semiparametric methods can thus provide predictions for individual patients. The main assumption behind these methods, particularly the CoxPH model, is that the effect of covariates is proportional over time. Intuitively, this means that if one individual has a higher risk than another at some point in time, this relative difference remains the same at all future time points.

- *Parametric* models assume that the event time follows a specific parametric distribution. This can simplify estimation and allow for more efficient inference when the assumed distribution is appropriate. Popular methods are Weibull time-to-event models and Accelerated Failure Time (AFT) models [13].

1.4 Challenges in Survival Analysis

Before turning to the details of survival analysis and the main contributions of this thesis, we outline the challenges that ML-based survival prediction models must address to support clinical decision-making in precision medicine. In addition to dealing with censored data, modern survival prediction models based on ML face other challenges that they must overcome to gain wider adoption: imagine developing a new survival model designed to predict whether hospitalized patients will experience specific medical emergencies, such as acute respiratory failure (ARF) or shock. How can we ensure that this model produces reliable (*i.e.*, trustworthy) predictions – ones that reflect the heterogeneity in the data, account for potential dependencies between competing events, and avoid imposing unrealistic assumptions (such as proportional hazards) on the outcome? Once we have trained the model, how do we fairly evaluate its performance in new cases, especially when many of them are censored? Censoring means that we often do not know the exact time or even the

occurrence of the event for some patients, which complicates the evaluation. As a result, it becomes both important and challenging to evaluate model performance in a meaningful way. The following sections take a closer look at the specific challenges this thesis addresses.

1.4.1 Uncertainty in Survival Prediction Tasks

Currently, the use of ML solutions for clinical decision-making is still in its early stages, as many clinicians remain reluctant to trust these opaque models due to a lack of understanding of how they arrive at their predictions [25]. This is particularly true for deep neural network–based models, which are typically difficult to interpret, provide limited information about predictive uncertainty, and may produce predictions that are sensitive to noise or inference errors [26]. Uncertainty estimation is also important when test data deviates from the training distribution, as out-of-distribution instances should have low credibility. That is, the model has made a prediction for some instance (*e.g.*, patient), but it is unsure about it because it has never seen such an instance before.

In survival prediction and within precision medicine, we are often interested in models that learn to compute a so-called individual survival distribution (ISD) based on the features of a patient [27]. As discussed, this distribution is a personalized curve that estimates the probability that an individual survives up to each future time point. Using the example from above, if a test instance differs significantly from the training data, we would expect a high level of uncertainty regarding its ISD. But traditional, deterministic approaches – *e.g.*, CoxPH [23, 24, 28] – only deliver a so-called risk score as a point prediction, which does not quantify the uncertainty about the prediction. We can subsequently estimate the survival function using, for example, the Breslow estimator [29], but this does not reflect the uncertainty in the weight estimates. This motivates the use of models that can capture the uncertainty in both the data distribution (aleatoric uncertainty) and in the weight estimates (epistemic uncertainty) [30] without a significant computational overhead. As a complementary, distribution-free alternative, conformal prediction can wrap an already trained model and provide prediction sets with guaranteed coverage, without assumptions about the model or its parameters.

RQ1. How can machine learning models, and deep survival models in particular, quantify uncertainty in individual predictions?

1.4.2 Competing Risks and Multiple Events

Survival prediction can be extended to handle situations involving multiple possible events – such as death from cancer, heart failure, or other causes. This includes settings with *competing risks* [31], where the occurrence of one event prevents the others from happening (*e.g.*, a patient can die from cancer or heart failure, but not both), as well as *multi-event* scenarios [32], where multiple (non-exclusive) events can occur for the same individual. For example, a patient with a chronic condition may experience several medical events over the course of their disease, or hospitalized patients may experience several types of medical emergencies, such as acute respiratory failure or shock, during their hospitalization.

In the current literature, some practitioners tackle multi-event problems by training separate single-event models [33, 34]. For example, a CoxPH model for cancer

and another for heart failure, treating each event independently. However, this approach has several limitations: (1) It ignores shared patterns between events, even though real-world events often have temporal or causal relationships. For example, some disease typically precedes death, for example [35]. Modeling each event in isolation fails to capture such dependencies. (2) Training separate models for each event is computationally expensive, especially for large datasets or many types of event. (3) Managing multiple models quickly becomes cumbersome as the number of events increases. Instead, a single multi-event model could address these issues by handling all events within one architecture, offering better scalability and potential for capturing inter-event relationships. Despite these advantages, there are relatively few multi-event survival models in the literature, and most existing approaches focus only on predicting risk scores – *i.e.*, the ranking of events within or across instances – rather than making time-to-event estimates.

RQ2. How can a single survival model predict the time to multiple events that may be competing or co-occurring, and use temporal ordering information when available?

1.4.3 Evaluation under Censoring

A fundamental challenge in survival analysis and survival prediction is censoring, where the exact event time is not observed for all individuals. For example, some patients may still be alive at the end of a study or may be lost to follow-up, meaning their true survival time is unknown. As a result, even if a model produces an estimate of the event time, there is often no ground-truth value to compare it against, and it is unknown how close the prediction is to the true outcome. This makes both model training and evaluation inherently more difficult than in standard regression settings. To draw valid conclusions from such incomplete data, and to evaluate predictions on new individuals, survival methods typically rely on the assumption of *independent* censoring. Under this assumption, censored individuals are considered representative of the population that remains under observation, and thus are assumed to have neither higher nor lower risk of future events. However, this assumption may not hold in practice. If censoring depends on unobserved factors – commonly referred to as *dependent censoring* – this can introduce bias in both parameter estimation and model evaluation.

The literature provides several metrics for evaluating time-to-event outcomes. Harrell's concordance index or simply C-index [36] is commonly used to assess how well predicted risk scores rank individuals, with the understanding that higher risk means the event will occur sooner. The Brier Score [37] measures the probabilistic accuracy of the predicted survival function, while the mean absolute error (MAE) evaluates the accuracy of the time-to-event estimates (*e.g.*, days, weeks, months) themselves [38]. These metrics rely on the assumption of independent censoring. When this assumption is violated – *i.e.*, when censoring depends on unobserved factors – the resulting estimates of model performance may be biased. Moreover, as argued in recent work on the evaluation of survival prediction models [38, 39], widely used metrics may shape model development in ways that do not necessarily reflect improvements in clinically meaningful survival predictions.

RQ3. How should survival predictions be evaluated when outcomes are censored, and how can evaluation bias be reduced under dependent censoring?

1.4.4 Applications Beyond Mortality Prediction

Survival analysis finds frequent use in clinical and epidemiological studies. For example, predicting the time to death in patients with amyotrophic lateral sclerosis (ALS) [40] or evaluating a new experimental treatment for cancer by comparing time-to-mortality predictions between a treatment group and a control group [41]. However, there exist other problems or applications where censored data are common, but practitioners still mainly rely on traditional classification or regression methods that do not handle censored data. We can choose to ignore the censored observations and analyze only cases with known outcomes, but doing so discards valuable information and can introduce bias into the model parameters [42]. In order to make better time-out-event predictions unrelated to mortality – such as when a bearing might fail or when an older adult is likely to experience a fall at home – we need to incorporate information from censored data into the analysis.

RQ4. How can survival prediction methods be applied to censored time-to-event outcomes beyond mortality, specifically time to fall in home care clients and time to degradation in ball bearings?

1.5 Thesis Overview

Chapter 2 provides the background. Chapters 3-6 address research questions RQ1-RQ4, respectively, and Chapter 7 concludes the thesis.

- **Background** (Chapter 2): We first go over the standard survival analysis framework and describe the problem of predicting the time to some event of interest. We state the traditional setting with a *single event* of interest and extend that to *competing risks* and *multi-events* scenarios. We look at the common assumptions behind censoring, describe the likelihood function to be maximized in continuous time (maximum likelihood is the standard way of learning time-to-event prediction models), and how to evaluate predictive accuracy. Lastly, we look at how to represent uncertainty in deep learning and relate that to survival analysis by examining various methods to quantify uncertainty.

- **Uncertainty in Machine Learning and Survival Analysis** (Chapter 3; RQ1): We next cover how to make deep survival models that can reason about the underlying uncertainty in the data (*i.e.*, aleatoric) and in the model (*i.e.*, epistemic). In Bayesian (probabilistic) models [43], epistemic uncertainty is often captured using distributions over model parameters, while aleatoric uncertainty is modeled through distributions over the observations. We present a novel method for uncertainty estimation in survival analysis, based on the Bayesian framework, and one application using a frequentist[1] approach, namely conformal prediction for a clinical classification task.

- **Competing Risks and Multiple Events** (Chapter 4; RQ2): We examine extensions of the traditional survival analysis framework that account for *competing risks* and *multi-event* settings. In competing risks problems, multiple possible events may occur, but the occurrence of one event prevents the others from

[1]Frequentist and Bayesian inference describe two popular modeling mindsets. Frequentists infer fixed but unknown distribution parameters based on a data sample. Bayesians infer model parameters as random variables which have some prior (before data) and posterior distributions (after data). See the book by Molnar [44] to learn more about the different ML mindsets.

happening. The goal is therefore to model the probability and timing of each mutually exclusive outcome. In contrast, multi-event problems involve events that are *not* mutually exclusive and may exhibit statistical dependence. Here, a multi-event survival model would predict the time until an instance (*e.g.*, a patient) experiences each of several different events. Such models allow information to be shared across related events, which can improve predictive performance. They also give practical advantages, including reduced computational overhead and simpler model management compared to maintaining separate models for each event. In this chapter, we present a method for multi-event survival prediction and motivate our approach through a concrete application.

- **Evaluation under Censoring** (Chapter 5; RQ3): Understanding how to evaluate survival models is crucial because traditional regression or classification metrics do not adequately handle censored data. In many cases, the true event time or event status at a given time point is unknown due to censoring in the test set. The problem becomes more challenging under dependent censoring, where variables related to the event of interest also influence the censoring mechanism. While collecting additional features may help reduce such dependence, model evaluation must still remain valid under dependent censoring and ideally be independent of the learned model itself. However, many commonly used survival metrics are biased in this setting [14, Ch. 1]. This chapter highlights evaluation methodology as a fundamental yet often overlooked aspect of survival analysis and surveys how current metrics shape model development. It also introduces a Brier score variant that supports dependent censoring and reduces evaluation bias compared with standard approaches.

- **Applications Beyond Mortality Prediction** (Chapter 6; RQ4): Although death is a common focus in survival analysis – not surprising given its name, which suggests studying how long something survives – the same methods and principles can also be used to predict other kinds of events. To demonstrate its broader potential, we present methods to predict fall risk in older adults, using the time to a personal alarm subscription as a proxy, and the time until ball bearings deviate from normal vibration behavior, as an early indicator of failure. Censoring is inherent to such problems, *e.g.*, many older adults do not receive an alarm within the follow-up period, and bearings are often taken out of service before they fail. The survival analysis framework allows us to leverage this information.

- **Conclusion** (Chapter 7): We answer the four research questions, summarize the thesis contributions, and outline directions for future research.

Table 1.1 lists the exact contributions of this thesis by their corresponding chapter, type of contribution and what specific challenge they target. The papers submitted or published during this PhD project are included in appendices A through I.

1.6 Preliminaries

Before moving on, we review some foundational knowledge that we assume the reader is familiar with. We also introduce the notation.

TABLE 1.1: Summary of the contributions in this PhD project.

Chapter	Type	Contribution	Research question
Chapter 3	Methodology	A probabilistic survival model (BaySurv) to estimate uncertainty in time-to-event outcomes using neural networks [45, 46].	RQ1
	Application	Prediction of temporomandibular joint (TMJ) involvement in patients with juvenile idiopathic arthritis (JIA) using clinical data [47].	
Chapter 4	Methodology	A multi-event survival model (MENSA), using a flexible number of Weibull distributions and temporal ordering information to predict multiple events [48].	RQ2
	Application	Prediction of significant functional decline across five domains in patients with ALS [49].	
Chapter 5	Methodology	A position paper highlighting the importance of proper evaluation methodology in survival analysis, including metric-objective mismatch and the consequences of dependent censoring [50].	RQ3
	Methodology	A copula-based Brier score for evaluating survival predictions under dependent censoring by modeling event-censoring dependence [51].	
Chapter 6	Application	Prediction of fall risk in community-dwelling elderly using survival analysis based on their home care usage [52].	RQ4
	Application	Prediction of time to degradation in ball bearings using survival analysis with censored observations and sensor data [53].	

1.6.1 Prerequisites

We assume the reader have a basic understanding of calculus, probability, statistics, and machine learning. Readers who want to implement or extend our methods should also know how to program and train neural networks using a framework such as TensorFlow [54] or PyTorch [55]. For an introduction to neural networks, we recommend Goodfellow et al. [17]. We use several specialized architectures and statistical methods, so we introduce the relevant concepts in the following chapters:

- Chapters 2 and 3 briefly introduce the Bayesian framework for learning probabilistic models. We provide a brief overview, but for those unfamiliar with Bayesian techniques, the book by Murphy [43] provides a good starting point, especially the first chapters. We also give a brief introduction to conformal prediction, but refer to Shafer and Vovk [56] for a more comprehensive review.
- Chapter 5 introduces copulas to capture the dependence structure between the event and censoring times. The reader should therefore be familiar with general probability theory, multivariate statistics, and dependency structures. Again, we provide a brief overview of the topic, but books by Emura and Chen [14] and Nelsen [57] give more details.

1.6.2 Notation

We usually let uppercase letters (*e.g.*, T) denote random variables (RVs) and lowercase letters (*e.g.*, t) denote deterministic quantities such as constants or specific realized values of random variables. The covariates of an instance are treated as a random vector X (written in plain font, since bold uppercase is reserved for matrices) with realization $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ is the covariate space. Moreover, we let bold uppercase letters (*e.g.*, $\boldsymbol{W}$) denote matrices, and bold lowercase letters (*e.g.*, $\boldsymbol{x}$, $\boldsymbol{\beta}$) denote vectors or column vectors of variables, features, coefficients, or instances. Scalar parameters (*e.g.*, ψ) are written in plain font, and θ denotes the collection of all parameters of a model, which may combine vectors, matrices, and scalars. We try to follow the traditional survival analysis literature and let the functions be uppercase or lowercase (*e.g.*, S represents the so-called survival function and f a probability density function); lowercase letters such as f or g may also denote parametric functions, such as neural networks. We use calligraphic letters to denote sets, with the exceptions of $\mathcal{L}$ (likelihoods and losses) and $\mathcal{N}$ (the normal distribution). For example, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$ denotes a dataset of N observed instances. In multi-event settings, the instance index is written as a superscript (*e.g.*, $\boldsymbol{x}^{(i)}$) to free the subscript for the event index. We use a caret notation $\widehat{\cdot}$ to indicate predicted or estimated values. When we refer to the "log" function, we always mean the natural logarithm. In the context of training neural networks, we write $\hat{\theta} := \arg\min_\theta \mathcal{L}_{\text{loss}}(\theta)$, where $\mathcal{L}_{\text{loss}}$ is a loss function (*e.g.*, the negative log-likelihood $\mathcal{L}_{\text{NLL}}$) with the parameter variable θ, while $\mathcal{L}$ without a subscript denotes a likelihood. We typically perform this minimization task using some variant of (stochastic) gradient descent, and we do not usually find a solution that reaches the global minimum in practice.

Chapter 2

Background

This chapter describes the main theory behind survival analysis, summarizes the state-of-the-art, recounts their shortcomings and hints at potential directions for improvement, which will be expanded upon in later chapters. We will start by examining the basic survival framework.

2.1 Survival Data

Many survival models aim to estimate the distribution of the time until an event of interest (typically death) occurs for an instance (*e.g.*, a patient). This time is modeled as a continuous, non-negative random variable (RV), $E \in \mathbb{R}_+$, representing the time from a given origin (*e.g.*, $t = 0$) to the occurrence of the event. In practice, however, E is not directly observed due to censoring; instead, we observe a censored version of this time, meaning that for some instances the exact time is unknown but constrained within certain bounds. Censored instances cannot simply be discarded, nor can their censoring times be treated as event times: both would bias estimates of the event-time distribution towards shorter times, since a censored instance is known to have remained event-free at least until it was censored. The types of censoring and the assumptions required for valid estimation are discussed in Section 2.2.

To formalize this setup, let $E \in \mathbb{R}_+$ denote the event time, $C \in \mathbb{R}_+$ the censoring time, $T \in \mathbb{R}_+$ the observed time, and $\Delta \in \{0, 1\}$ the event indicator. The observed time and event indicator are defined as

$$T := \min(E, C), \quad \Delta := \mathbb{1}[E \leq C]. \tag{2.1}$$

Thus, for each instance i, we observe a pair (t_i, δ_i), where t_i is either the event time or the censoring time, and δ_i indicates whether the event of interest was observed. We assume that each pair (t_i, δ_i) is drawn independently from the joint distribution $\Pr(T, \Delta)$, which is induced by $\Pr(E, C)$.

The distribution of E can be characterized by the cumulative distribution function (CDF), formally $F_E(t) = \Pr(E \leq t)$[1], or, equivalently, by the survival function, $S_E(t) = 1 - F_E(t) = \Pr(E > t)$. Note that $S_E(t)$ (respectively, $F_E(t)$) corresponds to the probability of being event-free (respectively, having experienced the event) by time t. When the full distribution of E is unknown, we rely on estimators constructed from observed data to estimate $S_E(t)$. This can be done nonparametrically using, for example, the Kaplan-Meier (KM) estimator [20].

[1] We use the subscript E to denote quantities related to the event time (*e.g.*, F_E, S_E, f_E) and the subscript C for those related to the censoring time (*e.g.*, S_C, f_C).

Definition 1 *The Kaplan-Meier (KM) estimator [20]:*

$$\hat{S}_E(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \quad \text{for} \quad t \leq t_{max} \tag{2.2}$$

where d_i is the number of events (e.g., *deaths) that happened at t_i, $n_i = \sum_{\ell=1}^{n} \mathbb{1}[t_\ell \geq t_i]$ is the number of instances known to have survived up to t_i, and $\mathbb{1}[\cdot]$ denotes the indicator function. $\hat{S}_E(t) = 1$ if no events occur up to t, and $\hat{S}_E(t)$ is undefined for $t > t_{max}$.*

Most often, however, we are interested in predicting $S_E(t)$ for an individual given a description of that individual. To do so, we represent each instance using a set of features or covariates – measurable characteristics such as age, biomarkers, or clinical test results – that capture the patient's condition. This turns survival analysis into a survival prediction problem, where the goal is to learn a mapping from covariates to a distribution over event times. By collecting data from many patients, we define a survival dataset $\mathcal{D}$ as

$$\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{N},$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the covariate vector for instance i, $t_i \in \mathbb{R}_+$ is the observed time, and $\delta_i \in \{0, 1\}$ indicates whether the event of interest was observed. Intuitively, censoring occurs when the exact event time is not observed, meaning that we only have partial information about when the event occurred.

For the remainder of this chapter, unless stated otherwise, we consider the survival time E conditional on covariates X taking values $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ (see Section 1.6.2). For a given realization $\boldsymbol{x} \in \mathcal{X}$, let $f_E(t \mid \boldsymbol{x})$ and $F_E(t \mid \boldsymbol{x})$ denote the conditional probability density function (PDF) and CDF, respectively, over some time horizon (*e.g.*, $t \in [0, t_{\max}]$). Either of these functions fully characterizes the conditional distribution $\Pr(E \mid X = \boldsymbol{x})$. In other words, we move from modeling a population-level distribution to modeling individual-specific risk profiles. With this in mind, we can now define one of the main prediction targets in survival analysis, namely the conditional survival function.

Definition 2 *The conditional survival function [6]:*

$$\begin{aligned} S_E(t|\boldsymbol{x}) &= Pr(\textit{survive beyond time } t \mid \textit{instance is } \boldsymbol{x}) \\ &= Pr(E > t \mid X = \boldsymbol{x}) \\ &= 1 - Pr(E \leq t \mid X = \boldsymbol{x}) \\ &= 1 - F_E(t|\boldsymbol{x}), \end{aligned} \tag{2.3}$$

where $t \geq 0$ and $\boldsymbol{x} \in \mathcal{X}$.

We refer to the conditional survival function $S_E(\cdot|\boldsymbol{x})$ simply as the survival function, as our notation already indicates that we are conditioning on $\boldsymbol{x}$. Given $\boldsymbol{x}_i$ is some feature vector for an instance i, this is also known as the individual survival distribution (ISD) [27]. We see that the true survival function $S_E(\cdot|\boldsymbol{x})$ is equal to 1 minus the CDF $F_E(\cdot|\boldsymbol{x})$, which means that:

1. The function $S_E(\cdot|\boldsymbol{x}) = 1 - F_E(\cdot|\boldsymbol{x})$ decreases monotonically from 1 to 0, because any CDF increases monotonically from 0 to 1.

2. Estimating $\widehat{S}_E(\cdot|\boldsymbol{x})$ is equivalent to estimating $\widehat{F}_E(\cdot|\boldsymbol{x})$, which implies that our goal is to estimate the conditional survival time distribution $\Pr(E \mid \boldsymbol{x})$.

3. While the full survival function provides the most complete description, it is often useful to summarize it with a single scalar quantity. To obtain a survival time estimate for $\boldsymbol{x}$, we can calculate that based on our estimate $\widehat{S}_E(\cdot|\boldsymbol{x})$ using, for example:

   - The *mean survival time*, which represents the expected duration of survival for an individual instance. It is computed as the area under the full survival function:

   $$\widehat{e}_i^{(\text{Mean})} = \mathbb{E}_{E\sim\widehat{S}_E(t|\boldsymbol{x}_i)}[\, E \,] = \int_0^\infty \widehat{S}_E(t \mid \boldsymbol{x}_i)\, \mathrm{d}t. \tag{2.4}$$

   This calculation involves integrating across the full follow-up horizon and relies on the assumption that the area under the survival function is finite, $\int_0^\infty S_E(t \mid \boldsymbol{x}_i)\, \mathrm{d}t < \infty$, which requires in particular that the survival function converges to zero. In practice, particularly with right-censored data or limited follow-up, the tail of the survival curve is often poorly estimated, which can introduce bias or substantial variability. As a result, the mean becomes sensitive to long-term survival behavior and is less robust when censoring is heavy or the distribution has a long tail.

   - The *restricted mean survival time (RMST)*, which represents the survival time up to a specific cutoff point t^*, calculated as the area under the survival curve up to t^*:

   $$\widehat{e}_i^{(\text{RMST})} = \mathbb{E}_{E\sim\widehat{S}_E(t|\boldsymbol{x}_i)}[\min(E, t^*)] = \int_0^{t^*} \widehat{S}_E(t \mid \boldsymbol{x}_i)\, \mathrm{d}t. \tag{2.5}$$

   The RMST is a truncated mean survival time, setting survival probability to zero beyond t^*. This makes it useful for limited follow-up and short-term survival analysis, offering stable and consistent estimation even when survival does not reach zero. However, it ignores long-term survival and "cure fractions"[2], and results depend on the chosen cutoff t^*, which may be arbitrary chosen. In clinical studies, it is often set to the maximum observed follow-up time. This choice can influence the interpretability and comparability across studies.

   - The *median survival time*, which is arguably the most used summary statistic in clinical and epidemiological studies. It is defined as the smallest time t at which the probability of surviving past t falls to 50%:

   $$\widehat{e}_i^{(\text{Median})} = \widehat{S}_E^{-1}(p = 0.5 \mid \boldsymbol{x}_i) = \inf\{t : \widehat{S}_E(t \mid \boldsymbol{x}_i) \leq 0.5\}. \tag{2.6}$$

   This offers a robust and easily interpretable estimate. The median is less sensitive to the tail behavior of the survival curve and outliers, making it particularly stable in practice. However, if the predicted survival curve remains above 50% throughout the study period (*e.g.*, due to extensive

[2] A "cure fraction" refers to the proportion of individuals who will never experience the event of interest, even with infinite follow-up. In such cases, the survival function does not converge to zero but instead approaches a positive constant, reflecting a subgroup that is effectively "cured" or no longer at risk.

censoring near the end), the median becomes undefined without extrapolation, limiting its applicability in such cases.

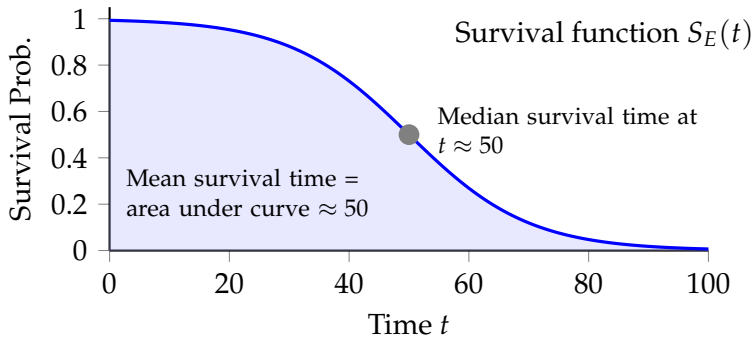


FIGURE 2.1: Example of a survival function with annotated median and mean survival times.

Figure 2.1 shows an example of the survival function and its corresponding median and mean survival times. Different time-to-event prediction aka. survival models make different assumptions on $S_E(\cdot \mid \boldsymbol{x})$ and often predict transformed variants of $S_E(\cdot \mid \boldsymbol{x})$ (*e.g.*, via hazard functions or cumulative hazard representations) rather than predicting $S_E(\cdot \mid \boldsymbol{x})$ directly. The next two prediction targets we discuss are both transformed versions of $S_E(\cdot \mid \boldsymbol{x})$.

Definition 3 *The hazard function [6]:*

$$h_E(t \mid \boldsymbol{x}) := -\frac{d}{dt} \log S_E(t \mid \boldsymbol{x}) = -\frac{\frac{d}{dt} S_E(t \mid \boldsymbol{x})}{S_E(t \mid \boldsymbol{x})} = -\frac{\frac{d}{dt}[1 - F_E(t \mid \boldsymbol{x})]}{S_E(t \mid \boldsymbol{x})} = \frac{f_E(t \mid \boldsymbol{x})}{S_E(t \mid \boldsymbol{x})}, \tag{2.7}$$

where $f_E(\cdot \mid \boldsymbol{x})$ is the PDF of the distribution $Pr(E \mid \boldsymbol{x})$. The hazard function is thus the instantaneous rate of event (e.g., *"death") conditioned on surviving up to time t.*

Like a probability density, a hazard is nonnegative but need not be bounded above by one. One of the most commonly-used survival models, the CoxPH model [23], is formulated in terms of the hazard function. If we know $h_E(\cdot \mid \boldsymbol{x})$, then we can recover $S_E(\cdot \mid \boldsymbol{x})$ since

$$\begin{aligned} h_E(t \mid \boldsymbol{x}) = -\frac{d}{dt} \log S_E(t \mid \boldsymbol{x}) \quad &\Leftrightarrow \quad \int_0^t h_E(u \mid \boldsymbol{x})\, du = -\log S_E(t \mid \boldsymbol{x}) \\ &\Leftrightarrow S_E(t \mid \boldsymbol{x}) = \exp\left(-\int_0^t h_E(u \mid \boldsymbol{x})\, du\right). \end{aligned} \tag{2.8}$$

Another commonly predicted target is the cumulative hazard function.

Definition 4 *The cumulative hazard function [6]:*

$$H_E(t \mid \boldsymbol{x}) := \int_0^t h_E(u \mid \boldsymbol{x})\, du. \tag{2.9}$$

From Equation 2.8, we see that $S_E(t \mid \boldsymbol{x}) = \exp(-H_E(t \mid \boldsymbol{x}))$, so if we know $H_E(\cdot \mid \boldsymbol{x})$, then we can recover $S_E(\cdot \mid \boldsymbol{x})$. Meanwhile, Equation 2.9 implies that $h_E(t \mid \boldsymbol{x}) = \frac{d}{dt} H_E(t \mid \boldsymbol{x})$, so if we know $H_E(\cdot \mid \boldsymbol{x})$, then we can recover $h_E(\cdot \mid \boldsymbol{x})$. Some survival models, such as Random Survival Forests [58], directly estimate the

cumulative hazard function. Note that while $S_E(\cdot \mid \boldsymbol{x})$ and $H_E(\cdot \mid \boldsymbol{x})$ are monotonic functions, $h_E(\cdot \mid \boldsymbol{x})$ need not be monotonic. Because we will often convert between $f_E(\cdot \mid \boldsymbol{x})$, $F_E(\cdot \mid \boldsymbol{x})$, $S_E(\cdot \mid \boldsymbol{x})$, $h_E(\cdot \mid \boldsymbol{x})$, and $H_E(\cdot \mid \boldsymbol{x})$ later in the thesis, we summarize the relationship between these functions in Summary 1.

Summary 1 *Assuming continuous time prediction targets, the following equations show how the PDF $f_E(\cdot \mid \boldsymbol{x})$, the CDF $F_E(\cdot \mid \boldsymbol{x})$, the survival function $S_E(\cdot \mid \boldsymbol{x})$, the hazard function $h_E(\cdot \mid \boldsymbol{x})$, and the cumulative hazard function $H_E(\cdot \mid \boldsymbol{x})$ are related [6]:*

$$
\begin{aligned}
f_E(t \mid \boldsymbol{x}) &= \frac{d}{dt} F_E(t \mid \boldsymbol{x}) = \frac{d}{dt}\big(1 - S_E(t \mid \boldsymbol{x})\big) = h_E(t \mid \boldsymbol{x}) S_E(t \mid \boldsymbol{x}), \\
F_E(t \mid \boldsymbol{x}) &= \int_0^t f_E(u \mid \boldsymbol{x})\, du = 1 - S_E(t \mid \boldsymbol{x}), \\
S_E(t \mid \boldsymbol{x}) &= 1 - F_E(t \mid \boldsymbol{x}) = \int_t^\infty f_E(u \mid \boldsymbol{x})\, du = e^{-H_E(t|\boldsymbol{x})} = e^{-\int_0^t h_E(u|\boldsymbol{x})\, du}, \\
h_E(t \mid \boldsymbol{x}) &= \frac{dH_E(t \mid \boldsymbol{x})}{dt} = -\frac{d}{dt} \log S_E(t \mid \boldsymbol{x}) = \frac{f_E(t \mid \boldsymbol{x})}{S_E(t \mid \boldsymbol{x})}, \\
H_E(t \mid \boldsymbol{x}) &= -\log S_E(t \mid \boldsymbol{x}) = \int_0^t h_E(u \mid \boldsymbol{x})\, du.
\end{aligned}
$$

Any of these functions fully specifies the conditional survival time distribution for the event of interest E.

2.2 Censoring Assumptions

We now elaborate on the censoring mechanism introduced in Section 2.1. Censoring is a defining characteristic of survival analysis and means that the exact event time is not observed for all instances. In a medical study as shown in Figure 2.2, some patients may experience the event (1st and 2nd instance), but some may withdraw from the study (3rd instance), experience a competing event (4th instance), which is often treated as a form of censoring in standard survival analysis, or remain alive at the end of the study (5th instance). For these patients, the exact event time is known only to exceed their last contact or the end of the study. This is known as *right-censoring*. Although datasets may also contain left- and interval-censored observations (6th–8th instances), right-censoring is the most common type in biomedical research and our primary focus in this thesis.

Beyond the type of censoring present in the data, survival analysis methods rely on assumptions about the relationship between the event time E and the censoring time C. In the simplest case, censoring is assumed to be independent of the event time, $E \perp\!\!\!\perp C$, implying that censored individuals are representative of the population at risk. When covariates are included, this assumption is typically relaxed to conditional independence, $E \perp\!\!\!\perp C \mid X$, meaning that any dependence between E and C is explained by the observed covariates. Figure 2.3 illustrates these assumptions using directed acyclic graphs (DAGs). In particular, dependent censoring arises when an unobserved variable influences both processes, inducing dependence that cannot be explained by the observed covariates. We now examine these censoring assumptions in more detail.

Random censoring: Random censoring assumes that the censoring time does not provide information about the event time [13, Ch. 1]. In a medical setting, this

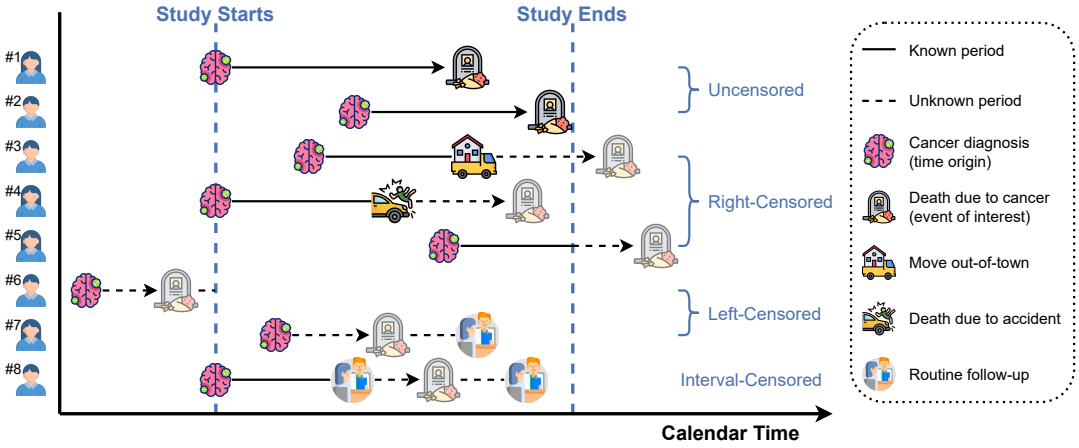


FIGURE 2.2: Illustration of different types of censoring in a hypothetical survival dataset with 8 patients. Each timeline starts at a shared time origin – the date of cancer diagnosis – and the outcome of interest is cancer-specific death. Solid lines denote observed time periods, while dashed lines indicate periods where the event time is unknown. A solid tombstone marks an observed event, whereas a translucent tombstone indicates a censored observation. For right-censored instances (3rd–5th), the true event time exceeds the observed time, meaning that we observe $(t_i, \delta_i = 0)$. Left- and interval-censored examples (6th–8th) illustrate settings where the event is only known to occur before or within a time interval. Competing events (*e.g.*, death due to an accident) are here treated as censoring, although they may be modeled separately in competing risks frameworks. Figure taken with permission from [59].

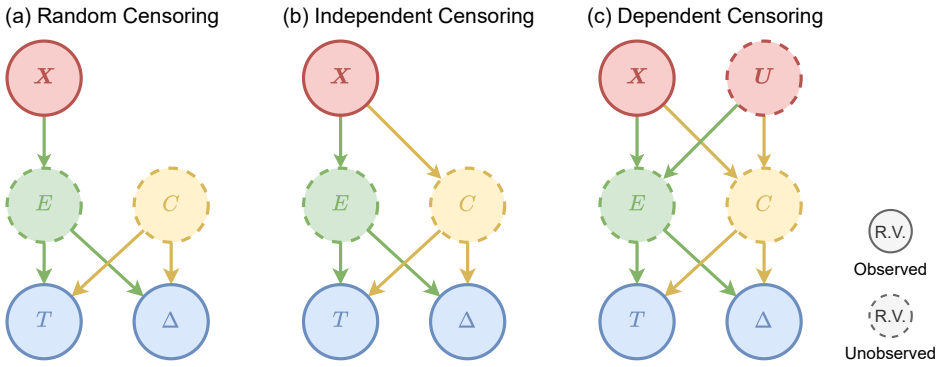


FIGURE 2.3: Comparison of three censoring assumptions illustrated using directed acyclic graphs (DAGs). Solid nodes denote observed variables, while dashed nodes indicate unobserved variables. Arrows represent potential dependencies between variables. (a) Random censoring: the censoring time C is unrelated to both the event time E and the covariates X. (b) Independent censoring: both E and C may depend on the observed covariates X, but are conditionally independent given X. (c) Dependent censoring: an unobserved variable U influences both E and C, inducing dependence that cannot be explained by X. Figure taken with permission from [59].

implies that the risk of event among censored individuals is the same as that of individuals who remain under observation in the risk set[3]. Formally, this corresponds

[3]The *risk set* at time t consists of all individuals who are still under observation and have not yet experienced the event of interest just prior to time t. These individuals are considered at risk of experiencing the event at time t, and thus contribute to the estimation of the hazard and related quantities.

to the assumption that the censoring time is independent of both the event time and the covariates, $C \perp\!\!\!\perp (E, X)$.

Definition 5 *Following the definition of Kleinbaum and Klein [13, Ch. 1], random censoring occurs when*

$$C \perp\!\!\!\perp (E, X), \tag{2.10}$$

which implies that

$$Pr(E > t \mid C \geq t) = Pr(E > t) \quad \textit{for all } t \textit{ with } Pr(C \geq t) > 0, \tag{2.11}$$

where E is the event time, C is the censoring time, and X are the observed covariates. In the absence of covariates, this reduces to $E \perp\!\!\!\perp C$.

Independent censoring: Independent censoring assumes that the censoring time does not provide information about the event time once the observed covariates are taken into account [13, Ch. 1]. In other words, within groups defined by the covariates, individuals censored at time t have the same risk of event as that of individuals who remain under observation in the risk set. This assumption is therefore also referred to as *conditional independent censoring*, as any dependence between the event and censoring processes is assumed to be explained by the observed covariates.

Definition 6 *Following the definition of Kleinbaum and Klein [13, Ch. 1], independent censoring occurs when*

$$E \perp\!\!\!\perp C \mid X, \tag{2.12}$$

which implies that

$$Pr(E > t \mid C \geq t, X) = Pr(E > t \mid X) \quad \textit{for all } t \textit{ with } Pr(C \geq t \mid X) > 0, \tag{2.13}$$

where E is the event time, C is the censoring time, and X are the observed covariates. That is, among subjects with the same covariates, those still under observation at time t are representative of all subjects with respect to their event times. This is a less restrictive assumption than random censoring.

Non-informative censoring: Non-informative censoring assumes that the censoring mechanism does not provide information about the parameters governing the survival distribution. In other words, the process determining when individuals are censored does not influence inference about the distribution of the event times [13, Ch. 1]. In practice, this assumption is often satisfied when censoring occurs independently of the event process, although the precise meaning of "non-informative" censoring varies somewhat across the survival analysis literature.

Definition 7 *Non-informative censoring occurs when the distribution of the event time provides no information about the distribution of the censoring time, and vice versa [13, Ch. 1]. In a parametric setting, this means that $Pr(E \mid X; \theta)$ and $Pr(C \mid X; \eta)$ are governed by functionally independent parameters θ and η.*

Dependent censoring: Dependent censoring arises when the censoring mechanism is related to the event process in a way that is not fully explained by the observed covariates. In such cases, even within groups defined by the covariates, individuals censored at time t are not representative of those who remain at risk at

time t in terms of their survival experience. As a result, the observed covariates do not fully account for the relationship between the event and censoring processes.

Definition 8 *Following Emura and Chen [14, Sec. 1.2], dependent censoring occurs when the dependence between the event time and the censoring time is not explained by the observed covariates,* i.e., *the independent censoring assumption [14, Ch. 2] is violated:*

$$E \not\perp\!\!\!\perp C \mid X. \tag{2.14}$$

In particular, dependent censoring is present whenever

$$Pr(E > t \mid C \geq t, X) \neq Pr(E > t \mid X) \quad \textit{for some } t \textit{ with } Pr(C \geq t \mid X) > 0, \tag{2.15}$$

where E is the event time, C is the censoring time, and X are the observed covariates.

Many standard techniques in survival analysis, for example, the KM estimator, and the semiparametric CoxPH model, give unbiased estimates under the assumption of random or independent censoring [13, Ch. 1], respectively. When this assumption is violated, as we saw with dependent censoring, these methods may give biased estimates of the survival probabilities and the hazard ratios. This is because the censored individuals are no longer representative of the full population at risk – their hazard of experiencing the event may differ systematically from those who remain uncensored. Consequently, if a significant proportion of censored individuals actually experienced the event (*e.g.*, died) shortly after censoring, the estimated survival probability at time t will be biased upwards, thereby overestimating the true survival probability. The magnitude of this bias depends on both how many such events are unobserved and how soon after censoring they occur.

2.3 Likelihood-based Approaches

Now that we have presented the standard framework for survival analysis and defined some common prediction targets, we look at the likelihood function commonly used to derive many survival models. Recall that we defined survival time E and censoring time C such that:

- $E = t_i$ and $C > t_i$ if $\delta_i = 1$,
- $E > t_i$ and $C = t_i$ if $\delta_i = 0$.

Combining these two events, the likelihood for the i-th instance is expressed as

$$\mathcal{L}_i = \Pr(E = t_i, C > t_i \mid \boldsymbol{x}_i)^{\delta_i} \Pr(E > t_i, C = t_i \mid \boldsymbol{x}_i)^{1-\delta_i} \tag{2.16}$$

Under the independent censoring assumption (Definition 6), $E \perp\!\!\!\perp C \mid X$, so the joint probabilities factorize and we have

$$\begin{aligned}
\mathcal{L}_i &= \left[\Pr(E = t_i \mid \boldsymbol{x}_i) \Pr(C > t_i \mid \boldsymbol{x}_i)\right]^{\delta_i} \left[\Pr(E > t_i \mid \boldsymbol{x}_i) \Pr(C = t_i \mid \boldsymbol{x}_i)\right]^{1-\delta_i} \\
&= \left[f_E(t_i \mid \boldsymbol{x}_i)\, S_C(t_i \mid \boldsymbol{x}_i)\right]^{\delta_i} \left[S_E(t_i \mid \boldsymbol{x}_i)\, f_C(t_i \mid \boldsymbol{x}_i)\right]^{1-\delta_i} \\
&= \left[f_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} S_E(t_i \mid \boldsymbol{x}_i)^{1-\delta_i}\right] \left[f_C(t_i \mid \boldsymbol{x}_i)^{1-\delta_i} S_C(t_i \mid \boldsymbol{x}_i)^{\delta_i}\right],
\end{aligned} \tag{2.17}$$

where $S_E(t \mid \boldsymbol{x}_i) = \Pr(E > t \mid \boldsymbol{x}_i)$ as in Definition 2, $f_E(t \mid \boldsymbol{x}_i) = -dS_E(t \mid \boldsymbol{x}_i)/dt$, and analogously $S_C(t \mid \boldsymbol{x}_i) = \Pr(C > t \mid \boldsymbol{x}_i)$ and $f_C(t \mid \boldsymbol{x}_i) = -dS_C(t \mid \boldsymbol{x}_i)/dt$. If,

in addition, censoring is non-informative (Definition 7), the second bracket does not depend on the parameters θ of the event-time distribution. It is therefore a constant factor with respect to θ and can be dropped when maximizing the likelihood over all N instances, which gives

$$\mathcal{L} \propto \prod_{i=1}^{N} f_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} S_E(t_i \mid \boldsymbol{x}_i)^{1-\delta_i}, \tag{2.18}$$

where $f_E(t_i \mid \boldsymbol{x}_i)$ is the PDF of the event time and $S_E(t_i \mid \boldsymbol{x}_i)$ is the conditional survival function of the event time. Note that the censoring terms $f_C(t_i \mid \boldsymbol{x}_i)$ and $S_C(t_i \mid \boldsymbol{x}_i)$ are absent in (2.18). If the event is observed, the likelihood contribution is the probability density of the event occurring at that time. If the observation is censored, the likelihood contribution is the probability of surviving beyond that time. The overall likelihood is the product of these contributions across all instances.

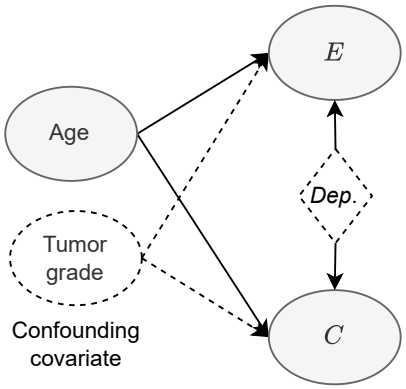


FIGURE 2.4: Residual dependence between event time E (cancer relapse) and censoring time C (dropout) denoted by the diamond "Dep", due to an unobserved confounding covariate – here tumor grade. Figure taken with permission from [51].

However, when E and C are no longer independent, we can no longer rely on this clean decomposition of the likelihood. Figure 2.4 shows how the independent censoring assumption fails to hold by omitting a confounding covariate, for example, *tumor grade*. In this example, since tumor grade relates to both the event of interest E (for example, cancer relapse) and censoring C (for example, dropout), the variation in this covariate induces changes in both E and C. That is, we want the model to predict the time until a patient experiences a cancer relapse, but dropout (which may be related to disease progression) acts as a censoring mechanism. Because tumor grade affects both relapse and dropout, advanced (resp., early) stages lead to shorter (resp., longer) event and censoring times.

Instead, we can introduce a so-called copula, denoted $\mathcal{C}(u_1, u_2)$[4], to model dependence between E and C. A copula links two or more random variables by specifying their dependence structure while preserving their marginal distributions [14, Ch. 1][5].

Figure 2.5 shows samples from a Clayton copula applied to the marginal CDFs, which exhibits *lower-tail dependence* – a tendency for early events (small values of E)

[4]Here, u_1 and u_2 represent marginal probabilities, typically $u_1 = S_E(t \mid \boldsymbol{x})$ and $u_2 = S_C(c \mid \boldsymbol{x})$, so the copula operates on probabilities rather than the raw variables themselves.

[5]Intuitively, a copula separates *what* each variable does on its own (its marginal distribution) from *how* variables move together (their dependence), and then recombines the two into a joint model.

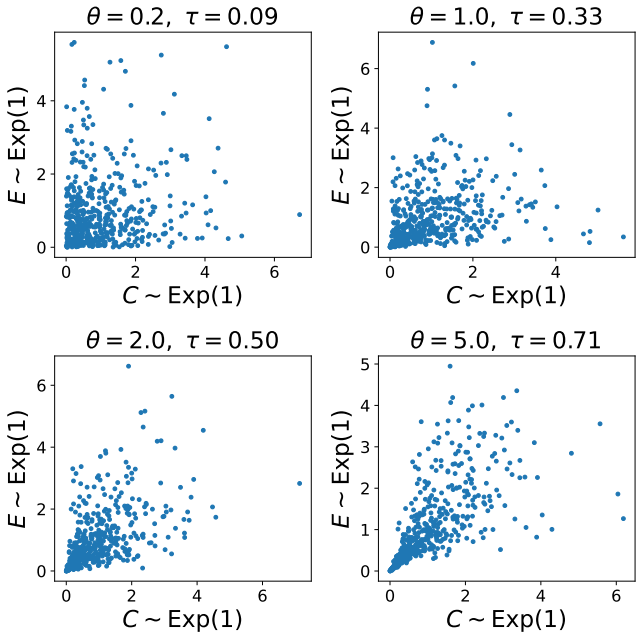


FIGURE 2.5: Scatter plots of (e_i, c_i), $i = 1, \ldots, 500$, simulated from an Archimedean Clayton copula [60] with parameter θ (reported via the corresponding Kendall's τ) and standard exponential marginals. Samples are generated by drawing $(u_{1i}, u_{2i}) \sim \mathcal{C}_\theta$ and applying the inverse exponential CDF. The structure is induced by the generator φ_θ, where $\mathcal{C}_\theta(u_1, u_2) = \varphi_\theta^{-1}(\varphi_\theta(u_1) + \varphi_\theta(u_2))$. Note that this construction applies the copula to the marginal CDFs; under the survival-copula convention of Lemma 1, the same dependence pattern corresponds to the rotated (survival) Clayton copula. Figure taken with permission from [51].

to occur alongside early censoring (small values of C). When the Clayton copula is instead applied to the marginal survival functions, as in Lemma 1, the dependence is concentrated at late times. The concept of a copula was first introduced by Abe Sklar in his study of probabilistic metric spaces [61]. In this work, Sklar defined copulas and proved what is now known as Sklar's theorem, which states that any joint cumulative distribution can be expressed as a copula applied to its marginal distributions. Among the various types of copulas, the Archimedean copula is widely used, and allows modeling of dependencies in arbitrarily high dimensions using a single parameter. An Archimedean copula is characterized by a convex generator function $\varphi : [0,1] \to [0,\infty]$, which is continuous, strictly decreasing, and satisfies the boundary conditions $\varphi(0) = \infty$ and $\varphi(1) = 0$ [14, Ch. 3]. Given a generator, the copula takes the form $C_\theta(u_1, u_2) = \varphi_\theta^{-1}(\varphi_\theta(u_1) + \varphi_\theta(u_2))$, where the parameter θ controls the strength of dependence; the generator $\varphi(u) = -\log u$ recovers the independence copula $C(u_1, u_2) = u_1 u_2$. Two common families in survival analysis are the Clayton copula [60], with $\varphi_\theta(u) = (u^{-\theta} - 1)/\theta$ and Kendall's $\tau = \theta/(\theta + 2)$, and the Frank copula [62], which models symmetric dependence without tail concentration [57]. Importantly, the dependence between E and C is not identifiable from right-censored data alone, since only the smaller of the two times is observed for each instance [63]; the copula family and its parameter must therefore be assumed or estimated under additional assumptions. Based on Sklar's theorem, we can modify Equation 2.17 to account for dependent censoring. To formalize this construction, we introduce Lemma 1.

Lemma 1 *Let E and C denote the event time and censoring time, respectively, with continuous marginal distributions, and suppose their joint survival function $S_{E,C}(t, c \mid \boldsymbol{x}) := Pr(E > t, C > c \mid \boldsymbol{x})$ can be expressed via a copula $\mathcal{C}(u_1, u_2)$ such that*

$$S_{E,C}(t, c \mid \boldsymbol{x}) = \mathcal{C}\big(S_E(t \mid \boldsymbol{x}),\, S_C(c \mid \boldsymbol{x})\big), \tag{2.19}$$

where $S_E(t \mid \boldsymbol{x})$ and $S_C(c \mid \boldsymbol{x})$ are the marginal survival functions. Then, the conditional survival functions of each time given the other satisfy

$$Pr(C > c \mid E = t, \boldsymbol{x}) = \partial_1 \mathcal{C}\big(S_E(t \mid \boldsymbol{x}),\, S_C(c \mid \boldsymbol{x})\big), \tag{2.20}$$

$$Pr(E > t \mid C = c, \boldsymbol{x}) = \partial_2 \mathcal{C}\big(S_E(t \mid \boldsymbol{x}),\, S_C(c \mid \boldsymbol{x})\big), \tag{2.21}$$

where $\partial_j \mathcal{C}(u_1, u_2) := \partial \mathcal{C}(u_1, u_2)/\partial u_j$ denotes the partial derivative of the copula with respect to its j-th argument.

Applying Lemma 1 to the general likelihood in Equation 2.16, *i.e.*, factorizing each joint probability into a marginal density and a conditional survival probability rather than a product of marginals, and taking the sum of logs, we obtain the likelihood under dependent censoring [64]:

$$\begin{aligned}\log \mathcal{L}(\mathcal{D}) = \sum_{i=1}^{N} \Big\{ &\delta_i \log f_E(t_i \mid \boldsymbol{x}_i) + \delta_i \log \partial_1 \mathcal{C}\big(S_E(t_i \mid \boldsymbol{x}_i), S_C(t_i \mid \boldsymbol{x}_i)\big) \\ &+ (1 - \delta_i) \log f_C(t_i \mid \boldsymbol{x}_i) + (1 - \delta_i) \log \partial_2 \mathcal{C}\big(S_E(t_i \mid \boldsymbol{x}_i), S_C(t_i \mid \boldsymbol{x}_i)\big) \Big\}.\end{aligned} \tag{2.22}$$

In Equation 2.22, the first term corresponds to the event log-likelihood, while the second captures the conditional probability of surviving censoring beyond t_i given $E = t_i$. The remaining terms describe the censoring contribution. For Archimedean copulas, these derivatives admit closed-form expressions and can be optimized via gradient-based methods.

Survival prediction models are often specified by parameterizing one of the functions in Summary 1, *e.g.*, the hazard $h_E(t \mid \boldsymbol{x}; \theta)$, where θ denotes the model parameters. We now consider two common likelihood-based models.

Example 1 ***(Exponential survival model [6])***. *Let $\mathcal{X} = \mathbb{R}^d$ and we define the hazard function as*

$$h_E(t \mid \boldsymbol{x}) := \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi), \tag{2.23}$$

with parameters $\theta = (\boldsymbol{\beta}, \psi) \in \mathbb{R}^d \times \mathbb{R}$. The hazard is constant in t. The cumulative hazard and survival functions are defined as

$$H_E(t \mid \boldsymbol{x}) = \int_0^t h_E(u \mid \boldsymbol{x})\, du = t \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi), \tag{2.24}$$

$$S_E(t \mid \boldsymbol{x}) = \exp\big(-H_E(t \mid \boldsymbol{x})\big) = \exp\Big(-t \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi)\Big). \tag{2.25}$$

Thus, $S_E(t \mid \boldsymbol{x})$ corresponds to an exponential distribution with rate parameter $\lambda(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi)$, which determines the constant event rate over time.

Recall from Summary 1 that (1) $f_E(t \mid \boldsymbol{x}) = h_E(t \mid \boldsymbol{x}) S_E(t \mid \boldsymbol{x})$, and (2) $S_E(t \mid \boldsymbol{x}) = \exp\Big(-\int_0^t h_E(u \mid \boldsymbol{x})\, du\Big)$. Using these identities, the likelihood in Equation 2.18 can

be rewritten as [6]:

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} f_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} S_E(t_i \mid \boldsymbol{x}_i)^{1-\delta_i}, \\ (1) &= \prod_{i=1}^{N} h_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} S_E(t_i \mid \boldsymbol{x}_i), \\ (2) &= \prod_{i=1}^{N} h_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} \exp\left(-\int_0^{t_i} h_E(u \mid \boldsymbol{x}_i)\, du\right). \end{aligned} \tag{2.26}$$

Next, we plug in $h_E(t \mid \boldsymbol{x}) = h_E(t \mid \boldsymbol{x}; \theta)$ to get the following likelihood function:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} h_E(t_i \mid \boldsymbol{x}_i; \theta)^{\delta_i} \exp\left(-\int_0^{t_i} h_E(u \mid \boldsymbol{x}_i; \theta)\, du\right). \tag{2.27}$$

We find θ by maximum likelihood estimation (MLE), typically by maximizing the log-likelihood using gradient-based methods, *i.e.*, $\hat{\theta} := \arg\max_\theta \log \mathcal{L}(\theta)$, which gives the same solution as maximizing $\mathcal{L}(\theta)$. The log-likelihood can be written as [6]:

$$\log \mathcal{L}(\theta) = \log \prod_{i=1}^{N} \left[h_E(t_i \mid \boldsymbol{x}_i; \theta)^{\delta_i} \exp\left(-\int_0^{t_i} h_E(u \mid \boldsymbol{x}_i; \theta)\, du\right)\right] \tag{2.28}$$

$$= \sum_{i=1}^{N} \left[\delta_i \log h_E(t_i \mid \boldsymbol{x}_i; \theta) - \int_0^{t_i} h_E(u \mid \boldsymbol{x}_i; \theta)\, du \right]. \tag{2.29}$$

In practice, the loss is often defined as the negative log-likelihood (NLL), averaged over the training data [6]:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{N} \log \mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left[\delta_i \log h_E(t_i \mid \boldsymbol{x}_i; \theta) - \int_0^{t_i} h_E(u \mid \boldsymbol{x}_i; \theta)\, du \right]. \tag{2.30}$$

Continuing with Example 1, where $\mathcal{X} = \mathbb{R}^d$, $\theta = (\boldsymbol{\beta}, \psi) \in \mathbb{R}^d \times \mathbb{R}$, and $h_E(t \mid \boldsymbol{x}; \theta) = \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi)$, we plug this parametric form of $h_E(t \mid \boldsymbol{x}; \theta)$ into Equation 2.30 to get the negative log-likelihood function for an exponential survival model [6]:

$$\begin{aligned}\mathcal{L}_{\text{Exp}}(\boldsymbol{\beta}, \psi) &= -\frac{1}{N} \sum_{i=1}^{N} \left[\delta_i (\boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi) - \int_0^{t_i} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi)\, du \right] \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left[\delta_i (\boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi) - t_i \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi) \right]. \end{aligned} \tag{2.31}$$

We then numerically minimize Equation 2.31 with respect to parameters $\boldsymbol{\beta}$ and ψ using some optimization method (*e.g.*, gradient descent) to obtain $(\hat{\boldsymbol{\beta}}, \hat{\psi})$.

Example 2 ***(Weibull survival model [6])***. *Suppose that $\mathcal{X} = \mathbb{R}^d$, and we set $h_E(t \mid \boldsymbol{x})$ to be*

$$h_E(t \mid \boldsymbol{x}; \theta) := t^{\phi} \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi + \phi) \quad \textit{for } t \geq 0,\ \boldsymbol{x} \in \mathcal{X}. \tag{2.32}$$

This is a generalization of the exponential model (from Example 1), which is a special case where $\phi = 0$. In this more general case, the collection of parameters is $\theta = (\boldsymbol{\beta}, \psi, \phi) \in \mathbb{R}^d \times \mathbb{R} \times (-1, \infty)$, and the survival function corresponds to a Weibull distribution. To see this, we calculate the cumulative hazard and survival functions:

$$\begin{aligned} H_E(t \mid \boldsymbol{x}) &= \int_0^t h_E(u \mid \boldsymbol{x}; \theta)\, du \\ &= \int_0^t u^{\phi} \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi + \phi)\, du \\ &= \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi + \phi) \frac{t^{\phi+1}}{\phi + 1}, \end{aligned} \tag{2.33}$$

$$\begin{aligned} S_E(t \mid \boldsymbol{x}) &= \exp(-H_E(t \mid \boldsymbol{x})) \\ &= \exp\left(- \exp(\boldsymbol{\beta}^\top \boldsymbol{x} + \psi + \phi) \frac{t^{\phi+1}}{\phi + 1} \right). \end{aligned} \tag{2.34}$$

Here, $S_E(t \mid \boldsymbol{x})$ corresponds to a Weibull distribution with shape parameter $\phi + 1$ and scale parameter $\left[(\phi + 1) \exp\big(- (\boldsymbol{\beta}^\top \boldsymbol{x} + \psi + \phi)\big)\right]^{1/(\phi+1)}$.

In terms of learning the model parameters for the Weibull model in Example 2, we can use a gradient-descent technique to minimize the loss from Equation 2.30, which for the Weibull model becomes [6]:

$$\begin{aligned} \mathcal{L}_{\text{Weibull}}(\boldsymbol{\beta}, \psi, \phi) &= -\frac{1}{N} \sum_{i=1}^{N} \left[\delta_i \log h_E(t_i \mid \boldsymbol{x}_i; \theta) - \int_0^{t_i} h_E(u \mid \boldsymbol{x}_i; \theta)\, du \right] \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left[\delta_i \left(\phi \log t_i + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi + \phi \right) \right. \\ &\qquad \left. - \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i + \psi + \phi) \frac{t_i^{\phi+1}}{\phi + 1} \right]. \end{aligned} \tag{2.35}$$

Similar to the exponential model, we obtain parameter estimates $(\hat{\boldsymbol{\beta}}, \hat{\psi}, \hat{\phi})$ by minimizing Equation 2.35 and plug the estimates into Equations 2.32, 2.33 and 2.34 to obtain the hazard, cumulative hazard and survival functions, respectively [6].

The exponential and Weibull models are special cases of the broader class of proportional hazards models, of which the most popular is the semiparametric CoxPH model [23] and its deep learning variant DeepSurv [28]. In general, proportional hazards models assume that the hazard function factorizes as

$$h_E(t \mid \boldsymbol{x}) = h_E^0(t; \theta)\, \exp\big(f(\boldsymbol{x}; \theta)\big) \quad \text{for } t \geq 0,\ \boldsymbol{x} \in \mathcal{X}, \tag{2.36}$$

where we have two functions to learn: the so-called baseline hazard function $h_E^0(\cdot; \theta) : [0, \infty) \to [0, \infty)$ and the log-partial hazard function $f(\cdot; \theta) : \mathcal{X} \to \mathbb{R}$, both of which have parameter θ. The factorization implies that, regardless of what the covariate instance $\boldsymbol{x}$ is, the hazard function $h_E(\cdot|\boldsymbol{x})$ must be proportional to the baseline hazard

$h_E^0(\cdot;\theta)$, which is why equation 2.36 is called the proportional hazards assumption. The factorization is also why CoxPH is commonly referred to as a semiparametric proportional hazards model.

The log-partial hazard function $f(\cdot;\theta)$ maps each instance $\boldsymbol{x} \in \mathcal{X}$ to a real-valued *risk score*. Broadly speaking, a risk score quantifies an instance's likelihood of experiencing the event of interest. A higher value of $f(\boldsymbol{x};\theta)$ implies that the survival time tends to be lower. Since all instances share the same time dependence captured by the baseline hazard $h_E^0(\cdot;\theta)$, under the proportional hazards assumption, the only difference between instances $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ is captured by comparing their log-partial hazard values $f(\boldsymbol{x};\theta)$ and $f(\boldsymbol{x}';\theta)$. Thus, proportional hazards models can be viewed as ranking instances via the scoring function $f(\cdot;\theta)$.

The standard maximum likelihood estimator $\hat{\theta}$ is obtained by maximizing the Cox partial log-likelihood:

$$\begin{aligned}\log \mathcal{L}(\mathcal{D} \mid \theta) &= \sum_{i:\delta_i=1} \log \frac{\exp(f(\boldsymbol{x}_i;\theta))}{\sum_{j:t_j \geq t_i} \exp(f(\boldsymbol{x}_j;\theta))} \\ &= \sum_{i:\delta_i=1} \left[f(\boldsymbol{x}_i;\theta) - \log \sum_{j:t_j \geq t_i} \exp(f(\boldsymbol{x}_j;\theta)) \right]. \end{aligned} \tag{2.37}$$

Examples 1 and 2 showed two proportional hazards models:

- In the exponential model, $\theta = (\boldsymbol{\beta}, \psi) \in \mathbb{R}^d \times \mathbb{R}$, $h_E^0(t;\theta) = \exp(\psi)$, and $f(\boldsymbol{x};\theta) = \boldsymbol{\beta}^\top \boldsymbol{x}$.
- In the Weibull model, $\theta = (\boldsymbol{\beta}, \psi, \phi) \in \mathbb{R}^d \times \mathbb{R} \times (-1, \infty)$, $h_E^0(t;\theta) = t^\phi \exp(\psi + \phi)$, and $f(\boldsymbol{x};\theta) = \boldsymbol{\beta}^\top \boldsymbol{x}$.

We call these models *parametric proportional hazards models* because we assume parametric forms for both the baseline hazard $h_E^0(\cdot;\theta)$ and the log-partial hazard $f(\cdot;\theta)$. Earlier, we showed how to estimate θ via maximum likelihood and use the fitted parameters to compute the hazard, cumulative hazard, and survival functions.

A major advantage of the CoxPH model – where $f(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{x}$ with both $\boldsymbol{\beta}, \boldsymbol{x} \in \mathbb{R}^d$ – is interpretability when the input features are themselves interpretable (as is typical for tabular data). In this setting, the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$ directly reflects feature importance: the j-th feature is weighted by β_j. A standard interpretive measure is the *hazard ratio*, defined as $\exp(\beta_j)$. To see this, consider $\boldsymbol{x} \in \mathbb{R}^d$ and $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ identical except that $\tilde{x}_j = x_j + 1$. Then,

$$\frac{h_E(t \mid \tilde{\boldsymbol{x}})}{h_E(t \mid \boldsymbol{x})} = \exp(\beta_j),$$

so increasing the j-th feature by one unit (holding others fixed) multiplies the hazard by $\exp(\beta_j)$. Thus, $\exp(\beta_j)$ quantifies the multiplicative change in risk associated with a one-unit increase in feature j.

2.4 Evaluation Metrics

The primary challenge in evaluating survival models is how to deal with censored instances. We are often given a training instance $\boldsymbol{x}_i$ with an observed time t_i instead of the actual event time e_i, which may be censored. Consequently, even if our survival model generates an estimate $\hat{e}_i$ for the true survival time e_i, the actual time

may be unknown (censored) and therefore cannot be directly compared. However, there exist ways to make such a comparison (indirectly) for the instances that are censored, and we will cover the common ones here.

Importantly, no single metric can be considered universally "best" on its own. In fact, many survival metrics have not been rigorously validated for all possible scenarios and some are not even proper scoring rules, which means that a better model does not always give a better evaluation result [65]. Therefore, we will review what aspect of model performance each metric measures and highlight its limitations. Figure 2.6 presents a taxonomy for evaluating survival outcomes [59], illustrated with four uncensored instances. These metrics can be grouped in two ways. First, by the *type of prediction*, as (a, b, c) focus on point predictions $\hat{e}_i$ of the time-to-event, (d, e, f) evaluate the single-time survival probabilities $\hat{S}_E(t^* \mid \boldsymbol{x}_i)$, and (g, h, i) assess the overall survival function for each individual $\hat{S}_E(\cdot \mid \boldsymbol{x}_i)$. Second, by the *aspect of evaluation*, as (a, d, g) evaluate only discrimination performance, (b, e, h) evaluate both discrimination and calibration performance (referred to absolute and squared errors), and (c, f, i) evaluate only calibration performance. Next, we will take a closer look at the three aspects of evaluations.

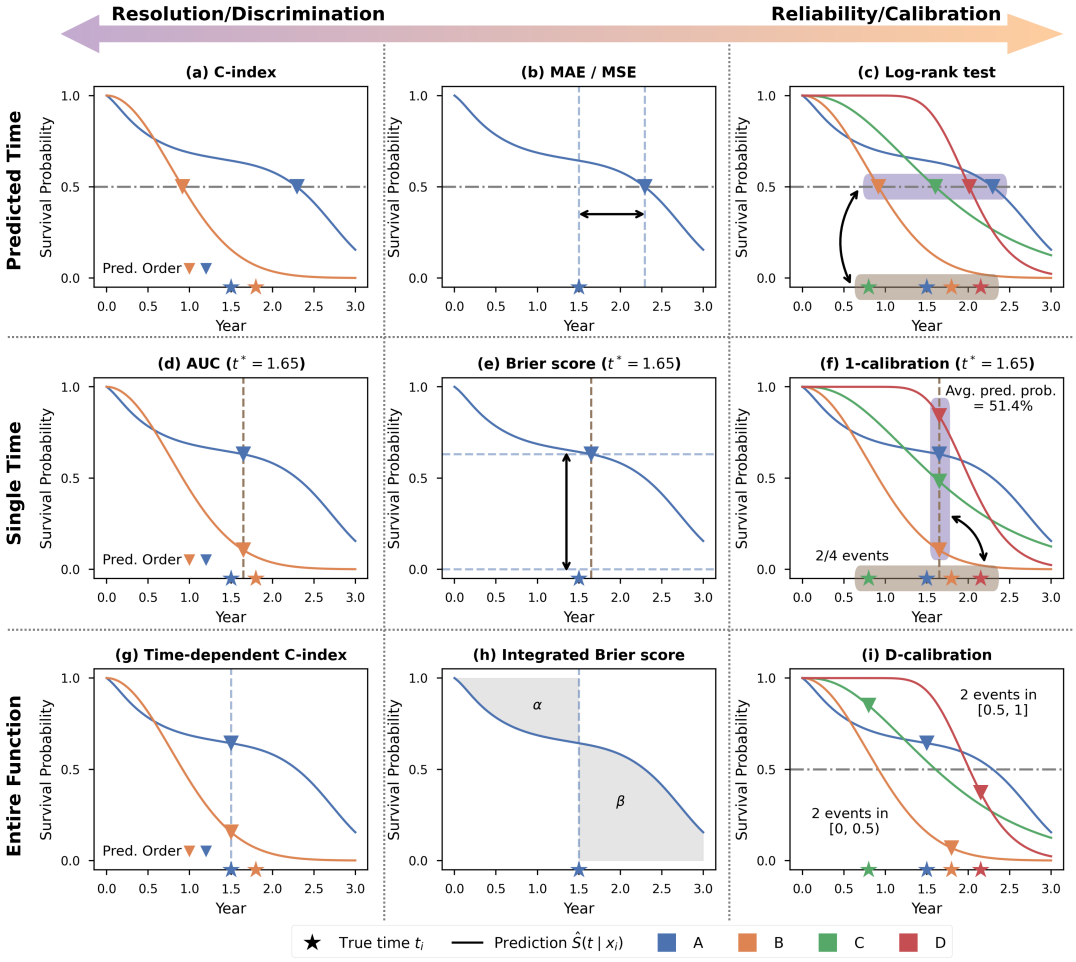


FIGURE 2.6: Taxonomy of evaluation metrics for survival analysis with four uncensored instances. Figure taken with permission from [59].

2.4.1 Discrimination

Discrimination metrics assess how well a model ranks individuals according to their risk of experiencing the event. This is particularly relevant when the goal is to prioritize high-risk individuals before the low-risk individuals – for example, ranking patients by severity to allocate scarce medical resources or schedule surgeries earlier for those most in need [66]. The most widely used metric for evaluating discriminative performance is Harrell's C-index [36], which measures the proportion of *concordant* pairs among all *comparable* pairs. A pair (i, j) is comparable if (1) individual i has an observed event time t_i that is strictly less than t_j, and (2) i experiences the event (*i.e.*, $\delta_i = 1$). For the visually inclined, Figure 2.7 shows the concept of comparable/not comparable pairs. A pair is *concordant* if the individual who experiences the event first also has the higher predicted risk score. Figure 2.6 (a) provides an illustrative example using two instances, A and B, and their predicted median survival times. The order of their true event times (stars) and the order of predicted median times (diamonds) are not aligned, indicating that this is a discordant pair. Similar to the area under the receiver operating characteristic curve (AUC) in binary classification, a C-index of 0.5 is considered low and can be obtained through "random guessing", so ideally we want a C-index between 0.5 and 1.

Definition 9 *Harrell's concordance index is defined as [36]:*

$$C_{Harrell}(\hat{\eta}(\cdot); \mathcal{D}) = \frac{\sum_{i,j} \delta_i \cdot \mathbb{1}[t_i < t_j] \cdot \mathbb{1}[\hat{\eta}_i > \hat{\eta}_j]}{\sum_{i,j} \delta_i \cdot \mathbb{1}[t_i < t_j]}, \tag{2.38}$$

where $\hat{\eta}_i := \hat{\eta}(\boldsymbol{x}_i)$ denotes the predicted risk score for the i-th instance and δ_i is the event indicator.

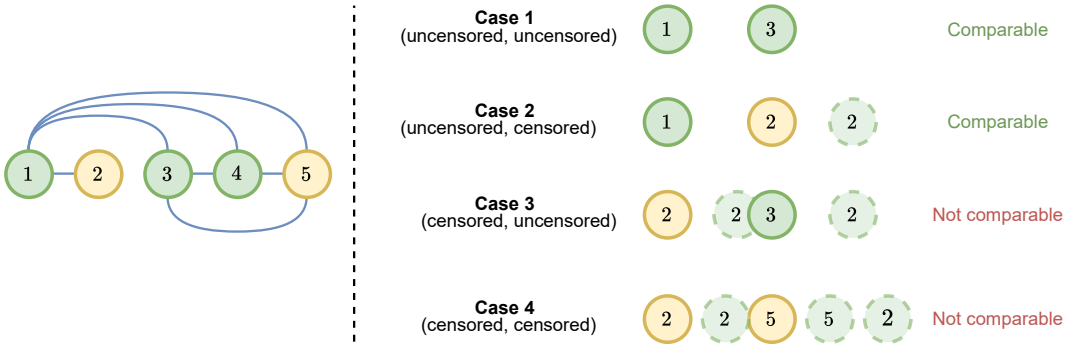


FIGURE 2.7: Comparable pairs under right-censoring, illustrated with five instances (1-5). The blue connections in the left panel indicate all comparable pairs. Solid circles indicate observed times: green for events and yellow for censoring. For censored instances, dashed transparent circles extend to the right, representing possible (unobserved) event times beyond the censoring point. Case 1: Both event times are observed exactly, so their ordering is clear and the pair is *comparable*. Case 2: All possible event times for the censored instance fall to the right, so the ordering is still known and the pair remains *comparable*. Case 3: The censored instance could fail either before or after the observed event, making the ordering ambiguous and the pair *not comparable*. Case 4: Either instance could experience the event first, so the pair is *not comparable*. Figure taken with permission from [59].

We note two key limitations to Harrell's C-index in 2.38. First, to compute it, we need to produce a risk score estimate for each instance. However, many survival models do not explicitly learn such a function, as they may learn the individual survival distribution, $\hat{S}_E(t^* \mid \boldsymbol{x}_i)$, directly. A simple solution is to define the risk score as the negative of the predicted *median survival time* ($-\hat{e}_i$) using Equations 2.4, 2.5 or 2.6 given the survival function. This allows for ranking the instances according to their estimated median (or mean) survival times, but according to Harrell Jr et al. [36], this substitution is only allowed when the predicted probabilities and times have a one-to-one mapping (*i.e.*, proportional hazard is satisfied). Though in many cases, the survival curves for different instances can intersect, violating the proportional hazard assumption. As a result, using predicted survival probabilities as risk scores is often not suitable [6]. Second, even under the random censoring assumption, Harrell's CI is biased because comparable pairs selected using the above criteria do not represent all possible pairs based on their true event times [67]:

$$\Pr(\delta_i = 1, t_i < t_j) \quad \neq \quad \Pr(e_i < e_j).$$

To overcome this limitation, Uno et al. [67] proposed a modified concordance index that incorporates inverse probability of censoring weighting (IPCW). The idea behind IPCW is to redistribute the weight of each censored instance to comparable instances that remain under observation, so that the observed data *stand in* for information lost to censoring [68, 69]. This adjustment has been shown to correct for the non-representative sampling induced by random censoring [70, 71].

Formally, the weight assigned to instance i at its observed time t_i is given by:

$$w_i^{\text{IPCW}} \;=\; \frac{1}{S_C(t_i)}, \tag{2.39}$$

which is the reciprocal of the probability of not being censored at time t_i, denoted by $S_C(t_i)$. Here, $S_C(t)$ represents the marginal survival function of the censoring time, which is typically estimated using the KM estimator (Definition 1). To avoid dividing by zero, we make the assumption that the probability of not being censored is strictly positive ($S_C(t) > 0$) within a prespecified horizon $0 \leq t < \tau$. Figure 2.8 provides a step-by-step illustration of IPCW. In the lower panel, eight instances are arranged chronologically (five events, three censored), each starting with an initial weight of 1. Whenever an instance is censored, its weight is redistributed equally across all later instances that are still *at risk*. In this example:

- Instance c_2 is censored, so its weight of 1 is divided into six equal parts, with each of the six subsequent instances receiving $\frac{1}{6}$.
- Instance c_5 enters with a total weight of $1 + \frac{1}{6}$ (its own initial 1 plus the fraction passed on from c_2). This total is then split evenly among the three later instances, giving each $\frac{1}{3}(1 + \frac{1}{6})$.
- Instance c_7 receives a cumulative weight of $1 + \frac{1}{6} + \frac{1}{3}(1 + \frac{1}{6})$ (its own, plus contributions from c_2 and c_5), which is then fully transferred to the single remaining instance.

Repeating this redistribution through the sequence gives the final weights shown, each equal to $1/\widehat{S}_C(t_i)$, *i.e.*, the inverse of the KM estimate from the upper panel. By up-weighting individuals who remain under observation, IPCW allows the analysis to proceed as though censoring were absent, thereby mitigating bias in survival estimates, prediction accuracy, and related measures.

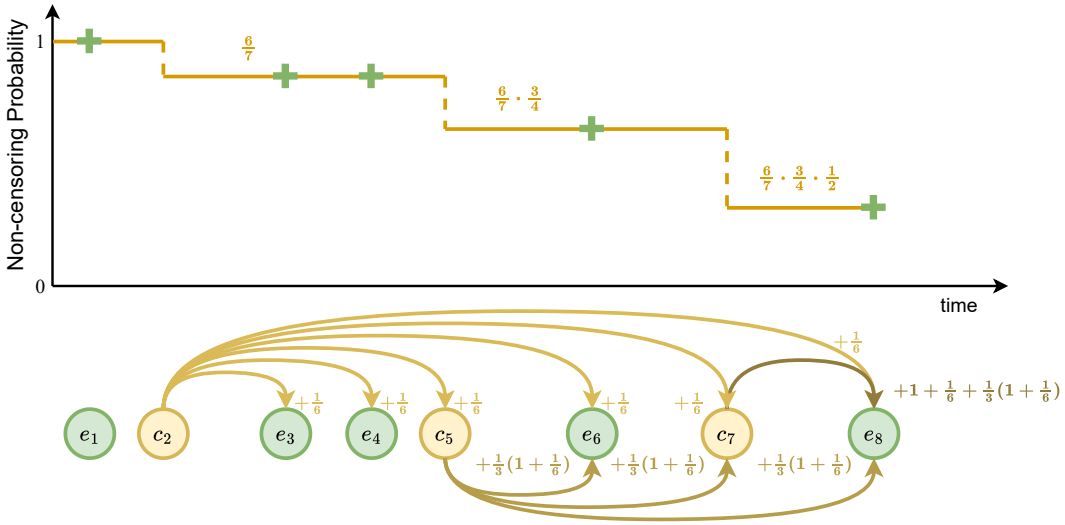


FIGURE 2.8: Example of inverse probability of censoring weighting (IPCW) [68, 69]. Top panel: KM estimate of the censoring survival function $\widehat{S}_C(t)$. Crosses denote observed events, and dashed vertical lines indicate the right-continuous drops of the KM curve occurring at censoring times. Bottom panel: the same timeline, with observations (circles) ordered by occurrence. Each circle is vertically aligned with its corresponding time in the upper panel. Arrows illustrate how the weight from a censored observation is reallocated to later ones, with arrow colors reflecting the magnitude of the transferred weights. Figure taken with permission from [59].

Let us return to Uno's C-index and define it formally. The method applies a squared IPCW adjustment at the earlier event time of each comparable pair, which is denoted $\left(\frac{1}{\widehat{S}_C(t_i)}\right)^2$. Because follow-up is restricted to the study period, a truncated form of Uno's C-index is often used, considering only those pairs for which the earlier event occurs before a specified horizon τ. In practice, τ is usually set near the maximum observed follow-up but not too close to the tail – for instance, at the 90th or 95th percentile of observed times – to avoid the instability introduced by heavy censoring at late times.

Definition 10 *Uno's concordance index is defined as [67]:*

$$C_{Uno}\left(\widehat{\eta}(\cdot), \widehat{S}_C(\cdot); \mathcal{D}, \tau\right) = \frac{\sum_{i,j} \delta_i \cdot \left\{\widehat{S}_C(t_i)\right\}^{-2} \cdot \mathbb{1}[t_i < t_j, t_i < \tau] \cdot \mathbb{1}\left[\widehat{\eta}_i > \widehat{\eta}_j\right]}{\sum_{i,j} \delta_i \cdot \left\{\widehat{S}_C(t_i)\right\}^{-2} \cdot \mathbb{1}[t_i < t_j, t_i < \tau]} \tag{2.40}$$

where $\hat{\eta}_i := \hat{\eta}(\boldsymbol{x}_i)$ denotes the predicted risk score for the i-th instance, δ_i is the event indicator, τ is a user-chosen truncation time, and $\widehat{S}_C(t)$ is the KM estimate of the survival function for the censoring distribution, i.e., *the probability of remaining uncensored beyond time t.*

Although IPCW is effective and can make Uno's C-index unbiased under the random censoring assumption [67], its weights are often approximated using the KM estimator. This estimation may be biased when the random censoring assumption

does not hold. As pointed out in Section 1.4, one of the challenges in survival analysis is evaluating predictive models under so-called *dependent censoring*. In Chapter 5, we will explore ways to deal with this phenomenon.

2.4.2 Absolute and Squared Errors

A good survival model should be both discriminative and calibrated. A class of evaluation metrics that measure both properties is based on absolute and squared errors. These are all negative scoring rules, meaning lower values indicate better performance. First, we will discuss mean absolute error (MAE) and mean squared errors (MSE). Similar to regression, these aim to estimate a real-valued outcome – the time to the event for each instance (*e.g.*, days, weeks, months) given an instance's description. We can calculate the MAE/MSE simply by excluding all censored individuals from the evaluation [72]. For example, for uncensored individuals, the MAE is computed as the average absolute difference between the true event time and the predicted time (Figure 2.6 (b)):

Definition 11 *The uncensored MAE is [59]:*

$$MAE_{uncensored}\left(\{\hat{e}_i\}_{i=1}^N; \mathcal{D}\right) = \frac{\sum_{i=1}^N \mathbb{1}[\delta_i = 1] \cdot |t_i - \hat{e}_i|}{\sum_{i=1}^N \mathbb{1}[\delta_i = 1]}, \tag{2.41}$$

where $\hat{e}_i$ can be derived using e.g., *the mean or median survival time.*

However, excluding all censored individuals may introduce a bias in our evaluation because the uncensored population is often not representative of the entire cohort. To incorporate censored instances into our evaluation, we can handle censoring similar to support vector censoring regression (SVCR) [73], which uses a one-sided (Hinge) absolute error function. The MAE computed with this type of error function is referred to MAE-hinge, formally $\text{MAE}_{\text{hinge}}(\hat{e}_i; t_i, \delta_i = 0) = \max(t_i - \hat{e}_i, 0)$. The MAE-hinge has been widely used in various studies [27, 52, 74], but has notable limitations: It underestimates the true error because it assigns a score of zero when the censoring time is earlier than the predicted survival time and because it assigns a score of $c_i - \hat{e}_i$ when the censoring time is later than the predicted time. As a result, in datasets with high censoring rates, poor models can achieve (unrealistic) low MAE-hinge scores, leading to misleadingly and overoptimistic evaluations [75].

To more accurately estimate error under censoring, Haider et al. [27] and Qi et al. [38] introduced two decensoring methods: MAE-margin and MAE-PO, respectively. Both of these methods try to "decensor" a censored instance: MAE-margin computes the expected survival time beyond the censoring time using a renormalized KM estimator, and MAE-PO uses so-called pseudo-observations [76], a jackknife estimator that also incorporates the KM survival function. Both methods thus adjust for censoring rather than simply ignoring the censored instances. Furthermore, they both weigh the errors using a confidence score based on the estimated decensored times. Notably, neither method guarantee strict consistency with the true MAE but they significantly outperform other MAE variants in simulated experiments with random and independent censoring [38].

Instead of evaluating the predicted time-to-event estimates, we can also evaluate the predicted survival probabilities. The Brier score (BS) measures the accuracy of the probabilistic predictions at a target time t^* [37].

Definition 12 *The Brier Score is defined as [37]:*

$$BS\Big(\widehat{S}_E(t^* \mid \cdot);\, \mathcal{D}\Big) = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathbb{1}\{t_i > t^*\} - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\Big)^2. \tag{2.42}$$

In Figure 2.6 (e), the Brier score calculates the squared error between the predicted probability (y coordinate of the blue diamond, *i.e.*, $S_E(t^* \mid \boldsymbol{x}_A)$) and the survival status of patient A at $t^* = 1.65$ (here the survival status is 0 because patient A already died at 1.5). To incorporate the censored instances, the Brier score typically uses IPCW to uniformly transfer a censored instance's weight (originally 1) to instances with known status at that time [77].

Definition 13 *The Brier score with IPCW weights is defined as [77]:*

$$BS_{IPCW}\Big(\widehat{S}_E(t^* \mid \cdot),\, \widehat{S}_C(\cdot);\, \mathcal{D}\Big) = \frac{1}{N}\sum_{i=1}^{N}\Bigg[\frac{\delta_i\, \mathbb{1}\{t_i \le t^*\}\, \widehat{S}_E(t^* \mid \boldsymbol{x}_i)^2}{\widehat{S}_C(t_i)} + \frac{\mathbb{1}\{t_i > t^*\}\,\big(1 - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\big)^2}{\widehat{S}_C(t^*)}\Bigg]. \tag{2.43}$$

where $\widehat{S}_C$ is the estimated survival function of the censoring distribution.

We see in Definition 12 and 13 that the Brier score depends on the choice of a specific time point t^*, which is often chosen by rule of thumb. And that different choices of t^* will always lead to different Brier scores. In practice, we often want to know an average performance over all prediction times, so we calculate the integral of the Brier score, as proposed by Graf et al. [77]. This is defined as the expectation of the Brier score over all observable time points.

Definition 14 *The integrated Brier score (IBS) is [77]:*

$$IBS\left(\hat{S}_E(\cdot \mid \cdot); \mathcal{D}, \widehat{S}_C(\cdot)\right) = \frac{1}{t_{max}} \cdot \int_0^{t_{max}} BS\left(\hat{S}_E(\tau \mid \cdot); \mathcal{D}, \widehat{S}_C(\cdot)\right) d\tau, \tag{2.44}$$

where the BS inside the integral can either be the vanilla BS or the IPCW-based BS, and t_{max} is usually the maximum event time in the training dataset.

Figure 2.6 (h) shows that the IBS measures the accuracy of the predicted probabilities over $t \in [0, 3]$ via the weighted squared error of the shaded regions ($\alpha + \beta$) – which is between the Heaviside function $\mathbb{1}[t > t_i]$ and the predicted $\hat{S}_E(t \mid \boldsymbol{x}_i)$.

2.4.3 Calibration

Calibration metrics measure the agreement between the empirical probabilities/frequencies and the expected probabilities/frequencies predicted by the model. In other words, they measure how well the prediction fits a set of observations. Here, we briefly motivate the concept of calibration, and then examine single-time calibration and distribution calibration methods.

Imagine that a meteorologist in Aarhus, Denmark, predicts a 50% chance of rain each day from June 1st to 10th. On June 11th, we review the actual weather over the past 10 days. Ideally, we would expect about 5 rainy days and 5 dry days. If we

observe 4 days of rain and 6 dry days, the forecast could still be considered fairly accurate. However, if all 10 days were either rainy or completely dry, we would have reason to doubt the reliability of these predictions. To assess the trustworthiness of the forecast, we may use a statistical test to compare observed and expected frequencies. We can formalize this using a goodness-of-fit test, where the null hypothesis is that the predictions are calibrated. If the p-value is less than 0.05, we reject this hypothesis and conclude that the forecast is not reliable (not calibrated). A high p-value indicates no evidence against calibration, but does not guarantee that the model is well calibrated.

This example shows how calibration can be assessed by comparing predicted probabilities with observed frequencies using a statistical test. This idea directly motivates the concept of single-time calibration (1-calibration) [78]: considering the same patient A and their predicted survival probability at $t^* = 1.65$, $\hat{S}_E(t^* = 1.65 \mid x_A) = 0.63$ (y coordinate of the blue diamond in Figure 2.6 (f)), this implies that for 100 identical patient As (in parallel universes), around 63 are expected to survive beyond 1.65 days. While we can use statistical tests to assess whether empirical outcomes align with predicted probabilities, we cannot observe multiple outcomes for the same patient A in practice. Instead, we group patients with similar predicted survival probabilities and compare their average observed outcomes to the average of their predicted probabilities. To calculate 1-Cal at t^* in practice, we first sort the predicted probabilities at t^* for all patients and put them into K groups, $G_1, \ldots, G_K$, as shown in Figure 2.6 (f) (using $K = 2$). Then, we check whether the average predicted probability and the observed survival probability in each group are statistically similar using the Hosmer-Lemeshow test [79]. Each group's observed survival probability at t^* can be determined using the KM estimator which is trained using the instances in the group [78]. This method remains asymptotically consistent under the assumption of random censoring.

However, instead of single-time calibration, often we want to measure the calibration of the entire survival curve over all times. Consider an oracle draws N realizations of $e_i \mid \boldsymbol{x}_i$ from the true conditional survival function $S_E(t \mid \boldsymbol{x}_i)$ – call them $\{e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(N)}\}$ – then the survival probability at observed time $\{S_E(e_i^{(n)} \mid \boldsymbol{x}_i)\}_n$ should be distributed across a standard uniform distribution $\mathcal{U}_{[0,1]}$ based on the probability integral transform [80]. Figure 2.9 shows this concept for some instance i using the PDF, CDF and survival function. Here, we first draw a single observation for the instance (the orange dot) from the true conditional survival function, then draw a new observation (the purple dot), and so forth, until we have 40 such draws (the gray dots). Ideally, the survival times should then be evenly distributed across the percentile intervals on the y-axis. We have four such intervals (1 to 0.75, 0.75 to 0.5, etc.), so there should be approximately 10 survival times per interval ($40/4 = 10$).

However, in practice, for each instance i, there is only one realization of $e_i \mid \boldsymbol{x}_i$, which means that we cannot check whether a model is calibrated in this way. To address this, Haider et al. [27] proposed *distribution calibration* (D-calibration), which holds when the predicted survival probabilities at event times across all instances in the dataset, *i.e.*, $\{\hat{S}_E(e_i \mid \boldsymbol{x}_i)\}_{i=1}^N$, follow a standard uniform distribution $\mathcal{U}_{[0,1]}$. A model is then considered D-calibrated if the proportion of predicted survival probabilities, $\{\widehat{S}_E(e_i \mid \boldsymbol{x}_i)\}_{i=1}^N$, that fall within any percentile interval, $[\rho_1, \rho_2]$, approximately matches the width of that interval, $\rho_2 - \rho_1$. To support censored instances, Haider et al. [27] proposed to "blur" each censored instance uniformly to the subsequent probability intervals after its predicted survival probability at observed time $\hat{S}_E(c_i \mid \boldsymbol{x}_i)$. Figure 2.6 (i) demonstrates such assessment using two percentile intervals, $[0, 0.5)$ and $[0.5, 1]$, where each interval contains two predictions (diamonds),

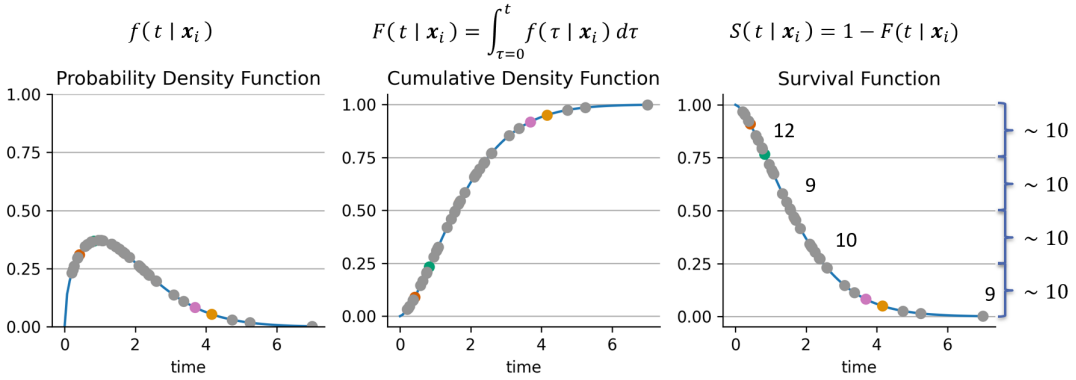


FIGURE 2.9: Illustration of distribution calibration, where an oracle draws N samples from the true conditional survival function. Figure taken with permission from [81].

indicating that the model is D-calibrated. The calibration level can be visualized using a histogram, a percentile-percentile (P-P) plot or via a chi-squared test.

We emphasize that discrimination and calibration are distinct aspects of model evaluation. In Figure 2.10, we present the predicted survival curves for four different patients (A, B, C, and D) generated by three separate models: $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. We see that $\mathcal{M}_1$ perfectly ranks the patients by their median predicted time (CI = 1), but achieves poor calibration, as one of the probability intervals contains all the patients (p-value = 0). In contrast, $\mathcal{M}_2$ fails to correctly rank patients as it gives the reverse order of predicted event times (CI = 0), but its survival probabilities are perfectly calibrated, as the predicted probabilities at event times are evenly distributed over the probability intervals (p-value > 0.05). Finally, $\mathcal{M}_3$ is able to perfectly rank the patients (CI = 1) and give calibrated survival curves (p-value > 0.05).

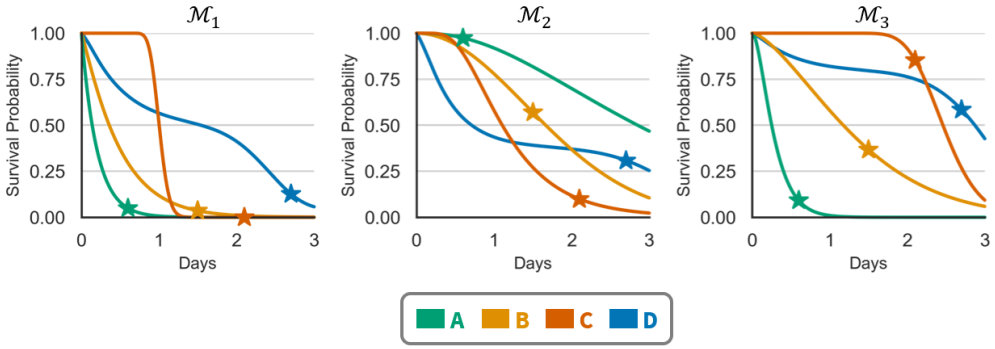


FIGURE 2.10: Predicted survival curves for four different patients generated by three separate models and their true event times (stars). In this example, $\mathcal{M}_1$ achieves *perfect* discrimination and *poor* calibration, $\mathcal{M}_2$ achieves *poor* discrimination and *perfect* calibration, and $\mathcal{M}_3$ achieves *perfect* discrimination and *perfect* calibration. Figure taken with permission from [81].

Lastly, we want to briefly mention *coverage calibration*, as introduced by Qi et al. [82]. Coverage calibration measures the agreement between the upper and lower bounds of predicted so-called credible intervals (CrIs) and the observed probability intervals. CrI is a concept from Bayesian learning, and is often described as the Bayesian analogue to a confidence interval. We will examine this concept in detail

in Section 2.6, however, they represent the range within which a parameter or outcome lies with a given posterior probability. For example, a 90% CrI should cover an individual's likelihood of an event with a probability of 90%. In practice, we calculate the observed coverage rates using CrIs with different percentages (ranging from 10% to 90%) and then use a chi-squared test to evaluate the calibration of the expected and observed coverage rates. To produce CrIs, we need a probabilistic survival model – that is, one that models the underlying uncertainty in its predictions. We will introduce such model in Chapter 3.

2.5 Competing Risks and Multiple Events

Up to this point, our analysis has focused on a single event of interest. However, this framework can be extended to accommodate multiple possible events, such as death due to cancer or death due to cardiovascular disease. This extension includes *competing risks*, where the occurrence of one event precludes the others (*e.g.*, a patient can die from either cancer or cardiovascular disease, but not both), as well as *multiple events* (multi-event) scenarios, where experiencing one event does not rule out the occurrence of others. Examples of multi-event scenarios include various medical emergencies in hospitalized patients, such as ARF, shock [83], and adverse outcomes after a bone marrow transplantation, such as relapse or death [84]. In competing risks, the goal is to model the probability and timing of the first event to occur for each individual, while in multi-event scenarios, the objective is to predict the sequence and timing of multiple events, from earliest to latest[6].

As a practical example, consider a clinical trial spanning $t \in [0, 100]$ days (see Figure 2.11), where a patient i, characterized by features $\boldsymbol{x}^{(i)}$, may experience four side effects: nausea (A), fatigue (B), fever (C) or headache (D).

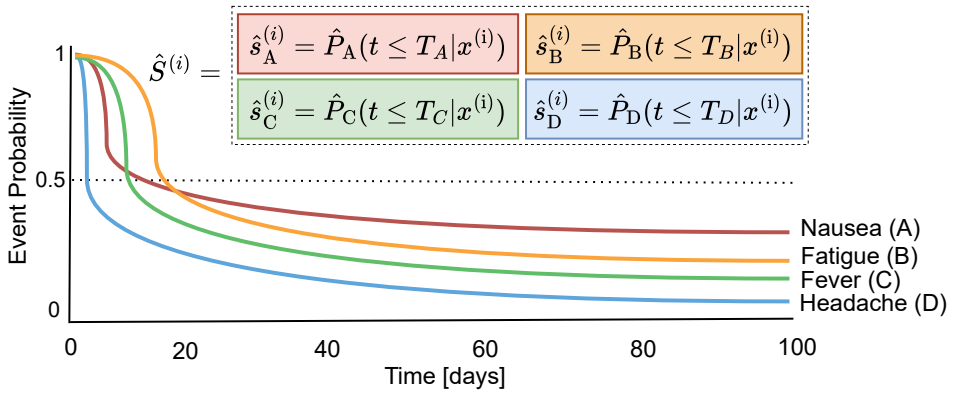


FIGURE 2.11: Outline of a multi-event survival prediction with four events. Note that we refer to $S(t)$ as the event probability instead of the survival probability in this case, but this is for pedagogical reasons, as the events we are predicting do not relate to mortality. Figure taken with permission from [48].

[6]In our terminology, we refer to *competing-risks* problems as those in which only one event can be observed per instance, and the event time is effectively a scalar: $t^{(i)}$. In contrast, *multi-event* problems involve multiple events per instance, and the event time is represented as a vector: $\boldsymbol{t}^{(i)}$. Cases involving *semi-competing risks* – where one event precludes another but not vice versa – fall under our definition of multi-event problems, as they require modeling multiple event times.

As discussed, the goal of survival prediction is to estimate the probability that each of these events occurs after a given time t – that is, the survival probability – within a finite time window, such as from 0 to 100 days. This can be formulated as a multi-event survival task, where a single model is trained to predict an individual survival distribution for each event, denoted $\hat{S}^{(i)}$. These distributions represent the probability that each of the four side effects occurs after t days from the baseline. Like in single-event survival modeling, we can extract useful prediction targets, such as the median predicted event time from each distribution. We can frame a multi-event problem using the multi-state framework proposed by Andersen and Keiding [32]. Here, we first consider the simplest case of a single event of interest and then extend the framework to competing risks and multiple events.

2.5.1 Single Event

Survival analysis typically models the time until a single event takes place [85, Ch. 11]. We can represent such problems using a two-state model (see Figure 2.12).

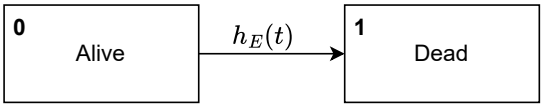


FIGURE 2.12: A single-event model with a single transition. Figure taken with permission from [48].

In the simplest case, there is one transient state (0, *e.g.*, "alive") and one absorbing state (1, *e.g.*, "dead"). Once an instance enters the absorbing state, no further transitions are possible. Recall that for a given instance, the observation can be represented by a random variable E, denoting the time from a chosen origin ($t = 0$) to the occurrence of the event. The distribution of E can be described by the cumulative distribution function $F_E(t) = \Pr(E \leq t)$ or, equivalently, by the survival function $S_E(t) = 1 - F_E(t) = \Pr(E > t)$. Here, $S_E(t)$ and $F_E(t)$ correspond to the probabilities of being in state 0 or state 1, respectively, at time t. In continuous time, E is often further characterized by the *hazard rate function*:

$$h_E(t) = -d \log S_E(t)/dt = \lim_{\Delta t \to 0} \frac{\Pr(t < E \leq t + \Delta t | E > t)}{\Delta t}, \tag{2.45}$$

where $S_E(t)$ is the *survival function*:

$$S_E(t) = \exp\left(-\int_0^t h_E(u)\, du\right). \tag{2.46}$$

Here, $h_E(t)$ denotes the transition intensity from state 0 to 1, *i.e.*, the instantaneous risk of moving from state 0 to state 1 at time t. This framework can be generalized to multiple state transitions using *marked point processes* [32]. Let $G(t)$, $t \in \mathbb{T}$ with $\mathbb{T} = [0, t_{\max}]$, denote a stochastic process with a finite state space $\mathcal{P} = \{0, \ldots, P-1\}$ and right-continuous sample paths, $G(t^+) = G(t)$. The process has an initial distribution $\pi_p(0) = [\Pr(G(0) = p) \mid p \in \mathcal{P}]$. A multi-state process $G(\cdot)$ generates a history $\mathcal{G}_t$, consisting of observations over $[0, t_{\max}]$ – for example, a patient may be healthy at study start ($\mathcal{G}_0$), become ill 10 days later ($\mathcal{G}_{10}$), and then the study ends after 100 days ($t_{\max} = 100$). Given this history, the transition probabilities between two states are defined as

$$\Pr_{p \to q}(s, t) = \Pr(G(t) = q \mid G(s) = p, \mathcal{G}_{s^-}), \tag{2.47}$$

for $p, q \in \mathcal{P}$, $s, t \in \mathbb{T}$ with $s \leq t$. The corresponding transition intensities, when they exist, are given by

$$h_{p \to q}(t) = \lim_{\Delta t \to 0} \frac{\Pr_{p \to q}(t, t + \Delta t)}{\Delta t}. \tag{2.48}$$

Some transitions may be impossible, *i.e.*, $h_{p \to q}(t) = 0$ for all t and given $p, q \in \mathcal{P}$. A state p is *absorbing* if $h_{p \to q}(t) = 0$ for all $t \in \mathbb{T}$ and all $q \neq p$; otherwise, p is *transient*. The state probabilities $\pi_p(t) = \Pr(G(t) = p)$ are

$$\pi_p(t) = \sum_{q \in \mathcal{P}} \pi_q(0) \Pr_{q \to p}(0, t), \tag{2.49}$$

where both $\Pr_{p \to q}(\cdot, \cdot)$ and $h_{p \to q}(\cdot)$ depend on the probability measure $\Pr(\cdot)$ and on the process history, though our notation omits this information.

2.5.2 Competing Risks

In a competing-risks setting, multiple events (*i.e.*, absorbing states) are possible, but only one event (*i.e.*, state transition) can occur for each instance (see Figure 2.13).

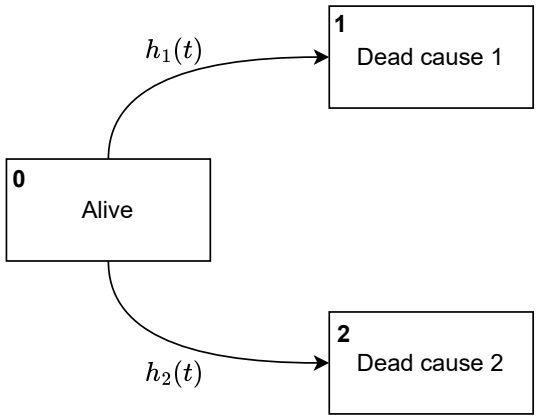


FIGURE 2.13: A competing-risks model for mortality from two causes.

A competing-risks model consists of one transient state (state 0) and $P - 1$ absorbing states, each corresponding to the occurrence of event k for $k \in \{1, \ldots, K\}$, with $K = P - 1$. If an instance is censored before experiencing any event, it remains in the transient state and is therefore event-free. Let T_p denote the (latent) time of transition to state p, let $E := \min_p T_p$ denote the time of the first event, *i.e.*, the time of leaving state 0, and let $J \in \{1, \ldots, P - 1\}$ denote the type of that event. Using our state notation, the transition intensities $h_p(t)$ for $p \in \{1, \ldots, P - 1\}$ correspond to the cause-specific hazard functions:

$$h_p(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq E < t + \Delta t, J = p \mid E \geq t)}{\Delta t}, \tag{2.50}$$

i.e., the instantaneous rate of experiencing event p at time t, given that no event of any type has occurred before t. In this model, all edges originate from state 0, the sole transient state, *i.e.*, $h_{p \to q}(t) = 0$ for all $p \neq 0$ and all q. The probability of remaining

in state 0 is given by the survival function

$$\Pr_{0\to 0}(0,t) = S_0(t) = \Pr(E > t) = \exp\left(-\int_0^t \sum_{p=1}^{P-1} h_p(u)\,du\right), \tag{2.51}$$

and the cumulative incidence function (CIF) for transitioning from state 0 to any absorbing state $p \neq 0$ is

$$\Pr_{0\to p}(0,t) = \Pr(E \leq t,\, J = p) = \int_0^t S_0(u^-)\, h_p(u)\,du, \quad p = 1,\ldots,P-1. \tag{2.52}$$

Concerning censoring in competing risks, we can see this as a form of dependent censoring [14, Ch. 1], as one risk (*e.g.*, event) dependently censors the other, and these risks are assumed to have some prior relationship. However, competing risks suffer from the so-called identifiability issue [63]: since events are mutually exclusive, the occurrence of the first event censors the others, so the joint distribution of the latent event times – and hence their dependence structure – cannot be identified from the observed data. Quantities defined in terms of the first event, such as the cause-specific hazards and the CIF above, remain identifiable without any assumption about this dependence. Given these challenges, multiple models have been proposed for competing-risks scenarios that built on the traditional likelihood-based approach we presented in Section 2.3 using, for example, the exponential (Example 1), Weibull (Example 2) or some other probability distribution to model the survival time. Here, we will give a brief description of these models, but refer to the corresponding works for more detail. Also, as was the case in the CoxPH and Weibull time-to-event models, these models assume that censoring is independent of the (competing) events given the features (Definition 6). Note that this is distinct from assuming independence *between* the competing events themselves, which models such as DeepHit do not require, since they model the observable time and type of the first event directly.

Lee et al. [86] proposed DeepHit, a discrete-time[7] model for competing risks, where the occurrence of one event precludes the occurrence of others. It assumes $\sum_{t=1}^{t_{\max}} \sum_{k=1}^{K} p_k[t] = 1$, where $t_{\max} \in \mathbb{N}_+$ is the number of discrete time bins (the maximum discretized time index) in the prediction horizon, and $K \in \mathbb{N}_+$ is the number of competing event types (causes). DeepHit employs a shared covariate layer across all events and cause-specific subnetworks for each competing event. The model outputs a probability distribution over each discrete time bin and each competing event via a softmax layer, enabling direct estimation of the joint distribution of survival times and events without assumptions about the underlying stochastic process. While it shares parameters across events through the common representation, the predicted probabilities sum to one over all time bins and events, so the model assumes every instance experiences some event within the prediction horizon and does not explicitly model individuals who remain event-free by $t_{\max}$.

Kim et al. [87] proposed MTLR-CR, an extension of the discrete Multi-Task Logistic Regression (MTLR) framework [88] tailored for competing-risks scenarios. MTLR is a discrete-time survival model that predicts the survival distribution through a sequence of dependent logistic regressors. By discretizing the time interval from $t = 0$ to $t_{\max}$ into B bins, MTLR trains B regressors to estimate the event probability at each

[7] In discrete-time survival analysis, follow-up time is divided into a finite set of intervals or time points, so the model predicts the probability of an event occurring in each interval rather than modeling event time as a continuous variable.

time bin t_B. In competing risks, MTLR-CR learns a sequences of logistic regressors for both the number of discrete bins B and the number of events K.

Nagpal et al. [89] introduced Deep Survival Machines (DSM), a parametric survival model that utilizes a mixture of predefined primitive distributions, such as mixtures of Weibull distributions. DSM first learns a shared feature representation across all competing risks, then estimates a distribution for each event via cause-specific maximum likelihood, treating other events as censored. However, beyond the shared representation, each cause-specific distribution is parameterized separately, so the model does not jointly model the events or share information between them beyond the common features.

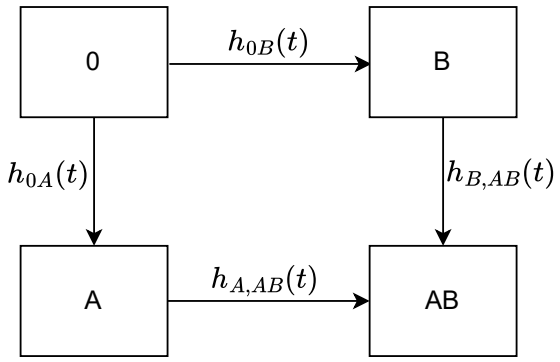


FIGURE 2.14: A multi-event model describing the joint behavior of A and B.

2.5.3 Multiple Events

In settings with multiple events, each instance i may experience multiple transitions, which we call events, and these are not necessarily mutually exclusive (see Figure 2.14). Such a model can, for instance, describe the joint behavior of two events A and B. If $h_{0B} = h_{A,AB}$ but $h_{0A} \neq h_{B,AB}$, then A is said to be *locally dependent* on B, whereas B is not locally dependent on A. The temporal ordering of events allows for an asymmetric notion of dependence, offering more information for causal inference than the symmetric association measures typically used in cross-sectional studies.

Multi-event analysis goes back as early as 1951, when Fix and Neyman [35] introduced the illness-death model, which has been used in multiple works since [90]. Unlike competing-risks, multi-event analysis does not necessarily suffer from the identifiability issue [63], as the events are not necessarily mutually exclusive. However, predicting the outcome of several (possibly related) events instead of just one poses another challenge: when events are in fact related, naive survival models that ignore the dependency structure between them can yield biased estimates. This may result in over- or underestimating event risks, misinterpreting temporal ordering, and reducing predictive performance, particularly when events share common risk factors or when the occurrence of one event modifies the hazard of others. Larson [91] adopted the popular CoxPH model [23] for multi-event survival analysis by modeling each event separately – an idea that has gained popularity since then [33, 34], but which fails to capture any dependencies between the events. As we are focused mainly on deep learning methods for survival analysis, existing approaches for multi-event survival analysis have generally been limited.

Most notably, Tjandra et al. [83] proposed a hierarchical multi-event model that learns the probability of event occurrence at different temporal resolutions. In practice, the model uses coarser predictions (*e.g.*, on a monthly basis) to iteratively guide predictions at finer resolutions (*e.g.*, weekly or daily). Given some complex prediction task over a long time horizon, the model decomposes this task into a sequence of simpler tasks that serve as intermediate guides. Thus, the hierarchical framework progressively predicts survival across multiple time granularities, starting with broader intervals (*e.g.*, months) and refining down to finer ones (*e.g.*, weeks). Formally, their model is a multi-task network that takes $\boldsymbol{x}$ as input and outputs $\hat{P}$, a set of probability distributions for each k event. Event-specific subnetworks, ϕ_k, estimate the probability distribution $\hat{p}_k$ hierarchically. At granularity m, the probability distribution $\hat{p}_{km}$ is constructed from $\psi_{km}(\Phi(\boldsymbol{x}))$ and the previous granularity level $\hat{p}_{k,m-1}$, ensuring that predictions at finer time scales are guided by coarser ones. There are two primary limitations to this approach. First, the hierarchical component scales poorly with larger time durations and has a significant parametric overhead. Second, the method is optimized solely for discriminative loss, which is only one facet of model performance in survival analysis (see the taxonomy in 2.6). As shown by Qi et al. [38], strong ranking performance does not necessarily translate into accurate time-to-event predictions, and vice versa. High discriminative ability (*e.g.*, a high concordance index) indicates that the model is good at ordering instances by relative risk, but it does not guarantee that the predicted survival times or event probabilities are calibrated or even accurate at all.

Before ending our introduction of multi-event models, we should also discuss how to evaluate such models. For single-event problems, as described in Section 2.4, we can evaluate the model's discrimination performance with regards to a single event using, for example, Harrell's concordance index [36]. However, in competing-risks and multi-event scenarios, now that we have multiple events of interest, our scope of evaluation also expands: we want to measure whether our model can (1) predict an accurate ordering of instances *within* an event, and (2) predict an accurate ordering of events *within* an instance. For this task, respectively, we can formulate a global C-index as the average of the C-indices (*e.g.*, Harrell's) across events, and the local C-index, as the average of the C-indices across individuals. Figure 2.15 shows an example of this with respect to the survival curve. Visually, global consistency is satisfied between instances if the instance-specific curves align with the observed events. Local consistency is satisfied within an instance if the event-specific curves align with the observed events.

Definition 15 *The global C-index is defined as [48, 83]:*

$$\mathcal{C}_{Global}(\hat{\eta}(\cdot); \mathcal{D}) = \frac{\sum_{k \in \mathcal{K}} \sum_{i,j \in \mathcal{D}_k} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \mathbb{1}[\hat{\eta}^{(i)} > \hat{\eta}^{(j)}] \cdot \delta^{(i)}}{\sum_{k \in \mathcal{K}} \sum_{i,j \in \mathcal{D}_k} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \delta^{(i)}}, \tag{2.53}$$

where $\hat{\eta}^{(i)} := \hat{\eta}(\boldsymbol{x}^{(i)})$ denotes the predicted risk score for the i-th instance, $\mathcal{K}$ is the set of all K distinct events with k corresponding to the k-th event, and $\mathbb{1}[\cdot]$ is the indicator function.

Definition 16 *The local C-index is defined as [48, 83]:*

$$\mathcal{C}_{Local}(\hat{\eta}_{k_1}(\cdot), \hat{\eta}_{k_2}(\cdot); \mathcal{D}) = \frac{\sum_{i \in N} \sum_{k_1, k_2 \in \mathcal{K}^{(i)}} \mathbb{1}[t_{k_1} < t_{k_2}] \cdot \mathbb{1}[\hat{\eta}_{k_1} > \hat{\eta}_{k_2}] \cdot \delta_{k_1}}{\sum_{i \in N} \sum_{k_1, k_2 \in \mathcal{K}^{(i)}} \mathbb{1}[t_{k_1} < t_{k_2}] \cdot \delta_{k_1}}, \tag{2.54}$$

where $\hat{\eta}_{k_1} := \hat{\eta}_{k_1}(\boldsymbol{x}^{(i)})$ *denotes the predicted risk score for event* k_1 *of instance* i, $\hat{\eta}_{k_2} := \hat{\eta}_{k_2}(\boldsymbol{x}^{(i)})$ *is the predicted risk score for event* k_2 *of the same instance,* $\mathcal{K}^{(i)}$ *is the set of all K distinct events, and* $\mathbb{1}[\cdot]$ *is the indicator function.*

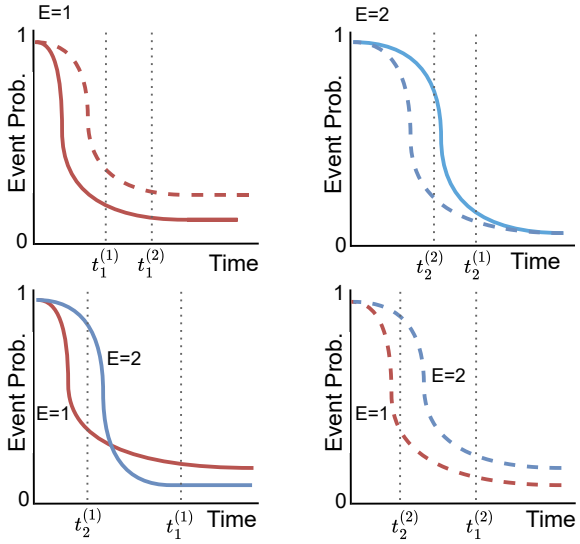


FIGURE 2.15: Global and local consistency. Red/blue lines represent survival functions $S_1(t)$ and $S_2(t)$ for event 1 and 2, respectively. Solid/dashed lines represent individuals $i = 1, 2$, respectively. Top-left: globally consistent at $t_1^{(1)}$ and $t_1^{(2)}$ since $S_1^{(1)}(t) < S_2^{(2)}(t)$ both times. Top-right: globally consistent at $t_2^{(2)}$ and $t_2^{(1)}$ since $S_2^{(2)}(t) < S_2^{(1)}(t)$ both times. Bottom-left: locally inconsistent at $t_2^{(1)}$ since $S_2^{(1)}(t) > S_1^{(1)}(t)$ and at $t_1^{(1)}$ since $S_1^{(1)}(t) > S_2^{(1)}(t)$. Bottom-right: locally inconsistent at $t_2^{(2)}$, since $S_2^{(2)}(t) > S_1^{(2)}(t)$. Figure taken with permission from [48].

2.6 Uncertainty Estimation

So far we have looked at models that rely on traditional maximum likelihood estimation (MLE) to estimate θ – that is, we find a single point estimate $\hat{\theta}$ that maximizes the likelihood of the observed data. After training the model, we compute $y_i = f_{\hat{\theta}}(\boldsymbol{x}_i)$, which is a prediction (for example, a risk score or survival probability) for an instance i given its description. However, standard deep learning training yields point estimates of the network weights without any measure of parameter uncertainty. Consequently, even when the model outputs a full survival distribution, it does not express how confident the model is in that prediction.

Consider an example of predicting relapse in cancer patients: Given a set of instances describing cancer patients, we train a (hypothetical) model to predict their

risk of relapse after surgery. When a medical doctor queries our model with the description of an early-stage patient, the model should return a prediction with rather high confidence. But what happens if the model has only been trained on late-stage cancer patients who underwent surgery, rather than on early-stage patients? Late-stage patients may differ from early-stage ones – for instance, in their biomarker profiles – and although the model has learned to recognize patterns in late-stage patients, it has never encountered an early-stage patient before. The patient would therefore fall outside of the data distribution in which the model was trained. This is an example of *out-of-distribution* test data. In such cases, ideally, we want the model to return a prediction (attempting to extrapolate far away from the observed data), but return an answer with the extra information that this instance lies outside of the data distribution. That is, the model should tell us the underlying *uncertainty* with respect to its prediction. Apart from out-of-distribution cases, several other factors can also contribute to uncertainty:

- Noisy data (the observed labels might be noisy, for example, as a result of measurement imprecision, leading to aleatoric uncertainty).
- Uncertainty in model parameters that best explain the observed data (many possible models might be able to explain a given dataset, in which case we might be uncertain which model parameters to choose for prediction).
- Uncertainty in the model architecture (some architectures may extrapolate or interpolate better than other architectures).

The latter two uncertainties can be grouped under *model uncertainty* (also referred to as *epistemic uncertainty*). The aleatoric uncertainty and epistemic uncertainty can then be used to induce *predictive uncertainty*, which represents the confidence that we have in a prediction. The healthcare domain and medical diagnosis are particularly important areas for uncertainty estimation due to the severe harm caused by predictive models that make confident but wrong predictions, *e.g.*, when our early-stage cancer patient receives a poor (unfair) prognosis because the model has only seen patients with late-stage cancers. As a general requirement, we want a model's predictive uncertainty to be aligned with the accuracy of its predictions, so that predictive uncertainty can be a fail-safe against incorrect predictions. If a given instance might result in an incorrect prediction because it is meaningfully different from the data in the training set, the model should express a high level of predictive uncertainty and flag the instance for further review by a medical expert.

Figure 2.16 shows a possible diagnosis pipeline where our (hypothetical) model explicitly reasons about its uncertainty. In this prediction task, we evaluate the model's predictive performance across different expert referral rates. For a given referral rate $\tau \in [0, 1]$, the model uses its predictive uncertainty to select the τ fraction of instances in the evaluation set whose predictions are most uncertain. We then refer these instances to a domain expert (*e.g.*, a medical professional) for further review, while assessing the model's predictions on the remaining $(1 - \tau)$ fraction of instances. By repeating this process over all referral rates and evaluating the model's performance on the retained instances, we estimate its reliability in a safety-critical downstream task, where we leverage predictive uncertainty alongside human expertise to reduce harmful predictions.

Next, we present two frameworks for uncertainty estimation: the Bayesian framework, which quantifies uncertainty from the model, parameters, and consequently the predictions; and conformal prediction, which provides calibrated prediction sets with guaranteed coverage levels.

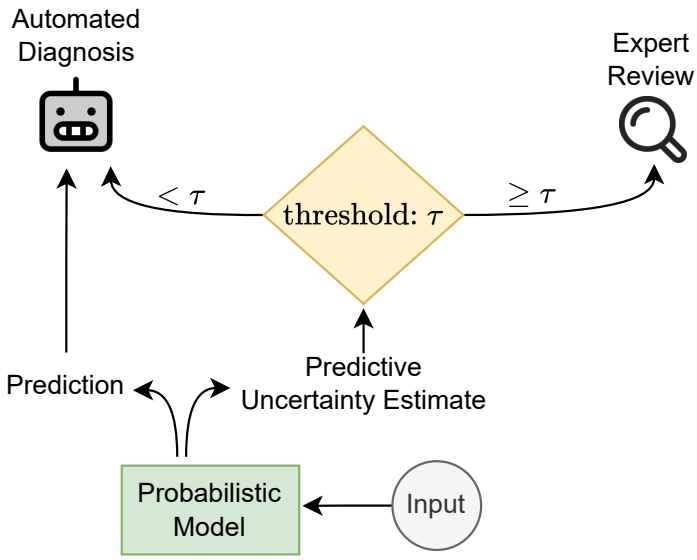


FIGURE 2.16: Automated diagnosis pipeline: For a given input, a model provides a prediction and a corresponding uncertainty estimate; if the uncertainty estimate is below a certain reference threshold τ (indicating a low degree of uncertainty) the diagnosis is processed without further review; otherwise, it is referred to a medical expert. Figure recreated from [92].

2.6.1 Bayesian Learning

Practitioners widely use a Bayesian approach to fit probabilistic model parameters and estimate their uncertainty [43]. This approach incorporates the epistemic uncertainty, which describes the uncertainty in model parameters. We will introduce Bayesian learning from a machine learning perspective, that is, the framework need not depend on a specific task. Later, we will show how it can be applied to survival analysis. One of the most famous formulae of Bayesian statistics is called Bayes' theorem.

Definition 17 *Bayes' theorem says that [43]:*

$$Pr(\theta \mid \mathcal{D}) = \frac{Pr(\mathcal{D} \mid \theta) Pr(\theta)}{Pr(\mathcal{D})} \tag{2.55}$$

where $Pr(\theta \mid \mathcal{D})$ is the posterior, i.e., *the probability of a certain value of a parameter θ given the data $\mathcal{D}$, $Pr(\mathcal{D} \mid \theta)$ is the likelihood, also called the inverse, $Pr(\theta)$ is the prior and $Pr(\mathcal{D})$ is the marginal likelihood or evidence.*

Note that in Bayesian learning, the parameter θ is not treated as a fixed value as in earlier sections; instead, it is associated with uncertainty described by a probability distribution. This distribution captures the likelihood of various possible values of θ, with the assigned probability representing the confidence in each value. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, Bayes' theorem allows us to update our beliefs and obtain the posterior distribution of the model parameter θ:

$$\Pr(\theta \mid \mathcal{D}) = \frac{\Pr(\mathcal{D} \mid \theta)\Pr(\theta)}{\Pr(\mathcal{D})} = \frac{\prod_{i=1}^{N} \Pr(y_i \mid \boldsymbol{x}_i, \theta)\Pr(\theta)}{\Pr(\mathcal{D})} \tag{2.56}$$

In non-Bayesian model fitting, we would instead use the training data $\mathcal{D}$ to find a single point estimate θ_{MLE} and predict with it:

$$\Pr(y \mid \boldsymbol{x}_{\text{test}}, \mathcal{D}) \approx \Pr(y \mid \boldsymbol{x}_{\text{test}}, \theta_{\text{MLE}}) \tag{2.57}$$

But with the Bayesian approach, we now get a whole distribution. Once we have learned the posterior distribution of θ, we can determine the predictive distribution:

$$\Pr(y \mid \boldsymbol{x}_{\text{test}}, \mathcal{D}) = \int \Pr(y \mid \boldsymbol{x}_{\text{test}}, \theta)\Pr(\theta \mid \mathcal{D})\, d\theta \tag{2.58}$$

Let us compare these two approaches. In Equation 2.57, we picked a distribution for the conditional probability distribution (CPD) with parameter θ and then used the data to determine θ_{MLE}. Given θ_{MLE}, all possible outcomes y follow the CPD $\Pr(y \mid \boldsymbol{x}_{\text{test}}, \theta_{\text{MLE}})$, so we can forget about the data after we compute θ_{MLE}. With the Bayesian approach, however, we do not work with a CPD that corresponds to a single (optimized) parameter value, but with many CPDs with weight $\Pr(\theta_i|\mathcal{D})$, where θ_i is a sample from the posterior distribution over model parameters given the data $\mathcal{D}$. Instead of making predictions based on a single model, we marginalize over the entire posterior distribution by integrating Equation 2.58. This is also referred to as marginalizing the likelihood over θ. If the model is simple, such as Bayesian linear regression, marginalization can be done analytically, but with more complex models, it becomes intractable. In such cases, an approximation is needed.

A popular choice for approximating the true posterior $\Pr(\theta|\mathcal{D})$ is by finding an approximate *variational* distribution $q_\lambda(\theta)$, parametrized by λ, whose structure is easy to evaluate. This is known as variational inference (VI). VI seeks to find the variational parameter λ so that $q_\lambda(\theta)$ gets as close as possible to the true posterior $\Pr(\theta|\mathcal{D})$. This is done by minimizing the so-called Kullback–Leibler (KL) divergence [93] with respect to λ, which intuitively measures the similarity between two distributions:

$$\text{KL}(q_\lambda(\theta) \parallel \Pr(\theta \mid \mathcal{D})) = \int q_\lambda(\theta) \log \frac{q_\lambda(\theta)}{\Pr(\theta \mid \mathcal{D})} d\theta. \tag{2.59}$$

However, minimizing $\text{KL}(q_\lambda(\theta) \parallel \Pr(\theta|\mathcal{D}))$ requires knowledge of the posterior $\Pr(\theta|\mathcal{D})$, which is typically intractable because it requires integrating over all possible parameter values $\Pr(\theta)$, which can be high-dimensional and complex. Still, we can find the optimal approximation by maximizing the so-called evidence lower bound (ELBO) with respect to the variational parameters defining $q_\lambda(\theta)$:

$$\begin{aligned}\mathcal{L}_{\text{VI}}(\lambda) := &\int q_\lambda(\theta) \log \Pr(\mathcal{D}, \theta) d\theta \\ &- \text{KL}(q_\lambda(\theta) \parallel \Pr(\theta)) \leq \log \Pr(\mathcal{D}) = \log \text{evidence}.\end{aligned} \tag{2.60}$$

This approach is computationally simpler because it relies on $\Pr(\theta, \mathcal{D})$ rather than the conditional distribution $\Pr(\theta \mid \mathcal{D})$. Here, $\mathcal{L}_{\text{VI}}(\lambda)$ is a lower bound of $\log \Pr(\mathcal{D})$ (*i.e.*, the log-evidence) because $\text{KL}(q_\lambda(\theta) \parallel \Pr(\theta \mid \mathcal{D})) \geq 0$. Maximizing the first term in Equation 2.60 (referred to as the expected log-likelihood) encourages $q_\lambda(\theta)$ to explain the data well, while minimizing the second term (referred to as the prior KL) encourages $q_\lambda(\theta)$ to be as close as possible to the prior. Let $q^*_\lambda(\theta)$ denote

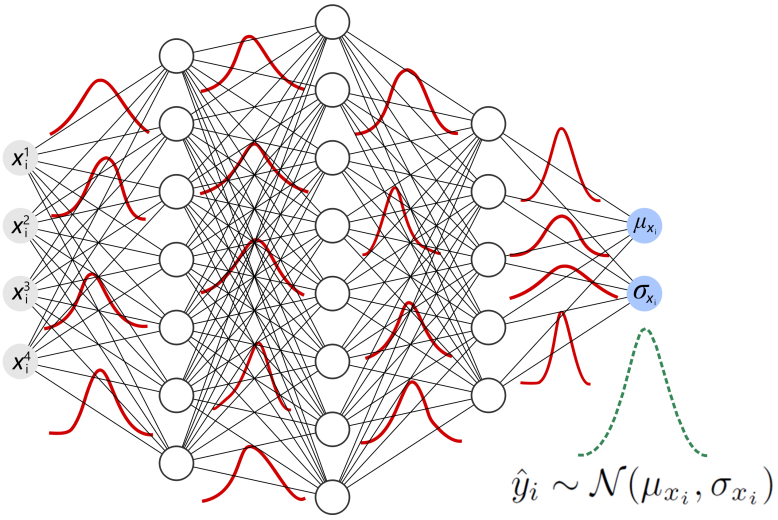


FIGURE 2.17: A BNN with three hidden layers. For a d-dimensional sample $\boldsymbol{x}_i$, the two output nodes provide the mean and variance of a Gaussian over the risk score, $f(\boldsymbol{x}_i; \theta) \sim \mathcal{N}(\mu_{x_i}, \sigma^2_{x_i})$. Figure taken with permission from [46]. ©2025 IEEE.

the minimum of the optimization objective (often a local minimum), then minimizing the KL divergence (or equivalently maximizing the ELBO $\mathcal{L}_{\text{VI}}(\lambda)$) allows us to approximate the predictive distribution as:

$$\Pr(y \mid \boldsymbol{x}_{\text{test}}, \mathcal{D}) \approx \int \Pr(y \mid \boldsymbol{x}_{\text{test}}, \theta) q^*_\lambda(\theta) d\theta =: q^*_\lambda(y \mid \boldsymbol{x}_{\text{test}}). \tag{2.61}$$

VI replaces the Bayesian modeling marginalization with optimization, *i.e.*, we replace the calculation of integrals with that of derivatives. But compared to the maximum likelihood approach often used in deep learning, in this setting, we optimize over distributions instead of point estimates. This gives us many of the advantages of Bayesian learning and results in probabilistic models that capture model uncertainty. Bayesian learning and modern neural networks have merged into what is called Bayesian Neural Networks (BNNs, or Bayesian NNs), which offer a probabilistic interpretation of deep learning models by inferring distributions over the models' weights. BNNs place a prior distribution over a neural network's weights, which induces a distribution over a parametric set of functions. Figure 2.17 is an illustration of a BNN, where sampling from the weight posteriors captures the epistemic uncertainty in the model, and sampling the output from a Gaussian captures the aleatoric uncertainty.

Letting $\theta = (\boldsymbol{W}, \boldsymbol{b})$ denote the weight matrices and bias vectors, we often place standard matrix Gaussian prior distributions over the weight matrices $\Pr(\boldsymbol{W}_i) = \mathcal{N}(0, \boldsymbol{I})$ for layer i, where $\boldsymbol{I}$ represents the identity matrix, and often assume a point estimate for the bias vectors for simplicity. Given $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, then variational inference in neural networks seeks to find an approximation $q_\lambda(\theta)$ to the posterior distribution over the parameters by maximizing Equation 2.60. Typically, we use the mean-field approximation, a common form of variational Bayes, where we assume that the unknown variables can be partitioned such that each partition is independent of the others: If $p = p_\Theta$ and $q = q_\Theta$ are distributions over parameters, and $q \in Q$ is the family of mean-field (*i.e.*, distributions that factorizes according to the

equation $q(\theta) = \prod q_i(\theta_i)$) Gaussian distributions, and the prior distribution over parameters p_θ is also a diagonal Gaussian, the resulting variational objective can be solved using stochastic VI and optimized using stochastic gradient methods [30, 94].

However, VI incurs a computational cost because it doubles the number of model parameters needed to capture uncertainty. Monte-Carlo Dropout (MCD) offers an alternative approach by approximating the posterior solution of VI while reducing the computational overhead. Gal and Ghahramani [26] formally established the connection between MCD and Bayesian Neural Networks (BNNs). In the following, we focus on the key advance relevant to the noncategorical case for the outcomes $\boldsymbol{y}$. Consider a deep Gaussian Process (GP) with a covariance function defined as:

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \int \Pr(\boldsymbol{w})\Pr(b)\sigma\big(\boldsymbol{w}'\boldsymbol{x} + b\big)\sigma\big(\boldsymbol{w}'\boldsymbol{y} + b\big)\mathrm{d}\boldsymbol{w}\mathrm{d}b, \tag{2.62}$$

where $\boldsymbol{W}_i$ is the weight matrix of dimension $K_i \times K_{i-1}$ for each layer $i = 1, \ldots, L$, and $\boldsymbol{b}_i$ is the corresponding bias vector. We denote $\boldsymbol{\omega} = \{\boldsymbol{W}_i\}_{i=1}^{L}$, where each row of $\boldsymbol{W}_i$ is distributed according to a standard multivariate normal distribution $\Pr(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, l^{-2}\boldsymbol{I})$. Let $\sigma(\cdot)$ be a general nonlinear activation function and $\Pr(b)$ some distribution on the elements of $\boldsymbol{b}_i$, and given some precision parameter $\tau > 0$, the predictive distribution of the deep Gaussian process is

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \int \Pr(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega})\Pr(\boldsymbol{\omega}|\mathcal{D})\mathrm{d}\boldsymbol{\omega}, \tag{2.63}$$

with

$$\Pr(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega}) = \mathcal{N}\Big(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{\omega}), \tau^{-1}\boldsymbol{I}\Big). \tag{2.64}$$

For the intractable posterior $\Pr(\boldsymbol{\omega}|\mathcal{D})$, we adopt a variational approximation $q(\boldsymbol{\omega})$ defined as:

$$\boldsymbol{W}_i = \boldsymbol{M} \cdot \mathrm{diag}\Big(\big[z_{i,j}\big]_{j=1}^{K_i}\Big), \tag{2.65}$$

$$z_{i,j} \sim \text{Bernoulli}(p_i),\ i = 1, \ldots, L,\ j = 1, \ldots, K_{i-1}, \tag{2.66}$$

where $\boldsymbol{M}_i$ represent the variational parameters and p_i denote specified probabilities. Gal and Ghahramani [26] demonstrate that minimizing the KL divergence between the variational posterior $q(\boldsymbol{\omega})$ and the true posterior of the deep GP corresponds to optimizing the following objective function using dropout:

$$\mathcal{L}_{\text{MCD}} = \frac{1}{N}\sum_{i=1}^{N} E(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) + \lambda \sum_{i=1}^{L}\big(||\boldsymbol{W}_i||_2^2 + ||\boldsymbol{b}_i||_2^2\big). \tag{2.67}$$

$\mathcal{L}_{\text{MCD}}$ is a standard non-probabilistic loss function that includes L_2 regularization controlled by the tuning parameter λ, along with the squared Euclidean loss $E(\cdot, \cdot)$. Dropout involves sampling a Bernoulli random variable with probability p_i for each input and each unit in every layer; a unit is dropped if the sampled variable equals zero. The dropout rate, defined as $1 - p_i$, represents the probability of dropping a unit. Applying dropout at every layer of a deep neural network (DNN) is equivalent to performing a Monte Carlo approximation of a probabilistic deep GP. In other words, dropout training minimizes the KL divergence between the variational posterior distribution and the true posterior of the deep GP, which models the uncertainty over the space of data-generating distributions. The predictive distribution

can then be approximated straightforwardly as follows:

$$\Pr(\boldsymbol{y}^\star|\boldsymbol{x}^\star) \approx \int \Pr(\boldsymbol{y}^\star|\boldsymbol{x}^\star, \boldsymbol{\omega})\, q(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}, \tag{2.68}$$

$$q(\boldsymbol{\omega}) = \text{Bernoulli}(z_1) \dots \text{Bernoulli}(z_L), \tag{2.69}$$

$$\Pr(\boldsymbol{y}^\star|\boldsymbol{x}^\star, \boldsymbol{\omega}) = \mathcal{N}\Big(\boldsymbol{y}^\star; \hat{\boldsymbol{y}}^\star(\boldsymbol{x}^\star, z_1, \dots, z_L), \tau^{-1}\boldsymbol{I}\Big), \tag{2.70}$$

and, *e.g.*, its mean through MC sampling:

$$\mathbb{E}_{\Pr(\boldsymbol{y}^\star|\boldsymbol{x}^\star)} \approx \frac{1}{S}\sum_{s=1}^{S} \hat{\boldsymbol{y}}^\star(\boldsymbol{x}^\star, \hat{z}_{1,s}, \dots, \hat{z}_{L,s}), \tag{2.71}$$

$$\hat{z}_{i,s} \sim \text{Bernoulli}(p_i). \tag{2.72}$$

Now, say we have trained a BNN by replacing the deterministic weights θ in a typical neural network by a probability distribution $\Pr(\theta)$, finding an (approximate) posterior distribution $q_\lambda(\theta)$ using *e.g.*, VI or MCD, and obtaining the posterior predictive distribution. In survival analysis, the prediction for patient i is an entire survival function $\hat{S}_E(\cdot \mid \boldsymbol{x}_i)$, *i.e.*, the ISD, so the posterior predictive is a distribution over survival curves. How do we communicate the uncertainty in this estimate? We can draw samples $\theta^{(s)} \sim q_\lambda(\theta)$, each of which yields a survival curve $\hat{S}_E^{(s)}(\cdot \mid \boldsymbol{x}_i)$. The pointwise mean of these curves gives the expected survival probability over time, and their spread quantifies the uncertainty (Figure 2.18, left). Summarizing each sampled curve by its median gives a distribution over the predicted survival time $\hat{e}_i$ (Figure 2.18, right).

To quantify the expected uncertainty, the Bayesian framework offers a tool called credible intervals (CrIs) [43, Ch. 4]. A credible interval consists of two values, called the lower and upper bounds, which enclose a certain percentage of the predictive distribution. The main difference between CrIs and their frequentist counterpart, confidence intervals (CI), is that credible intervals treat the parameter of interest as a random variable and the bounds as fixed values, while confidence intervals treat the parameter of interest as a fixed value and the bounds as random variables [43, Ch. 4]. If we, for example, compute a 95% probability bound, the Bayesian 95% CrI means there is a 95% probability that the value of the parameter (random variable) will fall within the interval given the observed data, while the frequentist 95% CI means if we take multiple samples of the input data, the computed interval will contain the true parameter (fixed value) in 95% of those repeated experiments. In this survival analysis, the CrIs can be simply interpreted as the bound containing a particular percentage of probable values, for instance, the 95% simultaneous CrIs for the ISD prediction are simply the central portions of the estimated posterior distribution that contains 95% of the curves [82].

Algorithm 1 gives a brief pseudocode example of how we can sample model weights, draw from the posterior distribution and estimate the mean, standard deviation and thus credible intervals of our predictions. Figure 2.18 shows the computed ISD and its CrIs for an individual i and also a histogram of the approximated underlying density of the predicted survival time using the median survival time. Here, the width of the credible interval/density function tell us the variance/uncertainty in the prediction – *e.g.*, a wide credible interval indicates high uncertainty.

While Bayesian learning provides distributions over model parameters and predictions to represent uncertainty, conformal prediction are concerned with providing

Algorithm 1 Estimating the uncertainty in $\hat{y}_i$ by random sampling.

Input: Trained BNN with posterior $q_\lambda(\theta)$, test input $\boldsymbol{x}_i$, number of samples S
Output: Mean prediction $\bar{y}$, standard deviation σ, and 95% credible interval CI_{95}

1: **for** $s = 1$ to S **do**
2: Sample model parameters: $\theta^{(s)} \sim q_\lambda(\theta)$
3: Sample survival prediction: $\hat{y}_i^{(s)} \sim \Pr(\hat{y}_i \mid \boldsymbol{x}_i, \theta^{(s)})$
4: **end for**
5: Compute mean prediction: $\bar{y}_i = \frac{1}{S}\sum_{s=1}^{S} \hat{y}_i^{(s)}$
6: Compute standard deviation: $\sigma_i = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{y}_i^{(s)} - \bar{y}_i)^2}$
7: Compute 95% credible interval: $\text{CrI}_{95} = [\hat{y}_{2.5\%}, \hat{y}_{97.5\%}]$
8: **return** $\bar{y}_i, \sigma_i, \text{CrI}_{95}$

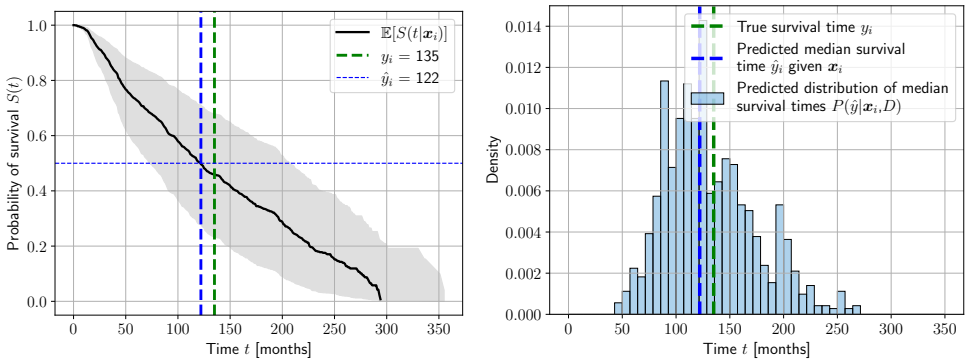


FIGURE 2.18: Individual survival distribution (ISD) for individual i. In the figure, y_i denotes the observed survival time (vertical green line), *i.e.*, t_i in our notation, and $\hat{y}_i$ denotes the median predicted survival time (vertical blue line), *i.e.*, $\hat{e}_i$. Left: mean individual survival function and its corresponding 90% CrI. Right: histogram (as an approximation of the underlying density) of the predicted median survival time. Figure taken with permission from Lillelund et al. [46]. ©2025 IEEE.

prediction intervals or sets that are calibrated to a desired confidence level. Conformal prediction is thus an alternative, nonparametric method of quantifying uncertainty. It is based on the idea of validating predictions using previously observed data, rather than relying on explicit assumptions about the underlying data distribution or model parameters. In the next section, we will give a brief introduction to this framework.

2.6.2 Conformal Prediction

Conformal prediction offers a different route to uncertainty quantification: rather than placing a distribution over model parameters, it post-processes the outputs of an already trained model so that its predictions come with a coverage guarantee [95, 96]. Instead of a single predicted label or value, the model returns a set of plausible outcomes – a set of classes in classification, or an interval in regression – constructed so that it contains the true outcome with a user-specified probability $1 - \alpha$, where $\alpha \in (0, 1)$ is the tolerated error rate (*e.g.*, $\alpha = 0.05$ for 95% coverage). This is attractive in clinical decision-making, where knowing which outcomes remain plausible for a patient is often more useful than a single point prediction. The

size of the set directly reflects the model's uncertainty: a small set indicates a confident prediction, whereas a large set signals that several outcomes remain plausible. Two properties distinguish conformal prediction from the Bayesian approach above. First, the guarantee requires no assumptions about the underlying data distribution or the model, only that the calibration and test data are exchangeable. Second, it holds for any finite calibration set rather than only asymptotically.

Formally, consider a binary classifier $\hat{f}$ that, for a patient with covariates $\boldsymbol{x}$, outputs estimated class probabilities $\hat{f}(\boldsymbol{x}) \in [0,1]^2$, *e.g.*, for the absence or presence of TMJ involvement (which we will see in Chapter 3). In split conformal prediction, we hold out a calibration set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ that is not used for training, and compute a nonconformity score for each calibration instance, for example $s_i = 1 - \hat{f}(\boldsymbol{x}_i)_{y_i}$, *i.e.*, one minus the probability the model assigns to the true class. Let $\hat{q}$ denote the empirical $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of $\{s_i\}_{i=1}^n$. For a new patient, the prediction set contains every class whose score does not exceed this threshold, $\Gamma(X_{\text{test}}) = \{y : 1 - \hat{f}(\boldsymbol{x}_{\text{test}})_y \leq \hat{q}\}$. If the calibration and test data are exchangeable (*e.g.*, i.i.d.), this set satisfies

$$1 - \alpha \leq \Pr\big(Y_{\text{test}} \in \Gamma(\boldsymbol{X}_{\text{test}})\big) \leq 1 - \alpha + \frac{1}{n+1}, \tag{2.73}$$

where the upper bound additionally requires the scores to be free of ties [96].

This guarantee is *marginal*: it holds on average over the calibration and test data, not for each patient or each class separately. In a binary clinical task, a prediction set therefore typically contains either one class, which is a confident prediction, or both classes, which flags an uncertain case that can be referred to a clinician, as in the pipeline in Figure 2.16; with some scores, the set can also be empty.

Chapter 3

Uncertainty in Machine Learning and Survival Analysis

This chapter explains the details of one method to train probabilistic neural networks for survival analysis, that can quantify the uncertainty in its predictions, and one method to predict involvement of the temporomandibular joint (TMJ) in children with juvenile idiopathic arthritis (JIA). In the first method, we take the Bayesian approach to uncertainty estimation, and in the second method, we use conformal prediction to provide calibrated predictive sets.

3.1 Uncertainty Estimation in Survival Models

3.1.1 Introduction

Variational Inference (VI) is a widely-used method for approximate Bayesian inference and uncertainty estimation in deep learning, but it incurs a significant computational overhead by doubling the number of trainable parameters. This becomes increasingly problematic in high-dimensional scenarios, prompting interest in alternative approaches like Monte Carlo Dropout (MCD) [26] and Spectral-normalized Neural Gaussian Processes (SNGP) [97]. Nevertheless, these alternatives are rarely applied in survival analysis, where VI continues to be the dominant method for training probabilistic neural networks, such as the work in [98] and [82]. Moreover, existing methods primarily focus on capturing epistemic uncertainty in model parameters, often overlooking the aleatoric uncertainty inherent in the data. In this section, we propose a method called BaySurv for training deep probabilistic survival models various learning techniques.

The work described in this section is published as [45] and [46].

3.1.2 Materials and Method

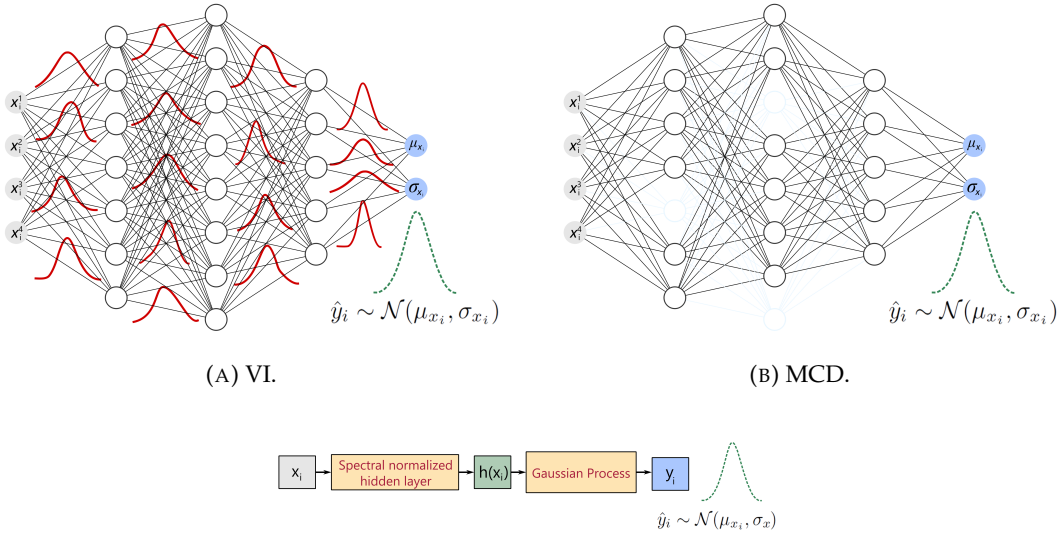
We use a multilayer perceptron (MLP) as the backbone network architecture. An MLP is a fully connected feedforward network; the number of hidden layers and neurons are hyperparameters tuned based on the data, see, for example, Magris and Iosifidis [99]. The adoption of an MLP is aligned with the literature, as it is capable of providing solid prediction performance [28, 89]. Using a backbone MLP as the foundation, we introduce three neural networks architectures for probabilistic survival analysis:

i. We train the model using VI by optimizing Equation (2.59), where weights and biases are treated as random variables. For each d-dimensional input $\boldsymbol{x}_i$, the two output nodes predict the mean and variance of a Gaussian over the

risk score, $f(\boldsymbol{x}_i;\theta) \sim \mathcal{N}\left(\mu_{\boldsymbol{x}_i}, \sigma^2_{\boldsymbol{x}_i}\right)$. Combined with the Breslow baseline hazard, each sample of $f(\boldsymbol{x}_i;\theta)$ yields a survival function, whose median we take as the predicted survival time $\hat{y}_i$ (denoted $\hat{e}_i$ in Chapter 2). This Bayesian approach captures epistemic uncertainty through the distribution over weights, while the Gaussian output accounts for aleatoric uncertainty inherent in the data. This distinguishes our VI model from prior VI-based survival models such as BNN-Cox [82], which capture only epistemic uncertainty. See Figure 3.1a.

ii. We train the model using MCD, a technique that approximates the Bayesian solution using VI. The architecture mirrors that of a standard deterministic MLP, but dropout is applied at every forward pass, including during inference. As in the VI model, the two output nodes correspond to the mean and variance of a Gaussian over the risk score $f(\boldsymbol{x}_i;\theta)$, with predictions sampled conditionally on the input $\boldsymbol{x}_i$ and a newly sampled dropout mask at each forward pass. We experiment with dropout rates $p \in \{0.1, 0.2, 0.5\}$, where p denotes the probability of dropping a unit, *i.e.*, $p = 1 - p_i$ in the notation of Section 2.6. See Figure 3.1b.

iii. We train using SNGP, which provides sample-wise uncertainty estimates without relying on stochastic sampling. Given a batch of inputs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, the Gaussian Process (GP) layer produces the corresponding risk scores $\{f(\boldsymbol{x}_1;\theta), \ldots, f(\boldsymbol{x}_N;\theta)\}$ along with a predictive covariance matrix. The model is trained by maximizing the likelihood objective defined in Equation (2.37). See Figure 3.1c.



(A) VI.

(B) MCD.

(C) SNGP.

FIGURE 3.1: The proposed architectures for computing $\hat{y}_i$, *i.e.*, the predicted survival time for individual i. (a): BNN with three hidden layers trained with VI, where $\boldsymbol{x}_i \in \mathbb{R}^4$ is a four-dimensional covariate vector. (b): MLP with three hidden layers trained with MCD. Transparent objects emphasize the random activation of the nodes under MCD. (c): MLP using the SNGP technique, with spectral normalized hidden layers and the output computed from a GP layer. Figure taken with permission from [46] ©2025 IEEE.

We train the networks by maximizing the Cox partial log-likelihood (Equation 2.37), and evaluate them on four real-world datasets (Table 3.1). We describe these

datasets further in [46]. As part of the preprocessing, we impute missing values using the sample mean for real-valued covariates and the mode for categorical covariates. We then apply z-score normalization to the data and one-hot encode categorical variables. We split the datasets into training (70%), validation (10%), and test (20%) sets using a stratification procedure by Qi et al. [82]. This ensures that the event times and the censoring rates are similar across all three subsets.

For all datasets and models, we tune hyperparameters using Bayesian optimization [100] over 10 iterations, evaluating performance on the validation set. Hyperparameters included number of training iterations, batch size, and network architecture, where applicable. The final models are configured using the hyperparameters that yield the highest average Antolini's concordance index [101]. For CoxPH-based models, we perform full-batch training on the training set. In contrast, neural network models are trained using stochastic minibatches drawn from the training set. We also apply early stopping to all neural network models, terminating training when the likelihood loss on the validation set no longer improves.

To evaluate prediction performance, we report Antolini's C-index (CI_{td}) [101], the mean absolute error using hinge loss (MAE_H) [102], the MAE using pseudo-observations (MAE_{PO}) [102], and the integrated Brier score (IBS) [77]. To evaluate calibration performance, we report the Integrated Calibration Index (ICI) [103], Distribution Calibration (D-Cal) [27] and Coverage Calibration (C-Cal) [82]. Both D-Cal and C-Cal are reported as p-values of χ^2 tests: D-Cal tests whether the predicted survival probabilities at the observed times are uniformly distributed, and C-Cal tests whether the observed coverage of the predicted CrIs matches their nominal levels (10% to 90%). We consider a model D- or C-calibrated if $p > 0.05$.

TABLE 3.1: Overview of the raw datasets.

Dataset	Sample size (N)	Censored (C%)	# Features (d)
METABRIC	1,902	42%	9
SEER	4,024	85%	28
SUPPORT	8,873	32%	14
MIMIC-IV	38,520	67%	91

3.1.3 Experiments and Results

Table 3.2 shows the prediction results of our baseline MLP, its probabilistic variants, and the literature benchmarks on the test sets. In the METABRIC dataset, all probabilistic models, especially VI, outperform the baseline MLP in CI_{td} and MAE_H, and all except MCD ($p = 0.2$) show lower MAE_{PO}. In the SEER dataset, probabilistic models achieve similar CI_{td}, while VI and MCD outperform the baseline in MAE_H. In the SUPPORT dataset, only MCD ($p = 0.5$) improves on the baseline in MAE_H and MAE_{PO}, while SNGP matches the baseline in CI_{td} and achieves the lowest IBS among our models. In the MIMIC-IV dataset, the differences between the models are smaller, but VI and MCD still perform better in MAE_H. Overall, MCD with $p = 0.5$ is the only model that improves MAE_H over the baseline on all four datasets.

Table 3.3 shows the calibration results. In the METABRIC dataset, all of our models are D-calibrated, with VI and MCD also C-calibrated. MCD outperforms the baseline in terms of ICI, while VI has a higher ICI despite being distribution and coverage calibrated. In the SEER dataset, all of our models are D-calibrated, but VI and MCD have higher (worse) ICI than the baseline (0.013-0.037 vs. 0.009). Among our models in the SUPPORT dataset, only the baseline and SNGP are D-calibrated, with VI and MCD performing poorly. A similar pattern holds in the MIMIC-IV dataset,

where again only the baseline and SNGP are D-calibrated. Varying the dropout rate makes no notable difference in calibration performance; in the SUPPORT and MIMIC-IV datasets, no MCD variant is D-calibrated regardless of p.

Table 3.4 compares two literature benchmark models and our proposed models in terms of space complexity and number of trainable parameters on the MIMIC-IV dataset. For a fair comparison, we configure all models with a single hidden layer with 32 nodes. Notice the doubling of the number of parameters when using VI as opposed to the baseline MLP and MCD, resulting in an increased overhead. Despite the fact that both are of $\mathcal{O}(d)$ complexity, doubling the number of trainable parameters can make a significant difference in high-dimensional models, so that VI is no longer an attractive solution.

TABLE 3.2: Prediction results across datasets. Bold indicates top 3 models (including ties). Table taken with permission from [46] ©2025 IEEE.

(a) METABRIC

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH	0.618	68.923	96.797	0.173
CoxNet	0.611	68.288	97.631	0.173
CoxBoost	0.638	69.976	102.745	0.187
RSF	0.641	67.242	98.728	0.175
DSM	**0.670**	70.361	100.83	0.170
DCM	0.637	74.187	104.532	0.179
BNN-Cox (VI)	**0.672**	65.777	94.184	**0.166**
BNN-MTLR (VI)	**0.655**	**59.027**	**89.387**	**0.163**
This work:				
Baseline (MLP)	0.604	68.668	95.540	0.168
+ VI	0.634	**63.776**	**91.843**	0.168
+ MCD ($p = 0.1$)	0.621	67.235	93.981	0.167
+ MCD ($p = 0.2$)	0.628	67.371	96.184	**0.166**
+ MCD ($p = 0.5$)	0.632	**65.651**	**92.353**	**0.164**
+ SNGP	0.631	67.378	93.434	**0.166**

(b) SEER

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH	**0.723**	100.849	196.898	**0.083**
CoxNet	**0.721**	95.253	197.066	0.083
CoxBoost	0.689	106.635	232.204	0.092
RSF	0.682	102.062	206.41	0.086
DSM	0.699	153.84	413.543	0.088
DCM	0.713	108.464	223.397	**0.082**
BNN-Cox (VI)	**0.716**	95.184	**202.151**	**0.081**
BNN-MTLR (VI)	0.702	**48.166**	**201.856**	0.085
This work:				
Baseline (MLP)	0.703	101.972	204.408	0.084
+ VI	0.699	**81.923**	**199.796**	0.084
+ MCD ($p = 0.1$)	0.702	91.202	203.457	0.084
+ MCD ($p = 0.2$)	0.691	89.103	203.323	0.085
+ MCD ($p = 0.5$)	0.692	**82.001**	207.956	0.087
+ SNGP	0.698	106.032	209.498	0.084

(c) SUPPORT

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH	0.568	409.947	754.978	0.201
CoxNet	0.568	409.915	754.960	0.201
CoxBoost	**0.612**	**366.928**	**721.443**	0.190
RSF	0.602	377.862	736.800	0.200
DSM	**0.626**	384.349	723.002	**0.185**
DCM	**0.629**	434.376	759.509	**0.185**
BNN-Cox (VI)	0.610	416.650	738.350	0.189
BNN-MTLR (VI)	0.606	**369.060**	**707.679**	0.186
This work:				
Baseline (MLP)	0.609	391.573	722.709	0.187
+ VI	0.531	414.133	768.599	0.208
+ MCD ($p = 0.1$)	0.600	416.175	738.567	0.193
+ MCD ($p = 0.2$)	0.567	400.064	751.343	0.202
+ MCD ($p = 0.5$)	0.605	**374.046**	**717.227**	0.192
+ SNGP	0.610	397.877	726.725	**0.184**

(d) MIMIC-IV

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH	0.726	257.364	286.516	**0.174**
CoxNet	0.727	255.802	283.951	**0.174**
CoxBoost	0.656	263.836	297.763	0.194
RSF	0.694	260.668	293.247	0.188
DSM	0.727	259.110	**280.361**	0.180
DCM	0.742	257.801	286.580	**0.174**
BNN-Cox (VI)	0.743	256.007	288.814	**0.169**
BNN-MTLR (VI)	0.744	**250.605**	**278.710**	**0.167**
This work:				
Baseline (MLP)	0.744	256.694	283.635	0.178
+ VI	0.743	254.735	282.345	0.180
+ MCD ($p = 0.1$)	**0.746**	255.815	283.672	0.177
+ MCD ($p = 0.2$)	**0.748**	**254.560**	283.156	0.175
+ MCD ($p = 0.5$)	**0.748**	**254.971**	**281.928**	0.177
+ SNGP	0.736	257.195	282.082	0.182

Figure 3.2 shows two ways in which uncertainty can be expressed along with a survival prediction for decision support purposes. For this, we selected a random instance from the METABRIC test set and drew 1000 samples from the predictive distribution to estimate its survival function. Although MCD and VI predict similar median survival times (in Figure 3.2, 122 and 128, respectively, where the actual time is 135), the latter predicts a greater amount of variance around the survival function

TABLE 3.3: Calibration results across datasets. D-Cal and C-cal are reported as a p-value; a model is considered calibrated if $p > 0.05$. Bold indicates top 3 models (including ties). Table taken with permission from [46] ©2025 IEEE.

(a) METABRIC

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH	0.022	**0.959**	-
CoxNet	**0.011**	**0.977**	-
CoxBoost	0.147	0.822	-
RSF	0.047	0.530	-
DSM	0.025	0.017	-
DCM	0.066	0.160	-
BNN-Cox (VI)	**0.017**	0.101	0.812
BNN-MTLR (VI)	0.044	0.438	0.763
This work:			
Baseline (MLP)	0.040	0.853	-
+ VI	0.062	0.059	0.796
+ MCD (0.1)	**0.017**	0.735	**0.849**
+ MCD (0.2)	0.020	0.689	**0.857**
+ MCD (0.5)	**0.017**	0.760	**0.865**
+ SNGP	**0.017**	**0.964**	-

(b) SEER

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH	0.011	0.999	-
CoxNet	0.013	**1.000**	-
CoxBoost	0.051	**1.000**	-
RSF	0.021	**1.000**	-
DSM	**0.009**	0.931	-
DCM	0.016	**1.000**	-
BNN-Cox (VI)	**0.009**	**1.000**	**0.489**
BNN-MTLR (VI)	0.056	0.016	0.486
This work:			
Baseline (MLP)	**0.009**	**1.000**	-
+ VI	0.027	0.678	**0.487**
+ MCD (0.1)	0.013	0.989	**0.487**
+ MCD (0.2)	0.014	0.983	**0.487**
+ MCD (0.5)	0.037	0.841	0.486
+ SNGP	0.011	**1.000**	-

(c) SUPPORT

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH	0.059	**0.983**	-
CoxNet	0.059	**0.983**	-
CoxBoost	**0.019**	0.708	-
RSF	**0.022**	0.532	-
DSM	0.065	0.000	-
DCM	0.090	**0.904**	-
BNN-Cox (VI)	0.073	0.096	0.952
BNN-MTLR (VI)	**0.042**	0.029	0.947
This work:			
Baseline (MLP)	0.069	0.229	-
+ VI	0.064	0.000	**0.977**
+ MCD (0.1)	0.077	0.000	0.938
+ MCD (0.2)	0.061	0.000	**0.979**
+ MCD (0.5)	0.050	0.000	**0.956**
+ SNGP	0.075	0.689	-

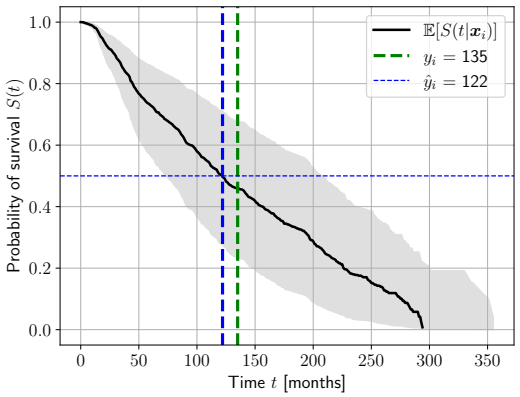
(d) MIMIC-IV

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH	0.037	0.549	-
CoxNet	0.033	**0.553**	-
CoxBoost	**0.016**	**0.997**	-
RSF	0.041	**0.935**	-
DSM	**0.006**	0.000	-
DCM	0.041	0.011	-
BNN-Cox (VI)	0.066	0.001	0.589
BNN-MTLR (VI)	0.057	0.060	0.583
This work:			
Baseline (MLP)	0.026	0.068	-
+ VI	0.096	0.000	0.583
+ MCD (0.1)	0.034	0.014	**0.600**
+ MCD (0.2)	0.038	0.013	**0.596**
+ MCD (0.5)	0.036	0.008	**0.597**
+ SNGP	**0.015**	0.230	-

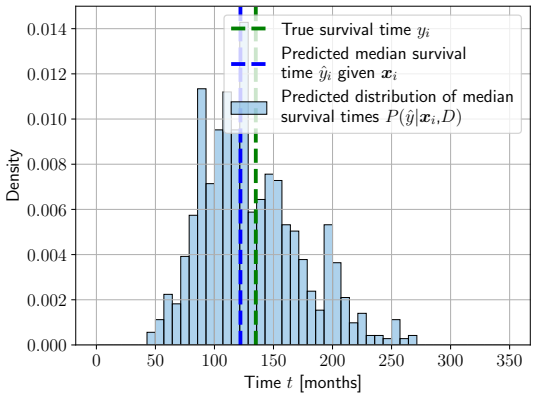
TABLE 3.4: Space complexity of the models and number of trainable parameters on the MIMIC-IV dataset. Table taken with permission from [46] ©2025 IEEE.

Model	Space complex.	#Params.
BNN-Cox (VI)	$\mathcal{O}(d)$	5570
BNN-MTLR (VI)	$\mathcal{O}(Kd)$	10718
This work:		
Baseline (MLP)	$\mathcal{O}(d)$	2849
+ VI	$\mathcal{O}(d)$	5700
+ MCD	$\mathcal{O}(d)$	2882
+ SNGP	$\mathcal{O}(d)$	3776

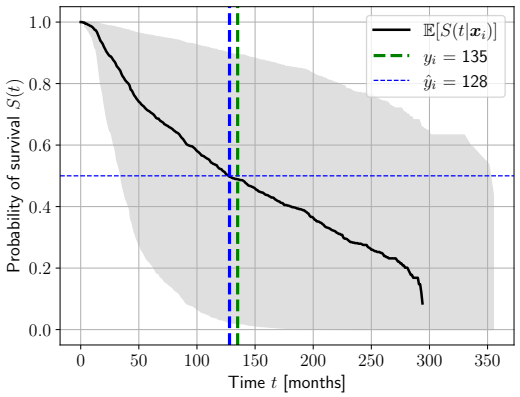
(left column) and also the histogram of the predicted survival time displays higher variance (right column). This example shows why it is important to communicate the uncertainty associated with a prediction.



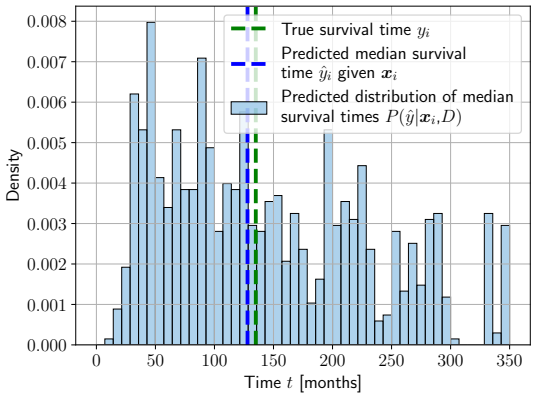
(A) Survival prediction using MCD ($p = 0.5$).



(B) Histogram of $\hat{y}_i$ using MCD ($p = 0.5$).



(C) Survival prediction using VI.



(D) Histogram of $\hat{y}_i$ using VI.

FIGURE 3.2: Model prediction on the METABRIC dataset using MCD and VI. For an individual i, we use the median of the predicted survival function as the predicted survival time $\hat{y}_i$. Although MCD and VI predict similar median survival times, VI has more variance around this particular prediction than MCD. Left column: mean individual survival function and its corresponding 90% CrI. Right column: histogram (as an approximation of the underlying density) of the predicted survival time. Figure taken with permission from [46] ©2025 IEEE.

3.1.4 Discussion

In this work, we developed BaySurv, a method for training probabilistic survival models, and implemented it with three techniques: VI, MCD, and SNGP. Unlike prior VI-based survival models, which capture only epistemic uncertainty, our VI and MCD variants also model aleatoric uncertainty through a Gaussian output. We compared these models to their deterministic counterparts in terms of prediction and calibration performance. Our main finding is that both MCD and SNGP offer compelling alternatives to VI in survival analysis. In our experiments, these methods can match VI's performance while reducing computational costs. Notably, the MCD

model frequently improved mean absolute error compared to state-of-the-art baselines and achieved calibration scores comparable to VI, whereas the SNGP model consistently gave D-calibrated survival curves across all datasets. These advances in uncertainty quantification and calibration are particularly valuable for precision medicine, where accurate and reliable individualized survival predictions can directly impact treatment decisions.

This work has limitations. First, it does not make use of high-dimensional data such as imaging or video, where computational efficiency is even more important than in tabular data. Second, the results highlight that the proposed models struggle to optimize both prediction accuracy and calibration simultaneously. Future research should focus on enhancing predictive performance without compromising calibration quality to better support clinical decision-making in precision medicine.

3.2 Prediction of TMJ Involvement in Children With JIA

3.2.1 Introduction

Early detection of temporomandibular joint (TMJ) involvement in children with juvenile idiopathic arthritis (JIA) poses a significant challenge for clinicians, as the condition is frequently asymptomatic in its initial stages [104]. In this context, TMJ involvement refers to findings that are considered to result from past or ongoing TMJ arthritis. There are several modalities to diagnose TMJ arthritis – *e.g.*, clinical (orthodontic) examinations, computed tomography (CT) scans, contrast-enhanced magnetic resonance imaging (MRI), and ultrasound [105]. Although clinical examinations are the most cost-effective modality to assess TMJ involvement of patients with JIA [106], [107] reported low sensitivity (0.47 for patients; 0.66 for joints) and poor agreement with MRI findings. While clinical exams detected all severe deformities, only 39% of early TMJ cases were correctly identified. This highlights the need for additional decision support when relying on clinical examinations to detect early TMJ involvement.

In this work, we propose a model to assess temporomandibular joint (TMJ) involvement in children under 16 diagnosed with juvenile idiopathic arthritis (JIA). The model relies solely on clinical examinations performed according to consensus-based standards tailored for this patient group. The goal is to support early diagnosis by providing clinicians with a decision-support tool that flags examinations with likely TMJ involvement, together with a calibrated measure of its uncertainty. We employ conformal prediction to rigorously quantify the uncertainty of the model's predictions.

The work described in this section is published as [47].

3.2.2 Materials and Method

We obtain a dataset containing 6154 longitudinal records of 1035 pediatric patients (<16 years old) over a 25-year period, recorded and provided by the Department of Dentistry and Oral Health, Aarhus University. Each patient underwent 2-17 clinical examinations, where the patient's symptoms, drug usage, general arthritis level, TMJ involvement and clinical findings were recorded (see Figure 3.3). The clinical variables examined ($d = 95$) at each visit are in agreement with evidence-based standards on orofacial examination in JIA [108]. Among the cohort of patients, 690 are girls and 345 are boys, reflecting the gender distribution commonly observed in patients with JIA within an open population [109]. All patients were diagnosed with

JIA prior to their first examination. However, not all patients initially showed signs of TMJ involvement; some developed it over time, while others were diagnosed with TMJ involvement after the examination was completed. The involvement of the TMJ was confirmed or refuted by an X-ray or MRI.

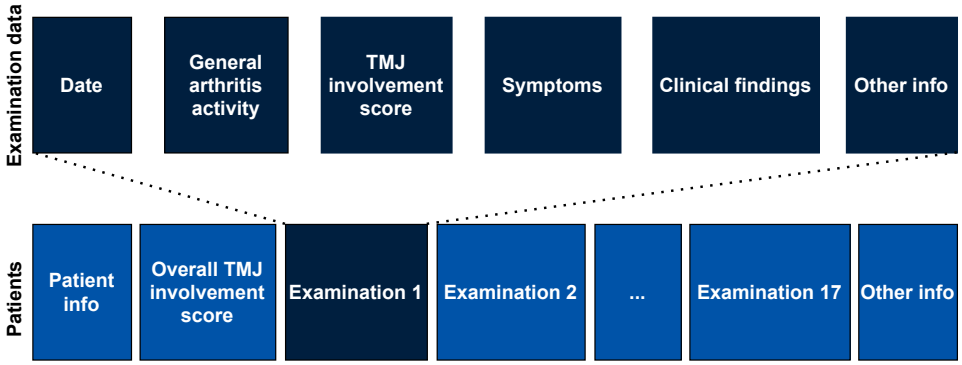


FIGURE 3.3: Each examination of a JIA patient includes a screening, where the patient's symptoms, drug usage and other clinical findings are recorded. The dataset contains historical examinations of 1,035 JIA patients spanning a 25-year period with 6,154 longitudinal records in total. Figure taken with permission from [47] ©2024 IEEE.

We implement a Random Forest model [110] to classify TMJ involvement (no involvement/involvement) in children with JIA. After preprocessing, we adopt an 80-10-10 data split; 80% is used for training, 10% is used for calibration, and 10% is used for testing. The split is performed at the record level, so examinations from the same patient can appear in more than one subset. We report the classifier's precision, sensitivity, and macro-averaged F1 score ($F1_m$) on the test set. The F1 and $F1_m$ are resp. defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}, \tag{3.1}$$

$$F1_m = \frac{F1_{\text{TMJ}_0} + F1_{\text{TMJ}_1}}{2}, \tag{3.2}$$

where precision and sensitivity are computed per class, treating that class as the positive class.

To quantify the uncertainties of the predictive model, we use the Model Agnostic Prediction Interval Estimator (MAPIE) library to train a conformal classifier [111]. This classifier wraps the original classifier and produces conformal prediction sets that have a guaranteed marginal coverage rate. Thus, the model output is then a set of classes instead of a single class. We set $\alpha = 0.1$, indicating that we can expect 90% of the prediction regions to cover the true outcome. However, the guarantee is only an average for the samples from the distribution of the calibration data.

3.2.3 Experiments and Results

Table 3.5 reports the predictive performance of the classifier. We adopt three sampling strategies (IID Data, Temporal Segmentation and Lagged Features) of the clinical dataset and five evaluation metrics for each class (precision, sensitivity, macro-averaged F1 score, coverage, and set size). We consider three sampling strategies:

- **Strategy 1: IID Data**

  - Assumes all clinical examinations are IID, ignoring temporal structure.
  - Achieves strong performance (macro F1 score of 0.87) using either all or expert-selected features.
  - But not practical, as it mixes early-stage (less evident involvement) and late-stage (clear involvement) data.

- **Strategy 2: Temporal Segmentation**
  - Splits data into 0–2, 2–5, and 5–15 year intervals since first examination.
  - Early-stage model (0–2 years, $N = 2682$) performs better than mid-stage (2–5 years).
  - Compared to the IID model with expert features (sensitivity 0.80), sensitivity drops by 12.5% relative in the early-stage model (to 0.70) and by 27.5% relative in the mid-stage model (to 0.58).
  - After 5 years, signs become clinically clearer due to disease progression and patient maturity.

- **Strategy 3: Lagged Features**
  - Transforms the dataset to supervised format by adding historical (lagged) features.
  - One-lag model: $N = 4983$, $d = 52$; two-lag model: $N = 3969$.
  - Slight performance gain: sensitivity improves from 0.8 to 0.82 (+2.5%) compared to IID with expert features.

TABLE 3.5: Prediction results on the test set. N is the number of samples in the dataset and d is the number of features. The predicted classes, TMJ_0 and TMJ_1, correspond to no and present TMJ involvement, respectively. Table taken with permission from [47] ©2024 IEEE.

Strategy	Dimensions	Class	Precision ↑	Sensitivity ↑	$F1_m$ ↑	Coverage ↑	Set size ↓
IID Data.	$N = 6006$	TMJ_0	0.88	0.93	0.8742	0.986	1.154
	$d = 95$	TMJ_1	0.88	0.81		0.899	1.207
	$N = 6006$	TMJ_0	0.88	0.93	0.8706	0.986	1.154
	$d = 26$	TMJ_1	0.88	0.80		0.899	1.207
Temp. Segmentation	$N = 2682$	TMJ_0	0.89	0.95	0.8455	0.969	1.036
	$d = 26$	TMJ_1	0.86	0.70		0.740	1.078
	$N = 1430$	TMJ_0	0.76	0.90	0.7451	0.988	1.233
	$d = 26$	TMJ_1	0.79	0.58		0.737	1.333
	$N = 1839$	TMJ_0	0.86	0.75	0.8144	0.967	1.565
	$d = 26$	TMJ_1	0.78	0.88		0.978	1.326
Lag. Features	$N = 4983$	TMJ_0	0.88	0.90	0.8603	0.949	1.233
	$d = 52$	TMJ_1	0.84	0.82		0.906	1.212
	$N = 3969$	TMJ_0	0.88	0.89	0.8578	0.961	1.301
	$d = 78$	TMJ_1	0.84	0.83		0.917	1.310

3.2.4 Discussion

The primary findings of this study show that the proposed model using the "Temporal Segmentation" strategy can classify patients early. That is, within a period of

two years (*i.e.*, from the first visitation and two years ahead) the model can classify patients with TMJ involvement with a precision of 0.86 and a sensitivity of 0.7. This means that out of all examinations predicted as showing TMJ involvement, 86% had actual TMJ involvement, and 70% of examinations with TMJ involvement were correctly identified. This is based on a dataset with 2682 samples and 26 clinical features. Compared to [107], who attempted to detect TMJ involvement within 1 month of the MRI for 33 patients, they report a 0.47 sensitivity (95% CI: 0.25-0.71) and a misdiagnosis rate of 61%. Therefore, using our model offers a 49% increase in sensitivity, although our dataset is also larger. The proposed method can be applied to current clinical examination results to help clinicians assess involvement of the temporomandibular joints in future patients, allowing for early and personalized follow-up.

Chapter 4

Competing Risks and Multiple Events

So far, we have only considered survival problems with a single event of interest. This chapter extends this idea to competing and multiple events by presenting a multi-event survival model, nicknamed MENSA, that can predict the time until an instance will experience one or more events. In practice, we model the conditional survival function for each event as a convex combination of Weibull distributions. This approach relaxes the proportional hazards assumption on the underlying marginal survival distribution for each event, addressing a key limitation of Cox-based methods.

Then, based on this multi-event framework, we present a practical method to predict the time until a patient with amyotrophic lateral sclerosis (ALS), a rare neurodegenerative disease, will experience significant functional impairment with regards to common physical functions. We apply this method to predict five event-specific individual survival distributions (ISDs) for each patient, each offering an interpretable and meaningful estimate of the likely timing of the respective event. Additionally, we make so-called counterfactual predictions, where we change certain covariates and study their effect on the predicted outcome.

4.1 MENSA: A Multi-Event Network for Survival Analysis

4.1.1 Introduction

Traditional time-to-event models have primarily been developed for either single-event or competing-risk settings, while multi-event scenarios have received comparatively little attention. Yet in many practical applications, patients may experience several non-exclusive events, which can even occur in semi-competing relationships. A common strategy is to train independent models for each event, but this approach fails to exploit dependencies and shared structures across events.

To address these limitations, we propose MENSA (Multi-Event Network for Survival Analysis). Unlike standard single-event models, MENSA jointly learns flexible time-to-event distributions for multiple possible events, whether competing or co-occurring. In practice, MENSA jointly models the distributions of the events, representing each as a convex combination of Weibull distributions. This formulation relaxes the strict proportional hazards assumption on the marginal survival distributions, which is a key limitation of Cox-based methods [23, 28, 46]. Our architecture incorporates a shared covariate layer that captures latent representations shared across all events and event-specific layers that capture representations specific to each event. In addition, rather than solely minimizing the negative log-likelihood (NLL) across events, MENSA optimizes a composite objective that includes both

the NLL and a novel trajectory-based likelihood term. This additional term encourages the model to respect known temporal orderings between events based on prior knowledge.

As an example, consider the illness-death scenario as introduced by Fix and Neyman [35] with two events, "illness (A)" and "death (B)": if both events are observed for patient i, then clearly event A must occur before event B. In this case, we want to maximize the probability of surviving event B at the time of event A – so if A occurs at 1 year, the probability of surviving B at 1 year should be very high. MENSA is based on the multi-state framework we introduced in Section 2.5, where an event is treated as a transition from one state to another – *e.g.*, from healthy to sick, or from alive to dead. This framework allows us to include the event-free (transient) state as an additional output, which is estimated jointly with the events of interest as an auxiliary term in the likelihood. The MENSA framework naturally extends to single-event and competing-risk scenarios, making it broadly applicable across diverse survival analysis tasks.

The work described in this section is published as [48].

4.1.2 Materials and Method

Before presenting our method, we first introduce some preliminaries. Section 2.1 defines the notation for single-event survival analysis, which we now extend to multiple events. Let $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{t}^{(i)}, \boldsymbol{\delta}^{(i)})\}_{i=1}^{N}$ be the dataset, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ is a d-dimensional vector of features (covariates), $\boldsymbol{t}^{(i)} \in \{1, 2, \ldots, T_{\max}\}^K$ is a vector of observed times for the i-th instance now across K multiple events, each at some time, and $\boldsymbol{\delta}^{(i)} \in \{0,1\}^K$ is vector of event indicators for each event. Moreover, let $e_k^{(i)} \in \{1, 2, \ldots, T_{\max}\}$ denote the event time and let $c_k^{(i)} \in \{1, 2, \ldots, T_{\max}\}$ denote the censoring time for the i-th instance for event k over the event horizon; these are realizations of the random event time $E_k^{(i)}$ and censoring time $C_k^{(i)}$. Thus $t_k^{(i)} = e_k^{(i)}$ if $\delta_k^{(i)} = 1$, otherwise, $t_k^{(i)} = c_k^{(i)}$ if $\delta_k^{(i)} = 0$. We define a multi-event survival problem as predicting, for each instance i, event k, and time $t \in \{1, \ldots, T_{\max}\}$, the survival probability $\Pr(E_k^{(i)} > t \mid \boldsymbol{x}^{(i)})$. This produces a matrix of survival functions $\boldsymbol{S}^{(i)} \in [0,1]^{T_{\max} \times K}$, where $\boldsymbol{S}_{t,k}^{(i)} = \Pr(E_k^{(i)} > t \mid \boldsymbol{x}^{(i)})$ and $\boldsymbol{S}^{(i)} = [\boldsymbol{s}_1^{(i)}, \boldsymbol{s}_2^{(i)}, \ldots, \boldsymbol{s}_K^{(i)}]$. Recall that the survival function is the probability that an event occurs after time t, *i.e.*, $\boldsymbol{s}_k^{(i)} = [\Pr(E_k^{(i)} > 1 \mid \boldsymbol{x}^{(i)}), \Pr(E_k^{(i)} > 2 \mid \boldsymbol{x}^{(i)}), \ldots, \Pr(E_k^{(i)} > T_{\max} \mid \boldsymbol{x}^{(i)})]$ describes the probability that instance i does not experience event k up to time $t \in \{1, \ldots, T_{\max}\}$.

In MENSA, we propose a multi-state architecture (Figure 4.1) in which events are represented as transitions between discrete states. Let $\mathcal{P} = \{0, 1, \ldots, P-1\}$ denote the finite set of states, where state 0 is the initial (event-free) state. Each event $k \in \{1, \ldots, K\}$ corresponds to a transition into a state $p_k \in \mathcal{P} \setminus \{0\}$, so that $P = K + 1$. For example, if event k denotes a *relapse*, then p_k is the absorbing state associated with *relapse*. This representation naturally accommodates sequential or co-occurring events – *e.g.*, a patient becoming ill and later dying.

For each instance i, we estimate a matrix of survival distributions $\boldsymbol{S}^{(i)} \in [0,1]^{T_{\max} \times P}$; one distribution for each state $p \in \mathcal{P}$, *i.e.*, the K event-specific distributions defined above plus one for the initial event-free state ($p = 0$). The marginal density function $f_p(t)$ for state p, which represents the probability of entering state p at time point t, is modeled as a convex combination of $\Psi \in \mathbb{N}_+$ Weibull distributions. This formulation offers two key advantages. First, it relaxes the proportional hazards assumption on the underlying marginal survival distribution, allowing the hazard function to vary flexibly over time. Second, it leverages the closed-form PDF and

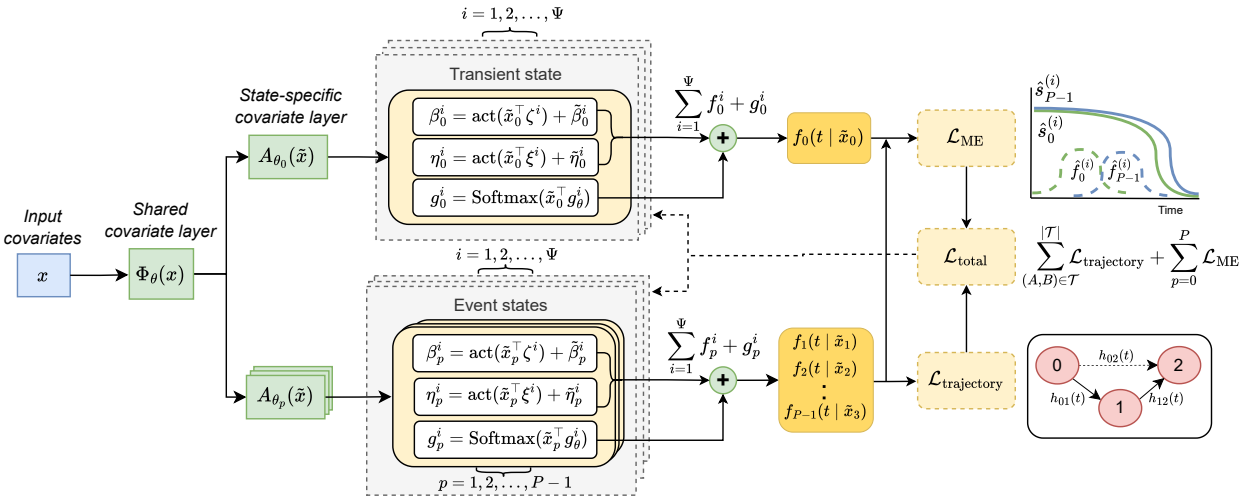


FIGURE 4.1: Overview of the MENSA architecture. The model first maps the input covariates through a shared MLP, yielding a common representation $\tilde{\boldsymbol{x}}$. This representation is then passed through event-specific layers, each producing $\tilde{\boldsymbol{x}}_p$ for state $p \in \mathcal{P}$. Each event-adapted representation parameterizes a mixture of Weibull distributions to predict a time-to-event density function $f_p(t)$, describing the transition time from the transient state (state 0) to event state p. Training is performed with a combined loss comprising a multi-event likelihood $\mathcal{L}_{\text{ME}}$, which evaluates the model's fit to event times, and a trajectory likelihood $\mathcal{L}_{\text{trajectory}}$, which enforces temporal ordering between events. Figure taken with permission from [48].

CDF of the Weibull distribution, respectively:

$$f(t;\beta,\eta) = \frac{\eta}{\beta}\left(\frac{t}{\beta}\right)^{\eta-1} \exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \tag{4.1}$$

$$F(t;\beta,\eta) = 1 - \exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \tag{4.2}$$

where $\beta > 0$ is the scale and $\eta > 0$ the shape parameter. We parameterize each such Weibull distribution by two covariate-dependent transformations, that is, $h_p(\boldsymbol{x}^{(i)}) = A_{\theta,p}(\Phi_\theta(\boldsymbol{x}^{(i)}))$, such that:

$$\log \beta_{\psi,p}(\boldsymbol{x}^{(i)}) = \tilde{\beta}_{\psi,p} + \text{act}\big(h_p(\boldsymbol{x}^{(i)})^\top \zeta_{\psi,p}\big), \tag{4.3}$$

$$\log \eta_{\psi,p}(\boldsymbol{x}^{(i)}) = \tilde{\eta}_{\psi,p} + \text{act}\big(h_p(\boldsymbol{x}^{(i)})^\top \xi_{\psi,p}\big), \tag{4.4}$$

where $\Phi_\theta(\cdot)$ is a shared covariate transformation, $A_{\theta,p}(\cdot)$ is a state-specific adapter applied to its output, $\text{act}(\cdot)$ denotes a non-linear activation function, $\psi \in \{1, 2, \ldots, \Psi\}$ indexes the Weibull mixture distributions, and $\zeta_{\psi,p}, \xi_{\psi,p} \in \mathbb{R}^{d_h}$ are learnable weight vectors associated with the scale and shape parameters, respectively, where d_h is the dimension of the hidden representation $h_p(\boldsymbol{x}^{(i)})$. The terms $\tilde{\beta}_{\psi,p}$ and $\tilde{\eta}_{\psi,p}$ are learnable biases that initialize the log-scale and log-shape, respectively, for each state p. We define $\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)}) = \big(\beta_{\psi,p}(\boldsymbol{x}^{(i)}), \eta_{\psi,p}(\boldsymbol{x}^{(i)})\big)$, and the corresponding distribution's

PDF and CDF for transitioning into state p at t given covariates $\boldsymbol{x}^{(i)}$ as:

$$f_p(t;\boldsymbol{x}^{(i)}) = \sum_{\psi=1}^{\Psi} g_{\psi,p}(\boldsymbol{x}^{(i)})\, f_{\psi,p}\big(t;\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)})\big), \tag{4.5}$$

$$F_p(t;\boldsymbol{x}^{(i)}) = \sum_{\psi=1}^{\Psi} g_{\psi,p}(\boldsymbol{x}^{(i)})\, F_{\psi,p}\big(t;\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)})\big), \tag{4.6}$$

where $g_{\psi,p}(\boldsymbol{x}^{(i)})$ are normalized mixture weights produced by a state-specific gating network, *i.e.*, a softmax over linear outputs of $h_p(\boldsymbol{x}^{(i)})$. The corresponding survival function is $S_p(t;\boldsymbol{x}^{(i)}) = 1 - F_p(t;\boldsymbol{x}^{(i)})$.

We then extend the standard survival likelihood (see Equation 2.17) to the proposed multi-state architecture, where each instance i may experience multiple events. Instead of indexing events directly, the likelihood is expressed in terms of state transitions. For each state $p \in \mathcal{P}$, we define an indicator-time pair $(\delta_p^{(i)}, t_p^{(i)})$, where $\delta_p^{(i)}$ indicates whether a transition into state p was observed, and $t_p^{(i)}$ is the observed or censored transition time. For $p \geq 1$, these are simply the indicator and time of the corresponding event, *i.e.*, $(\delta_k^{(i)}, t_k^{(i)})$ with $p = p_k$. For the event-free state, we set $\delta_0^{(i)} = 1$ if instance i experiences none of the K events and $\delta_0^{(i)} = 0$ otherwise, with $t_0^{(i)} = \max_k t_k^{(i)}$. The likelihood includes the state's density $f_p(t_p^{(i)} \mid \boldsymbol{x}^{(i)})$ if the transition was observed ($\delta_p^{(i)} = 1$), and its survival function $S_p(t_p^{(i)} \mid \boldsymbol{x}^{(i)})$ if censored ($\delta_p^{(i)} = 0$). State 0 thus acts as an auxiliary term during training; all predictions and evaluations use only the K event-specific distributions.

To mitigate class imbalance between states, we add nonnegative weights w_p, computed from the inverse frequency of transitions into each state and normalized such that $\sum_{p\in\mathcal{P}} w_p = |\mathcal{P}|$. Assuming conditional independence between states[1], the multi-event likelihood is then defined as:

$$\begin{aligned}\log \mathcal{L}_{\mathrm{ME}}(\mathcal{D}) \;=\; \sum_{i=1}^{N} \sum_{p\in\mathcal{P}} w_p \Big[& \delta_p^{(i)} \log f_p\left(t_p^{(i)} \mid \boldsymbol{x}^{(i)}\right) \\ & + (1-\delta_p^{(i)}) \log S_p\left(t_p^{(i)} \mid \boldsymbol{x}^{(i)}\right) \Big],\end{aligned} \tag{4.7}$$

where $f_p(\cdot)$ and $S_p(\cdot)$ are the conditional density and survival functions for state p, respectively.

When prior knowledge of temporal event ordering is available, it can be used to further guide the model. A classic example of this is the illness-death model [35], in which patients begin in a healthy state (state 0), may transition to illness (event A), and eventually die (event B). This produces four possible scenarios (Table 4.1): (1) neither event occurs, (2) only A occurs, (3) only B occurs, or (4) both occur, with A preceding B. In the last case, death before illness is not possible, and the likelihood must reflect this constraint. We incorporate such via an auxiliary likelihood term that maximizes the survival probability for event B at the time event A occurs, *i.e.*, $S_B(t_A)$. This enforces a low estimated probability of B before A and encodes temporal constraints that rule out implausible event sequences. Note that this

[1] That is, given the covariates, the transition to one state is independent of transitions to other states; any dependencies are assumed to be explained by the covariates, *i.e.*, $E_k \perp\!\!\!\perp E_{k'} \mid X \quad \text{for } k \neq k'$. We additionally assume independent censoring for each event (Definition 6).

TABLE 4.1: Possible scenarios for two events, A and B, and their log-likelihood contributions under conditional independence. δ_A and δ_B indicate whether each event is observed (1) or censored (0); $S_A(t)$ and $S_B(t)$ are survival functions; and $f_A(t)$ and $f_B(t)$ are probability density functions. If prior knowledge dictates that event A must precede event B when both occur, the joint-event case ($\delta_A = 1, \delta_B = 1$) additionally includes the trajectory term $\log S_B(t_A)$, highlighted in bold and weighted by λ in Equation (4.9).

#	δ_A	δ_B	Log-likelihood contribution
1	0	0	$\log S_A(t_A) + \log S_B(t_B)$
2	1	0	$\log f_A(t_A) + \log S_B(t_B)$
3	0	1	$\log S_A(t_A) + \log f_B(t_B)$
4	1	1	$\log f_A(t_A) + \log f_B(t_B) + \mathbf{\log S_B(t_A)}$

term enforces the ordering of events but does not model their dependence; in semi-competing settings, where the terminal event (*e.g.*, death) censors the non-terminal one, the independence assumptions above may be violated.

Formally, let $\mathcal{K} = \{1, \ldots, K\}$ denote the set of events, and let $\mathcal{T} = \{(A, B) \mid A, B \in \mathcal{K},\ A \text{ occurs before } B\}$ be the set of known event orderings. For each instance i, we consider all uncensored event pairs $(A, B) \in \mathcal{T}$. The proposed trajectory likelihood term then sums the survival probability of B at the time of A for these pairs, weighted by event indicators $\delta_A^{(i)}$ and $\delta_B^{(i)}$ (equal to 1 if the events occur for instance i, and 0 otherwise). It is written as:

$$\log \mathcal{L}_{\text{trajectory}}(\mathcal{D}) = \sum_{i=1}^{N} \sum_{(A,B)\in\mathcal{T}} \delta_A^{(i)} \delta_B^{(i)} \log S_B\big(t_A^{(i)} \mid \boldsymbol{x}^{(i)}\big), \tag{4.8}$$

where $\mathcal{T}$ is the set of trajectories. Finally, we have:

$$\log \mathcal{L}_{\text{total}}(\mathcal{D}) = (1-\lambda)\frac{1}{N}\log \mathcal{L}_{\text{ME}}(\mathcal{D}) + \lambda \cdot \frac{1}{N}\log \mathcal{L}_{\text{trajectory}}(\mathcal{D}), \tag{4.9}$$

where λ is a hyperparameter that controls the trade-off between the two losses.

4.1.3 Experiments and Results

We use an MLP as the backbone network, which maps input features to a shared representation that parameterizes the Weibull distributions. On top of this, small state-specific MLPs refine features for each prediction task. The backbone has a single hidden layer with ReLU6 activation, and event times are modeled using a mixture of Weibull distributions. We optimize using Adam [112] with fixed-size minibatches.

Across 10 runs, we follow [113] and split the data into stratified training (70%), validation (10%), and test (20%) sets using different random seeds. Missing values are imputed from the training set (mean for continuous, mode for categorical), numerical features are normalized, and categorical variables are one-hot encoded. Early stopping is applied if the validation likelihood does not improve for 20 consecutive epochs.

To evaluate discrimination, we report the global CI (CI_G) [83], which measures how well the model ranks instances within an event, and the local CI (CI_L) [83],

which measures how well the model ranks events within an instance (see also Definitions 15 and 16). We further report the time-dependent AUC, computed at the 25th, 50th, and 75th percentiles of observed event times in the test set and averaged across these points. To assess calibration performance, we report the mean absolute error using a margin loss (mMAE) [27], the integrated Brier score (IBS) [77], and the distribution calibration (D-calibration) score [27].

Table 4.2 presents the discrimination results of our proposed model alongside existing baselines on four multi-event datasets. Models marked with † are significantly and practically worse than MENSA (details on the statistical testing and minimum effect of interest (MEI) thresholds are provided in the appendix of [48]. Overall, MENSA outperforms several deep learning baselines across datasets and achieves performance comparable to strong tree-based methods such as GBSA and RSF. In contrast to DeepHit and Hierarchical, MENSA avoids unstable behavior, such as high variance or metric collapse. For these multi-event datasets, we further evaluate local discrimination using the local C-index, which measures the accuracy of within-patient event orderings and highlights the benefits of multi-event modeling. When MENSA outperforms competing methods on this metric, we attribute the gains to its multi-state architecture and, in datasets with known event orderings (Rotterdam and EBMT, where $\lambda > 0$), to the trajectory-based likelihood.

Table 4.3 reports calibration performance. MENSA achieves low IBS and mMAE – often matching or exceeding classical survival models (*e.g.*, CoxPH, RSF) – and demonstrates D-calibration comparable to DSM, though lower than DeepSurv on MIMIC-IV and EBMT. Compared to the Hierarchical model, our method shows consistently stronger performance across all datasets.

TABLE 4.2: Discrimination performance (mean ± SD.) across models on multi-event datasets, averaged over 10 runs. For all metrics, higher (↑) is better. Model types are abbreviated as SE = single event, CR = competing risks, and ME = multiple events. Bold denotes the top-3 models, including ties. † indicates models that are significantly and practically worse than MENSA, based on paired t-tests with Holm correction ($\alpha = 0.05$) and a minimum effect of interest (MEI) of 1.0. Table taken with permission from [48].

Model	Type	MIMIC-IV ($K=3$)			Rotterdam ($K=2$)			PRO-ACT ($K=4$)			EBMT ($K=5$)		
		CI_G	CI_L	AUC	CI_G	CI_L	AUC	CI_G	CI_L	AUC	CI_G	CI_L	AUC
CoxPH	SE	70.9±0.52	90.3±0.48	74.3±0.71	69.7±1.16	83.6±2.09†	75.1±1.48	70.6±0.47	**76.5**±0.55	76.7±0.91	54.9±1.04	66.3±1.16†	54.6±1.10
CoxNet	SE	71.0±0.53	90.4±0.45	74.4±0.71	69.8±1.16	83.8±2.05†	75.1±1.49	70.6±0.46	**76.5**±0.57	76.7±0.89	**55.1**±1.12	67.2±1.13†	54.5±1.17
Weibull	SE	71.2±0.54	**90.8**±0.43	74.8±0.70	69.7±1.18	87.9±1.99†	75.1±1.45	70.7±0.49	**77.1**±0.57	76.7±0.90	54.9±1.36	**68.0**±0.97	54.2±1.35
GBSA	SE	**71.9**±0.43	**91.4**±0.51	**75.2**±0.53	**75.3**±1.62	**90.5**±1.05	71.2±2.02†	**72.0**±0.31	75.6±0.66	**77.6**±0.44	**55.3**±0.93	67.2±0.82†	**55.4**±1.26
RSF	SE	69.1±0.48†	**90.8**±0.40	74.6±0.56	**70.7**±1.15	84.9±2.19†	**77.3**±1.74	**71.1**±0.51	75.6±0.83	**77.5**±0.60	55.0±0.75	64.6±0.92†	**55.6**±1.07
MTLR	SE	70.2±0.50†	88.3±0.95†	74.4±0.71	69.8±1.16	82.8±2.37†	**75.7**±1.85	70.7±0.53	**76.5**±0.53	**77.4**±0.83	53.6±1.04	61.6±2.38†	**55.4**±1.19
DeepSurv	SE	**71.8**±0.55	89.9±0.49	**75.8**±0.68	70.0±1.24	84.6±2.93†	75.4±1.75	**71.0**±0.59	76.0±0.57	76.7±1.00	**55.1**±1.02	67.6±0.82	54.9±1.22
DeepHit	CR	70.9±0.75	63.1±9.74†	75.1±0.94	66.5±1.09†	**90.2**±1.03	74.2±1.50	64.9±1.15†	65.3±3.35†	73.4±0.73†	52.4±1.29†	60.4±2.76†	55.2±1.46
DSM	CR	70.5±0.60	88.1±1.37†	74.5±0.83	**70.6**±1.26	86.0±2.81†	**76.4**±1.68	70.8±0.68	76.0±0.68	**77.4**±0.76	54.8±1.00	**68.0**±0.99	55.3±1.36
Hierarch.	ME	65.2±1.67†	85.3±1.73†	48.7±0.51†	67.8±2.06	55.7±8.11†	50.0±0.05†	70.8±0.60	73.6±1.42†	57.3±0.89†	54.7±1.33	44.5±1.12†	53.5±1.03†
MENSA (Ours)	ME	**71.4**±0.29	90.3±0.54	**75.2**±0.41	69.4±1.20	**90.1**±1.55	73.8±2.31	70.8±0.45	76.4±0.61	77.2±0.60	**55.1**±1.37	**68.8**±0.66	**55.7**±1.62

Table 4.4 compares the space complexity of baseline deep learning models and MENSA across single-event, competing-risks, and multi-event settings. For consistency, all models are configured with a single hidden layer of 32 nodes. By design, models that cannot learn events jointly must be trained K times in the multi-event scenario. For instance, DeepSurv exhibits a space complexity of $K\mathcal{O}(d)$, where d is the number of features, assuming training is performed sequentially rather than in parallel. This requirement can substantially inflate the number of trainable parameters in high-dimensional datasets with many events, making such approaches less appealing. In contrast, MENSA avoids this drawback, as it learns all events simultaneously with a space complexity of $\mathcal{O}(P\Psi + d)$.

TABLE 4.3: Calibration performance (mean ± SD.) across models on four multi-event datasets, averaged over 10 runs. For IBS and mMAE, lower (↓) is better, while for D-calibration, higher (↑) is better. D-calibration reports the number of run–event combinations (10 runs × number of events) in which a model is D-calibrated ($p > 0.05$). Bold denotes the top-3 models. † indicates models that are significantly and practically worse than MENSA, based on paired t-tests with Holm correction ($\alpha = 0.05$) and an MEI of 1.0 for IBS and mMAE. For MIMIC-IV, Rotterdam, and EBMT, mMAE is reported divided by 100, so the MEI corresponds to 100 days for these datasets; for PRO-ACT, mMAE is reported in days. Table taken with permission from [48].

Model	Type	MIMIC-IV ($K = 3$)			Rotterdam ($K = 2$)			PRO-ACT ($K = 4$)			EBMT ($K = 5$)		
		IBS	mMAE	D-Cal	IBS	mMAE	D-Cal	IBS	mMAE	D-Cal	IBS	mMAE	D-Cal
CoxPH	SE	15.3$_{\pm 0.51}$	10.8$_{\pm 5.77}$	30/30	19.6$_{\pm 1.53}$	23.2$_{\pm 1.15}$†	20/20	16.6$_{\pm 0.39}$	129.0$_{\pm 3.31}$†	36/40	20.4$_{\pm 0.68}$	48.6$_{\pm 0.71}$	49/50
CoxNet	SE	15.3$_{\pm 0.52}$	8.6$_{\pm 0.16}$	30/30	19.6$_{\pm 1.54}$	23.0$_{\pm 1.11}$	20/20	16.6$_{\pm 0.40}$	128.6$_{\pm 3.23}$†	36/40	**20.3**$_{\pm 0.68}$	**48.3**$_{\pm 0.70}$	49/50
Weibull	SE	15.6$_{\pm 0.42}$	11.3$_{\pm 0.14}$†	30/30	19.2$_{\pm 1.64}$	23.1$_{\pm 0.78}$	15/20	16.7$_{\pm 0.35}$	220.7$_{\pm 11.24}$†	36/40	**20.1**$_{\pm 0.98}$	59.4$_{\pm 0.86}$	18/50
GBSA	SE	16.7$_{\pm 0.63}$†	**6.9**$_{\pm 0.07}$	29/30	19.3$_{\pm 1.44}$	24.1$_{\pm 0.74}$†	20/20	16.7$_{\pm 0.42}$	**120.8**$_{\pm 2.14}$	35/40	20.4$_{\pm 0.75}$	**44.7**$_{\pm 0.44}$	50/50
RSF	SE	16.6$_{\pm 0.49}$†	**7.5**$_{\pm 0.08}$	27/30	**18.6**$_{\pm 1.29}$	**22.6**$_{\pm 1.05}$	18/20	**16.3**$_{\pm 0.41}$	170.1$_{\pm 22.18}$†	39/40	20.8$_{\pm 0.89}$	610.2$_{\pm 865.78}$	49/50
MTLR	SE	**14.9**$_{\pm 0.52}$	9.7$_{\pm 0.15}$†	30/30	**18.1**$_{\pm 1.65}$	23.0$_{\pm 0.72}$	19/20	**16.4**$_{\pm 0.42}$	**126.3**$_{\pm 4.56}$	40/40	**18.9**$_{\pm 1.25}$	68.9$_{\pm 1.00}$†	50/50
DeepSurv	SE	**14.9**$_{\pm 0.51}$	9.1$_{\pm 0.22}$	30/30	19.4$_{\pm 1.47}$	**22.7**$_{\pm 1.02}$	18/20	**16.3**$_{\pm 0.42}$	139.1$_{\pm 5.27}$†	37/40	20.5$_{\pm 0.73}$	**45.8**$_{\pm 0.63}$	49/50
DeepHit	CR	32.7$_{\pm 0.98}$†	20.0$_{\pm 0.31}$†	0/30	21.4$_{\pm 2.64}$†	27.3$_{\pm 0.79}$†	18/20	25.6$_{\pm 1.21}$†	196.5$_{\pm 4.00}$†	0/40	21.1$_{\pm 3.99}$	74.3$_{\pm 0.84}$†	50/50
DSM	CR	16.0$_{\pm 0.47}$†	17.0$_{\pm 1.92}$†	22/30	**18.3**$_{\pm 1.54}$	23.2$_{\pm 0.76}$†	15/20	16.5$_{\pm 0.35}$	231.5$_{\pm 15.51}$†	38/40	20.7$_{\pm 1.46}$	59.6$_{\pm 0.82}$	17/50
Hierarch.	ME	51.0$_{\pm 1.16}$†	101.9$_{\pm 233.61}$†	0/30	56.0$_{\pm 5.09}$†	43.5$_{\pm 0.76}$†	0/20	53.9$_{\pm 1.86}$†	280.5$_{\pm 5.47}$†	0/40	71.8$_{\pm 7.84}$†	92.2$_{\pm 0.80}$†	0/50
MENSA (Ours)	ME	**15.1**$_{\pm 0.59}$	**8.3**$_{\pm 0.09}$	21/30	18.8$_{\pm 1.59}$	**22.6**$_{\pm 0.68}$	18/20	16.6$_{\pm 0.41}$	**124.1**$_{\pm 2.14}$	39/40	22.0$_{\pm 2.33}$	63.1$_{\pm 2.88}$	10/50

TABLE 4.4: Comparison between the proposed model and state-of-the-art baselines in terms of space complexity. d: number of features, K: number of events, B: number of discrete time bins, M: levels of granularity, Ψ: number of mixture distributions, P: number of states. Table taken with permission from [48].

Model	Single-event	Competing risks	Multi-event
DeepSurv	$\mathcal{O}(d)$	$K\mathcal{O}(d)$	$K\mathcal{O}(d)$
DeepHit	$\mathcal{O}(d+B)$	$\mathcal{O}(d+KB)$	$K\mathcal{O}(d+B)$
MTLR	$\mathcal{O}(Bd)$	$\mathcal{O}(KBd)$	$K\mathcal{O}(Bd)$
DSM	$\mathcal{O}(\Psi+d)$	$\mathcal{O}(K\Psi+d)$	$K\mathcal{O}(\Psi+d)$
Hierarch.	$\mathcal{O}(MB+d)$	$\mathcal{O}(MKB+d)$	$\mathcal{O}(MKB+d)$
MENSA (Ours)	$\mathcal{O}(P\Psi+d)$	$\mathcal{O}(P\Psi+d)$	$\mathcal{O}(P\Psi+d)$

Figure 4.2 compares MENSA with deep learning baselines in terms of trainable parameters and floating-point operations (FLOPs) as a function of the number of features. FLOPs are computed using the fvcore library[2]. In the single-event setting, MENSA requires fewer parameters and FLOPs than Hierarchical, MTLR, and DeepHit as the number of covariates increases, while using slightly more resources than DeepSurv and DSM due to architectural differences. In the competing-risks scenario, MENSA outperforms all models except DSM, which has a marginally lower overhead; beyond 100 features, however, the differences become negligible. In the multi-event case, many baseline models (in their current implementations) must be trained separately for each event, requiring K separate trainings to make predictions for K events. Here, MENSA achieves a substantially lower parameter count and FLOPs than all baselines, including the Hierarchical model, which is itself designed for multi-event data. In total, MENSA is more computationally efficient – both in parameters and FLOPs – than five other deep learning baselines in the multi-event scenario.

[2]Meta Research's fvcore library, available at `https://github.com/facebookresearch/fvcore` (September 1, 2025).

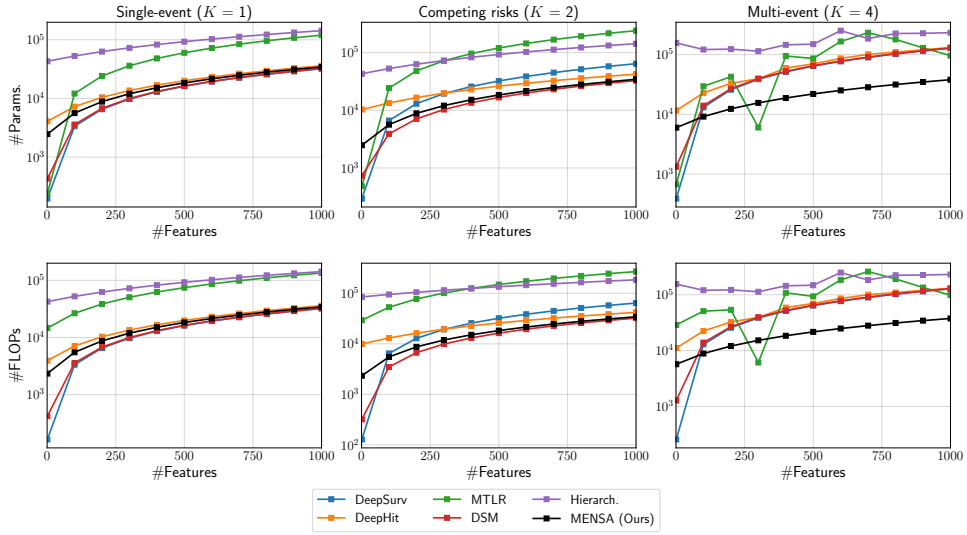


FIGURE 4.2: Number of features vs. number of trainable parameters (upper row) and number of floating-point operations (lower row) for single-event ($K = 1$), competing-risks ($K = 2$) and multi-event ($K = 4$) scenarios. Figure taken with permission from [48].

We perform two ablation studies that assess (1) the effectiveness of training events jointly in a single model and (2) the impact of the trajectory-based likelihood term. Table 4.5 presents the results for joint versus independent training. In both datasets, independent training slightly improves the global metrics (*e.g.*, global CI, AUC and mMAE) and D-calibration, but the local CI (within-patient event ordering) decreases, notably in Rotterdam (2.69% decrease) and marginally in PRO-ACT (0.19% decrease). Table 4.6 presents the results for training with/without the trajectory-based likelihood term. In Rotterdam, it improves local CI with only negligible shifts in global CI, AUC and mMAE – *i.e.*, small differences relative to the variability. In EBMT, it is consistently favorable for discrimination and prediction error, but D-calibration increase when the loss is removed. Overall, the loss appears to help inter-event predictions.

TABLE 4.5: Ablation study comparing joint vs. independent training of MENSA on two multi-event datasets. Results are averaged over 10 experiments, with changes shown relative (%) to the baseline. Table taken with permission from [48].

Dataset	Training	CI_G	CI_L	AUC	IBS	mMAE	D-Cal
Rotterdam	Joint	69.45±1.25	90.24±1.36	73.88±2.33	18.78±1.59	22.62±0.71	18/20
	Indep.	69.82±1.31 (+0.53)	87.81±1.74 (-2.69)	74.57±2.10 (+0.93)	18.84±1.67 (+0.32)	22.48±0.67 (-0.63)	20/20
PRO-ACT	Joint	70.87±0.50	76.07±0.51	77.12±1.20	16.60±0.52	124.37±3.44	36/40
	Indep.	70.98±0.64 (+0.15)	75.92±0.88 (-0.19)	77.29±0.75 (+0.23)	16.54±0.42 (-0.34)	122.78±2.33 (-1.27)	38/40

TABLE 4.6: Ablation study comparing training MENSA with/without the trajectory-based loss on two multi-event datasets. Results are averaged over 10 experiments, with changes shown relative (%) to the baseline. Table taken with permission from [48].

Dataset	$\mathcal{L}_{trajectory}$	CI_G	CI_L	AUC	IBS	mMAE	D-Cal
Rotterdam	With	69.45±1.25	90.24±1.36	73.88±2.33	18.78±1.59	22.62±0.71	18/20
	Without	69.54±1.22 (+0.13)	89.94±1.45 (-0.34)	74.16±2.09 (+0.38)	18.79±1.72 (+0.09)	22.58±0.71 (-0.18)	18/20
EBMT	With	55.26±1.24	68.73±0.56	55.66±1.47	22.13±2.30	62.98±2.97	9/50
	Without	54.93±1.15 (-0.6)	68.39±0.56 (-0.5)	55.53±1.43 (-0.24)	22.36±2.59 (+1.04)	63.76±3.90 (+1.25)	14/50

4.1.4 Discussion

We have presented MENSA, a novel method for multi-event survival analysis that jointly learns the time-to-event distribution for multiple events, which may be competing or co-occurring. Across four multi-event benchmark datasets, MENSA is competitive with strong deep learning baselines in both discrimination and calibration tasks, while achieving performance on par with boosted tree methods. Compared to the Hierarchical survival model – the main deep learning approach for multi-event survival prediction – MENSA delivers substantial gains across all evaluation metrics. In addition, MENSA is more computationally efficient in the multi-event setting, as single-event baselines must train a separate model for each event.

This work has limitations. First, MENSA assumes conditional independent censoring (see Definition 6), but this assumption may fail if important covariates are missing. Second, parametric survival models, such as MENSA, assume specific survival time distributions (such as Weibull), which often require more data to train efficiently. In contrast, semiparametric models, such as CoxPH and DeepSurv, make fewer assumptions about the distribution and are generally less sensitive to dataset size. Third, the current implementation of MENSA operates only on static covariates, which restricts its applicability in settings with time-varying covariates.

4.2 Prediction of Functional Decline in ALS Patients

4.2.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease that leads to the progressive loss of upper and lower motor neurons in the central nervous system. Most individuals with ALS die within 2-5 years of onset [114], primarily due to respiratory failure complications [115]. The exact causes of ALS remain unclear, and current treatments focus primarily on extending survival and enhancing the patient's quality of life. Due to its heterogeneous and unpredictable progression, as well as significant variation in clinical outcomes [116], determining the most effective treatment approach is often challenging. This uncertainty also complicates the timing of medical interventions, such as noninvasive ventilation, and the assessment of whether specific treatments can slow disease progression. As a result, there is an urgent need for tools that can accurately predict disease progression, support treatment decisions, and improve patient care.

In this work, we introduce a novel approach for predicting when an ALS patient is likely to experience each of five common types of functional decline, as defined by the ALSFRS-R protocol [117]. Specifically, we consider the onset of significant impairment (score ≤ 2) as an event with respect to five common functions: speaking, swallowing, handwriting, walking and breathing. These events are clinically distinct but not mutually exclusive, and may share underlying patterns in patient characteristics. For example, a patient with limb-onset ALS, Mr. Smith, may face walking and writing difficulties earlier than Mr. Johnson, who has bulbar-onset ALS and may instead experience earlier impairments with regards to speaking and swallowing. Our method aims to assist newly-diagnosed ALS patients by addressing critical clinical questions, such as:

- How much longer will I be able to speak clearly enough to be understood?
- How much longer will I be able to eat most solid foods without difficulty?

- How much longer will I be able to write a shopping list that my wife can read?
- How much longer will I be able to walk at a normal pace and keep my balance?
- How much longer will I be able to breathe normally during daily activities?

We formulate this task as a multi-event survival problem: given a patient instance (*e.g.*, a baseline profile of Mr. Smith), the model estimates the time until the occurrence of each of the specified functional decline events. These predictions are expressed as individual survival distributions (ISDs) [27], which represent the probability of each event occurring at any given time after the baseline. An overview of the proposed approach is illustrated in Figure 4.3.

The work described in this section is published as [49].

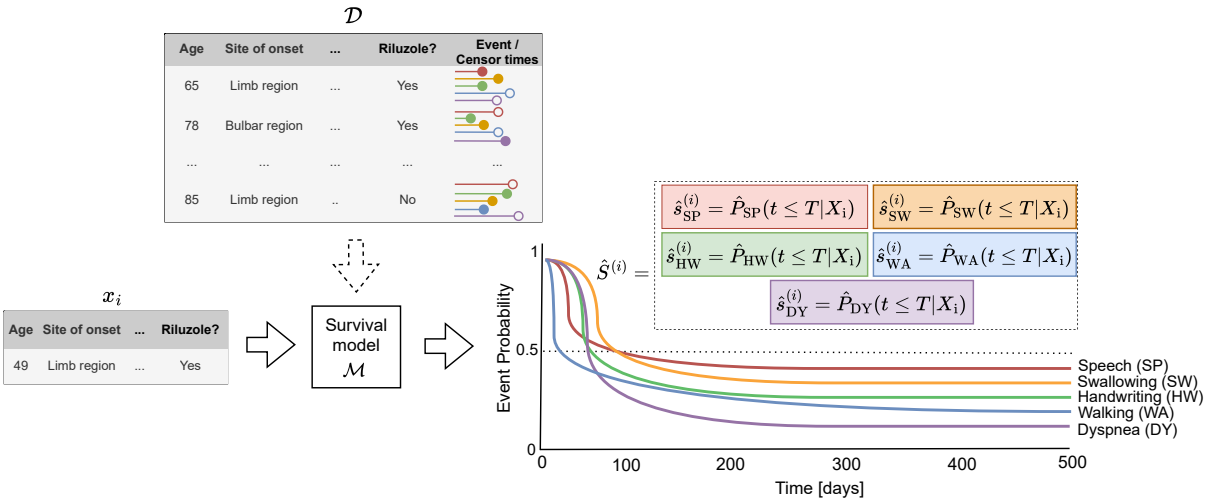


FIGURE 4.3: Overview of the proposed approach. Historical ALS patient data are compiled into a dataset $\mathcal{D}$ consisting of N training instances and d covariates, along with time-to-event information – represented as filled circles for events and hollow circles for censored cases – measured from the patient's initial visit (the "baseline time"). We focus on five distinct but related events: *Speech*, *Swallowing*, *Handwriting*, *Walking*, and *Dyspnea*. Using covariates recorded at baseline and the corresponding event outcomes, we train a survival model $\mathcal{M}$ to estimate an ISD for each of these five events for a new patient $\boldsymbol{x}_i$, denoted by $\hat{S}^{(i)}$. Each ISD describes the probability of the corresponding event occurring after t days from baseline, for all $t > 0$. These ISDs can be used to estimate expected event times for $\boldsymbol{x}_i$, such as by identifying the point at which each survival curve crosses the 50% threshold. Figure taken with permission from [49].

4.2.2 Materials and Method

For our empirical analyses, we use the publicly-available Pooled Resources Open-Access ALS Clinical Trials (PRO-ACT) dataset [118], the largest ALS dataset to date, comprising data from over 11,600 patients enrolled in 23 clinical trials. The dataset includes demographic details, lab results, medical records, and family history. During each trial, patients undergo regular follow-up visits until the trial ends or the patient can no longer participate due to illness or another cause. At each visit, various covariates are recorded, including the total ALSFRS-R score, age, sex, region

of onset (*e.g.*, bulbar or limb), mean FVC score, time since diagnosis, progression rate[3], and Riluzole usage. For our study, we use only the covariates collected at each patient's baseline visit.

To define our events of interest, we use the ALSFRS-R scale recorded at each follow-up visit. We focus on five specific functional decline events: *Speech*, *Swallowing*, *Handwriting*, *Walking*, and *Dyspnea*. We annotate an event as occurring when a patient receives a score of 2 or less on the corresponding ALSFRS-R item at any post-baseline visit. Censoring may occur due to death, inability to perform the assessment, or early withdrawal from the trial. To concentrate on early-stage progression and reduce the effects of high censoring, we restrict follow-up to a maximum of 500 days. Patients scoring 2 or lower at baseline on any of the five assessments are excluded, as the corresponding loss of function occurred before study entry. We also remove patients with no ALSFRS-R records, yielding a final dataset of 3,220 patients. Table 4.7 provides a summary of the dataset, and Figure 4.4 illustrates the distribution of event and censoring times.

TABLE 4.7: Overview of the dataset. "SP", "SW" "HA", "WA" and "DY", are the *Speech*, *Swallowing*, *Handwriting*, *Walking* and *Dyspnea* events, respectively. Event distribution indicates the percentage of uncensored instances of each type in the dataset. N: sample size, d: number of covariates, K: number of event types, T: observed time (of the uncensored instances) for any event in days. Table taken with permission from [49].

Dataset	N	d	K	T_{min}	T_{max}	T_{mean}	Event distribution (%):
PRO-ACT	3,220	8	5	1	498	130	SP: 37.8, SW: 31.9 HA: 50.3, WA: 60.6 DY: 27.0

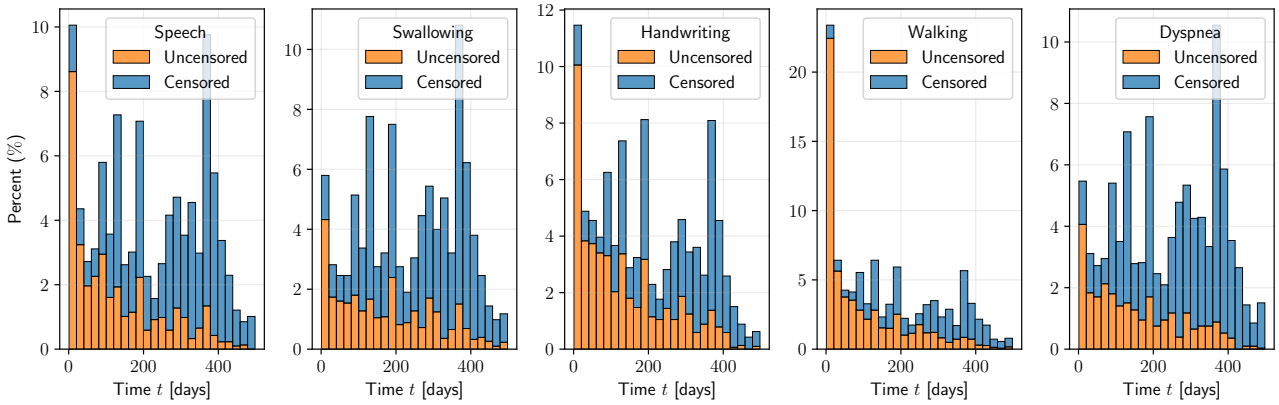


FIGURE 4.4: Distribution of uncensored and censored times in the PRO-ACT dataset for the five events. Figure taken with permission from [49].

We implement several survival models to validate the feasibility of our approach. As single-event models, we implement the traditional CoxPH model [23], Random Survival Forests [58], DeepSurv [28] and MTLR [88]. As a multi-event model, we implement our MENSA model [48]. We evaluate the survival models in the PRO-ACT dataset with different random seed to ensure variability. In each experiment, we follow [113] and first split the dataset into training, validation, and test sets in a stratified manner with respect to event indicators and event times across the K

[3]Calculated as $(48 - \text{ALSFRS-R score})/(\text{days since diagnosis}/30)$

events given a random seed. This allocates 70% of the data to the training set, 10% to the validation set and 20% to the test set. After splitting, we impute missing values using sample mean for real-valued covariates or mode for categorical covariates based on the training set only. We encode categorical covariates using a one-hot encoding strategy and apply a z-score data normalization to speed up the training process and improve stability. No outlier detection is performed.

4.2.3 Experiments and Results

We present our empirical results (Table 4.8 and Figure 4.5) in terms of three key aspects with respect to performance: discrimination, absolute/squared error, and calibration. Below, we discuss the results and critically reflect on their implications.

Discrimination: In Table 4.8, *Speech* and *Swallowing* have the best Harrell's/Uno's C-indices (around 74-76% and 70-71%). *Handwriting* and *Walking* are harder to discriminate (around 67-69%). RSF excels at *Dyspnea* with 78.23%/78.95%. DeepSurv, MTLR, and MENSA consistently perform well, with MENSA dipping for *Walking* and *Dyspnea*.

Absolute/squared error: In Table 4.8, *Swallowing* shows consistently low IBS (12-13) and mMAE (110-150 days for all models except DeepSurv, at 179 days). For *Dyspnea*, IBS is moderate (11.9–13.8), but RSF has low IBS and high mMAE (approx. 850$\pm$471 days), unlike other models (approx. 200-400 days). This suggests the squared error of the survival probabilities does not mimic the absolute error of the time-to-event estimates. MENSA achieves reasonable mMAE across all events (119-189 days) and performs well for *Swallowing*, similar to MTLR.

Calibration: In Table 4.8, *Walking* is a consistent calibration challenge, with predictions deviating from observed frequencies. RSF shows perfect D-calibration (10/10) across all events, highlighting robust performance. DeepSurv and MTLR also perform well, achieving D-calibration in most experiments. MENSA is D-calibrated in all 10 runs for *Swallowing* and *Dyspnea*, but slightly lower for *Speech* (9/10), *Handwriting* (9/10), and *Walking* (6/10).

Figure 4.5 compares the prediction error, measured in days, between covariate-based models and the population-level Kaplan–Meier (KM) estimator [20]. The KM curves are calculated per event using both observed and censored times, without accounting for individual covariate information. Across all five events, covariate-based models generally produce more accurate time-to-event estimates than the KM estimator, although performance varies between models. This underscores the value of incorporating patient-specific features when predicting functional decline, rather than relying solely on population-level trends.

Our method can predict personalized disease trajectories as K ISDs, where $K = 5$ corresponds to the five events. Figure 4.6 illustrates these ISDs for a representative patient (*e.g.*, Mr. Smith), covering the period from baseline ($t = 0$) up to 500 days. Mr. Smith is a 72-year-old male with limb-onset ALS, who entered the study with early speech difficulty and leg weakness, but retained functionality in all five areas (ALSFRS-R scores above 2). As an illustrative scenario, suppose his neurologist anticipates that walking impairment will manifest early. The model supports this, predicting *Walking* as his earliest event (median of approximately 80 days), which could, for example, prompt early provision of a walking aid. Indeed, Mr. Smith required walking support after approximately 45 days. The model is less accurate for *Speech*, which Mr. Smith experienced first (after approximately 7 days) but which is predicted at approximately 300 days. Since he entered the study with speech difficulty, this event occurred almost immediately after baseline, illustrating that the

TABLE 4.8: Prediction results, averaged over 10 independent experiments (± SD). D-calibration counts the number of times the respective model was D-calibrated. Harrell's CI, Uno's CI and IBS results are multiplied by 100 for better readability. Table taken with permission from [49].

Model	Event	Harrell's CI ↑	Uno's CI ↑	IBS ↓	mMAE ↓	D-Cal ↑
CoxPH	Speech	74.52±1.39	70.44±3.34	14.41±0.43	206.85±9.88	(10/10)
	Swallowing	75.75±1.53	70.96±4.58	12.58±0.72	146.15±25.50	(10/10)
	Handwriting	67.57±0.92	63.75±2.29	16.73±0.70	141.85±7.40	(10/10)
	Walking	68.67±1.20	67.15±1.40	16.60±0.50	145.04±4.50	(7/10)
	Dyspnea	68.54±1.71	66.99±6.30	13.79±0.41	235.46±27.57	(10/10)
RSF	Speech	75.05±1.13	70.08±2.75	14.69±0.44	164.51±19.12	(10/10)
	Swallowing	75.53±1.46	69.65±2.98	13.02±0.74	111.71±11.46	(10/10)
	Handwriting	67.38±1.39	63.70±2.05	16.87±0.57	146.97±6.88	(10/10)
	Walking	68.44±1.47	66.68±1.54	16.78±0.56	145.98±4.63	(10/10)
	Dyspnea	78.23±1.07	78.95±4.01	11.89±0.33	850.45±471.39	(10/10)
DeepSurv	Speech	74.18±3.05	70.77±3.57	14.36±0.89	232.15±28.48	(10/10)
	Swallowing	75.61±1.37	70.67±5.52	12.48±0.65	178.77±39.68	(10/10)
	Handwriting	67.12±1.08	63.72±2.15	16.81±0.84	151.83±18.59	(9/10)
	Walking	69.53±0.93	68.44±1.11	16.23±0.51	148.17±6.60	(8/10)
	Dyspnea	71.72±3.99	70.61±7.60	13.11±0.61	384.96±284.17	(10/10)
MTLR	Speech	74.22±1.67	70.17±3.23	13.95±0.44	179.35±5.34	(10/10)
	Swallowing	75.13±1.43	70.63±5.41	12.34±0.59	122.93±19.52	(10/10)
	Handwriting	67.32±1.11	63.71±2.42	16.72±0.66	140.94±6.29	(9/10)
	Walking	68.97±1.16	67.71±1.06	16.30±0.51	149.93±10.99	(9/10)
	Dyspnea	66.54±2.22	65.90±6.85	13.50±0.50	203.36±24.57	(10/10)
MENSA	Speech	74.21±1.11	70.56±3.47	14.53±0.41	181.81±7.92	(9/10)
	Swallowing	74.94±1.36	70.85±5.20	12.61±0.56	119.16±7.12	(10/10)
	Handwriting	66.46±0.90	63.80±2.65	17.10±0.66	141.97±5.70	(9/10)
	Walking	68.63±1.29	66.94±1.50	17.04±0.61	149.50±6.49	(6/10)
	Dyspnea	66.56±1.77	64.40±7.07	13.78±0.30	189.43±24.28	(10/10)

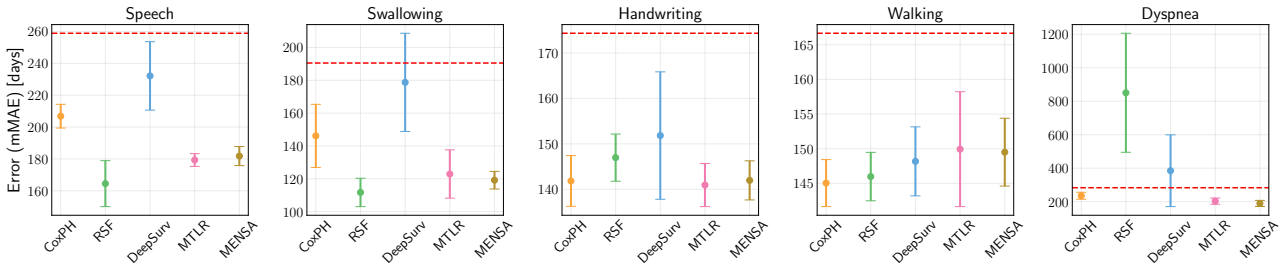


FIGURE 4.5: The mMAE (in days) as a function of covariate-based models (x-axis) in the PRO-ACT test set, with error bars representing empirical 95% confidence intervals. The horizontal dashed line is the KM estimator [20]. Lower is better. Figure taken with permission from [49].

median is only a point summary of the ISD and that individual event times can fall far in the tail of the predicted distribution.

Our method also supports individual-level counterfactual predictions [119], which explore hypothetical scenarios by altering specific patient covariates, to assess their predicted impact on functional decline. Such predictions can provide insights into

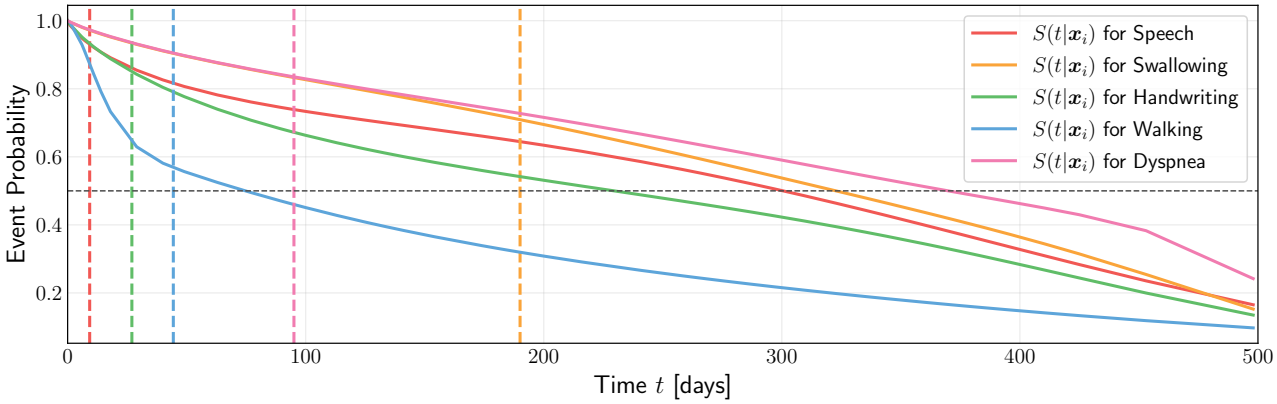


FIGURE 4.6: Predicted ISDs for Mr. Smith (i). The point where the survival curve intersects the dashed horizontal line at 50% indicates the predicted median time of event – so this predicts, for example, *Walking* at approximately 80 days, *Handwriting* at 240 days, and *Speech* at 300 days. The dashed vertical lines are Mr. Smith's actual time of event for the respective events – so *Speech* at approximately 7 days, *Handwriting* at 35 days, and *Walking* at 45 days. Figure taken with permission from [49].

disease mechanisms and help evaluate potential treatment strategies. Using MENSA, we demonstrate counterfactual predictions for Mr. Smith, based on test-set data from PRO-ACT. These predictions reveal how modifying individual risk factors could influence event probabilities. To validate broader trends, we stratify the test set based on the same covariate used in the counterfactual analysis and apply the log-rank test [120] to assess differences in survival outcomes between strata. We correct for multiple comparisons using the Bonferroni method, adjusting the significance level to $\alpha_{\text{adjusted}} = 0.01$ for the five events ($\alpha = 0.05/5$).

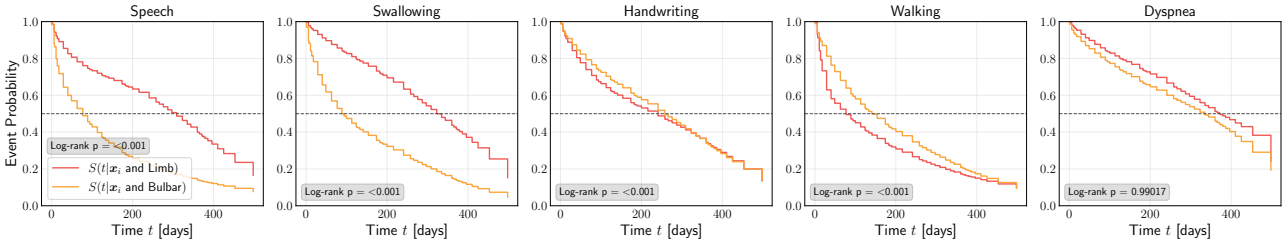


FIGURE 4.7: Predicted ISDs for Mr. Smith based on whether he has limb-onset or bulbar-onset. Figure taken with permission from [49].

As an example of a counterfactual prediction, we examine the *site of onset* covariate, indicating where the ALS symptoms first appear. Bulbar-onset and limb-onset are the most common types, with bulbar-onset generally linked to faster disease progression [121]. While Mr. Smith was diagnosed with limb-onset ALS, we explore how his predicted disease course would differ if he had been diagnosed with bulbar-onset ALS instead. Figure 4.7 shows his predicted ISDs for both cases. The ISDs for *Speech* and *Swallowing* differ significantly, suggesting a worse prognosis with bulbar-onset, whereas the ISDs for *Handwriting*, *Walking*, and *Dyspnea* are more similar. More results are available in [49].

4.2.4 Discussion

This study had three main objectives: (1) to evaluate whether covariate-based survival models can more accurately predict individual time to functional decline in ALS patients compared to population-level estimators, (2) to precisely estimate risk scores and event times, and (3) to investigate how counterfactual scenarios impact the likelihood of functional decline. Regarding the first objective, personalized predictions are crucial for tailoring treatment decisions. We demonstrate that covariate-based models – even relatively simple ones – consistently outperform the KM estimator in predicting time to functional decline. Unlike previous approaches, our method provides actual predicted times (in days) when the decline is likely going to happen. For the second objective, we thoroughly assess five survival models using metrics for discrimination, absolute/squared error, and calibration. Importantly, we find that strong discriminative ability does not always correspond to accurate time-to-event predictions, underscoring differences between ranking performance and absolute error. Lastly, for the third objective, we employ counterfactual predictions to reveal how modifying patient characteristics influences the risk and type of functional decline – a novel application in ALS research. For instance, for our example patient Mr. Smith, the predicted risk of functional decline for all five events showed no significant differences whether or not he was taking Riluzole.

The main limitation of this work is that all evaluated survival models assume independent censoring (see Definition 6). This means that the event time is assumed independent of the censoring time given the patient's covariates. However, ALS is a highly heterogeneous disease and progresses quickly. At this point, no study or dataset can fully capture all disease mechanisms, so there are likely to be unobserved confounders present. Clinical trials in ALS also face high dropout rates [122]. This phenomenon is likely associated with disease progression and subsequent functional decline; however, because functional decline is typically unobserved in patients who withdraw from the study, such dependent censoring may introduce a bias in the results.

Chapter 5

Evaluation under Censoring

Censoring is a defining characteristic of survival data, where the true event time may only be partially observed. Since censored observations appear in both training and test sets, evaluation metrics must properly account for censoring to yield unbiased estimates of predictive performance. However, in practice, evaluation metrics are often applied without verifying their underlying censoring assumptions or their alignment with the predictive objective.

This chapter addresses these challenges from two perspectives. First, we analyze common pitfalls in survival model evaluation, focusing on mismatches between modeling objectives, evaluation metrics, and censoring assumptions. Second, we introduce a new evaluation metric that enables evaluation under dependent censoring. Specifically, we propose a dependent Brier score that models the joint distribution of event and censoring times using an Archimedean copula.

5.1 Common Pitfalls in Model Evaluation

5.1.1 Introduction

Evaluating survival models is complicated by censoring, which causes event times to be only partially observed for some instances. Because censored observations appear in both training and test sets, evaluation metrics must both reflect the intended prediction objective and properly account for censoring in order to provide unbiased estimates of predictive performance. In practice, however, these considerations are often overlooked and leads to two frequent shortcomings in the literature:

- Evaluation metrics are often misaligned with the modeling objective. In a survey, we find that discrimination-based measures such as the C-index are commonly used as default metrics even when the goal is accurate time-to-event prediction or well-calibrated survival probabilities.
- Censoring assumptions required for these metrics are rarely stated or validated. Many studies implicitly assume random or independent censoring without justification, which can lead to biased estimates of predictive performance.

While the limitations of the C-index are well documented [39], its use remains widespread. We did a structured review of 92 methodological and application-oriented survival analysis papers published between 2023 and 2025 (see the appendix of [50]) which shows that approximately 72% rely on evaluation metrics that do not align with their stated objectives or do not explicitly address the censoring assumptions required for valid evaluation. Figures 5.1 and 5.2 summarize these findings.

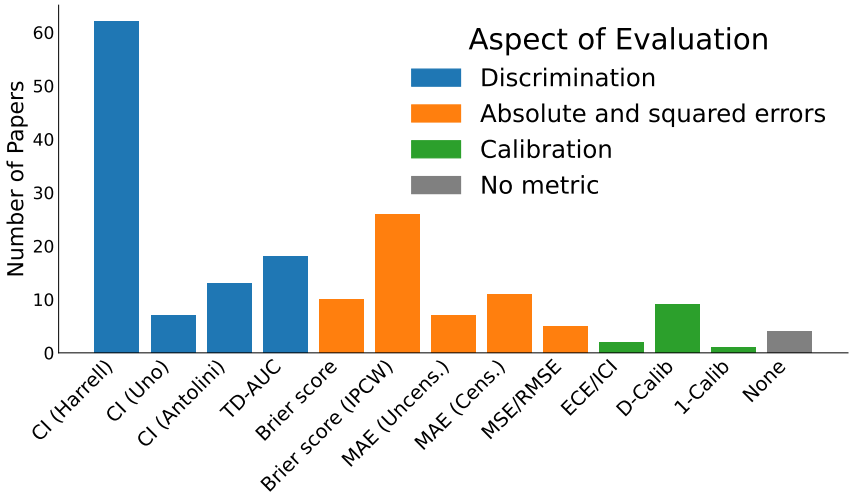


FIGURE 5.1: Overview of metric usage in recent survival analysis publications (2023-2025). Papers that report several metrics are counted once for each metric. Most papers rely on metrics that only assess a model's discrimination (*i.e.*, ranking) ability, not how well it can make accurate time-to-event estimates or provide calibrated survival probabilities. Figure taken with permission from [50].

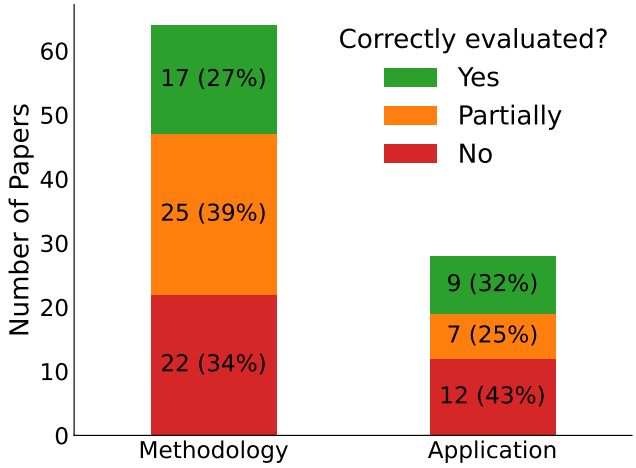


FIGURE 5.2: Overview of metric correctness in recent survival analysis publications. Each paper is labeled **Yes**, **Partially**, or **No** depending on whether the reported evaluation metrics align with the stated modeling objective and whether censoring assumptions are stated, justified or corrected for to make those metrics valid. We find that 73% of surveyed methodology papers and 68% of application-focused papers do not fully satisfy this criteria. Figure taken with permission from [50].

We therefore argue that a substantial amount of recent survival analysis research is evaluated incorrectly due to a systematic mismatch between modeling objectives, evaluation metrics, and the censoring assumptions those metrics rely on. To make

this mismatch explicit, we introduce a ladder hypothesis of model-metric consistency, which shows that the development of evaluation metrics lags behind the development of new survival models. Through controlled experiments, we demonstrate that commonly used metrics can become increasingly biased as censoring deviates from the common assumption of randomness. Our goal is not only to highlight these issues but also to provide practical guidance for selecting evaluation metrics that are aligned with both the research objective and the censoring mechanism.

The work described in this section is published as [50].

5.1.2 Objective-Metric Alignment

Censoring is not a fixed mechanism and may arise in different ways across datasets. After a model has been trained on censored data, it is typically evaluated on held-out test sets that also contain censored observations. In many real-world survival datasets, the proportion of censored instances is substantial. For example, in a 23-year prostate cancer study, approximately 68-80% of men did not die from prostate cancer and were therefore censored [123]. Our analysis suggests that the community does not always carefully consider how survival models are evaluated or which criteria are used to assess them. We illustrate this issue with three examples.

Steinberg et al. [124] introduced *MOTOR*, a foundation model trained on 55M patient records to predict time-to-event outcomes. Although the objective was accurate event-time prediction, the evaluation relied primarily on C-indices [36, 125], which measure only a model's ability to rank patients by risk. A model may achieve a high C-index while systematically calculating event times wrongly, since rankings can remain correct even when predicted times are inaccurate.

Zisser and Aran [126] proposed a Transformer-based survival model to predict patient-specific survival probabilities for acute lymphoblastic leukemia. Despite targeting time-to-event prediction, the model was evaluated exclusively using Harrell's C-index, creating a mismatch between the modeling objective and the evaluation metric.

Liu et al. [127] introduced HACSurv, a copula-based survival model designed to capture dependence between competing risks and censoring. However, the method was evaluated using standard metrics such as Antolini's C-index [101] and the integrated Brier score with inverse probability of censoring weighting (IPCW) [77], which assume random or independent censoring. This creates a mismatch between the modeling assumptions and the evaluation setting.

5.1.3 Aspects of Evaluation and Common Metrics

To help guide metric selection better, we propose five desiderata for evaluating survival models. These desiderata are desirable properties and formalize what it means for an evaluation metric to be appropriate for a given scientific goal.

D1 Be a proper scoring rule. A proper scoring rule is a loss function that is optimized when the predicted probabilities match the true underlying probability distribution (see the appendix of [50] for a formal definition). Such metrics reward truthful probabilistic predictions: models that better reflect the data-generating process should receive better scores.

D2 Be interpretable. A metric should be expressed in domain-relevant, human-understandable units (*e.g.*, days, weeks, months; or probabilities) and described using common, non-technical terminology. Its value should have a direct and

intuitive connection to model behavior, without requiring additional transformations or assumptions. This allows practitioners and non-technical stakeholders to readily understand what is being measured and how differences in the metric reflect differences in model performance.

D3 Be model-agnostic. A metric should not depend on model parameters or on properties of the covariate distribution $\Pr(X)$. This improves comparability and robustness: two models that produce identical predictions should receive identical scores. Metrics that depend on model-specific quantities may instead produce different evaluations simply because one model has more parameters or a different structure.

D4 Be sensitive to miscalibration. A metric should detect systematic discrepancies between predicted and observed survival probabilities. This allows one to identify models that consistently overestimate or underestimate risk. For example, a model might predict a 70% probability of surviving the next year for a cancer patient, while in reality only 50% of such patients survive.

D5 Be robust to censoring. A metric should handle censoring in a manner consistent with how censoring arises in the data. If censoring is random, such that censored individuals are equally likely to experience the event as those remaining under observation, the metric may rely on this assumption. Conversely, if censoring is dependent, the metric should relax this assumption and remain valid under the induced dependence.

TABLE 5.1: Overview of commonly-used survival metrics and which desiderata they satisfy. A checkmark (✓) indicates that the metric fully satisfies the desideratum, a partial mark (∼) indicates partial satisfaction, and a cross (✗) indicates that the metric does not satisfy the desideratum. For the interpretability desideratum, we rank the metrics from 1 (most interpretable) to 7 (least interpretable), assigning better ranks to metrics whose units and values are more directly understandable. The log-likelihood (LL) is included merely for comparison, as it is typically not used as a final evaluation metric. Table taken with permission from [50].

Desiderata	CI (Harrell)	CI (Uno)	CI (Antolini)	IBS	MAE	D-Cal	LL
Proper scoring rule (D1)	✗	✗	✗	✓	∼	✗	✓
Interpretable (D2)	#2	#3	#5	#6	#1	#4	#7
Model-agnostic (D3)	∼	∼	∼	✓	✓	✓	✗
Sensitive to miscalibration (D4)	✗	✗	✗	✓	✗	✓	✓
Robust to censoring (D5)	✗	∼	✗	∼	✗	✗	∼

Table 5.1 summarizes several commonly used survival evaluation metrics and which of the desiderata introduced above they satisfy. Rather than reintroducing the metrics (see Chapter 2), here we briefly describe the aspects of model performance they evaluate and how they relate to the proposed criteria.

Discrimination. Discrimination metrics assess how well a model ranks individuals by risk. The most widely used example is Harrell's C-index, with variants such as Uno's C-index and Antolini's C-index. While these metrics measure a model's ability to correctly order individuals by risk, they do not evaluate the accuracy of predicted event times or survival probabilities. Consequently, they are not proper scoring rules and are insensitive to calibration. Some variants partially correct for censoring under random or independent censoring assumptions, but all commonly used C-index variants remain biased under dependent censoring.

Absolute and squared error metrics. Error-based metrics evaluate the discrepancy between predicted and observed outcomes. The integrated Brier score (IBS) measures the accuracy of predicted survival probabilities over time and is a proper scoring rule under random censoring when estimated with IPCW. The mean absolute error (MAE) evaluates the accuracy of predicted event times. IBS is sensitive to miscalibration because it penalizes inaccurate survival probabilities, whereas MAE evaluates only point predictions. Both metrics typically rely on assumptions about the censoring mechanism and can become biased under dependent censoring.

Calibration metrics. Calibration metrics assess whether predicted survival probabilities match observed event frequencies. For example, 1-Calibration evaluates calibration at a specific time horizon, while D-Calibration assesses calibration across the entire survival curve. These metrics are sensitive to miscalibration and generally model-agnostic, but they are not proper scoring rules and, when estimated using KM-based methods, are not robust to dependent censoring.

Together, these comparisons illustrate that commonly used metrics emphasize different aspects of predictive performance and satisfy the proposed desiderata to varying degrees. As a result, the choice of evaluation metric should depend on both the modeling objective and the censoring assumptions one is willing to defend.

5.1.4 Model-Metric Consistency

We introduce a guiding principle called the *Ladder Hypothesis of Model-Metric Consistency*. The central idea is that a survival model should be evaluated using metrics whose validity relies on the same (or fewer) censoring assumptions as the model itself. If the assumptions underlying the evaluation metric are stronger than those of the model, the resulting performance estimates may be biased and therefore unreliable.

> **Ladder Hypothesis of Model-Metric Consistency.** A model can be considered trustworthy only when it is evaluated with metrics whose own validity rests on the same or weaker set of statistical assumptions that also govern the model. Conversely, an evaluation metric is itself trustworthy only when any auxiliary models or estimators on which it relies satisfy those same assumptions.

Figure 5.3 illustrates this idea as an assumption ladder. As survival models are developed under progressively weaker censoring assumptions (moving from random to independent and eventually dependent censoring), the evaluation metrics used in practice have not evolved at the same pace. This creates a gap between modeling assumptions and evaluation procedures, leading to systematic model-metric mismatch.

To illustrate this phenomenon, we conduct a controlled simulation in which a predictive model (CoxPH) is held fixed while the censoring mechanism is varied. Event times are generated from the same covariate-dependent distribution across all settings, while censoring is introduced under three regimes: random censoring, independent censoring, and dependent censoring with increasing levels of dependence. We then evaluate the same predicted survival curves using several commonly used metrics, including variants of the C-index and the IBS. Oracle metrics computed using the true event times provide a reference for the model's actual performance. Figures 5.4 and 5.5 show that the estimation error of these metrics increases as censoring deviates from the assumptions under which the metrics were derived.

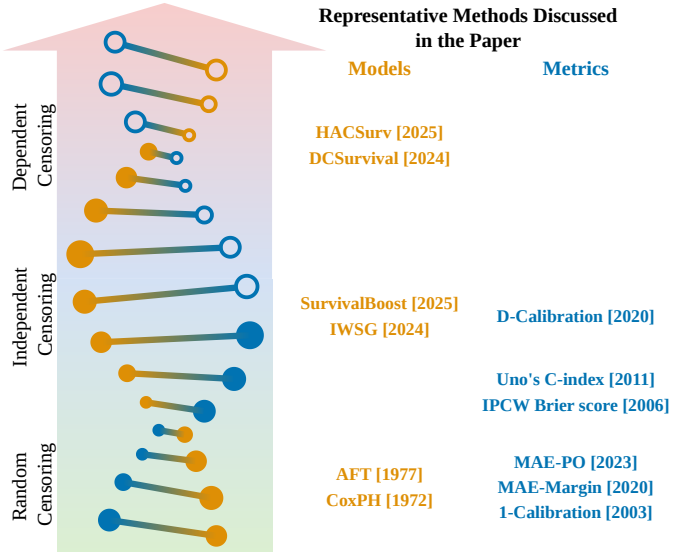


FIGURE 5.3: The double-helix ladder, showing how survival models (orange) and evaluation metrics (blue) have developed as the censoring assumptions become weaker (moving upward). Filled dots mark settings where valid methods are available, and hollow dots mark open gaps. Models have begun to address dependent censoring, whereas metrics have not yet fully addressed independent censoring. The growing gap between the two strands illustrates a systematic mismatch between models and metrics along the ladder. Figure taken with permission from [50].

In particular, both discrimination and error-based metrics become increasingly biased under dependent censoring. Since errors are measured relative to the oracle metrics, which capture genuine changes in the fitted model's performance, this bias reflects the evaluation itself rather than the model. This illustrates that apparent improvements or degradations in performance may reflect evaluation bias rather than genuine differences in predictive ability.

These findings highlight a broader issue in survival analysis: models and evaluation metrics are often developed under different censoring assumptions. Historically, many classical survival models and evaluation procedures were derived under random censoring [50]. Modern methods increasingly relax this assumption and instead rely on independent censoring given covariates. At the same time, some recent work attempts to model dependent censoring directly using approaches such as copula-based models or causal inference techniques. Despite this progress in modeling, most commonly used evaluation metrics-including the C-index, Brier score, MAE, and calibration-based tests—remain valid only under random or independent censoring assumptions. As a result, models designed to handle dependent censoring are often evaluated using metrics that are themselves biased in that setting [127]. This limitation motivates the development of evaluation metrics that remain valid under dependent censoring, which is the focus of the next section. Detailed experimental results and methodological analysis supporting this argument are provided in the appendix of [50].

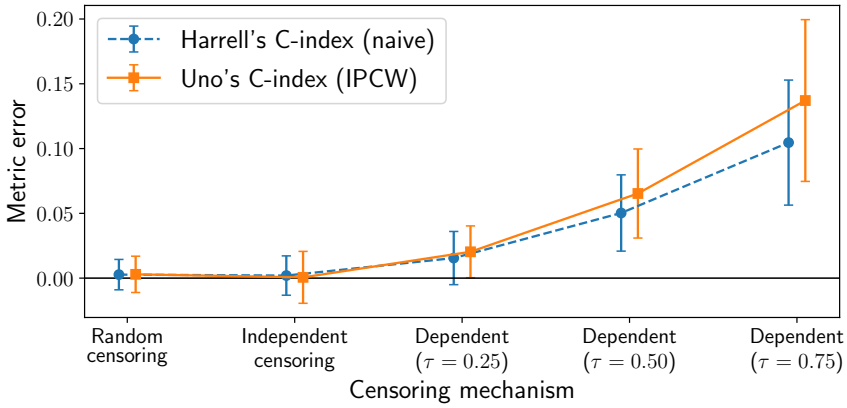


FIGURE 5.4: Mean (± SD.) metric error of the C-index averaged over 100 random seeds. Censoring is random, independent, or dependent via a Clayton copula with increasing dependence strength (Kendall's τ), while the model class (CoxPH) is held fixed. Errors are measured relative to the oracle C-index computed from fully observed event times. The bias increases as censoring deviates from independence. Figure taken with permission from [50].

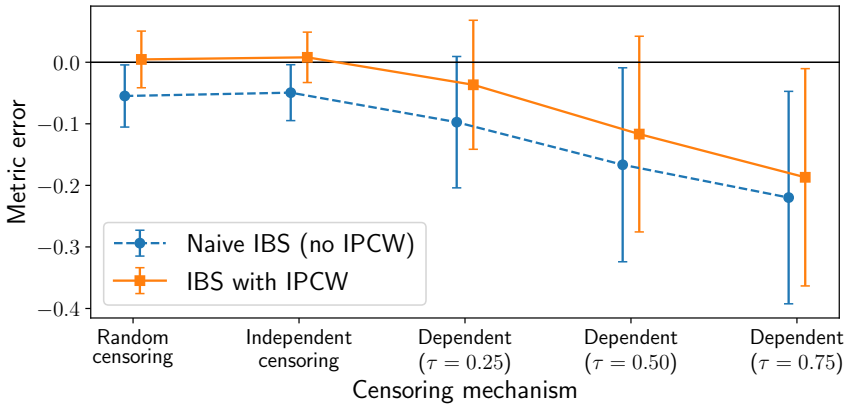


FIGURE 5.5: Mean (± SD.) metric error of the integrated Brier score (IBS) averaged over 100 random seeds under increasing dependence between event and censoring times. Errors are measured relative to the oracle IBS. As dependence increases, the IPCW-based estimator becomes increasingly biased because its KM-based censoring correction assumes random censoring. Figure taken with permission from [50].

5.1.5 Practical Recommendations

We conclude with some practical recommendations for model evaluation.

Recommendation 1 (Stick to your goal). In application-oriented studies, the primary evaluation metric should reflect the study objective. Discrimination-based objectives (*e.g.*, prioritizing patients for treatment) call for ranking metrics such as the C-index or AUC variants. Objectives focused on accurate time-to-event prediction favor error-based metrics such as MAE or MSE, while decisions based on predicted survival probabilities are best assessed using calibration-oriented measures. When multiple metrics address the same aspect of performance, the desiderata in Table 5.1 can guide the choice.

Recommendation 2 (Evaluate from multiple angles). In methodology-oriented

studies (*e.g.*, developing a new general-purpose algorithm), we recommend reporting metrics from multiple evaluation aspects rather than relying on a single headline score. Many modern survival models predict full survival curves, but these curves are only partially observed and may be unreliable in the tail due to censoring. In addition to curve-based metrics, reporting well-chosen scalar summaries – such as landmark survival probabilities, predicted median survival, or restricted mean survival time – can provide more stable and practically meaningful results.

Recommendation 3 (Strive for consistency). Evaluation results should always be interpreted relative to the censoring assumptions underlying the chosen metric. We therefore recommend explicitly stating these assumptions and, ideally, selecting metrics whose validity aligns with them. When such alignment cannot be guaranteed, researchers should acknowledge the limitation and perform robustness checks or sensitivity analyses to ensure that conclusions are not artifacts of the evaluation procedure itself.

5.2 Overcoming Dependent Censoring in Model Evaluation

5.2.1 Introduction

As we have discussed, survival analysis differs from standard regression because outcomes may be censored: for some individuals, we only observe that the event time exceeds a certain value. Most survival models and evaluation metrics rely on the assumption of independent censoring (Definition 6). Under this assumption, the censoring time does not convey additional information about the event time beyond the observed covariates [13]. However, this assumption is often violated in practice when censoring and event times share unobserved causes, leading to dependent censoring (Definition 8). While recent work has begun addressing dependent censoring in model estimation [64, 127, 128], its implications for model evaluation remain largely unexplored. In particular, widely used metrics such as the Brier score rely on the IPCW method (see Definition 13), typically estimated using KM or Cox-based models. These approaches implicitly assume independent censoring, and when the assumption fails the resulting weights can be misspecified, which can lead to biased performance estimates and misleading model comparisons. Figure 5.6 illustrates how dependent censoring can distort the KM estimate beyond the censoring time.

In this work, we propose a dependent Brier score that supports dependent censoring. Our approach models the joint distribution of event and censoring times using an Archimedean copula, enabling correction of the weighting mechanism without altering the marginal survival predictions. We further establish large-sample guarantees, proving consistency and asymptotic normality of the proposed estimator. To evaluate the method in realistic settings, we introduce a semi-synthetic data generation framework that induces controlled dependent censoring while preserving the original covariate structure. Across 12 survival datasets and multiple feature-selection regimes, our approach substantially reduces estimation error relative to standard IPCW-based evaluation. To our knowledge, this is the first survival evaluation metric that supports dependent censoring.

The work described in this section is published as [51].

5.2.2 Materials and Method

We introduce a modified Brier score that relaxes the standard independent censoring assumption by explicitly modeling dependence between event and censoring times

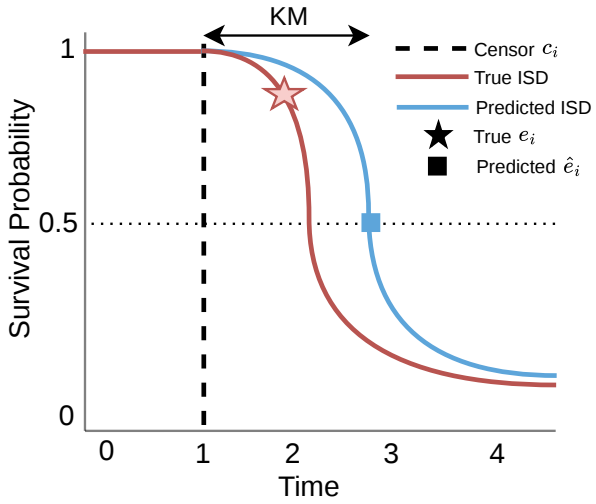


FIGURE 5.6: Evaluation under dependent censoring, where a patient is censored at time c_i shortly before the event time e_i due to factors related to the event but not observed in the data. This induces bias in the predicted ISD and the estimated event time $\widehat{e}_i$ obtained using the KM estimator, since the censored patient is treated as comparable to those still at risk. The true and KM-estimated survival curves coincide on the interval $[0, c_i]$, but diverge afterward due to the dependence between the event and censoring times. Figure taken with permission from [51].

using an Archimedean copula. The proposed approach has two main components: (1) the copula-based version of the KM estimator, which is used to obtain *margin-time* estimates for censored observations, and (2) a weighting scheme based on the censoring time. We first present the copula-based estimator for $\widehat{S}_E(t)$ and then describe how it can be used to compute margin-time estimates that ultimately lead to the proposed evaluation metric.

The Copula-Graphic estimator. Zheng and Klein [129] extended the nonparametric KM estimator to dependent censoring by introducing the parametric Copula–Graphic (CG) estimator. This relaxes the independent censoring assumption underlying KM by modeling dependence through a copula. Later, Rivest and Wells [130] derived an explicit form of the CG estimator under an *assumed copula* from the Archimedean family (*e.g.*, the Clayton [60] and Frank [62] copulas). In this formulation, dependence is governed by a single parameter θ through a generator function φ_θ [14]. Given an assumed Archimedean copula, the marginal survival function $\widehat{S}_E(t)$ can be estimated using the CG estimator:

$$\widehat{S}_E(t) = \varphi^{-1}\left(\sum_{i:\, t_i \le t,\, \delta_i = 1} \left[\varphi\left(\tfrac{n_i - 1}{n}\right) - \varphi\left(\tfrac{n_i}{n}\right) \right] \right), \tag{5.1}$$

where $\varphi(\cdot)$ denotes the generator of an Archimedean copula, and $n_i = \sum_{\ell=1}^{n} I(t_\ell \ge t_i)$ represents the number of individuals at risk at time t_i. If no event occurs before time t, then $\widehat{S}_E(t) = 1$, while the estimator is undefined for $t > t_{\max}$ [14]. When the copula corresponds to independence (*i.e.*, $\varphi(u) = -\log u$), the CG estimator reduces to the standard KM estimator on the interval $[0, t_{\max}]$.

Margin time. Once an estimate of the marginal survival curve $\widehat{S}_E(t)$ is available, it can be used to infer the expected event time for censored individuals. This

quantity, known as the *margin time*, represents the expected event time conditional on surviving beyond the censoring time. In other words, it provides the expected total event time for an individual whose event has not occurred prior to t_i. For an instance censored at time t_i, the margin time is computed as [27]:

$$e^{\text{margin}}\left(\widehat{S}_E, t_i\right) = \mathbb{E}[E \mid E > t_i] = t_i + \frac{\int_{t_i}^{\infty} \widehat{S}_E(t)\, dt}{\widehat{S}_E(t_i)}, \tag{5.2}$$

where $e^{\text{margin}}\left(\widehat{S}_E^{\text{CG}}, t_i\right)$[1] denotes the conditional expectation of the event time based on the CG survival estimate $\widehat{S}_E^{\text{CG}}$. As the copula parameter approaches independence ($\theta \to 0$), the CG estimator converges to the KM estimator, and the corresponding margin time coincides with the KM-based margin. Thus, the Brier score requires specification of a copula family and its parameter, and corrects for marginal dependence between E and C through their joint copula model, but does not require modeling interactions with covariates X.

Uncertainty weighting. Early censoring provides limited information about the eventual event time, whereas observations censored later in the study period offer stronger constraints on when the event might occur. To reflect this uncertainty, we introduce a weighting scheme that down-weights censored observations according to how early censoring occurs. A simple proxy for an individual's progress in the event process is the estimated cumulative distribution function, $\widehat{F}_E(c_i) = 1 - \widehat{S}_E(c_i)$, where $\widehat{S}_E(c_i) = \widehat{S}_{\text{CG}}(c_i)$. We therefore define a per-instance weight $w_i = \widehat{F}_E(c_i)$: observations censored early (small $\widehat{F}_E(c_i)$) receive smaller weights, while those censored later obtain weights close to one. These weights are applied only to censored observations with imputed event times; uncensored instances are assigned $w_i = 1$.

Definition. This leads to the proposed dependent Brier score, which is free from IPCW and supports dependent censoring:

$$\begin{aligned} \text{BS}_{\text{Dep}}\left(\mathcal{D}; \widehat{S}_E(\cdot), t^*\right) &= \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \left(\mathbb{1}[\tilde{e}_i > t^*] - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\right)^2, \\ \text{where} \quad \tilde{e}_i &= \begin{cases} e_i, & \delta_i = 1, \\ e_i^{\text{margin(CG)}}(t_i), & \text{otherwise.} \end{cases} \end{aligned} \tag{5.3}$$

To calculate the expectation of Equation 5.3 over all times, we simply take the integral from $t = 0$ to $t = t_{\max}$:

$$\text{IBS}_{\text{Dep}}\left(\mathcal{D}; \widehat{S}_E(\cdot)\right) = \frac{1}{t_{\max}} \cdot \int_0^{t_{\max}} \text{BS}_{\text{Dep}}\left(\mathcal{D}; \widehat{S}_E(\cdot), \tau\right) d\tau. \tag{5.4}$$

Theoretical properties. We briefly summarize the main theoretical guarantees of the proposed estimator; detailed assumptions and proofs are provided in the appendix of [51].

First, the margin time admits a natural probabilistic interpretation as the conditional expectation of the event time given survival beyond a landmark,

$$e^{\text{margin}}(c) = \mathbb{E}[E \mid E > c],$$

provided $S_E(c) > 0$ and the tail integral is finite. This interpretation justifies the use of margin time as a summary of the expected prognosis after the censoring time.

[1] For brevity, we write $e^{\text{margin}}(t_i)$ when the survival estimator is fixed.

Second, the plug-in estimator $\widehat{e}^{\text{margin}}(c)$ is consistent for a large-sample target $e^{*,\text{margin}}(c)$ induced by the assumed Archimedean copula model of Rivest and Wells [130]. In particular, if the CG estimator $\widehat{S}_E$ converges uniformly to S_E^* on a finite time horizon and $S_E^*(c) > 0$, then $\widehat{e}^{\text{margin}}(c)$ converges to $e^{*,\text{margin}}(c)$. When the copula model is correctly specified, this target coincides with the true margin time.

Finally, the estimator admits a first-order asymptotic expansion obtained via the delta method. Combining this expansion with the functional central limit theorem for the CG estimator yields an asymptotically normal limit with convergence rate $n^{-1/2}$. This result enables principled uncertainty quantification for margin-time estimates.

5.2.3 Experiments and Results

Synthetic experiments. We first validate the dependent Brier score on simulated data, where the true event times are known. For each random seed, we draw a fixed set of covariates $\boldsymbol{x}_i \sim \text{Unif}([0,1]^d)$ with $d = 10$. Both the event and the censoring times are generated from Weibull proportional hazards models whose risk functions are linear in the covariates, $g(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$, with fixed shape and scale parameters. We control the dependence between the two times through Kendall's τ: they are independent for $\tau = 0$, whereas for $\tau > 0$ a copula couples them, and the dependent metric is given the true copula family and parameter. We repeat the experiment for 10 random seeds, each with $N = 10{,}000$ samples and a fixed 70%/30% train/test split. Since the event times and the fitted CoxPH model are kept identical across all censoring settings, any change in evaluation error can be attributed to the censoring mechanism and not to the predictive model. We compute the oracle IBS on the test set from the event times before censoring is applied. Further details are given in the appendix of [51].

Figure 5.7 reports the estimation error for increasing values of Kendall's τ at a censoring rate of roughly 50%. As the dependence grows, the error of IBS-IPCW becomes larger and more variable, whereas IBS-Dep stays relatively stable and attains a lower or similar error. This is expected: once the independent censoring assumption no longer holds, the marginal inverse weights used by IPCW cannot correct for the selection induced by censoring, which introduces both bias and variance. The degradation is more pronounced under the Clayton copula. Because the copula links the survival functions S_E and S_C, the lower-tail dependence of the Clayton copula concentrates the dependence at late times, where both survival probabilities are small. This is also where few instances remain at risk and the inverse weights $1/\widehat{S}_C(t)$ are largest, which plausibly explains why the misspecified weights distort the estimate more under this copula.

Figure 5.8 instead fixes $\tau = 0.5$ and varies the censoring rate. Heavier censoring increases the error of both metrics, but the error of IBS-IPCW grows much faster, especially once more than about 80% of the instances are censored. In this regime, IPCW must rely on a small number of uncensored instances that receive very large inverse weights, which amplifies both bias and variance when censoring is dependent. IBS-Dep is less affected, because it accounts for the joint structure of (E, C) instead of relying on extreme reweighting.

Finally, Figure 5.9 examines how IBS-Dep behaves when the assumed copula does not match the one that generated the data. We study three cases: (1) a *wrong family*, where data generated with a Clayton copula are evaluated assuming a Frank copula (and vice versa) with the same Kendall's τ; (2) a *wrong dependence strength*, where the family is correct but data generated with $\tau = 0.5$ are evaluated with a

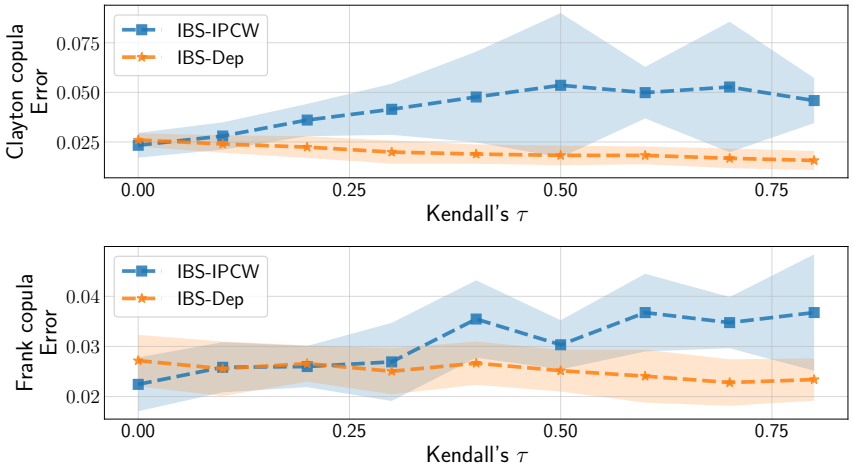


FIGURE 5.7: Absolute estimation error (mean ± SD over 10 experiments) of IBS-IPCW and IBS-Dep for increasing dependence between event and censoring times, measured by Kendall's τ. The censoring rate is held at 50% and the copula is assumed known. Lower is better. Figure taken with permission from [51].

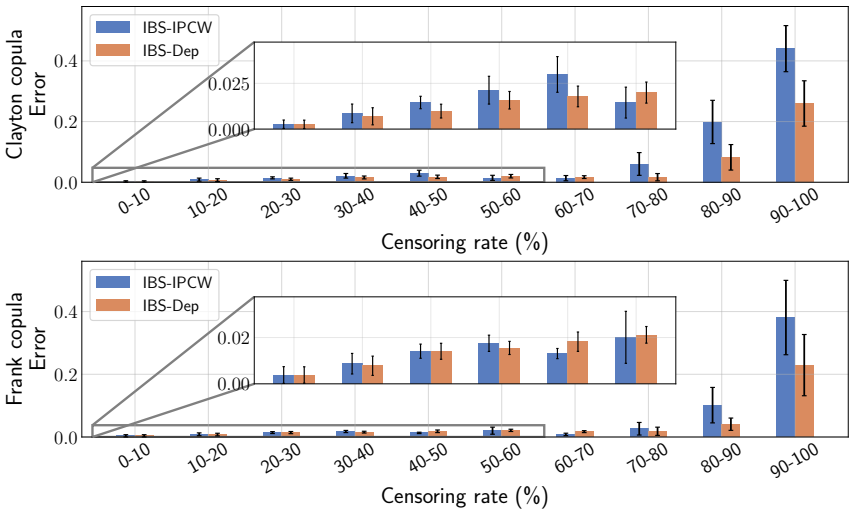


FIGURE 5.8: Absolute estimation error (mean ± SD over 10 experiments) of IBS-IPCW and IBS-Dep for increasing censoring rates, with Kendall's τ fixed at 0.5 and the copula assumed known. Lower is better. Figure taken with permission from [51].

range of assumed values τ'; and (3) a *wrong copula type*, where data generated with a (non-Archimedean) Gaussian copula are evaluated assuming a Clayton or Frank copula with the same τ. In all three cases, IBS-Dep is less affected by dependent censoring than IBS-IPCW whenever dependence is present ($\tau > 0$), even though the assumed copula is imperfect; without dependence, IBS-IPCW is slightly more accurate, as expected. We note, however, that IBS-Dep does not necessarily attain its lowest error at the true τ. The metric has a slight upward bias, which a stronger assumed dependence can partly offset. These experiments should therefore be read as a test of robustness, not as a way to recover the true dependence parameter.

Semi-synthetic experiments. Real datasets do not provide ground-truth event times due to censoring. To enable oracle evaluation, we construct semi-synthetic

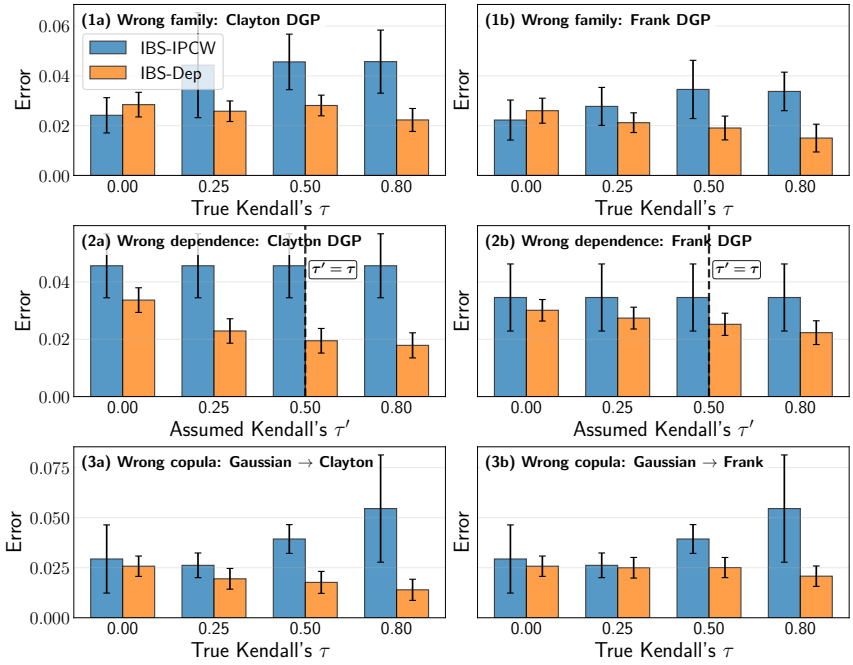


FIGURE 5.9: Absolute estimation error (mean ± SD over 10 experiments) of IBS-IPCW and IBS-Dep when the copula is misspecified, with the censoring rate held at 50%. From top to bottom, the rows correspond to a wrong copula family, a wrong dependence strength, and a wrong copula type. In the top two rows, the left and right columns use Clayton and Frank data-generating copulas, respectively; in the bottom row, data generated with a Gaussian copula are evaluated assuming a Clayton (left) or Frank (right) copula. Figure taken with permission from [51].

datasets (see Figure 5.10) in which event and censoring mechanisms are learned from a real dataset and used to resample outcomes while preserving the covariate distribution. Dependent censoring is induced through feature-selection strategies that omit information influencing both event and censoring processes. This produces datasets with realistic covariates while allowing oracle evaluation using the sampled event times.

We evaluate on 12 datasets spanning healthcare, clinical studies, and industry applications, including WHAS, METABRIC, GBSG, SUPPORT, MIMIC-IV, and several SEER cohorts. Dataset statistics are summarized in Table 5.2; additional preprocessing details are provided in the appendix of [51]. For FLCHAIN, the *chapter* feature (cause of death, defined only for deceased patients) was retained as a covariate. We consider three feature-selection regimes: (1) the original feature set, (2) top-k features selected by permutation importance, and (3) random subsets of features. Copula parameters are estimated by maximizing the likelihood under dependent censoring (see Equation 2.22). Survival predictions are produced using five learners: CoxPH, GBSA, RSF, DeepSurv, and MTLR.

Table 5.3 reports mean estimation errors on the semi-synthetic datasets, measured as the absolute deviation from the oracle IBS computed using the true event times. Across all 12 datasets and four feature-selection strategies, the proposed IBS-Dep metrics reduce estimation error relative to IBS-IPCW on average. The CG-based estimator achieves a 4-8% reduction, while the uncertainty-weighted variant yields

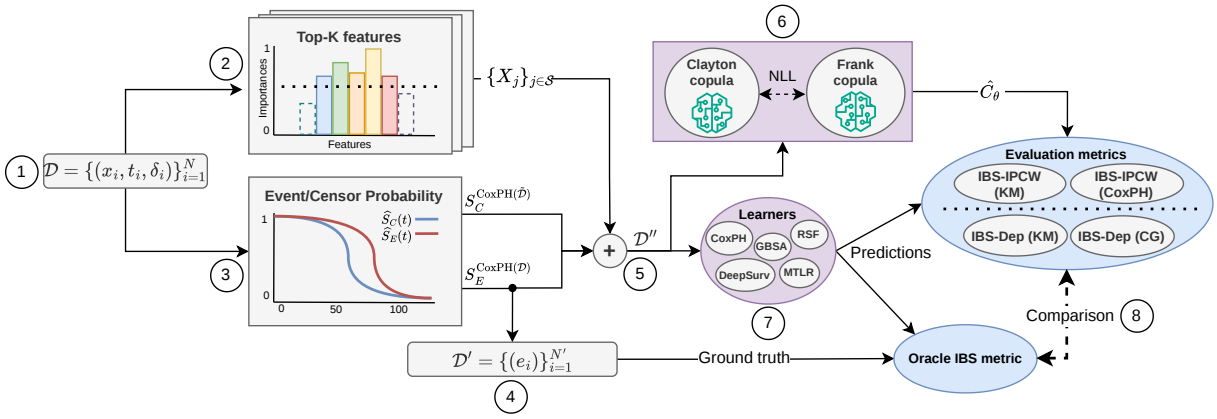


FIGURE 5.10: Overview of the evaluation pipeline. (1) The starting point is a real-world dataset $\mathcal{D}$. (2) A feature selection strategy (*e.g.*, top-k) produces a reduced set of features. At the same time, (3) we fit two CoxPH models to learn the event-time and censoring-time distributions, and (4) draw uncensored event times from the event distribution, which are kept as $\mathcal{D}'$ for oracle evaluation. (5) The semi-synthetic dataset $\mathcal{D}''$ is formed by pairing the selected features with event and censoring times drawn from the fitted distributions. We then (6) fit two Archimedean copula models to obtain $\widehat{C}_\theta$ and (7) train five survival learners on $\mathcal{D}''$. (8) The learners' test-set predictions are scored with IBS-IPCW, using KM or CoxPH estimates of the censoring distribution, and with IBS-Dep, using KM- or CG-based estimates under $\widehat{C}_\theta$, and each score is compared with the oracle IBS computed from $\mathcal{D}'$. Figure taken with permission from [51].

TABLE 5.2: Overview of the raw datasets. †Number of features before (after) one-hot encoding.

Dataset	#Instances	#Features†	%Censored
WHAS	500	14 (14)	57.0%
GBSG	686	8 (8)	56.4%
METABRIC	1,902	9 (9)	42.1%
Churn	1,958	12 (25)	52.4%
NACD	2,396	48 (48)	36.4%
FLCHAIN	7,871	9 (31)	72.5%
SUPPORT	8,873	14 (27)	32.0%
Employee	14,999	8 (17)	76.2%
MIMIC-IV	38,520	91 (103)	66.7%
SEER-brain	73,703	10 (10)	40.1%
SEER-liver	82,841	14 (14)	37.6%
SEER-stomach	100,360	14 (14)	43.4%

the largest gains (12-16%). The improvement becomes more pronounced under aggressive feature removal (*e.g.*, Random 25%). This does not imply that the KM-based IPCW estimator itself becomes more biased, since its bias depends on the joint distribution of (E, C) rather than the learner's covariates. Instead, removing informative features increases prediction error; under dependent censoring, IPCW applies misspecified inverse weights to these larger squared errors, amplifying the resulting bias. In addition, weaker covariates leave more unexplained dependence between

E and C. The conditional IPCW baseline, IPCW (CoxPH), depends strongly on the selected feature set: it reduces error by 4.4% and 7.9% for the Top-5 and Top-10 strategies, respectively, but has almost no effect for the Original strategy (0.4%) and increases error by 1.6% under the Random 25% strategy. This highlights a key limitation of conditional IPCW, since an uninformative feature set can yield a misspecified censoring model and invalidate the conditional independent censoring assumption. Only the uncertainty-weighted variant improves on conditional IPCW under all four strategies. The ablation study further highlights the role of modeling dependence explicitly. Combining the dependence-aware functional with KM yields modest gains of roughly 1-3% over IBS-IPCW, whereas replacing KM with the CG estimator increases the reduction to 4-8%. This indicates that modeling the conditional event distribution under dependent censoring, rather than relying solely on marginal KM weights, substantially improves evaluation accuracy.

TABLE 5.3: Mean absolute estimation error of IBS-based metrics across 12 datasets and 4 feature selection strategies. Results are averaged over 5 survival learners and 10 independent experiments. Relative (%) change is shown with respect to the KM-based IBS-IPCW baseline. Lower error is better; bold marks the lowest error per column within each strategy. *IPCW (CoxPH)* denotes IPCW with a CoxPH censoring model, corresponding to conditional independent censoring. *Dep (KM)* denotes the proposed metric using the KM estimator (assuming independence) instead of the CG estimator as an ablation study, and *UW* denotes uncertainty weighting. Full results, including variances, are in the appendix of [51]. Table taken with permission from [51].

	Metric	WHAS	METAB.	Churn	GBSG	NACD	FLC.	SUPP.	Emp.	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stom.)	Average
Original	IPCW (KM)	0.073	0.061	0.117	**0.125**	0.092	0.156	0.108	0.152	0.026	0.077	0.104	0.119	0.101
	IPCW (CoxPH)	0.106	0.051	**0.115**	0.129	0.148	0.162	0.109	**0.100**	0.039	**0.055**	0.097	**0.095**	0.100 (-0.4%)
	Dep (KM)	**0.055**	**0.047**	0.120	0.130	0.091	0.135	0.111	0.148	**0.023**	0.089	0.118	0.125	0.099 (-1.4%)
	Dep (CG)	**0.055**	**0.047**	0.120	0.126	0.091	0.135	0.109	0.154	0.024	0.084	0.106	0.112	0.097 (-4.0%)
	Dep (CG)^{UW}	0.063	0.049	0.124	0.134	**0.082**	**0.090**	**0.093**	0.152	0.024	0.067	**0.092**	0.097	**0.089** (-11.8%)
Top-5	IPCW (KM)	0.071	0.062	0.121	**0.124**	0.103	0.133	0.109	0.155	0.027	0.076	0.109	0.124	0.101
	IPCW (CoxPH)	0.065	0.061	**0.114**	0.126	0.098	0.138	0.109	**0.103**	0.027	0.083	0.109	0.128	0.097 (-4.4%)
	Dep (KM)	**0.055**	**0.048**	0.120	0.130	0.098	0.108	0.112	0.149	0.026	0.090	0.120	0.126	0.098 (-2.8%)
	Dep (CG)	**0.055**	**0.048**	0.119	0.126	0.097	0.108	0.109	0.156	0.027	0.085	0.111	0.116	0.096 (-4.9%)
	Dep (CG)^{UW}	0.065	0.049	0.125	0.134	**0.089**	**0.066**	**0.094**	0.154	**0.025**	**0.068**	**0.096**	**0.101**	**0.089** (-12.2%)
Top-10	IPCW (KM)	0.071	0.061	0.119	**0.125**	0.098	0.143	0.109	0.152	0.026	0.077	0.103	0.119	0.100
	IPCW (CoxPH)	0.082	0.051	**0.111**	0.130	0.093	0.150	0.109	**0.101**	0.031	**0.055**	0.097	0.099	0.092 (-7.9%)
	Dep (KM)	**0.054**	**0.047**	0.121	0.131	0.095	0.123	0.111	0.147	**0.023**	0.090	0.118	0.124	0.099 (-1.4%)
	Dep (CG)	**0.054**	**0.047**	0.121	0.126	0.095	0.123	0.109	0.153	0.024	0.085	0.107	0.111	0.096 (-4.0%)
	Dep (CG)^{UW}	0.063	0.049	0.125	0.134	**0.086**	**0.078**	**0.093**	0.152	0.025	0.068	**0.093**	**0.097**	**0.088** (-11.7%)
Rand. 25%	IPCW (KM)	0.090	0.061	0.129	0.153	0.109	0.140	0.142	0.161	**0.016**	0.111	0.128	0.132	0.114
	IPCW (CoxPH)	0.091	0.056	0.122	0.157	0.111	0.147	0.142	**0.110**	0.017	0.111	0.197	0.133	0.116 (+1.6%)
	Dep (KM)	**0.081**	**0.049**	0.126	0.153	0.101	0.140	0.145	0.137	0.031	0.113	0.127	0.135	0.111 (-2.6%)
	Dep (CG)	**0.081**	**0.049**	**0.117**	**0.147**	0.099	0.140	0.141	0.120	0.032	0.102	0.109	0.120	0.105 (-8.4%)
	Dep (CG)^{UW}	0.095	0.050	0.131	0.161	**0.091**	**0.032**	**0.125**	0.147	0.021	**0.089**	**0.098**	**0.106**	**0.096** (-16.4%)

Table 5.4 evaluates whether a censored metric identifies the same best-performing models as the oracle. We first compute the oracle IBS for each model using the uncensored event times and identify the top-3 models, then assess whether each censored metric recovers at least two of them. IBS-IPCW meets this criterion most consistently overall, while the dependent metrics perform comparably on several datasets but worse on FLCHAIN, Employee, and MIMIC-IV. Thus, although the copula-based IBS substantially reduces absolute estimation error under dependent

censoring, this does not necessarily translate into better separation between closely performing models. IBS-IPCW may therefore remain reasonable for within-dataset ranking when its bias affects competing models similarly. However, it does not estimate the true IBS under dependent censoring and may give unfair rankings when models differ in how they handle censoring.

TABLE 5.4: Ranking performance of IBS-IPCW and the proposed dependent metrics across 12 datasets and 4 feature selection strategies. Each entry counts the number of runs in which a metric identified at least two of the three best survival learners under the oracle IBS, pooled over 10 experiments for each of the four feature selection strategies (40 runs in total). *IPCW (CoxPH)* denotes IPCW with a CoxPH censoring model and *UW* denotes uncertainty weighting. Higher is better; bold marks the best result per dataset. Table taken with permission from [51].

Metric	WHAS	METAB.	Churn	GBSG	NACD	FLC.	SUPP.	Emp.	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stom.)
IPCW (KM)	38/40	**40/40**	**37/40**	**40/40**	39/40	**39/40**	**37/40**	33/40	**33/40**	37/40	**39/40**	36/40
IPCW (CoxPH)	39/40	**40/40**	**37/40**	**40/40**	**40/40**	**39/40**	36/40	**35/40**	29/40	**38/40**	**39/40**	38/40
Dep (KM)	**40/40**	39/40	34/40	38/40	38/40	18/40	35/40	12/40	1/40	**38/40**	38/40	**40/40**
Dep (CG)	**40/40**	39/40	29/40	37/40	36/40	18/40	34/40	11/40	1/40	36/40	30/40	35/40
Dep (CG)^{UW}	39/40	**40/40**	32/40	35/40	38/40	35/40	31/40	13/40	13/40	36/40	35/40	35/40

5.2.4 Discussion

Our results show that accounting for dependent censoring during evaluation can substantially affect estimated model performance. Across synthetic and semi-synthetic experiments, the proposed IBS-Dep consistently reduced estimation bias relative to IBS-IPCW when event and censoring times were dependent. On average, the copula-based estimator achieved reductions of 4-8%, while the uncertainty-weighted variant improved accuracy by 12-16%. These gains were observed across 12 datasets with diverse sample sizes, feature dimensions, and censoring rates, suggesting that dependent censoring is a practically relevant phenomenon in many survival applications. While prior work has primarily focused on handling dependent censoring during model estimation, evaluation metrics have largely remained tied to independence assumptions. Our results highlight that this mismatch can bias reported performance and motivate the need for evaluation procedures that explicitly account for dependence.

This work has limitations. First, the method assumes a stable dependence structure; under distribution shift, the copula parameter may need to be re-estimated, motivating adaptive or time-varying dependence models. Second, the true copula is not identifiable from standard survival data because only one of the event or censoring times is observed for each individual [63], making some degree of model misspecification unavoidable. Third, likelihood-based estimation of the copula parameter may become unstable in small samples where the dependence signal is weak. More broadly, extending dependence-aware corrections to other commonly used evaluation metrics, such as concordance and calibration measures, remains an important direction for future work.

Chapter 6

Applications Beyond Mortality Prediction

While survival prediction and survival analysis is often associated with predicting or studying the time to death, the methodology is broadly applicable to a wide range of time-to-event problems beyond mortality. In this chapter, we explore two such applications that illustrate the versatility of survival analysis. First, we introduce a method to predict fall risk in community-dwelling older adults based on their home care usage. For this, instead of focusing narrowly on older adults that have experienced a fall and excluding those have not, survival analysis allows us to consider both groups – individuals whose event was observed during follow-up and individuals whose event was not (censored). Since falls are often underreported, we use personal alarm subscriptions as an observable proxy for fall risk. This goes beyond standard regression and binary classification techniques and we thus avoid biased estimates that could otherwise overestimate or underestimate the true risk. Second, we expand outside the realm of precision medicine and propose a method to predict the time-to-failure (also called the remaining useful life, RUL) in ball bearings. Recently, there has been a renewed interest in using ML methods to predict RUL in mechanical components, but few studies have explicitly addressed the challenge of handling censored data.

6.1 Prediction of Fall Risk in Home Care Clients

6.1.1 Introduction

Falls among elderly individuals represent a significant and widespread issue in society [131]. Research has shown that simple strength training exercises and daily outdoor activities can lower the incidence of falls [132, 133], yet Danish municipalities have typically employed reactive approaches to address the problem. To shift towards a more proactive strategy and improve the identification of at-risk individuals, ML has been incorporated within the broader paradigm of precision medicine to enable more individualized prediction of fall risk, particularly in hospitalized adults with positive outcomes [134–136]. However, few studies have focused on community-dwelling elderly, and many proposed methods are invasive, such as placing cameras in participants' homes, which can be perceived as intrusive and a breach of privacy [137–139]. The few noninvasive studies available typically use binary classification techniques [140, 141], which are limited in their ability to classify individuals at risk across more nuanced categories. Additionally, these studies often rely on fall reports, which are frequently underreported or inaccurately reported among older adults [142].

In this work, we propose a novel noninvasive method for predicting fall risk within six months using only home care data, without requiring any direct involvement from the individual. Unlike existing approaches, our method offers three key contributions. First, it employs survival analysis to handle censored observations, addressing a limitation in prior ML-based approaches [135, 143]. Second, it gives time-to-event estimates rather than simple binary groups, as seen in previous works [140, 141, 144]. And third, it uses personal alarm subscriptions as a proxy for falls, thereby reducing biases associated with self-reported fall data [142, 145].

The work described in this section is published as [52].

6.1.2 Materials and Method

We use electronic health record (EHR) data from a large Danish municipality, comprising two datasets. The first dataset includes 229,850 home care observations collected over 52 weeks in 2021 and covers a total of 6,398 citizens. Each observation records the amount of care delivered (in minutes), the type of service provided, and the number of social or healthcare workers involved. Services are labeled according to FSIII (Fællessprog III, or Common Language III), a standardized classification used across all Danish municipalities. The data are divided into two six-month cohorts. The second dataset contains all citizens who subscribed to a personal alarm in 2021 in the municipality.

We merged the two datasets and excluded individuals who had subscribed to a personal alarm prior to the study period. We then constructed a window that spanned 26 weeks (six months, from $t = 1$ to $t = 26$) and aligned all citizens' start times at $t = 1$ within this window. For each individual, we defined an event time or censoring time based on the event indicator $\delta \in \{0, 1\}$, where $\delta = 1$ indicates a personal alarm subscription during the window and $\delta = 0$ indicates either dropout or no subscription. We restricted the cohort to individuals born between 1930 and 1970 who received at least 100 minutes of home care per week.

The final dataset consists of 2,542 observations from 1,499 citizens, with a censoring rate of 95% (2,416 censored, 126 events) and 49 nonzero covariates. No feature scaling, outlier removal, or imputation was applied, as there were no missing values. Table 6.1 provides an overview of the dataset.

TABLE 6.1: Overview of the dataset after preprocessing. Table taken with permission from [52].

Number of observations (N)	2542
Number of covariates	49
Number of citizens	1499
Average age (2021)	77 ($\pm$ 10) years
Censoring rate	95% (2416/126)
Follow-up period	6 months (2 cohorts)
Study period	Jan. 1st to Dec. 31st 2021

To validate our method, we implement several survival models: the CoxPH, CoxPH with LASSO regularization (CoxPH ℓ_1), CoxPH with Ridge regularization (CoxPH ℓ_2), Random Survival Forests (RSF), and CoxBoost. To assess model performance, we report Harrell's C-index (CI_H), Uno's C-index (CI_U), the integrated Brier score (IBS), mean absolute error (MAE) computed with a hinge loss, and D-calibration. Models are trained using either all covariates or a subset selected by feature selection methods such as low-variance thresholding (LowVar), SelectKBest

(SKB), or recursive feature elimination (RFE), with the number of selected features fixed at $K = 10$.

We employ stratified nested cross-validation with 5 outer folds and 5 inner folds. Stratification maintains consistent distributions of event times and censoring rates across training and test splits. Within each fold, feature selection and hyperparameter tuning are performed jointly to avoid information leakage.

6.1.3 Experiments and Results

Table 6.2 presents the results obtained from cross-validation, averaged across 5 folds. The CoxPH model achieved a C-index (CI_H) of 0.53, indicating that it correctly ranked 53 out of 100 pairs. When a LASSO penalty term was applied to the CoxPH model, the CI_H increased to 0.61, suggesting that reducing the number of covariates improved discriminative performance in this case. The CoxBoost model, combined with low-variance thresholding, yielded the most concordant risk predictions, achieving a mean CI_H of 0.64 (95% CI = 0.57-0.72). It is worth noting that certain types of home care are provided sporadically, resulting in many covariates having predominantly zero values. This sparsity can complicate the model's ability to distinguish between important and non-essential covariates, potentially leading to poor performance on new data. We observe that restricting the feature space through LASSO or a feature selection algorithm tends to yield better average discrimination results.

At first glance, the IBS results appear encouraging, since lower values indicate better survival curve estimation. However, the high censoring rate in this dataset leads many predicted survival curves to exhibit only a slight decline over the follow-up period. While this pattern is consistent with a low observed event incidence, it limits the practical usefulness of the predictions: the estimated survival probabilities for the two groups – citizens with and without alarm subscriptions – remain very similar over time. As a result, the model provides little meaningful differentiation in risk between the groups, despite achieving favorable IBS values. The RSF model demonstrated the best performance in predicting the survival curve, achieving a mean IBS of 0.033 (95% CI = 0.031-0.035). It also achieved the lowest MAE among the models, 83.5 weeks (95% CI = 79.1-87.9), using Recursive Feature Elimination (RFE) for covariate selection. However, this error is more than three times the 26-week follow-up window: because of the high censoring rate, most predicted survival curves do not drop below 0.5 within the window, so the predicted median event time must be extrapolated far beyond it. The MAE-hinge measures the absolute difference between actual and predicted survival times where predictions beyond the censored time incur no penalty. Consequently, errors stem from censored samples with overly pessimistic predictions (less than the censoring time) and uncensored samples with overly optimistic predictions (greater than the event time). Given the 95% censoring rate, the model tends to overestimate the event time for the few individuals who actually experience it, limiting its effectiveness in accurately predicting *when* the event will occur in this population. Finally, D-calibration assessed via Pearson's chi-squared test indicated that CoxPH with LASSO, RSF, and CoxBoost all produced predicted survival curves for which D-calibration could not be rejected, although with few events per fold the test has limited power.

Figure 6.1 shows the most influential covariates and their corresponding values

TABLE 6.2: Experimental results from 5-fold cross-validation. For D-calibration, we report the number of D-calibrated mean survival curves across the five folds according to a Pearson's chi-squared test. Bold marks the best result per column within each metric, including ties. Table taken with permission from [52].

Metric	Model	Feature selector			
		None	LowVar	SKB10	RFE10
$CI_H \uparrow$	CoxPH	0.53±0.04	0.53±0.04	0.59±0.06	0.60±0.05
	CoxPH (ℓ_1)	0.61±0.04	0.61±0.04	0.61±0.04	0.61±0.04
	CoxPH (ℓ_2)	0.52±0.03	0.53±0.04	0.58±0.04	0.55±0.11
	RSF	**0.63±0.04**	0.62±0.04	**0.62±0.04**	0.62±0.04
	CoxBoost	**0.63±0.07**	**0.64±0.05**	**0.62±0.03**	**0.63±0.06**
$CI_U \uparrow$	CoxPH	0.52±0.04	0.52±0.04	0.57±0.06	0.58±0.05
	CoxPH (ℓ_1)	0.60±0.04	0.60±0.04	0.60±0.04	0.60±0.04
	CoxPH (ℓ_2)	0.51±0.04	0.51±0.04	0.56±0.05	0.54±0.10
	RSF	**0.62±0.05**	0.62±0.05	**0.61±0.04**	**0.62±0.04**
	CoxBoost	**0.62±0.07**	**0.64±0.05**	**0.61±0.03**	**0.62±0.06**
IBS ↓	CoxPH	0.037±0.001	0.036±0.001	**0.033±0.002**	0.034±0.002
	CoxPH (ℓ_1)	**0.034±0.001**	**0.034±0.001**	0.034±0.001	0.034±0.002
	CoxPH (ℓ_2)	0.037±0.001	0.037±0.000	0.034±0.001	0.035±0.002
	RSF	**0.034±0.001**	**0.034±0.001**	**0.033±0.001**	**0.033±0.001**
	CoxBoost	**0.034±0.001**	**0.034±0.001**	0.034±0.001	0.034±0.001
MAE ↓	CoxPH	212.4±111.4	211.3±109.1	106.2±33.3	118.1±46.3
	CoxPH (ℓ_1)	87.5±1.3	87.5±1.3	87.4±1.4	108.6±43.0
	CoxPH (ℓ_2)	296.8±109.4	224.3±113.5	115.7±29.7	117.1±46.6
	RSF	**84.9±2.0**	**85.0±1.7**	**84.0±2.8**	**83.5±3.2**
	CoxBoost	87.9±1.1	87.9±1.0	87.8±1.0	87.8±1.1
D-calib. ↑	CoxPH	4/5	4/5	**5/5**	**5/5**
	CoxPH (ℓ_1)	**5/5**	**5/5**	**5/5**	**5/5**
	CoxPH (ℓ_2)	3/5	3/5	4/5	**5/5**
	RSF	**5/5**	**5/5**	**5/5**	**5/5**
	CoxBoost	**5/5**	**5/5**	**5/5**	**5/5**

as determined by SHAP[1] [146], based on the RSF model. Covariates are listed along the y-axis, while the x-axis represents their SHAP values, which indicate the covariate's contribution to the predicted risk. The color gradient shows whether a covariate's value is low (blue) or high (red) for a given observation. According to the SHAP analysis, the top five contributing covariates are *Personal hygiene*, *SUL §138*, *Excretions*, *RH Nutrition*, and *Medication administration*. The SHAP values suggest a weak positive correlation between the risk of an alarm subscription and the number of minutes received under *Personal hygiene*, whereas covariates such as *Excretions* and *SUL §138* exhibit a stronger and more consistent positive association with risk. In this context, *SUL §138* refers to a provision in Danish health legislation under which a community nurse is assigned to a citizen following a referral from their general practitioner. Interestingly, some covariates – such as *Cleaning*, *Medication administration*, and *RH Personal hygiene* – show both positive and negative SHAP values across the dataset. While this may initially seem counterintuitive, since greater reliance on home care might be expected to signal higher risk, the observed relationships reflect correlations with the likelihood of receiving a personal alarm. Elderly individuals receiving substantial home care often benefit from frequent supervision, including

[1]SHAP (SHapley Additive exPlanations) is a method for interpreting machine learning models by assigning each feature a contribution value that reflects its impact on a prediction. It is based on Shapley values from cooperative game theory.

daily visits from nurses or care staff, which may reduce their need for a personal alarm despite their underlying frailty.

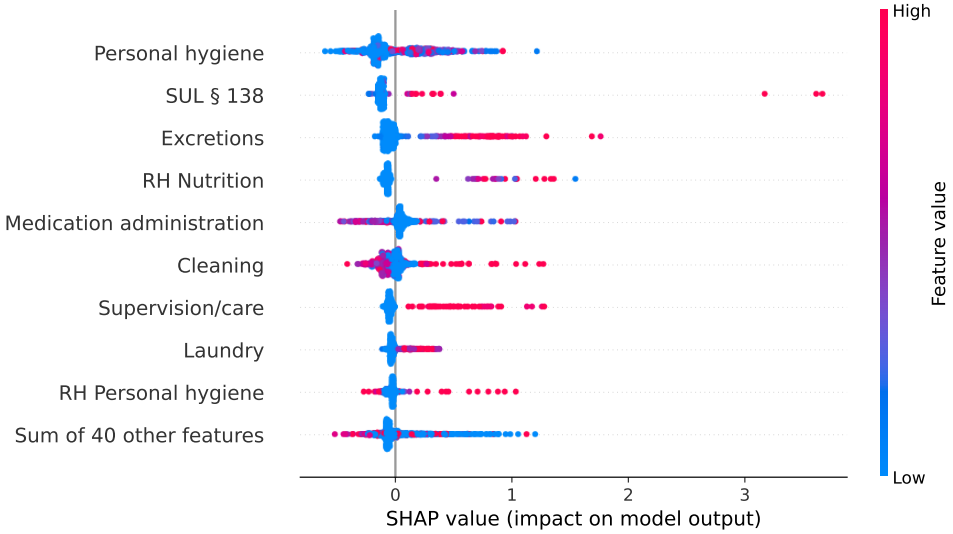


FIGURE 6.1: SHAP summary plot. The covariates are on the y-axis and the SHAP-value is on the x-axis. Covariates are ranked based on their importance, which is given by the mean of their absolute SHAP values (higher positioning means more important). For each covariate, one point corresponds to a single citizen. Its position on the x-axis represents the impact that covariate had on the model's output, which corresponds to the risk across citizens. The color indicates whether the citizen had a low (blue) or high value (red) of that covariate. Figure taken with permission from [52].

Figure 6.2 presents the predicted individual survival and hazard functions for 20 randomly chosen citizens, as estimated by the RSF model. Solid lines correspond to individuals who did not subscribe to a personal alarm within six months, while dotted lines indicate those who did. In the top row, we observe that most individuals have a survival probability (*i.e.*, probability of no alarm) between 0.90 and 0.98 at $t = 26$, and a corresponding cumulative hazard between 0.01 and 0.10. However, a few individuals display notably lower survival probabilities and higher hazards, reflecting their specific covariate profiles. In this sample, only one person subscribed to a personal alarm during the study period, yet their predicted survival trajectory is similar to those who did not. We attribute this to the high censoring rate and the resulting limited number of observed events, which restrict the model's ability to learn distinct risk patterns. This highlights the challenge of accurately predicting event timing based solely on individual survival curves in heavily censored settings.

6.1.4 Discussion

In this study, we implemented several survival models to predict individual fall risk over a six-month period. The models were trained on a dataset comprising 2,542 home care observations from 1,499 citizens (59% female, 41% male) with a mean age of 77 years (SD. = 10). Using 5-fold cross-validation, CoxBoost emerged as the top-performing model, achieving a mean Harrell's concordance index of 0.64 (95% CI: 0.57-0.72), which corresponds to an average ranking accuracy 14 percentage points above chance-level. Additionally, CoxBoost produced survival curves that were D-calibrated across all folds, as confirmed by Pearson's chi-squared test.

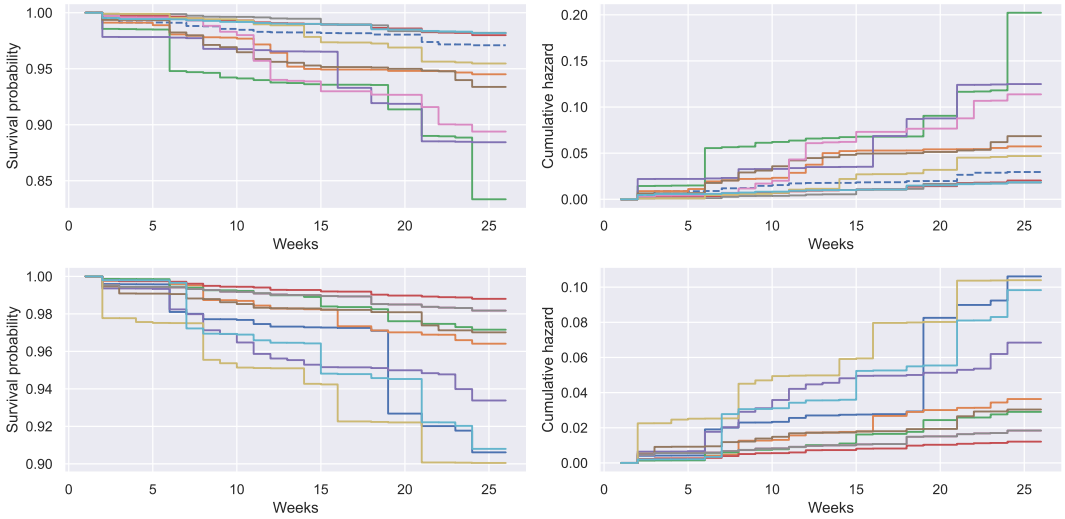


FIGURE 6.2: Predicted individual survival (left) and hazard (right) functions for 20 randomly selected citizens from the test set using the RSF model. Solid lines correspond to citizens who did not subscribe to a personal alarm within six months, while dotted lines indicate those who did. Figure taken with permission from [52].

The study has two important limitations. First, personal alarm subscriptions are only a proxy for fall risk, since alarms can also be given for other reasons, such as anxiety or a need for a secure environment. Second, many citizens contribute observations to both six-month cohorts, and the cross-validation folds were not grouped by citizen, so the same individual may appear in both training and test folds, which can make the reported performance optimistic.

Our proposed tool is intended as a whitelisting method to identify individuals who may benefit from fall prevention programs. Since supervised strength training is generally safe and beneficial for older adults [147, 148], offering such interventions broadly carries minimal risk. Currently, many elderly individuals are only offered prevention after experiencing multiple falls ($\geq$ 2), as prior falls are a primary eligibility criterion [149]. Moreover, access to these services is often influenced by socioeconomic status [150]. By providing an objective, data-driven risk assessment, our model can support earlier and more equitable access to fall prevention, helping reduce health disparities.

6.2 Prediction of Survival Time in Ball Bearings

6.2.1 Introduction

Ball bearings, which are widely used in rotating machinery, are particularly prone to degradation caused by factors such as contamination, poor lubrication, and improper mounting [151]. A common strategy is to monitor vibration signals and use frequency-based analysis to detect early signs of wear. In recent years, machine learning methods, especially neural networks, have been used to predict the remaining useful life of bearings from such sensor data. A key metric for assessing bearing health is the remaining useful life (RUL), defined as the time from the current inspection until the bearing reaches the end of its service life or fails. In this setting, RUL estimation can be viewed as estimating the remaining lifetime, where both accuracy

and proper uncertainty quantification are essential. Reliable RUL estimates support early fault detection and allow engineers to plan maintenance proactively, thereby reducing the risk of unexpected downtime [152–158].

However, many existing approaches treat the task as a standard regression problem and ignore censored observations, even though such cases (*i.e.*, bearings that have not failed within a certain time window) are relatively common, as noted by Ferreira and Gonçalves [159]. If we simply ignore these censored instances, it could lead to inefficient use of data and biased model estimates [42]. While survival analysis naturally accommodates censored data, its use in bearing prognostics and RUL prediction remains limited.

In this work, we propose a survival method for RUL prediction in ball bearings (see Figure 6.3). We first process vibration signal data in the frequency domain to identify the time point at which a bearing begins to deviate from its normal operating behavior, which typically occurs well before catastrophic failure. This deviation is used as an event definition to annotate bearings in two benchmark datasets. Using these annotations, we train survival models that estimate the probability of failure over time rather than producing a single point estimate of the remaining useful life. This approach allows the models to leverage censored observations and provide progressive risk estimates that can support maintenance decisions.

The work described in this section is published as [53].

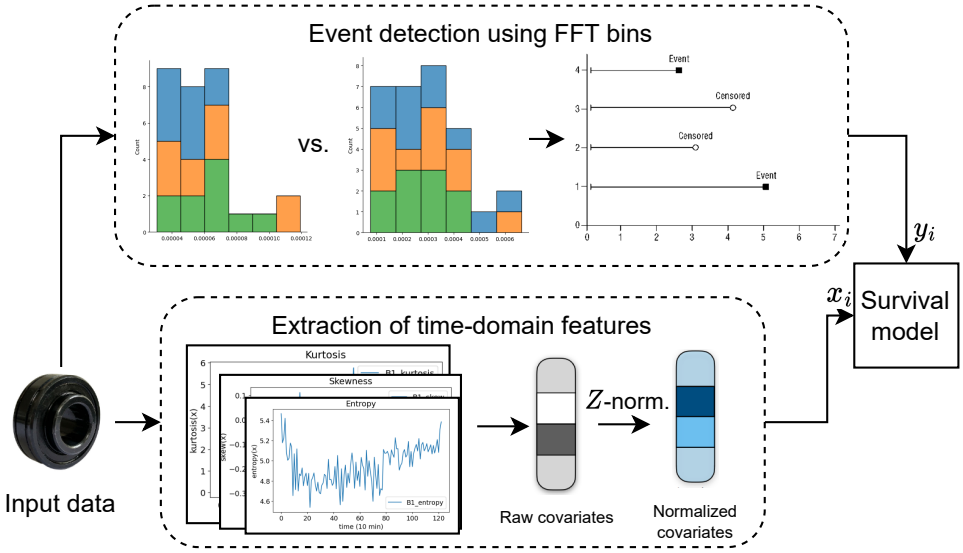


FIGURE 6.3: Outline of the proposed solution. Raw data is used for event detection (top) and feature extraction (bottom) before training a survival model. Figure taken with permission from [53].

6.2.2 Materials and Method

We consider two public bearing datasets: XJTU [155] and PRONOSTIA [160]. In both datasets, a single bearing under test is subjected to radial load during rotation while vibration signals are recorded using two piezoelectric accelerometers mounted orthogonally (horizontal and vertical). The signals are sampled at 25.6 kHz and stored as raw time-series measurements. In both experimental setups, bearings are operated under high loads to accelerate degradation and eventual failure. The XJTU dataset contains experiments on 15 deep-groove ball bearings under three operating conditions, with five bearings per condition. In this work, we use bearings operating

at a C/P ratio of 1.1, a radial force of 12.0 kN, and a rotation speed of 2100 RPM. The PRONOSTIA dataset contains 6 deep-groove ball bearings tested under three operating conditions with two bearings per condition; we use bearings operating at a C/P ratio of 0.8, a radial force of 5.0 kN, and a rotation speed of 1500 RPM.

We extract features from vibration signals recorded along the two accelerometer axes. For each signal frame $\boldsymbol{x} = (x_i)_{i=1}^N$, a set of standard statistical features is computed. These features summarize key properties of the vibration signal and have been widely used in bearing monitoring. Table 6.3 shows the 12 time-domain features used in this work.

TABLE 6.3: Statistical features computed from signal frame $\boldsymbol{x} = (x_i)_{i=1}^N$. Table taken with permission from [53].

Absolute mean	$\bar{x} = \frac{1}{N}\sum_{i=1}^N \lvert x_i\rvert$
Standard deviation	$\sigma = (\frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^2)^{1/2}$
Skewness	$SK = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^3/\sigma^3$
Kurtosis	$K = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^4/\sigma^4$
Entropy	$H = -\sum_{i=1}^N P(x_i)\log P(x_i)$
Root-mean-square	$RMS = (\frac{1}{N}\sum_{i=1}^N x_i^2)^{1/2}$
Max value	$max = \max(\lvert\boldsymbol{x}\rvert)$
Peak-to-peak	$P2P = \max(\lvert\boldsymbol{x}\rvert) - \min(\lvert\boldsymbol{x}\rvert)$
Crest factor	$CR = max/RMS$
Clearance factor	$CL = x_p/(\frac{1}{N}\sum_{i=1}^N \lvert x_i\rvert^{1/2})^2$
Shape factor	$SH = RMS/\bar{x}$
Impulse	$IM = max/\bar{x}$

To construct survival outcomes, we detect when a bearing deviates from normal operation using a signal-processing procedure on vibration data. Specifically, signals are transformed via the Hilbert transform and FFT, and changes in their frequency distributions are tracked using KL divergence and standard deviation. When these exceed predefined thresholds, the bearing is considered degraded, defining the event time (see [53] for details).

Both datasets contain few bearings, so we augment the data through preprocessing. Vibration signals from two accelerometer axes are treated as independent observations, and constrained bootstrapping [161] is applied to generate additional samples. Each bearing lifetime is further divided into 20 time slices, yielding covariates paired with time-to-event or censoring indicators. To simulate incomplete data, 20% of bearings are randomly censored before the event. All covariates are standardized using z-score normalization. Finally, we split by bearing: for XJTU, three are used for training and two for testing; for PRONOSTIA, one for training and one for testing. We evaluate several survival models commonly used for time-to-event prediction (see [53] for details). For each dataset and model, hyperparameters are tuned over 10 iterations using 5-fold cross-validation on the training bearings only. The hyperparameter configuration achieving the highest average concordance index (Antolini's) on the validation folds is selected to train the final model, which is then evaluated on the test bearings.

6.2.3 Experiments and Results

Table 6.4 summarizes the predictive performance of the evaluated survival models. We report training time, Harrell's concordance index (CI) [36], Antolini's time-dependent concordance index (CI_{td}) [101], and the integrated Brier score (IBS) [77]. On the XJTU dataset, RSF and DeepSurv achieve the highest Harrell's CI, and RSF and CoxBoost the highest CI_{td}, while DSM provides the best predictive accuracy in terms of IBS. On the PRONOSTIA dataset, DSM, DeepSurv, and Weibull AFT perform best overall.

TABLE 6.4: Performance metrics on the test sets. N: total sample size, C: percentage of censored data, d: number of covariates. Table taken with permission from [53].

(A) XJTU ($N = 1000$, $C = 20\%$, $d = 12$).

Model	T_{train}	CI ↑	CI_{td} ↑	IBS ↓
CoxPH	**0.094**	0.648	0.642	0.271
RSF	0.988	**0.702**	**0.769**	0.423
CoxBoost	2.791	0.59	**0.769**	0.261
DeepSurv	9.148	0.687	0.679	0.245
DSM	22.456	0.66	0.628	**0.214**
WeibullAFT	0.483	0.664	0.675	0.216

(B) PRONOSTIA ($N = 400$, $C = 20\%$, $d = 12$).

Model	T_{train}	CI ↑	CI_{td} ↑	IBS ↓
CoxPH	**0.016**	0.626	0.625	0.34
RSF	0.265	0.715	0.701	0.383
CoxBoost	0.768	0.74	0.701	0.223
DeepSurv	2.392	0.745	0.705	**0.197**
DSM	9.737	0.739	**0.748**	0.294
WeibullAFT	0.489	**0.764**	0.691	0.236

Figure 6.4 illustrates inference results on the XJTU dataset. The leftmost panel shows the estimated survival functions across models together with the KM estimate as a reference. The predicted survival functions lie above the KM estimate, *i.e.*, they predict lower risk; note, however, that the KM curve is estimated on all five bearings, whereas the model curves are averaged over the two test bearings. The center panel shows predicted survival functions for two groups of bearings defined by the RMS feature using a CoxPH model. The separation between the curves indicates that the models capture meaningful differences in degradation behavior between the groups. The rightmost panel shows simulated survival times and the corresponding hazard functions for these groups. We see a clear separation in predicted survival times between the two groups, indicating that bearings with lower values of RMS tend to operate longer.

6.2.4 Discussion

In this work, we proposed a survival analysis method for predicting RUL in ball bearings. We first identify the time at which a bearing begins to deviate substantially from its original frequency characteristics, defining this point as the event of interest. We then train several survival models to estimate failure risk based on covariates extracted from sensor readings. This application is particularly relevant in manufacturing and engineering settings, as survival models naturally accommodate

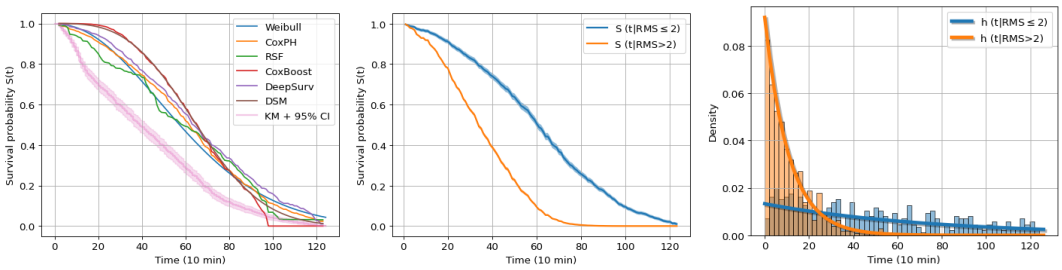


FIGURE 6.4: Model inference on the XJTU dataset. Left & center: survival function for different estimation approaches and for two separated groups based on RMS value (Solid lines: mean values; Shadow area: 95% conf. int.). Right: normalized survival times (bins) and hazard functions (solid lines) of the two RMS groups. Figure taken with permission from [53].

censored observations and produce probabilistic predictions with associated uncertainty rather than single point estimates. The results should, however, be interpreted with care: the event is defined by our detection algorithm rather than by an actual failure, censoring was simulated rather than observed, and the evaluation relies on very few physical bearings (two test bearings for XJTU and one for PRONOSTIA), with the larger sample sizes resulting from data augmentation. We therefore encourage further research building on this framework.

Chapter 7

Conclusion

7.1 Concluding Remarks

This thesis addresses four research questions in survival prediction within precision medicine: how to quantify uncertainty in individual predictions (RQ1), predict multiple related events (RQ2), evaluate predictions under censoring (RQ3), and apply survival methods beyond mortality prediction (RQ4). Although each question motivates a distinct contribution, they share a central concern: how to make useful time-to-event predictions when outcomes are complex or only partly observed.

For RQ1, we developed BaySurv, which trains probabilistic survival models with one of three techniques: variational inference (VI), Monte Carlo dropout (MCD), or a spectral-normalized neural Gaussian process (SNGP) (Chapter 3). The VI and MCD variants capture both epistemic uncertainty, through the distribution over weights, and aleatoric uncertainty, through a Gaussian output over the risk score. Our main finding is that MCD and SNGP are viable alternatives to VI: on MIMIC-IV, MCD matched VI in concordance (0.748 vs. 0.743) with roughly half the trainable parameters (2,882 vs. 5,700), and SNGP produced D-calibrated survival functions in all four datasets, compared with two for VI. As a complementary, frequentist approach, we applied conformal prediction to a classification task, predicting temporomandibular joint involvement in children with juvenile idiopathic arthritis. With a target coverage of 90%, the prediction sets covered the true class for about 90% of examinations with TMJ involvement under the IID strategy, but for only 74% in the early temporal segment, illustrating that the coverage guarantee holds on average across all examinations rather than for each class separately. These results show that uncertainty can be quantified efficiently, but not that uncertainty estimates alone make predictions trustworthy in clinical practice.

For RQ2, we developed MENSA, which models each event as a transition between states, represents each event-specific time-to-event distribution as a mixture of Weibull distributions, and learns all events in a single network with shared and event-specific layers (Chapter 4). To use temporal ordering information when it is available, we introduced a trajectory-based likelihood term that encourages the predicted survival of a later event to remain high at the time of an earlier one. Across four multi-event datasets, MENSA was competitive with strong deep learning and tree-based baselines, consistently outperformed the Hierarchical multi-event model, and required fewer parameters than training one model per event. Our ablations show that both components matter: joint training improved the within-patient ordering of events on Rotterdam (local C-index 90.2 vs. 87.8 with independent training), and the trajectory term improved the local C-index on both Rotterdam and EBMT, as well as the global C-index, IBS, and mMAE on EBMT, although these gains were small relative to their variability. The ALS application illustrates how predicting several aspects of functional decline gives a fuller account of disease progression

than a single event time.

For RQ3, we showed that meaningful evaluation depends on matching the metric to the prediction objective and examining the assumptions used to handle censoring (Chapter 5). A metric that measures ranking, for example, cannot by itself establish the accuracy of predicted event times or survival probabilities, yet approximately 72% of the 92 surveyed studies used metrics that did not align with their objectives or did not address the censoring assumptions those metrics require. To reduce evaluation bias under dependent censoring, we developed a copula-based Brier score that imputes margin times for censored instances using the Copula-Graphic estimator. Across 12 semi-synthetic datasets, its uncertainty-weighted variant reduced the estimation error of the integrated Brier score by 12-16% on average relative to KM-based IPCW, and it improved on conditional IPCW under all four feature selection strategies. However, lower estimation error did not translate into better identification of the top-performing models, and the metric still requires assumptions about the dependence between event and censoring times.

For RQ4, we applied survival methods to two problems beyond mortality (Chapter 6): predicting the time to a personal alarm subscription, used as a proxy for fall risk in older adults, and predicting the time until ball bearings deviate from normal vibration behavior, used as an early indicator of failure. In the home care data, most citizens did not subscribe to an alarm during follow-up and were therefore censored; in the bearing data, censoring was simulated. In the home care setting, the best model ranked citizens by risk 14 percentage points better than chance (C-index 0.64), but could not accurately predict when an alarm subscription would occur, largely because of the 95% censoring rate. In the bearing setting, the models reached C-indices of up to 0.70 on XJTU and 0.76 on PRONOSTIA, although based on very few physical bearings. These applications show how survival analysis can make time-to-event predictions for outcomes beyond mortality, while also illustrating that the conclusions depend on how well the observed endpoint represents the outcome of actual interest.

Taken together, the thesis contributes methods for representing uncertainty, predicting multiple events, and evaluating predictions when censoring complicates the observed data. Its applications show how these methods and the broader survival analysis framework can address varied time-to-event problems. The following section considers directions for further work, particularly where current methods and evaluation practices leave important questions unresolved.

7.2 Future Work

7.2.1 Interpretability

Interpretability and explainability have become important topics for the machine learning community [44]. Broadly speaking, researchers are pursuing two main strategies. The first approach aims to develop models that are inherently interpretable, *e.g.*, extensions of the traditional CoxPH model [23] or shallow tree-based models such as decision trees. The second approach aims to learn an arbitrarily complex black-box model and use some post-hoc technique to come up with an explanation of the black box model. In the following, we discuss both strategies in the context of survival analysis.

Interpretable survival analysis. The first strategy is to build models that are inherently interpretable. As discussed in Section 2.3, the CoxPH model is interpretable by design. It models the log-partial hazard function as $f(x;\theta) := \theta^\top x$, where $\theta \in \mathbb{R}^d$

and $x \in \mathcal{X} \subseteq \mathbb{R}^d$. In this formulation, the parameters of θ directly specify the contribution of each feature, with zero weights indicating that the corresponding features are ignored by the model. Interpretation can be further refined to focus on a subset of features. For instance, suppose each input can be decomposed as $x = (u, v)$, where $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$, we may choose to interpret the model only with respect to u, treating v as unrelated or uninterpreted variables. In this scenario, the log-partial hazard function can be written as [6]:

$$f(x; \theta) := \psi^\top u + g(v; \phi), \tag{7.1}$$

where $\psi \in \mathbb{R}^{d_1}$ is a parameter vector and $g(\cdot; \phi) : \mathbb{R}^{d_2} \to \mathbb{R}$ is a model with parameters ϕ. The full parameter vector is then $\theta = (\psi, \phi)$. This formulation, known as the partially linear CoxPH model $g(\cdot; \phi)$, has known statistical guarantees [162].

Several deep survival models incorporate interpretability by design. For example, Chapfuwa et al. [163], Nagpal et al. [164], Manduchi et al. [74], Jeanselme et al. [165], and Shen and Chen [166] propose clustering-based approaches that learn clusters in a supervised fashion for time-to-event prediction. These clusters can highlight subgroups with similar survival patterns (*e.g.*, hazard rates) and offer insights into how specific features or covariates relate to outcomes. For instance, clustering might reveal that patients with a certain genetic profile face elevated or reduced survival risk.

Another approach to interpretability uses survival trees – analogous to regression trees but adapted to censored data and often able to estimate individual survival distribution (ISDs). We consider survival trees as interpretable because they partition data based on feature values, detailing how various factors influence survival. Examples include single survival trees and Optimal Survival Trees (OST) [167, 168]. As long as the depth and number of leaves remain manageable, interpretation is straightforward. Ensembles such as Random Survival Forests (RSF) [58] support non-linear interactions and are non-parametric, but their interpretability is largely lost because predictions are averaged over many trees.

Explaining black box models. The second strategy is to train an arbitrarily complex deep survival model, that is not interpretable, and then apply some post-hoc explanation tool afterward. LIME [169] and SHAP [146] are model-agnostic attribution methods that explain a data point's prediction as the sum of feature effects [44], and we already used SHAP in Chapter 6. These two tools, although not originally designed for survival analysis, have already been adapted more specifically to survival models as SurvLIME [170] and SurvSHAP [171], respectively. The main advantage of the post-hoc method is that we are free to decide on a model architecture and the choice of explanation method. For example, if we are visualizing feature effects of an RSF model with SurvLIME, we can change the underlying model and still use the same type of explanation method. Or, if we no longer want to use SurvLIME, we can use SurvSHAP without having to change the RSF model. On the other hand, if we are using a linear CoxPH model and interpret the coefficients, switching to a tree-based regressor will also change the means of explanation.

Exploring how to effectively apply explanation methods into deep survival models and how such techniques can be evaluated is an interesting direction for future research.

7.2.2 Empirical Evaluation

In our position paper (Chapter 5), we identify two major limitations in current evaluation practices. First, evaluation metrics are often misaligned with the stated objective: naive discrimination (*i.e.*, ranking-based) measures are frequently used by default, even when the goal is accurate time-to-event or probabilistic prediction. Second, these metrics are commonly applied without accounting for censoring or appropriately correcting for it. This points to two areas in evaluation (*i.e.*, choosing the right metric and adjusting for censoring) where current practices leave room for improvement.

Beyond the C-index. In a recent survey, Zhou et al. [172] reported that over 80% of survival analysis studies published in leading journals use the C-index as their primary evaluation metric. This widespread reliance is unsurprising: many classical survival models output a scalar risk score $M(x) \in \mathbb{R}$ for each input $x \in \mathcal{X}$ (typically $\mathbb{R}^d$), with the interpretation that a higher score indicates higher risk. So that if $M(z) > M(y)$, the model predicts that instance z will experience the event before y. The C-index then, conveniently, measures how often these rankings agree with the observed ordering of events. However, this widespread reliance on the C-index is largely driven by the limitations of traditional models rather than the needs of practitioners. Modern survival models – particularly ISD models [27] – can estimate, for each x, the full survival function $\{S(t \mid x) \mid t \geq 0\}$, which enables us to predict quantities such as event probabilities within a time window, median survival time, or the expected time-to-event.

Despite this capability, evaluation practices have not kept pace. Our review of 92 methodological and applied survival analysis papers published between 2023 and 2025 (see [50]) shows that approximately 72% of studies rely on evaluation metrics that are misaligned with their stated objectives or do not address the censoring assumptions those metrics require. In other words, while many works aim to predict when an event will occur, they continue to evaluate models primarily on how well they rank risk. Instead, as a direction for future work, we suggest a shift in focus from merely measuring ranking performance to explicitly evaluating the predicted time-to-event estimates or survival probabilities. In many cases, having an actual time-to-event estimate (*e.g.*, days, weeks, months) is more useful than a relative risk estimate without any physical interpretation. For example, for a newly diagnosed ALS patient, relative risk may be less informative than knowing the expected time until loss of mobility or respiratory function. Given such estimates, standard error metrics (Section 2.4), such as the mean absolute error (MAE), can be used by comparing predicted times to observed or imputed survival times.

Beyond random censoring. Many survival models and evaluation metrics assume random (or at least independent) censoring, which simplifies modeling, but does not always reflect the true data generation process. In clinical or industrial settings, censoring is often what we call informative; for example, patients leave a study because they get too sick, or ball bearings are decommissioned because they are already showing early signs of failure or degradation, even if they have not yet failed. Models should therefore be evaluated under a range of censoring mechanisms, including different forms of dependent censoring (*e.g.*, early, late, and symmetric patterns), to assess how robust and generalizable they are across realistic data-generating scenarios. A broader and looser assumption than random censoring is independent censoring, which means that the event times and censoring times are independent given the observed covariates. Under this assumption, we can extend the KM estimator (used for weighting or "de-censoring") to a conditional estimator

that incorporate observed features to capture the dependency between event and censoring times. An example is the CoxPH model [23], which adjusts for covariates, and in turn allows evaluation metrics to remain valid under independent censoring [173].

However, if unobserved confounders influence both the censoring and event, even a well-specified conditional estimator may fail to fully account for these (unobserved) dependencies. This scenario is called dependent censoring, which we described in detail in Section 2.2 and proposed a solution for in Chapter 5. Building on this idea, an important direction for future work is the development of dependence-aware variants of other commonly-used evaluation metrics. Most standard metrics in survival analysis – such as Harrell's C-index, Uno's C-index, and D-calibration – implicitly rely on independence between censoring and event times. As a result, their validity may be compromised under dependent censoring, which may lead to biased or misleading model assessments (see Figures 5.4 and 5.5 in particular). Extending our copula-based framework to these metrics is a natural next step. A general strategy is to systematically replace independence-based correction mechanisms, such as the KM estimator, with dependence-aware alternatives, such as copula-based estimators. This suggests a modular approach to evaluation design: rather than proposing entirely new metrics, one can adapt existing and well-understood metrics by modifying only the censoring correction step. Such a framework would enable consistent and interpretable evaluation across many scenarios.

Looking forward, we argue that progress in survival prediction should increasingly depend on improving evaluation methodology, rather than solely developing more complex models. Researchers need to take a closer look at the assumptions behind the metrics they use and ask whether those assumptions really hold. As a community, we should prioritize developing evaluation methods that are aligned with the underlying data assumptions before claiming new modeling advances.

7.2.3 Foundation Models

Deep learning is in the midst of a paradigm shift into the "foundation model" (FM)-era, characterized by models with millions, or even billions, of parameters. Recent text and image foundation models (FMs) are indeed very impressive, and have taken over the machine learning community by storm [174]. So far, language has been the predominant modality used to train these models, with little attention given to tabular data, despite it being the dominant modality in many fields. We will spend the first part of this section motivating foundation models for tabular data, inspired by Breugel and Schaar [175], and later discuss some interesting foundation models for regression and survival problems that have great potential in risk prediction and time-to-event applications.

FMs for tabular data. In a 2024 position paper, Breugel and Schaar [175] argued that now is the time to start developing tabular foundation models, or what they call Large Tabular Models (LTMs). LTMs have the potential to transform how science and machine learning handle tabular data – moving away from treating datasets as isolated entities analyzed in a vacuum, toward a paradigm where each dataset is understood and modeled in the context of related datasets. This contextualization can unlock deeper insights and improve generalization across studies. Large language models (LLMs) are already widely used as tools, and vision FMs (*e.g.*, Saharia et al. [176]) can generate synthetic images or embed real images for many downstream tasks. LTMs can play a similar role for tabular data, where deep learning models

often struggle to put up as impressive performances as tree-based or boosting models. Recent benchmarking papers find that XGBoost [177], a tree-based model, is still among the top performers for supervised learning on tabular data. Breugel and Schaar [175] attribute this to several factors: tabular data are often discontinuous, feature spaces are heterogeneous and sometimes uninformative, and there is a lack of spatial invariances (*i.e.*, no meaningful local neighborhoods between features) that could guide the choice of a prior. Additionally, the scale of data, model architectures, and task complexity in tabular domains differs significantly from those in vision and language domains. There exist, in fact, settings in which neural networks seem to outperform XGBoost rather consistently. For instance, when the number of samples increases [178], the number of target classes increases [179], and when modern regularization techniques are applied [180]. These observations suggest that the scaling advantages demonstrated in other modalities may also extend to tabular data [174]. However, these larger model sizes and data scenarios remain largely unexplored, as most benchmarks focus primarily on smaller datasets and models.

PFNs (Prior-Data Fitted Networks) is a relatively new and promising approach in the context of foundation models for tabular data [181]. PFNs are neural networks designed to approximate Bayesian posterior inference over a distribution of tasks. Instead of performing posterior inference for each task individually (*e.g.*, using variational inference), a PFN learns a direct mapping from a small training dataset and a test input to the posterior predictive distribution of the test label. Formally, the PFN approximates

$$\Pr(y_{\text{test}} \mid x_{\text{test}}, \mathcal{D}_{\text{train}}) \approx \text{PFN}(\mathcal{D}_{\text{train}}, x_{\text{test}}), \tag{7.2}$$

where $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ is some small training dataset, and x_{test} is a test instance. A PFN is trained on a large number (millions) of tiny, synthetic datasets generated from known distributions in a supervised fashion. The intuition is this: instead of training a new model every time we have a small dataset, we train one big model ahead of time that learns how to quickly solve many types of small problems. Later, when we give it a new small dataset for some downstream task, it can make predictions instantly, as if it had done Bayesian inference from scratch – without needing to run an expensive training algorithm.

TabPFN [182, 183] is a new and promising model for tabular problems. The approach involves first generating a large collection of synthetic tabular datasets and training a Transformer-based neural network [184] to solve these synthetic prediction tasks. Then, leveraging so-called in-context learning (ICL), the trained model is applied to new, unseen real-world datasets by providing training samples as context, enabling the model to predict labels for the unseen data in a single pass. This allows TabPFN to adapt to new datasets rapidly and without additional gradient updates, and unlike traditional supervised learning pipelines that require model training or adaptation for each new task. TabPFN gives state-of-the-art performances on datasets with up to 10,000 samples and 500 features, beating baselines such as XGBoost and Random Forest.

FMs for survival analysis. In survival analysis, foundation models have also started to emerge. Conceptually, one may replace the base neural network in models such as DeepSurv [28] or MENSA [48] with a Transformer architecture pretrained on large amounts of raw or synthetic data using self-supervised learning [185]. The resulting model can then be adapted to downstream tasks through, e.g., fine-tuning, potentially improving data efficiency and robustness to distribution shifts [186].

MOTOR (Many Outcome Time Oriented Representations) is a recently published foundation model for time-to-event prediction [124]. It combines a Transformer

backbone with a piecewise exponential survival head [187], which approximates the hazard using P piecewise-constant intervals. For patient i, event j, task k, and interval p, MOTOR predicts a hazard λ_{ijkp}, yielding the standard piecewise exponential likelihood

$$\mathcal{L}(U_{ijkp} \mid \lambda_{ijkp}) = \left(\lambda_{ijkp} \exp(-\lambda_{ijkp} U_{ijkp})\right)^{\delta_{ijkp}} \left(\exp(-\lambda_{ijkp} U_{ijkp})\right)^{1-\delta_{ijkp}}. \tag{7.3}$$

This permits non-constant hazards while retaining closed-form survival and likelihood functions. MOTOR's main novelty lies in scale: it is pretrained on 55M patient records containing 9B clinical events and evaluated across 19 tasks and three healthcare datasets. Pretraining enables a 143M-parameter model to transfer effectively to smaller downstream datasets, but at substantial computational cost. Although the reported gains over deep learning and tree-based baselines are impressive, comparisons between FM and non-FM approaches remain difficult when architectures, training procedures, and other factors differ substantially [6]. Thus, observed gains may partly reflect representation learning and model scale rather than survival-specific modelling choices.

Still, MOTOR and foundation models in general represent an exciting avenue for future research in survival prediction, with many open questions yet to be answered. Very recent work has also begun extending the PFN paradigm directly to survival analysis. SurvivalPFN [188] and SurvPFN [189] pretrain on synthetic censored survival tasks and perform prediction through in-context inference, thereby avoiding task-specific model training at inference time. These approaches are particularly interesting because they combine the amortized inference perspective of PFNs with the statistical structure of censored time-to-event data. However, both methods are still recent, and it remains to be seen how well such models generalize across datasets, censoring mechanisms (*e.g.*, dependent censoring), sample sizes, and clinical domains. Together with models such as MOTOR, they nevertheless suggest that foundation models may become an important direction for future survival modelling.

Appendix A

Uncertainty Estimation in Deep Bayesian Survival Models

Abstract: Bayesian methods can express uncertainty about their predictions, but have seen little adaptation in survival analysis using neural networks. Proper uncertainty estimation is important in high-risk domains, such as the healthcare or medical field, if machine learning methods are to be adopted for decision-making purposes, however, uncertainty estimation is a known shortcoming of neural networks. In this paper, we introduce the use of Bayesian inference techniques for survival analysis in neural networks that rely on the Cox proportional hazard assumption, for which we discuss a new flexible and effective architecture. We implement three architectures: a fully-deterministic neural network that acts as a baseline, a Bayesian model using variational inference, and one using Monte-Carlo Dropout. Our comprehensive experiments show that on the WHAS500 dataset, Bayesian techniques improve predictive performance over the state-of-the-art neural networks and on the larger SEER and SUPPORT datasets provide comparable performance. In all experiments, training with Monte Carlo Dropout is significantly faster than training with variational inference. Our Bayesian models additionally provide quantification of both aleatoric and epistemic uncertainty, which we exhibit by plotting 95% confidence intervals around the survival function and showing a probability density function of the survival time. Our work motivates further work in leveraging uncertainty for survival analysis using neural networks.

Authors: Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen.

Status: Published in the proceedings of the 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2023.

Uncertainty Estimation in Deep Bayesian Survival Models

Christian Marius Lillelund, Martin Magris and Christian Fischer Pedersen
Department of Electrical and Computer Engineering, Aarhus University
Finlandsgade 22, 8200 Aarhus N, Denmark
Emails: cl@ece.au.dk, magris@ece.au.dk and cfp@ece.au.dk

***Abstract*—Bayesian methods can express uncertainty about their predictions, but have seen little adaptation in survival analysis using neural networks. Proper uncertainty estimation is important in high-risk domains, such as the healthcare or medical field, if machine learning methods are to be adopted for decision-making purposes, however, uncertainty estimation is a known shortcoming of neural networks. In this paper, we introduce the use of Bayesian inference techniques for survival analysis in neural networks that rely on the Cox proportional hazard assumption, for which we discuss a new flexible and effective architecture. We implement three architectures: a fully-deterministic neural network that acts as a baseline, a Bayesian model using variational inference, and one using Monte-Carlo Dropout. Our comprehensive experiments show that on the WHAS500 dataset, Bayesian techniques improve predictive performance over the state-of-the-art neural networks and on the larger SEER and SUPPORT datasets provide comparable performance. In all experiments, training with Monte Carlo Dropout is significantly faster than training with variational inference. Our Bayesian models additionally provide quantification of both aleatoric and epistemic uncertainty, which we exhibit by plotting 95% confidence intervals around the survival function and showing a probability density function of the survival time. Our work motivates further work in leveraging uncertainty for survival analysis using neural networks.**

***Index Terms*—uncertainty estimation, neural networks, survival analysis, variational inference, MC Dropout**

I. Introduction

The Cox's proportional hazards model [1] has long been the standard approach for survival analysis in many healthcare applications, but recent advances in machine learning research have made neural networks (NNs) a powerful tool for survival regression [2]–[4]. Such networks have proven to provide solid performance in terms of ranking and prediction accuracy in survival analysis applications, but traditional maximum-likelihood-based methods notoriously perform poorly when data is sparse [5] and proper uncertainty estimation is a known shortcoming of NNs. In healthcare and biomedical applications, confidence in the prognostic outcome of, e.g., a cancer diagnosis, is essential for medical doctors to make the decision on individual treatment. Bayesian models can express uncertainty about their predictions, specifically aleatoric and epistemic. The former refers to the inherent noise in the data, and the latter to the lack of knowledge of the model [6].

This work was supported by the PRECISE project under Grant Agreement No. AAL-2021-8-90-CP by the European AAL Association.

Bayesian Neural Networks (BNNs) are stochastic NNs trained using Bayesian inference, e.g., variational inference (VI) [7] and Monte-Carlo Dropout (MCD) [5]. A BNN was adopted for survival analysis by [8], but their work used pseudo survival probabilities instead of censored ones.

In this work, we propose BNNs as a tool for uncertainty estimation in survival analysis models. To the authors' best knowledge, no prior work has explored the use of BNNs for Cox survival analysis. We consider both VI and MCD as estimation approaches, and, to the end of comparing Bayesian and non-Bayesian approaches, we introduce a feasible network architecture and furthermore adopt five existing reference models. We observe agreement and juxtaposition between deterministic point estimates and the Bayesian ones, indicating the overall validity of our setup. In-depth analyses unveil an advantage of adopting a Bayesian framework, especially in small-sized datasets. The intrinsic probabilistic dimension of BNNs naturally allows for recovering uncertainty estimates, estimating predictive errors and plotting confidence bands. This provides a much more comprehensive description and understanding of the data and model under consideration. Source code is available at: https://github.com/thecml/UE-BNNSurv

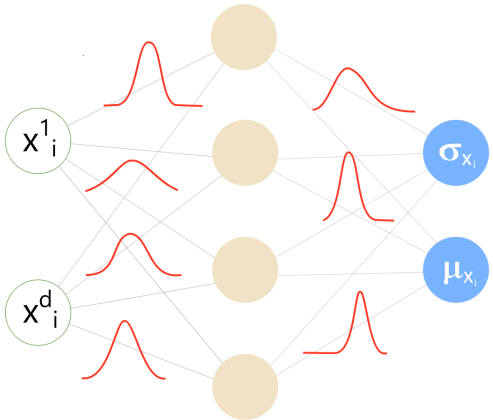


Fig. 1: Illustration of the BNN architecture with a single hidden layer, where weights and biases are treated as random variables. The output nodes provide the mean and standard deviation of a risk score, $\hat{r}_i$, as Gaussian samples, i.e., $\hat{r}_i \sim \mathcal{N}(\mu_{\boldsymbol{x}_i}, \sigma_{\boldsymbol{x}_i})$ for a d-dimensional sample $\boldsymbol{x}_i$.

II. Fundamentals

A. Elements of survival data analysis and notation

Modelling the probability that an event occurs at time T later than t, i.e., the *survival* probability $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$, is a major task in survival analysis. A main ingredient is the so-called hazard function $h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T > t)/\Delta t$. This corresponds to the death *rate* at an instant after time t, giving survival past that time [9]. The hazard function is related to the survival function through $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density associated with T, $f(t) := \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t)/\Delta t$, i.e., the instantaneous rate of death at time t. In this view, $h(t)$ is the density of T conditional on $T > t$, and the functions $S(t)$, $h(t)$, $f(t)$, all correspond to equivalent ways of describing the distribution of T, formalizing, e.g., the intuition that higher values for $h(t)$ correspond to higher death probabilities.

B. Cox Proportional Hazard Model

For the task of fitting a regression model to survival times, let $\mathcal{D} = \{(y_i, \delta_i, \boldsymbol{x}_i,)\}_{i=1}^N$ be the data, where i denotes the i-th individual. With T_i being the event time and C_i the censoring time, $y_i = \min(T_i, C_i)$, and $\delta_i = 1$ if $T_i \leq C_i$ (zero otherwise). We denote $\boldsymbol{x}_i$ as a vector of d covariates.

Cox's Proportional Hazards (CoxPH) model [1] assumes a conditional individual hazard function of the form $h(t|\boldsymbol{x}_i) = h_0(t)\exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$. The risk score is denoted as $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$. In [1] f is set to a linear function of the covariates, i.e., $f(\boldsymbol{\theta}, \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\theta}$, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is derived by numerically maximizing the (partial) log-likelihood:

$$\begin{aligned}\log p(\mathcal{D}|\boldsymbol{\theta}) &= \sum_{i:\delta_i=1} \log \frac{h(y_i|\boldsymbol{x}_i)}{\sum_{j:T_j \geq T_i} h(y_i|\boldsymbol{x}_j)}, \\ &= \sum_{i:\delta_i=1} f(\boldsymbol{\theta}, \boldsymbol{x}_i) - \sum_{j:T_j \geq T_i} f(\boldsymbol{\theta}, \boldsymbol{x}_j). \end{aligned} \quad (1)$$

Different estimators have been derived for the baseline hazard function, e.g., [10], enabling the estimation of the survival function as $\hat{S}(t) = \hat{S}_0(t)^{\exp(\boldsymbol{x}_i\hat{\boldsymbol{\theta}})}$, with $\hat{S}_0(t) = \exp(-\int_0^t \hat{h}_0(t)\mathrm{d}t)$. A purely non-parametric approach, not using covariates' information, is provided by the Kaplan-Meier estimator, used as a reference in Fig. 2. For more details, see, e.g., [9].

C. Variational inference

We denote the data as $\mathcal{D}$, the likelihood as $p(\mathcal{D}|\boldsymbol{\theta})$ and prior distribution on the parameter of interest $\boldsymbol{\theta}$ as $p(\boldsymbol{\theta})$. The target of Bayesian inference is the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})/p(\boldsymbol{\theta})$. The distribution of unobserved (new) data conditioned on $\mathcal{D}$, known as predictive distribution, is obtained from the posterior through marginalization over the parameter space Θ, i.e., $p(\boldsymbol{x}_{\text{new}}|\mathcal{D}) = \int_\Theta p(\boldsymbol{x}_{\text{new}}|\mathcal{D}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}$.

The direct computation of the posterior is challenging as, in general, the term $p(\mathcal{D})$ is intractable. Sampling methods do not scale well in high dimensions and are time-consuming: VI approximates the true and unknown posterior distribution with a distribution $q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})$ chosen within a class of tractable parametric distributions $\mathcal{Q}$. Hereafter $\mathcal{Q}$ is the class of multivariate Gaussians with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\zeta} \equiv \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. VI seeks for the best approximation in $\mathcal{Q}$ to $p(\boldsymbol{\theta}|\mathcal{D})$ by minimizing the Kullback-Leibler (KL) divergence from $q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathcal{D})$, i.e., by recovering the variational parameter $\boldsymbol{\zeta}^\star$ via minimizing:

$$\mathrm{KL}(q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) := \mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\zeta}}}\left[\frac{\log p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}|\mathcal{D})}{\log q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})}\right], \quad (2)$$

commonly attained via the Bayes-by-backprop method [11]. A NN estimated with Bayesian inference is commonly referred to as Bayesian Neural Network (BNN), and our approach provides a BNN-based extension of the non-Bayesian and fully-linear standard Cox model.

D. Monte-Carlo Dropout

Optimizing Eq. (2) requires sampling, estimating stochastic gradients, and involves several caveats and difficulties [6]. Though Monte-Carlo Dropout (MCD) has been historically developed as a computationally efficient method for regularization, it has been shown to have a connection with Bayesian inference. Indeed, NNs with dropout applied at every layer are equivalent to approximate VI in deep Gaussian Processes [12]. Dropout optimizes the KL objective between an approximate distribution and the posterior of a deep Gaussian Process as a mixture of Gaussians [13]. Such a setup is limited compared to VI, yet its simplicity makes it attractive and a widely adopted baseline for comparison with alternative Bayesian approaches. So far, within NN applications in survival analysis, MCD has not been adopted for uncertainty quantification.

III. Experiments and results

A. Network architecture

For the purpose of comparability in the results, consistency in the hyperparameters, and evaluation fairness, we propose a multi-layer perceptron (MLP) backbone network architecture. A MLP is fully-connected feedforward NN; the number of hidden layers and neurons is a hyperparameter tuned based on the data, see, e.g., [6]. Using an MLP is aligned with the literature, as capable of providing solid predictive performance [2]–[4].

The proposed architecture is depicted in Fig. 1. There are three configurations: (i) baseline: a fully-deterministic NN obtained by removing the σ_x neuron and optimizing the objective Eq. (1). (ii) Bayesian model trained with VI, learning the variational approximation by optimizing Eq. (2). (iii) Bayesian model trained with MCD (25% dropout rate).

In relation to Sec. II-C, Eq. (2) adopts the Cox likelihood (1) where $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ corresponds to the MLP neural network and $\boldsymbol{\theta}$ to its parameter (collection of weight and biases), the objective of the Bayesian inference. Whereas the epistemic uncertainty in the model is captured by the Bayesian framework, to capture the aleatoric uncertainty, the network's outputs are sampled from a Gaussian [14], Fig. 1.

B. Datasets and models

For our empirical analyses, we use the openly-available WHAS500 [15], SEER [16] and SUPPORT [17] datasets; they differ in the number of samples and the percentage of censored data (see Tab. I). After imputing missing values by sample mean for real-valued covariates or mode for categorical covariates, applying a z-score data normalization and one-hot encoding categorical covariates, we adopt a 70%-30% train-test split for training and testing.

We implement the Cox model [1], two related traditional survival models CoxNet [18], RSF [19], and three models based on a NN architecture, DSM [3], DeepSurv [2], and the MLP discussed in Sec. III-A. For all datasets and models, we use Bayesian optimization [20] to tune the respective models' hyperparameters over ten iterations using 5-fold cross-validation. This includes number of iterations, batch size and network architecture if applicable, and is done solely on the 70% training split. We use the hyperparameters leading to the highest average concordance-index (Harrell's) on the validation folds.

C. Results

Table I reports the predictive performance of our baseline MLP and its VI and MCD variants, and literature benchmarks. We adopt four evaluation metrics: Harrell's concordance-index (CI_H) [21], Uno's concordance-index (CI_U) [22], and the negative log-likelihood (NLL). For VI and MCD, such measures are constructed based on predictive means over 100 posterior draws. Model were trained until convergence (no improvement was seen in test CI_H).

Concerning the baseline MLP models and their VI and MCD variants, on the smallest WHAS500 dataset we observe that modelling both aleatoric and epistemic uncertainty by adopting Bayesian methods improves the ranking and accuracy performance in terms of CI_H, CI_U and NLL. In the mid-sized SEER dataset, the effect of the different estimation approaches for the MLP reduces, and all the measures are generally aligned with each other, whereas for the SUPPORT dataset, maximum likelihood can take full advantage of the considerable size of the data and outperforms, consistently although slightly, Bayesian methods.

With respect to the existing models, our work outperforms NN-based solutions on the small dataset (high CI_H, CI_U and low IBS) and we see an advantage in combining the MLP architecture with Bayesian inference for ranking and accuracy in predicting the survival function. In the mid-sized dataset, the performance of all the models is similar, whereas in the large dataset, it seems that methods based on NNs clearly outperform the traditional CoxPH, CoxNet and RSF in terms of CI_H, CI_U and NLL, though behave similarly in terms of IBS.

Regardless, the use of Bayesian methods does not deteriorate performance metrics in any of our experiments. This demonstrates that aleatoric and epistemic uncertainty can be included in survival analysis using NNs at no additional cost, besides the increase in training time for VI. The leftmost panel in Fig. 2 shows that our methods provide consistent survival function estimates that align with the Kaplan-Meier estimator. VI and MCD additionally capture the aleatoric and epistemic uncertainty in the confidence intervals, which provides a clear predictive advantage and facilitates better adaptation for decision support. In the central panel, observe that grade 1 and 2 tumors have a statistically different impact on survival times approximately after 25 years, a detail that non-probabilistic methods could not capture. This is an ideal application for modelling the underlying uncertainty in, e.g., different treatments. In the rightmost panel, by following [23], we plot the survival time of a randomly selected individual from the test set: the distribution of the exponential random samples with a rate parameter randomly sampled from the MCD predictive distribution is quantitatively different than the one obtained based on the predictive mean, which shows lighter tails, and thus a higher likelihood of early-experiencing the event compared to the non-Bayesian MLP estimation.

TABLE I: Performance metrics on the test sets. N: total sample size, C: pct. of censored data, d: number of covariates.

(a) WHAS500 ($N = 500$, $C = 57\%$, $d = 14$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.05s	0.806	0.785	0.152	0.849
CoxNet [18]	0.04s	0.807	0.785	0.149	0.832
RSF [19]	0.07s	0.785	0.754	0.167	
DSM [3]	2.78s	0.785	0.776	0.195	
DeepSurv [2]	0.24s	0.774	0.758	0.185	
This work:					
Baseline (MLP)	10.67s	0.774	0.761	0.154	0.890
+ VI	2m 53s	0.801	0.776	0.142	0.855
+ MCD	6.47s	0.798	0.772	0.160	0.884

(b) SEER ($N = 4024$, $C = 85\%$, $d = 28$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.23s	0.739	0.726	0.105	0.582
CoxNet [18]	0.23s	0.740	0.727	0.105	0.582
RSF [19]	2.05s	0.724	0.714	0.113	
DSM [3]	2.09s	0.738	0.737	0.115	
DeepSurv [2]	0.55s	0.746	0.746	0.113	
This work:					
Baseline (MLP)	24.92s	0.743	0.745	0.107	0.587
+ VI	6m 15s	0.741	0.733	0.109	0.589
+ MCD	26.81s	0.741	0.726	0.110	0.588

(c) SUPPORT ($N = 8873$, $C = 32\%$, $d = 14$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.36s	0.583	0.586	0.222	2.862
CoxNet [18]	0.10s	0.581	0.585	0.224	2.865
RSF [19]	8.81s	0.569	0.571	0.232	
DSM [3]	4.61s	0.603	0.606	0.222	
DeepSurv [2]	0.70s	0.618	0.617	0.221	
This work:					
Baseline (MLP)	58.72s	0.615	0.617	0.210	2.836
+ VI	40m 28s	0.606	0.607	0.310	2.840
+ MCD	60.87s	0.604	0.606	0.216	2.845

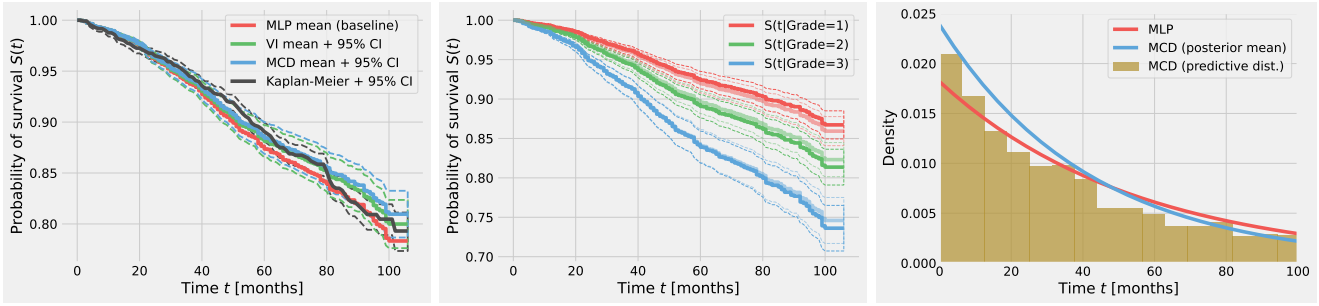


Fig. 2: Model inference on the SEER dataset. Left & center: survival function for different estimation approaches and for various grades of the tumor severity in breast cancer cases (Solid lines: VI; Opaque lines: MCD; Dashed lines: 95% conf. int.). Right: probability density function of the survival time, see Sec. III-C for details.

IV. Conclusion

We adopt variational inference and Monte-Carlo Dropout to provide Bayesian estimation of survival probability and risk in Cox models using neural networks. By proposing a suitable and effective neural network architecture, we perform extensive experiments over six models and three datasets. These show that Bayesian techniques can in situations where data is sparse increase predictive performance of survival models compared to other neural network approaches. We encourage further work within our framework, which we have shown to be effective in providing an immediate quantification of combined aleatoric and epistemic uncertainty. Such information can support risk-aware decision-making in high risk domains, such as the healthcare or medical domain.

References

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[2] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, 2018.

[3] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, 2021.

[4] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[5] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016.

[6] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: an algorithmic survey," *Artificial Intelligence Review*, pp. 1–51, 2023.

[7] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[8] D. Feng and L. Zhao, "Bdnnsurv: Bayesian deep neural networks for survival analysis using pseudo values," *Journal of Data Science*, vol. 19, no. 4, pp. 542–554, 2021.

[9] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*, 2nd ed. Spinger, 2021.

[10] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review*, pp. 45–57, 1975.

[11] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, 2015, p. 1613–1622.

[12] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 1050–1059.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] M. Valdenegro-Toro and D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 1508–1516.

[15] D. W. Hosmer, S. Lemeshow, and S. May, *Applied survival analysis: regression modeling of time-to-event data*, 2nd ed. Wiley-Interscience, 2008.

[16] L. A. Gloeckler Ries, M. E. Reichman, D. R. Lewis, B. F. Hankey, and B. K. Edwards, "Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) program," *Oncologist*, vol. 8, no. 6, pp. 541–552, 2003.

[17] W. A. Knaus, "The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults," *Annals of Internal Medicine*, vol. 122, no. 3, pp. 191–203, 1995.

[18] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of statistical software*, vol. 39, no. 5, pp. 1–13, 2011.

[19] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[20] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[21] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[22] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[23] P. C. Austin, "Generating survival times to simulate cox proportional hazards models with time-varying covariates," *Statistics in Medicine*, vol. 31, no. 29, pp. 3946–3958, 2012.

Appendix B

Predicting Survival Time of Ball Bearings in the Presence of Censoring

Abstract: Ball bearings find widespread use in various manufacturing and mechanical domains, and methods based on machine learning have been widely adopted in the field to monitor wear and spot defects before they lead to failures. Few studies, however, have addressed the problem of censored data, in which failure is not observed. In this paper, we propose a novel approach to predict the time to failure in ball bearings using survival analysis. First, we analyze bearing data in the frequency domain and annotate when a bearing fails by comparing the Kullback-Leibler divergence and the standard deviation between its break-in frequency bins and its break-out frequency bins. Second, we train several survival models to estimate the time to failure based on the annotated data and covariates extracted from the time domain, such as skewness, kurtosis and entropy. The models give a probabilistic prediction of risk over time and allow us to compare the survival function between groups of bearings. We demonstrate our approach on the XJTU and PRONOSTIA datasets. On XJTU, the best result is a 0.70 concordance-index and 0.21 integrated Brier score. On PRONOSTIA, the best is a 0.76 concordance-index and 0.19 integrated Brier score. Our work motivates further work on incorporating censored data in models for predictive maintenance.

Authors: Christian Marius Lillelund, Fernando Pannullo, Morten Opprud Jakobsen, and Christian Fischer Pedersen.

Status: Presented at the AAAI Fall Symposium 2023: Survival Prediction Algorithms, Challenges & Applications (SPACA). Preprint: arXiv:2309.07188.

Predicting Survival Time of Ball Bearings in the Presence of Censoring

Christian Marius Lillelund*, Fernando Pannullo, Morten Opprud Jakobsen, Christian Fischer Pedersen

Department of Electrical and Computer Engineering, Aarhus University
Finlandsgade 22, 8200 Aarhus N, Denmark
{cl,202102261,morten,cfp}@ece.au.dk

Abstract

Ball bearings find widespread use in various manufacturing and mechanical domains, and methods based on machine learning have been widely adopted in the field to monitor wear and spot defects before they lead to failures. Few studies, however, have addressed the problem of censored data, in which failure is not observed. In this paper, we propose a novel approach to predict the time to failure in ball bearings using survival analysis. First, we analyze bearing data in the frequency domain and annotate when a bearing fails by comparing the Kullback-Leibler divergence and the standard deviation between its break-in frequency bins and its break-out frequency bins. Second, we train several survival models to estimate the time to failure based on the annotated data and covariates extracted from the time domain, such as skewness, kurtosis and entropy. The models give a probabilistic prediction of risk over time and allow us to compare the survival function between groups of bearings. We demonstrate our approach on the XJTU and PRONOSTIA datasets. On XJTU, the best result is a 0.70 concordance-index and 0.21 integrated Brier score. On PRONOSTIA, the best is a 0.76 concordance-index and 0.19 integrated Brier score. Our work motivates further work on incorporating censored data in models for predictive maintenance.

Introduction

Ball bearings find extensive use in various rotary machines, but are susceptible to defects, such as contamination wear, poor lubrication, and improper mounting (Kim, An, and Choi 2017). Monitoring wear in bearings plays an important role in predictive maintenance programs, and engineers have traditionally used frequency-based vibration analysis tools to assess the severity of emerging mechanical damages (Randall 2021; Randall and Antoni 2011). Predictive maintenance should be done as soon as a bearings' operating characteristics start to deviate significantly from its normal operating state in order to avoid actual failure and eventual breakdown of the machinery. To this end, researchers have applied various machine learning (ML) algorithms, particularly neural networks, to train models that can predict the remaining useful life (RUL) of bearings (Guo et al. 2017; Zheng et al. 2018; Al Masry et al. 2019; Wang et al. 2020, 2022; Xu et al. 2022). This strategy has shown initial success and high predictive performance, but support for censored observations is generally overlooked in the field and only few studies have attempted to build models that support it (Widodo and Yang 2011b,a; Hochstein et al. 2013; Wang et al. 2022). Given the rarity ball bearings failures, censored observations are common, and simply ignoring them can lead to a loss of efficiency and introduce estimation bias (Stepanova and Thomas 2002). Survival analysis is a type of regression that can leverage censored data (Gareth et al. 2021). Current works in bearing prognostics using survival analysis do not provide any evaluation of predictive accuracy or ranking performance (Wang et al. 2022), and the evaluated methods, such as the nonparametric Kaplan-Meier (KM) estimator (Kaplan and Meier 1958), are often mainly descriptive, too simplistic or do not model the relationship between covariates and outcome (Widodo and Yang 2011a).

In this paper, we propose a novel approach for predictive maintenance of ball bearings using survival analysis (see Fig. 1). First, we process the raw bearing data in the frequency-domain to identify at which point in time a bearing starts to deviate significantly from its normal frequency characteristics. This happens well in advance of actual bearing failure, at a point where the faulty component can be readily replaced by the maintenance staff. This is our event detection algorithm, that we use to annotate bearings in two datasets (XJTU and PRONOSTIA) by the time this deviation occurs. The algorithm is a simple distance metric, and using it for forecasting is computationally expensive and prone to error. Therefore, we use the annotated data to train several survival models, that can predict the probability of failure as a progressive estimate over time, contrary to traditional RUL regression, that only offer a mere point estimate of the failure time. This method also enables us to quantify survival probabilities between groups of bearings by their time-domain features.

We compare our work to the method proposed by Xu et al. (Xu et al. 2022) to estimate the RUL of bearings in the PRONOSTIA dataset. However, since our methodology, the annotation algorithm we have used and the bearing data differ from (Xu et al. 2022) and others, we perform an indirect comparison. Source code is available at: https://github.com/thecml/ball-bearing-survival

*Corresponding author.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

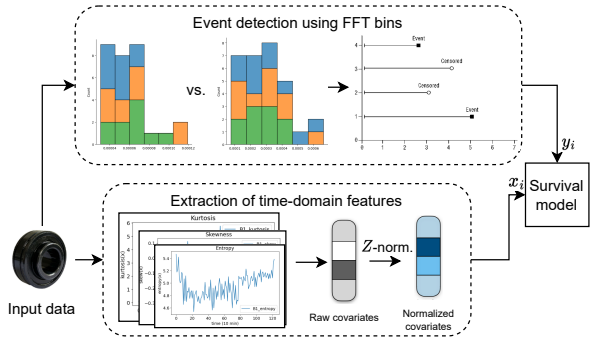


Figure 1: Outline of the proposed solution. Raw data is used for event detection (top) and feature extraction (bottom) before training a survival model. Illustrations partly from (Nagpal, Li, and Dubrawski 2021).

Fundamentals

Elements of Ball Bearings

Ball bearings are used to carry rotating loads by separating elements in motion with two bearing races carried by steel or ceramic balls. Ball bearings can support either axial or radial loads, or a combination.

The health of a bearing can be tracked by monitoring the presence of defects in the bearing components. Bearing defects excite the bearing components and the resulting vibrations can be measured. By following the development of energy in selected frequency bands, ballpass frequency outer race (BPFO), ballpass frequency inner race (BPFI), fundamental train frequency (FTF), ball spin frequency (BSF), and shaft frequency (FS), each relating to a bearing's components, the bearing's health and remaining useful lifetime can be tracked (Wang et al. 2017; Randall 2021). During infant bearing defects, the vibration signals will be weak and amplitude modulated with eigenfrequencies related to the bearing components, necessitating use of wideband Piezo or MEMS accelerometers (1–20 kHz) to capture any vibrations.

Elements of Survival Analysis

Survival analysis is a form of regression modeling that studies the time to an event, which can be partially observed (i.e., censored). Survival data contain the observed covariates, the time to event, and a label indicating whether the event occurred or was censored. We treat the survival time as discrete and limit the time horizon to a finite duration, denoted as $T = \{T_0, ..., T_{max}\}$, with T_{max} representing the predefined maximum time horizon (e.g., 1 year). Within this framework, we consider bearing failure as the event of interest, and assume that exactly one such event will occur eventually for each observation (e.g., a bearing will eventually fail, but from only one cause). However, not all events of interest are always observed due to factors such as bearings being decommissioned before failure occurs or simply running problem-free after T_{max}, resulting in right-censored data.

Survival analysis models the probability that an event occurs at time T later than t, which is denoted as the survival probability $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$. To estimate $S(t)$, we use the so-called hazard function, $h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T > t)/\Delta t$, which corresponds to the failure rate at an instant after time t, given survival past that time (Gareth et al. 2021). The relationship between the survival and hazard function is given by $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density associated with T, $f(t) := \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t)/\Delta t$, which is the instantaneous rate of failure at time t. In this regard, $h(t)$ is the density of T conditional on $T > t$, and the functions $S(t)$, $h(t)$, $f(t)$, are equivalent ways of describing the distribution of T, which formalizes the intuition that higher values for $h(t)$ correspond to higher failure probabilities. In order to fit a regression model to survival times, Cox's Proportional Hazards (CoxPH) is a popular choice (Cox 1972). The model assumes a conditional individual hazard function of the form $h(t|\boldsymbol{x}_i) = h_0(t) \exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$, where i denotes the i-th individual, $\boldsymbol{x}_i$ is a vector of d covariates and $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ is a linear function of the covariates.

Methods and Materials

Datasets

We consider two bearing datasets in this study, XJTU (Wang et al. 2020) and PRONOSTIA (Nectoux et al. 2012). For both datasets, a single bearing under test is subjected to radial load during rotation, and vibrations are captured. The bearing is instrumented with two piezoelectric accelerometers, mounted horizontally and vertically, and sampled at 25.6 kHz. The resulting data are saved as raw comma-separated values. In both experiments, the bearings are run to failure under very high load (C/P$\leq$4) to accelerate degradation. This significantly increases the risk of initiating bearing defects but also comes with the risk of local flash heating and subsequent uncontrolled damage of a bearing.

XJTU: The dataset is generated from 15 deep grove ball bearings. Three types of tests are conducted with different loads and speeds, and five bearings are used in each test. In our experiments, we use five bearings with the following characteristics: a operating condition (C/P) of 1.1, a radial force of 12.0 kN and a rotating speed of 2100 RPM.

PRONOSTIA: The dataset is generated from 6 deep grove ball bearings. Three types of tests are conducted with different loads and speeds, where two bearings are used in each test. In our experiments, we use two bearings with the following characteristics: a C/P of 0.8, a radial force of 5.0 kN and a rotating speed of 1500 RPM.

Feature extraction

Features come from accelerometer sensors placed on the ball bearings at a angle of 90 degrees (referred to as the X-axis and Y-axis). Feature extraction starts by discretizing into bins a sequence of raw samples, i.e., $\boldsymbol{x} = (x)_{i=1}^{N}$. We then apply the expressions in Tab. 1 to each of these bins and obtain a total of twelve time-domain features for the whole lifetime of a bearing.

Absolute mean	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert$
Standard deviation	$\sigma = (\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2)^{1/2}$
Skewness	$SK = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3/\sigma^3$
Kurtosis	$K = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4/\sigma^4$
Entropy	$H = -\sum_{i=1}^{N} P(x_i)\log P(x_i)$
Root-mean-square	$RMS = (\frac{1}{N}\sum_{i=1}^{N} x_i^2)^{1/2}$
Max value	$max = \max(\lvert \boldsymbol{x} \rvert)$
Peak-To-Peak	$P2P = \max(\lvert \boldsymbol{x} \rvert) - \min(\lvert \boldsymbol{x} \rvert)$
Crest factor	$CR = max/RMS$
Clearance factor	$CL = x_p/(\frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert^{1/2})^2$
Shape factor	$SH = RMS/\bar{x}$
Impulse	$IM = max/\bar{x}$

Table 1: Features expressed in relation to a signal frame $\boldsymbol{x} = (x)_{i=1}^{N}$.

Event detection

In this paper, we propose an event detection algorithm to label ball bearing data to be used in a survival model (see Fig. 2). Collecting the raw bearing data, performing the Hilbert transformation, and applying the fast Fourier transform (FFT), produces five main bins that make up a probability density function (PDF) for appropriate time windows. This allows us to observe different conditions of the bearing through time and enables us to identify the behavior and characteristic that describe a bearing in a good or bad state.

After preprocessing, we apply the Kullback-Leibler divergence and standard deviation (SD) formulas in conjunction to detect changes in the PDF over time that possibly can identify the event of interest (see Fig. 2). We consider two distributions: T_0 is the reference distribution and T_i is one over a moving time window. In this context, the KL divergence is a statistical method that measures the difference between two probability distributions, denoted $D_{\mathrm{KL}}(P\|Q)$, as a measure of how one probability distribution P is different from the probability distribution Q. This measures changes in entropy related to the distribution shape. We also compute the SD of T_0 and T_i, assuming they are Gaussian. By comparing the two PDFs' by SD and KL divergence from the break-in phase to the end of the bearing's lifetime, the progression of the PDF discrepancies are obtained. The KL is illustrated as a solid blue line in Fig. 3. The KL and SD progressions are similar, but not the same, because of changes to the underlying mechanics of the bearing throughout its lifetime. The KL progression line tends to reach a peak level (key point #1 in Fig. 3), i.e., high entropy and toward diversification, which we use as a threshold (red dotted line in Fig. 3, the highest point of the blue line plus 10%). Later, the threshold will be used to establish the event. After the break-in phase, the line reaches a plateau (key point #2 in Fig. 3), i.e., low entropy and toward equalization, and then begins to rise again when the bearing starts to degrade, i.e., high entropy and towards diversification between the PDFs, and then failure occurs (key point #3 in Fig. 3). We establish a KL-threshold and a SD-threshold and whichever of two are crossed last indicates an event.

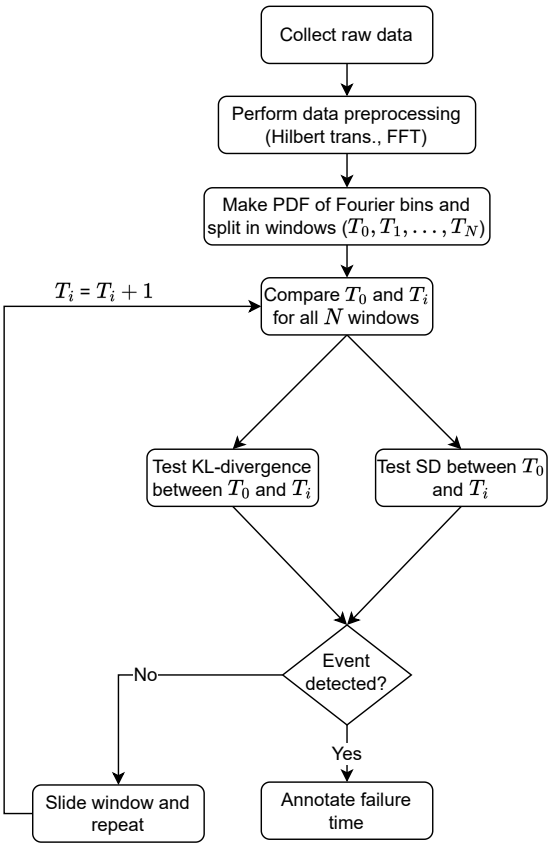


Figure 2: Flowchart of proposed event detection algorithm. The entropy or deviation between the signal at time T_0 and time T_i is continuously compared over the lifetime of the bearing.

Data preprocessing

The XJTU and PRONOSTIA datasets contain only few actual bearings, so the number of data points and events is low. To remedy this, we perform data augmentation to create synthesized versions of the two datasets. This process is done independently for each bearing. First, we up-sample (double) the data points by treating each bearing's X-axis and Y-axis as independent bearings. Second, we do Constrained Bootstrapping as in (Zelterman, Le, and Louis 1996). Third, we divide the timeseries into 20 slices of the same size covering the entire lifetime, and for each slice assign a time to event or censoring, and the covariates at that time. This essentially re-samples a single event several times, while preserving the correlation between covariates and event time. We introduce a censoring rate of 20% in both datasets by simulating that some bearings did not experience the event.

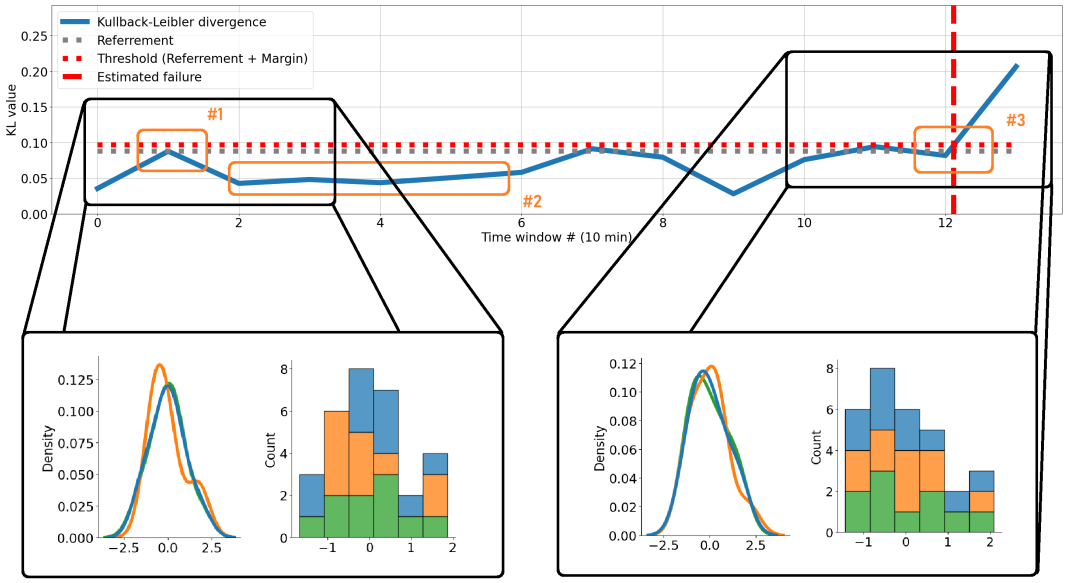


Figure 3: Top: Event detection based on KL divergence. The key point #1 is the break-in period, #2 is the plateau and #3 is the break-out period. The solid blue line is the KL divergence. The dotted gray line is the maximum entropy reached during the break-in period. The dotted red line is the threshold given by the maximum entropy plus 10%. Bottom left: FFT bins of the signal at break-in time. Bottom right: FFT bins of the signal at failure time.

Survival models

The following models are evaluated:

Cox Proportional Hazards (CoxPH): (Cox 1972) is the most commonly used regression model for survival data. It assumes a conditional individual hazard function. The risk score is estimated as a linear function of covariates and parameters, found by maximizing the partial log-likelihood.

Random Survival Forest (RSF): (Ishwaran et al. 2008) RSF is an ensemble of survival trees, where the data are recursively partitioned based on some splitting criterion, and similar data points based on the event of interest are put on the same node.

CoxBoost: (Hothorn et al. 2005) CoxBoost is an extension of traditional gradient boosting that supports survival data by minimizing the weighted empirical risk function as a least-squares problem.

DeepSurv: (Katzman et al. 2018) DeepSurv is a neural network that uses a Cox likelihood function to compute a relative risk score, which quantifies the likelihood of experiencing an event.

Deep Survival Machines (DSM): (Nagpal, Li, and Dubrawski 2021) DSM is a neural network that estimates the conditional survival function as a mixture of primitive distributions, either Weibull or Log-Normal.

Weibull AFT: (Lee and Wang 2013) The Accelerated Failure Time (AFT) model estimates the survival function using the Weibull distribution. It is a generalization of the exponential distribution, but does not assume a constant hazard rate, which allows for a broader application.

Experiments and Results

Setup

For our empirical analyses, we use the openly available XJTU and PRONOSTIA datasets, which differ in the number of bearings, operating conditions, and the number of samples. After preprocessing, we apply a z-score data normalization to all covariates and do a train-test split based on the number of bearings: for XJTU, we select three bearings for training and two for test, and for PRONOSTIA, we select one bearing for train and one for test. For all datasets and models, we tune the following hyperparameters over ten iterations using 5-fold cross-validation:

- CoxPH: Number of iterations and tolerance.
- RSF: Number of estimators, tree depth and split criterion.
- CoxBoost: Learning rate, tree depth, split criterion.
- DeepSurv: Iterations, learning rate and batch size.
- DSM: Iterations, learning rate and batch size.
- Weibull AFT: Penalizer coefficient.

Tuning is done solely on the training bearings. We use the hyperparameters leading to the highest average concordance index (Antolini's) on the validation folds to adjust the final models. Models are then evaluated on the test bearings.

Results

Table 2 shows the predictive performance of the survival models. We report the training time, Harrell's concordance index (CI) (Harrell, Lee, and Mark 1996), Antolini's time-dependent concordance index (CI_{td}) (Antolini, Boracchi, and Biganzoli 2005) and the integrated Brier score (IBS)

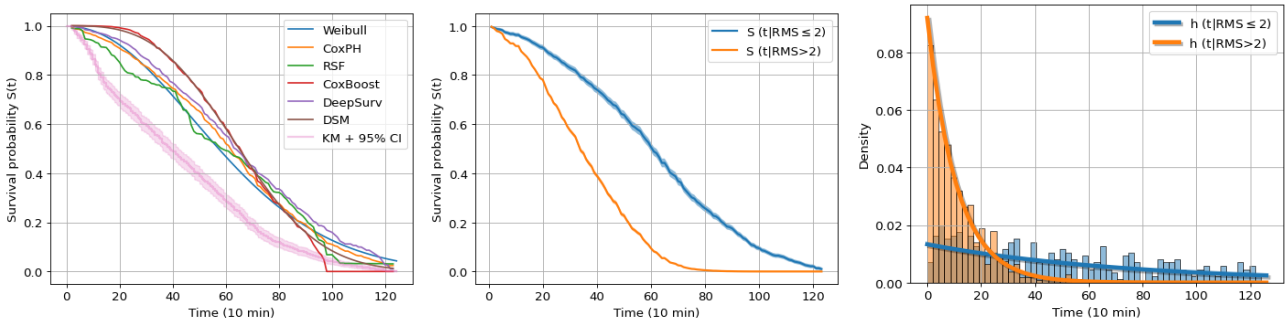


Figure 4: Model inference on the XJTU dataset. Left & center: survival function for different estimation approaches and for two separated groups based on RMS value (Solid lines: mean values; Shadow area: 95% conf. int.). Right: normalized survival times (bins) and hazard functions (solid lines) of the two RMS groups.

(Graf et al. 1999). Concerning the results on the XJTU dataset (see Tab. 2a), we observe that RSF, CoxBoost and DeepSurv perform the best in terms of ranking (CI) and DSM in terms of predictive accuracy. A CI of 0.70 means that 70 out of a 100 comparable pairs are ranked correctly. An IBS of 0.214 equals a 78.6% accuracy in predicting the survival function over all available times. Concerning the results on the PRONOSTIA dataset (see Tab. 2b), DSM, DeepSurv and WeibullAFT performs the best.

To compare our work to SOTA, (Xu et al. 2022) report a mean 70% accuracy (95% CI: 0.58-0.87) on eleven PRONOSTIA bearings. In our work, DeepSurv (Katzman et al. 2018) performs the best in terms of predicting the survival function, yielding a 0.197 integrated Brier score on the test set. This translates to a predictive accuracy of 80.3%, however, the results are not directly comparable, as (Xu et al. 2022) use simple least-squares regression as opposed to survival regression, employ a different training/test split and only report the estimation error at the end of the bearing lifetime, instead of at available times.

Figure 4 shows inference results on the two XJTU test bearings. We see that all evaluated models provide consistent estimates of the survival function (leftmost panel). The Kaplan-Meier is used as a reference here and estimated on the full dataset (five bearings). In the central panel, we split the test data into two groups by the root mean square (RMS) covariate value, $RMS \leq 2$ and $RMS > 2$, and compute the mean survival curve of the two separately using a Cox proportional hazards model. This shows a clear separation in predicted survival probability, already from $t = 10$, to convergence at $t = 120$. The shadowed area represents a 95% confidence interval, which is wider for the bearings in the first group. The plot presents an interesting application for predictive maintenance, as it can be used as decision support for maintenance personnel. In the rightmost panel, by following (Austin 2012), we computed 500 simulated survival times using a Cox model for the two RMS groups, and plotted their respective hazard function and survival time; the two hazard functions are quantitatively distinct, and the expected survival time for $RMS > 2$ is lower than for $RMS \leq 2$.

Conclusion

We have proposed survival analysis as a method to predict the risk of failure in ball bearings. First, we identify the point in time when a bearing starts to diverge significantly from its original frequency characteristic, which marks our event of interest, and second, we train several survival models to estimate the risk of failure given a set of covariates sampled from the time-domain. We find the application of such methods interesting in manufacturing and engineering, as they inherently support censored data and provide a probabilistic prediction and confidence estimation instead of mere point estimates, thus we encourage further work within our framework.

Acknowledgments

This work was supported by the PRECISE project under Grant Agreement No. AAL-2021-8-90-CP by the European AAL Association.

(a) XJTU ($N = 1000$, $C = 20\%$, $d = 12$).

Model	T_{train}	CI ↑	CI_{td} ↑	IBS ↓
CoxPH	**0.094**	0.648	0.642	0.271
RSF	0.988	**0.702**	**0.769**	0.423
CoxBoost	2.791	0.59	**0.769**	0.261
DeepSurv	9.148	0.687	0.679	0.245
DSM	22.456	0.66	0.628	**0.214**
WeibullAFT	0.483	0.664	0.675	0.216

(b) PRONOSTIA ($N = 400$, $C = 20\%$, $d = 12$).

Model	T_{train}	CI ↑	CI_{td} ↑	IBS ↓
CoxPH	**0.016**	0.626	0.625	0.34
RSF	0.265	0.715	0.701	0.383
CoxBoost	0.768	0.74	0.701	0.223
DeepSurv	2.392	0.745	0.705	**0.197**
DSM	9.737	0.739	**0.748**	0.294
WeibullAFT	0.489	**0.764**	0.691	0.236

Table 2: Performance metrics on the test sets. N: total sample size, C: pct. of censored data, d: number of covariates.

References

Al Masry, Z.; Schaible, P.; Zerhouni, N.; and Varnier, C. 2019. Remaining useful life prediction for ball bearings based on health indicators. *MATEC Web of Conferences*, 261: 02003.

Antolini, L.; Boracchi, P.; and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24): 3927–3944.

Austin, P. C. 2012. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29): 3946–3958.

Cox, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2): 187–202.

Gareth, J.; Daniela, W.; Trevor, H.; and Robert, T. 2021. *An introduction to statistical learning: with applications in R*. Spinger, 2nd edition.

Graf, E.; Schmoor, C.; Sauerbrei, W.; and Schumacher, M. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18): 2529–2545.

Guo, L.; Li, N.; Jia, F.; Lei, Y.; and Lin, J. 2017. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, 240: 98–109.

Harrell, F. E.; Lee, K. L.; and Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15: 361–87.

Hochstein, A.; Ahn, H.-I.; Leung, Y. T.; and Denesuk, M. 2013. Survival analysis for HDLSS data with time dependent variables: Lessons from predictive maintenance at a mining service provider. In *Proceedings of 2013 IEEE International Conference on Service Operations and Logistics, and Informatics*, 372–381.

Hothorn, T.; Bühlmann, P.; Dudoit, S.; Molinaro, A.; and Van Der Laan, M. J. 2005. Survival ensembles. *Biostatistics*, 7(3): 355–373.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random Survival Forests. *The Annals of Applied Statistics*, 2(3): 841–860.

Kaplan, E. L.; and Meier, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282): 457–481.

Katzman, J.; Shaham, U.; Bates, J.; Cloninger, A.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1).

Kim, N.-H.; An, D.; and Choi, J.-H. 2017. *Prognostics and Health Management of Engineering Systems*. Springer International Publishing.

Lee, E. T.; and Wang, J. W. 2013. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Incorporated, 4th edition.

Nagpal, C.; Li, X.; and Dubrawski, A. 2021. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8).

Nectoux, P.; Gouriveau, R.; Medjaher, K.; Ramasso, E.; Chebel-Morello, B.; Zerhouni, N.; and Varnier, C. 2012. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *IEEE International Conference on Prognostics and Health Management*, 1–8.

Randall, R. B. 2021. *Vibration-based condition monitoring*. Wiley, 2nd edition.

Randall, R. B.; and Antoni, J. 2011. Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, 25(2): 485–520.

Stepanova, M.; and Thomas, L. 2002. Censoring issues in survival analysis. *Operations Research*, 50(2): 277–289.

Wang, B.; Lei, Y.; Li, N.; and Li, N. 2020. A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Transactions on Reliability*, 69(1): 401–412.

Wang, J.; Li, C.; Han, S.; Sarkar, S.; and Zhou, X. 2017. Predictive Maintenance Based on Event-Log Analysis: A Case Study. *IBM Journal of Research and Development*, 61(1): 11–121.

Wang, Y.; Zhao, J.; Yang, C.; Xu, D.; and Ge, J. 2022. Remaining useful life prediction of rolling bearings based on Pearson correlation-KPCA multi-feature fusion. *Measurement*, 201: 111572.

Widodo, A.; and Yang, B.-S. 2011a. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications*, 38(3): 2592–2599.

Widodo, A.; and Yang, B.-S. 2011b. Machine health prognostics using survival probability and support vector machine. *Expert Systems with Applications*, 38(7): 8430–8437.

Xu, W.; Jiang, Q.; Shen, Y.; Xu, F.; and Zhu, Q. 2022. RUL prediction for rolling bearings based on Convolutional Autoencoder and status degradation model. *Applied Soft Computing*, 130: 109686.

Zelterman, D.; Le, C. T.; and Louis, T. A. 1996. Bootstrap techniques for proportional hazards models with censored observations. *Statistics and Computing*, 6(3): 191–199.

Zheng, C.; Liu, W.; Chen, B.; Gao, D.; Cheng, Y.; Yang, Y.; Zhang, X.; Li, S.; Huang, Z.; and Peng, J. 2018. A Data-driven Approach for Remaining Useful Life Prediction of Aircraft Engines. In *2018 21st International Conference on Intelligent Transportation Systems*, 184–189.

Appendix C

Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis

Abstract: The present study has designed, implemented, and evaluated a machine learning model that can predict fall risk and fall occurrence in community-dwelling elderly based on their home care usage. A dataset consisting of 2542 weekly home care records for 1499 citizens (59% female, 41% male) with a mean age of 77 years (SD 10 years) was collected from a large municipality in Denmark. The data were recorded between January 1, 2021, and December 31, 2021. The dataset was divided into two cohorts. Subsequently, five machine learning-based survival analysis models were trained and evaluated using cross-validation. The CoxBoost model showed the best discriminative performance with a mean 0.64 (95% CI 0.57–0.72) Harrell's concordance index, indicating better ranking than chance-level by 14% on average. However, the model could not accurately predict when the next fall would take place. The proposed method enables professionals to assess individual fall risk by using home care records from an Electronic Health Record (EHR) system. This facilitates the initiation of targeted fall-prevention programs for those at highest risk. Additionally, it is expected that a risk-based approach can lead to a lower number needed to treat (NNT), indicating greater effectiveness of health interventions.

Authors: Christian Marius Lillelund, Michael Harbo, and Christian Fischer Pedersen.

Status: Published in the Journal of Public Health, vol. 34, pp. 905–917, 2026.

Journal of Public Health
https://doi.org/10.1007/s10389-024-02317-9

ORIGINAL ARTICLE

Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis

Christian Marius Lillelund[1] · Michael Harbo[2] · Christian Fischer Pedersen[1]

Received: 24 January 2024 / Accepted: 18 June 2024
© The Author(s) 2024

Abstract

Aim The present study has designed, implemented, and evaluated a machine learning model that can predict fall risk and fall occurrence in community-dwelling elderly based on their home care usage.

Subjects and methods A dataset consisting of 2542 weekly home care records for 1499 citizens (59% female, 41% male) with a mean age of 77 years (SD 10 years) was collected from a large municipality in Denmark. The data were recorded between January 1, 2021, and December 31, 2021. The dataset was divided into two cohorts. Subsequently, five machine learning-based survival analysis models were trained and evaluated using cross-validation.

Results The CoxBoost model showed the best discriminative performance with a mean 0.64 (95% CI 0.57–0.72) Harrell's concordance index, indicating better ranking than chance-level by 14% on average. However, the model could not accurately predict when the next fall would take place.

Conclusions The proposed method enables professionals to assess individual fall risk by using home care records from an Electronic Health Record (EHR) system. This facilitates the initiation of targeted fall-prevention programs for those at highest risk. Additionally, it is expected that a risk-based approach can lead to a lower number needed to treat (NNT), indicating greater effectiveness of health interventions.

Keywords Fall risk prediction · Time-to-event · Survival analysis · Machine learning · Community-dwelling elderly · Home care · Personal alarms

Introduction

Falls among elderly citizens constitute a significant and widespread problem in society (Franse et al. 2017). Studies have shown that simple strength training exercises and daily outdoor activities can reduce the incidence rate of falls (Salminen et al. 2009; Klaperski-van Der Wal et al. 2023), however, the current efforts of Danish municipalities have traditionally focused on reactive approaches. To adopt a proactive approach and narrow down the search criteria for eligible adults, machine learning has been adopted to identify people at risk of falling. This is done primarily in hospitalized adults with good results (Lindberg et al. 2020; Patterson et al. 2019; Nakatani et al. 2020), but few studies have focused on community-dwelling elderly. In this setting, most proposed methods are invasive, where researchers, for example, place cameras around participants in their homes, which can be seen as intrusive and a violation of privacy (Dubois et al. 2019; Yang et al. 2019; Wilmink et al. 2020). The few noninvasive studies that exist have only studied binary classification methods (Chelli and Patzold 2019; Dormosh et al. 2022), which cannot distinguish individuals at risk beyond a few classes. They also rely on fall reports to determine if someone has fallen or not, but falls in older adults are often underreported or incorrectly reported (Hoffman et al. 2018).

In this observational study, we propose a noninvasive method to predict individual fall risk using survival analysis. Survival analysis is a form of regression modeling that supports censored data, i.e., where the event of interest is partially observed. Conventional machine learning methods

Christian Marius Lillelund
cl@ece.au.dk

Michael Harbo
michael@digirehab.dk

Christian Fischer Pedersen
cfp@ece.au.dk

[1] Department of Electrical and Computer Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark

[2] DigiRehab A/S, Brovej 20A, 8800 Viborg, Denmark

Published online: 30 July 2024

do not support censoring and can thus produce biased predictions if censored observations are ignored (Stepanova and Thomas 2002a). We use personal alarm subscriptions as our event of interest (Hoffman et al. 2018). Our method is based on data from a large Danish municipality, where personal alarms are given to citizens who have been evaluated as being at risk of falling. The municipality participated in a signature project funded by the Danish government on fall prevention with the corresponding university from 2020 to 2022. The data consist of 2542 weekly home care observations for 1499 citizens over a period of one year, which we divide into two cohorts with a 6-month follow-up period. We train five different survival models and establish an evaluation pipeline using cross-validation to access model performance. The goal is to offer a noninvasive decision support tool that can help healthcare professionals select the right candidates for fall prevention programs. In summary, our contributions are the following:

- We propose a novel and noninvasive predictive model to assess fall risk in home care clients using survival analysis.
- We use personal alarm subscriptions as the event of interest to avoid the bias and noise that surround fall reports.
- We demonstrate the effectiveness of our approach using different models and feature selection algorithms.

Related work

Work in fall-risk assessment and prevention using machine learning can generally be split into two categories: sensor data studies, where sensors or motion trackers are placed on or in the vicinity of the individual, and electronic health record (EHR) studies, where stored data about an individual's historical traits are used. We denote the former as invasive methods, since they require active involvement of the individual to operate, whereas the latter is noninvasive, since a model trained on EHR data can be used unobtrusively and without involving the individual. In this section, we provide a brief overview of the latest research in noninvasive methods and motivate our approach.

Noninvasive methods do not require ubiquitous or pervasive monitoring to work, and rely solely on historical information about the individual. Kuspinar et al. (2019) developed a predictive algorithm to predict fall risk based on home care data from a cohort of more than 80,000 adults between 2002 and 2014; however, the authors only report odds ratios for different groups of fallers without any assessment of predictive performance. In addition, the follow-up period is only 90 days (3 months) and all observed falls are self-reported, which may lead to underreporting or recall bias. The fall assessment is based on data from a Resident Assessment Instrument-Home Care (RAI-HC) evaluation, a standardized assessment scheme, which would have to be performed in person with the patient, making their method not entirely noninvasive. Lo et al. (2019) used a random forests algorithm on home care data extracted from the Outcomes and Assessment Information Set (OASIS) database to predict fall risk, which showed initial promise (ROC scores between 0.66 and 0.68), however, their model can only classify people into two categories (faller or not a faller) and thus cannot predict neither the time to fall nor individual risk scores. Other data sources in addition to home care have also been explored for noninvasive fall assessment, such as clinical notes (Santos et al. 2020; Fu et al. 2022) and fall reports (Dos Santos et al. 2019).

Recently, Dormosh et al. (2022) developed a fall-prediction model using logistic regression based on hospital registration data. The authors found significant correlations between a range of health record features and the risk of falling, but had to rely on cumbersome manual labeling to identify fallers from clinical notes. The main limitation of their work is that older people tend not to report falls, unless medical attention is required (Stevens et al. 2012), which is consistent with the observation that many falls go unreported (Hoffman et al. 2018). Furthermore, reported falls may lack detail or may not be caused by a lack of strength or mobility, but by other factors (e.g., alcoholism). Another limitation of Dormosh et al. (2022) is the use of binary logistic regression, which does not provide information about the time to the next fall, cannot rank individuals by risk, nor exploit any temporal information in the data. Logistic regression also struggles with bad expressiveness (e.g., interactions must be added manually).

We propose a novel, noninvasive method that can predict fall risk within 6 months based solely on home care data and without involving the individual of interest. Our method has the following novelties: (a) it is based on survival analysis, which can handle the problem of incomplete observations, something not addressed in previous approaches based on machine learning (Patterson et al. 2019; Ye et al. 2020). (b) Our model captures the time-to-event information not used in previous works (Chelli and Patzold 2019; Fu et al. 2022; Dormosh et al. 2022), which rely on binary classification. (c) Instead of falls, we use personal alarm subscriptions as the event of interest, which addresses the issue of biases in fall reports (Stevens et al. 2012; Hoffman et al. 2018).

Materials and methods

Survival analysis

Survival analysis is a form of regression that models the time until some event takes place, which can be partially observed

Table 1 Key demographics for Denmark and the municipality, where the data came from (2021, Statistics Denmark)

Demographic	Area	Statistic
Population	Denmark	5,867,412
	Municipality	221,114
Population (DOB. 1930-1970)	Denmark	2,307,176 (38.2% of total)
	Municipality	74,684 (33.7% of total)
Gender distribution* (DOB. 1930-1970)	Denmark	46.8%, 50.2%
	Municipality	47.8%, 52.1%
Marital status† (DOB. 1930-1970)	Denmark	12.3%, 56.8%, 11.3%, 16.6%
	Municipality	13.4%, 57.8%, 11.9%, 16.7%

*Gender distribution is males/females
†Marital status is percentage of single, married, widowed and divorced

(i.e., censored). It has found important use in many domains, such as healthcare informatics (Zhu et al. 2016; Kim et al. 2019), econometrics (Stepanova and Thomas 2002b), and predictive maintenance (Lillelund et al. 2023). We define a survival problem by a sequence of observations represented as triplets, $(\boldsymbol{x}_i, t_i, \delta_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature/covariate vector for some observation i, $t_i \in \mathbb{R}$ is the time of censoring or the time of event depending on which occurred first, and $\delta_i \in \{0, 1\}$ is the binary event indicator. If $\delta_i = 0$, then $t_i = c_i$, where c_i is the time of censoring, otherwise, if $\delta_i = 1$, then $t_i = e_i$, where e_i is the time of event. A survival model can predict the probability that some event occurs at time T later than t, i.e., the survival probability, $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$. To estimate $S(t)$, we use the so-called hazard function:

$$h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T > t)/\Delta t, \tag{1}$$

which corresponds to the event rate at a point after t, assuming the individual survived past that time (Gareth et al. 2021, Ch. 11). The hazard function is related to the survival function through $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density associated with T, $f(t) := \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t)/\Delta t$, that is, the instantaneous rate of event at time t. In this regard, $h(t)$ is the probability density of T conditional on $T > t$, and the functions $S(t)$, $h(t)$, $f(t)$, all correspond to equivalent ways of describing the distribution of T, formalizing, e.g., the intuition that higher values for $h(t)$ correspond to higher event probabilities.

The Cox proportional hazards (CoxPH) model is a popular regression model for survival analysis. It assumes a conditional individual hazard function of the form $h(t|\boldsymbol{x}_i) = h_0(t) \exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$. The risk score is denoted as $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$. In Cox (1972), f is set to a linear function of the covariates, that is, $f(\boldsymbol{\theta}, \boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\theta}$, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is derived by numerically maximizing the partial Cox log-likelihood.

The dataset

A large Danish municipality has provided the two datasets used in this study from their EHR system. The first dataset consists of 229,850 home care observations over 52 weeks (2021) divided into two cohorts (6 months) and covers home care for 6398 citizens in total. Each observation contains the amount of care delivered in minutes, the type of care and the number of social or health workers who provided the service. The type of care follows a standardized naming scheme across all Danish municipalities, FSIII (English: Common Language III. Danish: Faellessprog III). The second dataset contains a list of citizens who have subscribed to a personal alarm system in 2021. First, we merged the two datasets and removed individuals who had already subscribed to a personal alarm prior. Second, we created a window that spanned 26 weeks (6 months, from $t = 1$ to $t = 26$) and fixed a starting time for citizens at $t = 1$ in that window. For each observation, we assigned the time of event or censoring given the event indicator $\delta \in \{0, 1\}$; $\delta = 0$ if the citizen had dropped out or not subscribed yet, $\delta = 1$ if the citizen had subscribed. We included only citizens who received at least 100 min of home care per week and were born between 1930 and 1970. The final dataset has 2542 observations for 1499 citizens, a censoring rate of 95% (2416/126) and 49 nonzero covariates. No outlier detection or feature scaling was done. There was no missing values. Columns with only zero values were removed (Tables 1, 2 and 3).

Table 2 Key statistics of the dataset after preprocessing

Property	Statistic
Number of observations (N)	2542
Number of covariates	49
Number of citizens	1499
Average age (2021)	77 (± 10) years
Censoring rate	95% (2416/126)
Follow-up period	6 months (2 cohorts)
Study period	January 1 to December 31, 2021

Results

Setup

We implement the plain vanilla Cox model (CoxPH) (Cox 1972), the Cox model with LASSO regularization (CoxPH ℓ_1) (Simon et al. 2011), the Cox model with Ridge regularization (CoxPH ℓ_2) (Simon et al. 2011), Random Survival Forests (RSF) (Ishwaran et al. 2008) and the Cox model using boosting (CoxBoost) (Hothorn et al. 2005). To evaluate our approach, we report Harrell's concordance-index (CI_H) (Harrell et al. 1996), Uno's concordance-index (CI_U) (Uno et al. 2011), the integrated Brier score (IBS) (Graf et al. 1999), the mean absolute error (MAE) (Qi et al. 2024) using hinge loss, and D-calibration (Haider et al. 2020). See Appendix A for more details on the performance metrics. We train the models using either all covariates or only those selected by a feature selector, e.g., low-variance thresholding (LowVar), SelectKBest (SKB) or Recursive Feature Elimination (RFE). We set $K = 10$, i.e., the desired number of covariates will be 10.

To estimate the generalization error of each model and feature selector, we run stratified nested cross-validation with five outer loops and five inner loops. Stratification ensures that the event times and censoring rates are consistent across the training and test sets. Using nested cross-validation, feature selection and hyperparameter optimization are performed together within each fold. Appendix B reports the best obtained hyperparameters by highest CI_H.

Model performance

Table 4 shows the predictive and calibration performance from cross-validation, averaged over five folds. The CI (CI_H and CI_U) measures how well a model can predict risk scores that match the order of events, i.e., people with higher risk should experience the event before people with lower risk. The plain vanilla CoxPH model obtained a CI_H of 0.53, which means that it has ranked 53 out of 100 pairs correctly. A CI of 0.5 indicates chance-level ranking. A CoxPH model with a LASSO penalty term obtained a CI_H of 0.61, thus reducing the number of covariates leads to better discriminative performance in this case. CoxBoost provided the most concordant risk predictions with a mean CI_H of 0.64 (95% CI 0.57–0.72) using low-variance thresholding. We note that multiple types of home care are provided only on a sporadic basis. This means that the dataset has many covariates with predominant zero values, which in turn can make it difficult for the model to identify which covariates are important and which are not. This phenomenon can lead to poor perfor-

Table 3 Descriptive statistics of covariates ranked in descending order by observational count

Covariate name	Number of obs.	Mean
Personal hygiene	1254	102 (±249)
Cleaning	1225	31 (±51)
Medication administration	783	29 (±67)
Laundry	468	4 (±10)
Nutrition	413	25 (±90)
Excretions	352	34 (±180)
Medication dispensing	342	3 (±8)
RH Personal hygiene	266	14 (±63)
Prepare/serve food	254	12 (±46)
Supervision/care	188	3 (±21)
Compression therapy	183	5 (±21)
Everyday activities	157	3 (±22)
SUL § 138	140	2 (±9)
Shopping	108	1 (±6)
RH Nutrition	82	3 (±23)
RH Excretions	81	5 (±42)
Mobility	77	8 (±68)
RH Prepare/serve food	66	4 (±31)
Catheterization and care	66	2 (±15)
RH Practical interventions in relation to children	64	3 (±42)
RH Mobility	53	2 (±27)
Examinations and measurement of values	40	1 (±16)
RH Cleaning	39	1 (±7)
RH Everyday activities	35	1 (±9)
Stoma care	33	2 (±19)
Treatment and care of skin problems	30	1 (±13)
Relief at home	29	1 (±11)
Nutritional intervention	26	1 (±10)
Circulation treatment	21	0 (±4)
0-Benefit, SEL	15	2 (±60)
Wound care	15	0 (±3)
Rehabilitation	13	0 (±5)
RH Laundry	13	0 (±2)
Respiratory treatment	9	0 (±5)
RH Shopping	7	0 (±1)
SEL COVID-19, Protective equipment	5	0 (±3)
General 0-benefit	4	0 (±4)
Oxygen therapy	4	0 (±6)
Tube feeding	4	1 (±30)
Care with the use of personal aids	3	0 (±2)
RH Supervision/care	3	0 (±2)
Treatment and care of gastrointestinal problems	2	0 (±1)

Table 3 continued

Covariate name	Number of obs.	Mean
SEL COVID-19 Vaccine	2	0 (±0)
Fluid per os (Note: 'per os' means by mouth)	1	0 (±0)
Transfer and mobilization	1	0 (±4)
TCNH Assistance with measurements	1	0 (±0)
Guidance	1	0 (±0)
Personal care	1	0 (±2)

The dataset consists of 2542 samples, where each sample includes the amount of home care provided in 1 week for different types of care for one individual. The mean is reported in minutes per week (± standard deviation)

mance on new data. We see that restricting the feature space using LASSO or a feature selection algorithm gives better CI results on average.

On the surface, the results of the integrated Brier score look promising, as a lower number indicates a better estimate of the survival curve. However, the high censoring in this dataset means that for many cases, the predicted survival curve only has a slight decline over the follow-up period. This reflects the cumulative incidence function for the event, but does not offer much valuable application, since the discrepancies in the survival probabilities between the two groups (citizens that receive an alarm and citizens that do not) are few. RSF was best at predicting the survival curve with a mean integrated Brier score of 0.033 (95% CI 0.031–0.035). RSF was also best at predicting the time to event with a mean MAE of 83.5 (95% CI 79.1–87.9) using RFE to select the best covariates. The MAE is the absolute difference between the actual and predicted survival times (e.g., the median of the curve) using hinge loss; if the prediction is higher than the censored time, the loss is not penalized. This means that the error comes from censored samples with an

Table 4 Experimental results from 5-fold cross-validation. For D-calibration, we report the number of D-calibrated mean survival curves across the five folds according to a Pearson's χ^2 test

Metric	Model	Feature selector			
		None	LowVar	SKB10	RFE10
$CI_H \uparrow$	CoxPH	0.53±0.04	0.53±0.04	0.59±0.06	0.60±0.05
	CoxPH (ℓ_1)	0.61±0.04	0.61±0.04	0.61±0.04	0.61±0.04
	CoxPH (ℓ_2)	0.52±0.03	0.53±0.04	0.58±0.04	0.55±0.11
	RSF	**0.63±0.04**	0.62±0.04	0.62±0.04	0.62±0.04
	CoxBoost	0.63±0.07	**0.64±0.05**	**0.62±0.03**	**0.63±0.06**
$CI_u \uparrow$	CoxPH	0.52±0.04	0.52±0.04	0.57±0.06	0.58±0.05
	CoxPH (ℓ_1)	0.60±0.04	0.60±0.04	0.60±0.04	0.60±0.04
	CoxPH (ℓ_2)	0.51±0.04	0.51±0.04	0.56±0.05	0.54±0.10
	RSF	**0.62±0.05**	0.62±0.05	0.61±0.04	**0.62±0.04**
	CoxBoost	0.62±0.07	**0.64±0.05**	**0.61±0.03**	0.62±0.06
IBS ↓	CoxPH	0.037±0.001	0.036±0.001	0.033±0.002	0.034±0.002
	CoxPH (ℓ_1)	**0.034±0.001**	**0.034±0.001**	0.034±0.001	0.034±0.002
	CoxPH (ℓ_2)	0.037±0.001	0.037±0.000	0.034±0.001	0.035±0.002
	RSF	**0.034±0.001**	**0.034±0.001**	**0.033±0.001**	**0.033±0.001**
	CoxBoost	**0.034±0.001**	**0.034±0.001**	0.034±0.001	0.034±0.001
MAE ↓	CoxPH	212.4±111.4	211.3±109.1	106.2±33.3	118.1±46.3
	CoxPH (ℓ_1)	87.5±1.3	87.5±1.3	87.4±1.4	108.6±43.0
	CoxPH (ℓ_2)	296.8±109.4	224.3±113.5	115.7±29.7	117.1±46.6
	RSF	**84.9±2.0**	**85.0±1.7**	**84.0±2.8**	**83.5±3.2**
	CoxBoost	87.9±1.1	87.9±1.0	87.8±1.0	87.8±1.1
D-calib. ↑	CoxPH	4/5	4/5	**5/5**	**5/5**
	CoxPH (ℓ_1)	**5/5**	**5/5**	**5/5**	**5/5**
	CoxPH (ℓ_2)	3/5	3/5	4/5	**5/5**
	RSF	**5/5**	**5/5**	**5/5**	**5/5**
	CoxBoost	**5/5**	**5/5**	**5/5**	**5/5**

Best performing models highlighted in bold

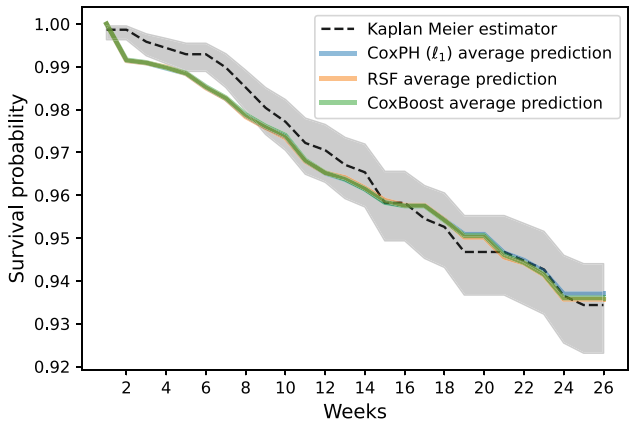


Fig. 1 Mean predicted survival functions by the CoxPH, RSF and CoxBoost models, versus the unbiased Kaplan–Meier estimator on the test set. The survival functions are predicted over the entire event horizon from $t = 1$ to $t = 26$

overly pessimistic prediction (less than the censoring time) and uncensored samples with an overly optimistic prediction (more than the event time). With a 95% average censoring rate, the model overshoots the true event time for the few individuals who experience the event; thus, our model is poor at predicting *when* the event occurs in the sample population. We used Pearson's χ^2 goodness-of-fit test to assess D-calibration, and find that, on average, CoxPH with LASSO, RSF and CoxBoost gave predicted survival curves that were calibrated with respect to the actual survival distribution.

Model explainability

After cross-validation, we adopt a 70–30% train-test split, configure the models with sensible default parameters and retrain them for plotting purposes. Figure 1 shows the mean predicted survival function for all samples in the test set by the CoxPH using LASSO, RSF and CoxBoost models. As a reference, we have included the Kaplan–Meier (KM) (Kaplan and Meier 1958) estimator with a 95% confidence interval. In the present study, we assume censoring to be non-informative: consider two citizens with the same risk factors, both yet to experience the event of interest by time t. One individual is lost to follow-up at that time, i.e., their event time at t is right-censored, and the other continues in the study. Under the assumption of noninformative censoring, we assume these two individuals have the same subsequent risk of experiencing the event, and knowing that one of them is censored does not add additional information. Under this assumption, the KM is unbiased, regardless of the proportion of censoring. Similarly, the hazard ratios of a Cox's proportional hazards model represent a good estimate of relative risk and are unbiased regardless of the amount of censoring. The predicted survival probabilities and the KM align well in this plot. All models fall within the KM's confidence interval after $t = 8$, however, a slight overestimation of risk is seen between $t = 0$ and $t = 7$.

Figure 2 shows the most important covariates and covariate values according to SHAP (Lundberg and Lee 2017). These values were obtained from a Random Survival Forests model. The covariates are on the y-axis and the SHAP value on the x-axis. The color indicates if the covariate value is low (blue) or high (red) for a particular observation. According to the SHAP values, the most important covariates are: "Personal hygiene", "SUL §138", "Excretions", "RH Nutrition" and "Medication administration". The SHAP values exhibit a slight positive correlation between risk and the number of minutes received in the category "Personal hygiene", but a strictly positive correlation for "Excretions" and "SUL §138". In this context, "SUL §138" refers to a section of the Danish health legislation that requires the

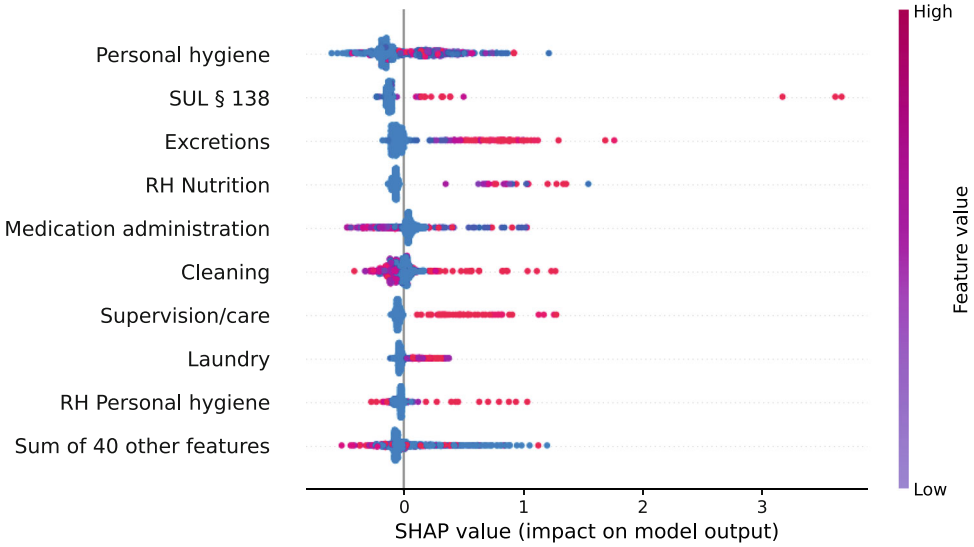
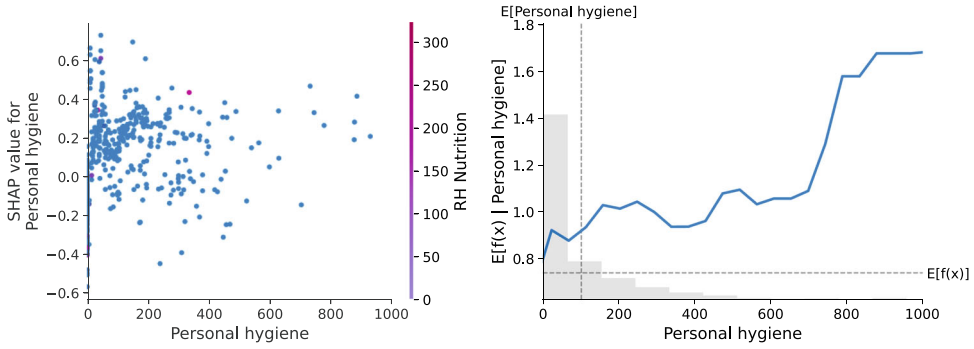


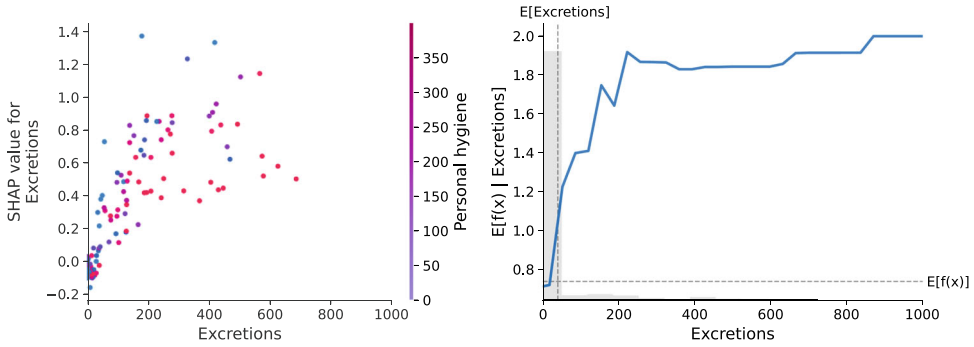
Fig. 2 SHAP summary plot. The covariates are on the y-axis and the SHAP-value is on the x-axis. Covariates are ranked based on their importance, which is given by the mean of their absolute SHAP values (higher positioning means more important). For each covariate, one point corresponds to a single citizen. Its position on the x-axis represents the impact that covariate had on the model's output, which corresponds to the risk across citizens. The color indicates whether the citizen had a low (*blue*) or high (*red*) value of that covariate

municipality to provide an individual with a community nurse by referral from their general practitioner. We see that some covariates have both positive and negative correlations (e.g., "Cleaning", "Medication administration" and "RH Personal hygiene"). These may seem counterintuitive at first, as we would usually associate increased reliance on home care with an increased risk of falling, but we study purely the correlations between care and the risk of getting a personal alarm. Elderly people who depend on a lot of home care have fewer falls, as they are attended to very often and many have daily check-ins by a nurse or a care practitioner; hence, they have less need for a personal alarm.

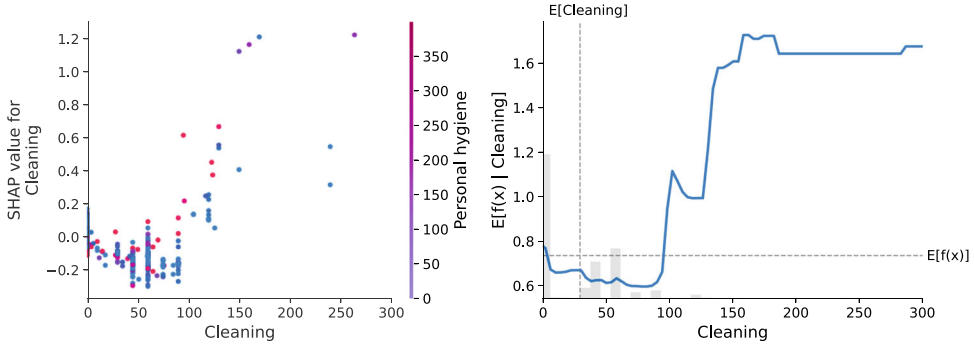
Figure 3 is a SHAP feature dependence plot and partial dependence plot for three of the most important covariates ("Personal hygiene", "Excretions" and "Cleaning"). These provide more detail on the exact correlations between risk and received care. A partial dependence plot is created by marginalizing the model output over the distribution of the covariates in a set of irrelevant covariates, C, and returning a function that depends only on covariates in a relevant covariate set, S, interactions included. This tells us, for given values of covariates S, what the average marginal effect on the prediction is, i.e., how the outcome changes when a specific independent variable changes. In Fig. 3a on the left, we see a vague, but positive correlation between minutes and SHAP value for the "Personal hygiene" covariate, which is also explained by the partial dependence plot on the right. The vertical gray bars in the plot show the data distribution. In Fig. 3b, we plot the "Excretions" covariate and see a much steeper increase in both SHAP value and estimated risk from



(a) "Personal hygiene"



(b) "Excretions"



(c) "Cleaning"

Fig. 3 *Left*: SHAP dependence plots of the respective covariate. The color indicates whether the covariate has a low (*blue*) or high (*red*) value. *Right*: Partial dependence plot of the same covariate

0 to 200 minutes, but it increases only slightly after that. Lastly, Fig. 3c shows that "Cleaning" has a negative trend between 0 and 100 minutes, but rises suddenly from 100 and onward.

Figure 4 shows predicted individual survival and hazard functions for 20 randomly selected citizens using a random survival forests model. Solid lines represent citizens who did not subscribe to a personal alarm within 6 months. The dotted lines represent the citizens who subscribed to one. We see in the upper row that most citizens reach a survival probability (probability of no alarm) at $t = 26$ between 0.98 and 0.90 and a cumulative hazard between 0.01 and 0.10, but some citizens have a significantly lower survival probability and associated cumulative hazard. In this figure, only one citizen in this random draw subscribed to a personal alarm during the study, but the individual has similar predicted survival probabilities over the event horizon as the remaining cohort. This tells us that predicting the time to event from these curves alone is difficult.

Discussion

Applicability

The proposed method represents a change from a reactive mindset to a proactive mindset and is applicable in a municipality setting, where efficient management of resources and healthcare personnel is crucial. By accurately identifying those citizens who are at the highest risk of falling, the municipality can allocate its resources more effectively by providing personalized home care, physical therapy, and other healthcare services. This is important, as the consequences of falls can often be more severe for the elderly due to existing health problems and reduced mobility (Franse et al. 2017). Wang et al. (2023) find that exercise appears to be particularly effective for people with higher fall rates. In their study, exercise was more effective in trials with a higher prospective fall rate (32% reduction in falls and 442 prevented falls in 1000 people over one year). However, trials with a lower prospective fall rate showed a comparatively reduced effectiveness, resulting in a mere 12% reduction in the number of falls and preventing only 64 falls in 1000 people in the same period. At first glance, it appears to be a sound strategy to offer exercise and physical therapy to any older adult over 65 years of age, but a broad preventive approach is very expensive if training is to be supervised, which multiple studies advocate for (Donat and Özcan 2007; Youssef and Shanb 2016). Moreover, we expect that the number needed to treat (NNT), as a measure of the effectiveness of a health intervention, will likely be lower (better) with a more targeted and risk-based approach. When prevention and intervention are focused on the most relevant citizens, the number of people who need to be trained to prevent a single fall is reduced.

Model performance

Concerning the predictive performances, all models obtained concordance index scores between 0.51 and 0.64, indicating that the proposed method can rank individuals by risk statistically better than chance-level ranking (0.5). This was validated by Harrell's and Uno's concordance index using cross-validation across five folds. However, the proposed method cannot accurately predict when the next fall would occur for the few citizens who experienced the event within the 6-month cohort given the current dataset and event horizon. This was confirmed by observing a relative high mean absolute error between the actual event time and the predicted event time. We attribute this to the high censoring rate and the lack of qualitative covariates for this specific task. Regarding the calibration performance, the CoxPH model predicted D-calibrated survival curves in all five cross-validation folds when using the LASSO regularization technique, i.e., zeroing out coefficients and thus preventing overfitting. The two

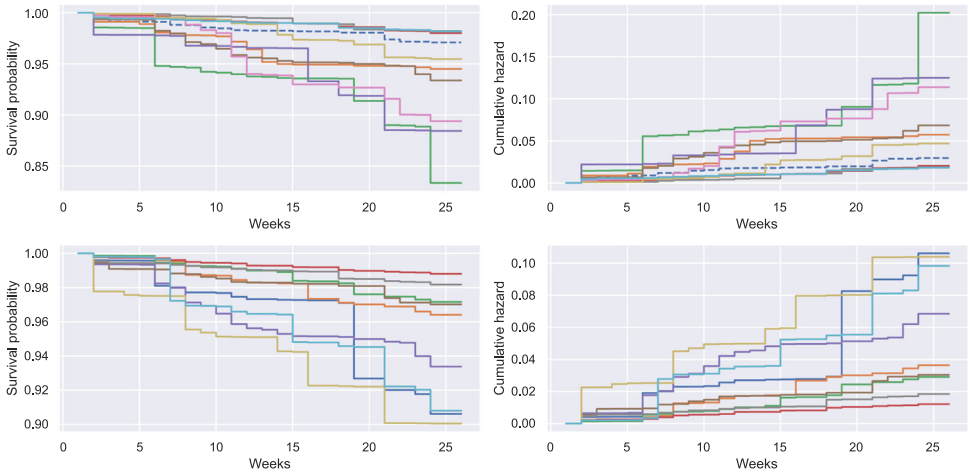


Fig. 4 Predicted individual survival (*left*) and hazard (*right*) functions for 20 randomly selected citizens from the test set using the RSF model. Solid lines represent citizens who do not receive an alarm within 6 months. *Dotted lines* represent citizens that do

ensemble-based and gradient-boosting models predicted D-calibrated survival curves in all folds as well.

Limitations

Our method is trained only on EHR data from a single municipality. Although being one of the largest in Denmark, the correlations we have found between home care usage and personal alarm subscriptions may not translate to other municipalities or government institutions, let alone in Denmark. Although personal alarms are an objective arrangement to avoid future falls, not all personal alarms in Denmark are given to people at risk of falling or who have had falls. They can be subscribed to a person in need of a secure environment or individuals suffering from anxiety. When adopting a machine learning model for fall prediction, it should be trained and evaluated in the population that it is used in. Another limitation is the high censoring rate. Ninety-five percent of the people in this study did not receive a personal alarm during either cohort. We regard this kind of censoring as noninformative, i.e., the presence of censoring contains no information about the actual time to event. Under this assumption, the proposed Cox proportional hazards models are not biased (Leung et al. 1997). However, the high censoring rate means that using the median of the survival function as the predicted time to event often requires a great deal of extrapolation outside the event horizon, and these estimates hereby carry a lot of uncertainty. Without more predictive features (e.g., physical strength tests, previous falls) or a longer study period, our method is not able to predict the time to the next fall accurately.

Ethical concerns

The present study did not include human trials and all data were completely anonymized before any data processing or machine learning took place. A contract was signed between the data owner and Aarhus University to help ensure privacy and confidentiality in the handling, processing, and dissemination stages. The proposed tool is intended as a whitelisting tool that can highlight potential candidates for fall prevention programs. There are no side effects if more people are selected for fall prevention, as strength training for elderly people is always beneficial if it is done correctly and properly supervised (Donat and Özcan 2007; Youssef and Shanb 2016). Today, most of the elderly often experience multiple falls at home before being offered fall prevention, and the socioeconomically advantaged elderly have easier access to health care resources than the disadvantaged elderly (McMaughan et al. 2020). A data-driven model can ensure that risk assessment and eligibility for a fall prevention program are independent of socioeconomic status. In this way, it will contribute to reducing health inequality, as all citizens, regardless of background, will have access to relevant and timely prevention offers.

Conclusions

Based on 2542 home care observations for 1499 citizens (59% female, 41% male) with a mean age of 77 years (SD 10 years), we have trained a selection of machine learning-based survival analysis models to predict the per-individual risk of falling over 6 months. Using 5-fold cross-validation, the best model in terms of ranking performance was CoxBoost with a mean Harrell's concordance index of 0.64 (95% CI 0.57–0.72). This concordance index indicates better ranking performance than chance-level by 14% on average. The CoxBoost model also produced survival curves that were D-calibrated across all five folds according to a Pearson's χ^2 test. Our method can be used as a decision support tool to choose the right candidates for a fall prevention program and can provide a cost-effective and timely way to assess fall risk among community-dwelling elderly, which can potentially improve their quality of life and reduce the burden on the healthcare system.

Appendix A: Performance metrics

CI_H : Harrell's concordance index measures a model's ability to rank individuals by risk. It is defined as the ratio of all comparable pairs in which the predictions and outcomes are concordant. A pair is considered comparable if we can determine who has the event first. It is defined as (Harrell et al. 1996):

$$CI = \frac{\sum_{i,j} 1_{t_i<t_j} \cdot 1_{\eta_i>\eta_j} \cdot \delta_i}{\sum_{i,j} 1_{t_i<t_j} \cdot \delta_i}, \tag{2}$$

where η_i represents the risk score of individual i.

CI_U: Uno's concordance index measures an estimator's ranking ability by comparing relative risk scores across all comparable pairs like Harrell's, however, it does not depend on the distribution of censoring times in the test data. Instead, it introduces censoring weights based on the estimated censoring cumulative distribution computed from the survival times in the training data. This has been shown to be preferable to Harrell's in presence of high censoring (Uno et al. 2011).

BS/IBS: The Brier score is the mean squared error between the actual binary outcome and the probability of the outcome at some time point t, i.e., the accuracy of the survival prediction. The metric has a probabilistic outcome between 0 and 1, and all outcomes for a single individual must sum up to one. Consider a sample of N observations,

$\boldsymbol{y}_i$ $(i = 1, 2, ..., N)$, the predicted binary outcome is $\hat{y}_i(t_k)$, where t_k is some measurement time on the event horizon of K discretized times, $[t_0, t_1, t_2, \ldots, t_K]$, and the actual outcome is $y_i(t_k)$, then the Brier score is Graf et al. (1999):

$$BS(t_k) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i(t_k) - y_i(t_k))^2, \tag{3}$$

where the actual outcome $y_i(t_k)$ for each observation can only be 1 or 0. The integrated Brier score is then the accuracy of the predicted survival curve at all available times $t_0 \leq t_k \leq t_K$ over the interval $[t_0;t_K]$, i.e., $IBS = \int_{t_0}^{t_K} BS(t)\, dw(t)$, where $w(t) = t_k/t_K$ is used as a weighting function (Graf et al. 1999).

MAE: The mean absolute error is the absolute difference between the actual and predicted survival times (e.g., median of the survival function). Given some individual survival distribution, $S(t \mid \boldsymbol{x}_i) = \Pr(T > t \mid \boldsymbol{X} = \boldsymbol{x}_i)$, we compute the predicted survival time $\hat{t}_i$ as the median survival time as in Qi et al. (2024):

$$\hat{t}_i = \text{median}\ (S(t \mid \boldsymbol{x}_i)) = S^{-1}(\tau = 0.5 \mid \boldsymbol{x}_i), \tag{4}$$

where $\tau \in [0, 1]$ denotes the quantile probability level. To calculate the MAE, we use hinge loss function that supports censored observations; if the predicted survival time is smaller than the censored time, the loss is the time of censoring minus the predicted time. If the predicted survival time is equal to or greater than the censored time, the loss is zero. Thus, the MAE and MAE-Hinge are, respectively:

$$\text{MAE}\ (\hat{t}_i, t_i, \delta_i = 1) = \mid t_i - \hat{t}_i \mid, \tag{5}$$

$$\text{MAE-Hinge}\ (\hat{t}_i, t_i, \delta_i = 0) = \max\ (t_i - \hat{t}_i,\ 0). \tag{6}$$

D-calibration: Distribution calibration is a statistical test to evaluate the calibration performance of the survival curve $S(t)$ (Haider et al. 2020), i.e., should we trust the probability predictions suggested by $S(t)$? For each uncensored citizen $\boldsymbol{x}_i$, we can observe the time of the event t_i, and also determine the probability $P(S(t|\boldsymbol{x}_i(t)))$ for that time based on $S(t)$. By example of Haider et al. (2020), if $S(t)$ is D-calibrated, we expect 10% of the citizens to fall in the [100%, 90%] interval and subsequently another 10% to fall in the [90%, 80%] interval. For all individual distributions d_i, the set $\{S_i(d_i)\}$ should therefore be uniformly distributed between 0 and 1, so that each 10% interval covers 10% of D. We adopt Pearson's χ^2 goodness-of-fit test to examine if the proportion of citizens in each interval is uniformly distributed given mutually exclusive and equal-sized intervals, as in Haider et al. (2020).

Table 5 Best experimental hyperparameters from the cross-validation step

Model	Parameters
CoxPH	max_iter=100
	tol=1e-09
CoxPH (ℓ_1)	max_iter=100000
	tol=1e-06
	n_alphas=100
CoxPH (ℓ_2)	max_iter=100
	tol=0.1
Random Survival Forests (RSF)	max_depth=5
	n_estimators=100
	min_samples_split=0.1
	min_samples_leaf=0.1
Gradient-boosted Cox model (CoxBoost)	max_depth=5
	n_estimators=100
	learning_rate=0.1
	dropout_rate=0.2
	subsample=0.8
	max_features='log2'
	min_samples_split=0.5
	min_samples_leaf=0.1

Appendix B: Hyperparameter optimization

Table 5 shows the best obtained hyperparameters from cross-validation on average. The following hyperparameters were evaluated:

- CoxPH: The maximum number of iterations and the tolerance for the optimizer.
- CoxPH (ℓ_1): The maximum number of iterations, the tolerance for the optimizer and the number of alphas.
- CoxPH (ℓ_2): The maximum number of iterations and the tolerance for the optimizer.
- RSF: The maximum depth, the number of estimators, and the minimum number of samples required for splitting internally and at a leaf node.
- CoxBoost: The maximum depth, the number of estimators, the learning rate, the dropout rate, the subsampling ratio, the feature selection strategy and the minimum number of samples required for splitting internally and at a leaf node.

Appendix C: Baseline survival function

We used the Breslow estimator to estimate the baseline survival function $\hat{S}_0(t)$ on the 70% training dataset (1779 observations, 95% censoring) over the entire event horizon

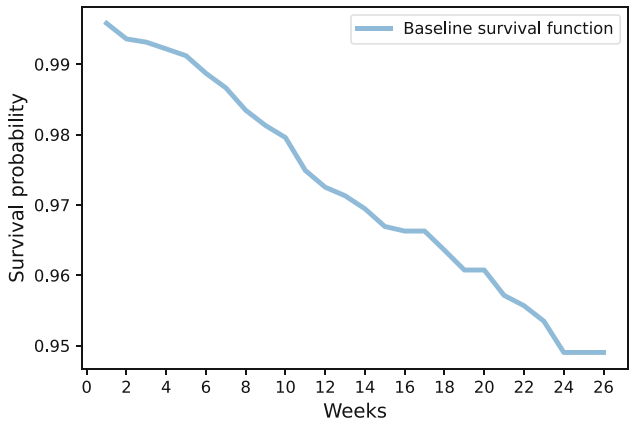


Fig. 5 Baseline survival function $\hat{S}_0(t)$ (Breslow 1975)

(see Fig. 5) (Breslow 1975). Setting the coefficients $\beta_i = 0$, the baseline can be estimated without estimating the parameters of the model, and vice versa in a proportional hazards model.

Appendix D: Cox regression coefficients

We train a plain vanilla Cox proportional hazards model on the 70% training dataset (1779 observations, 95% censoring)

Table 6 CoxPH hazard ratios for all covariates in order of log(HR)

Covariate	log(HR)	HR
Wound care	0.020782	1.020999
SUL § 138	0.020256	1.020463
RH Everyday activities	0.010345	1.010399
Laundry	0.008768	1.008806
Medication dispensing	0.008712	1.008751
RH Cleaning	0.006411	1.006431
Compression therapy	0.005066	1.005079
RH Nutrition	0.004498	1.004508
RH Personal hygiene	0.003692	1.003699
Treatment and care of skin problems	0.003496	1.003502
Everyday activities	0.001319	1.001320
Personal hygiene	0.000965	1.000965
RH Practical interventions in relation to children	0.000787	1.000787
Circulation treatment	0.000695	1.000695
Cleaning	0.000546	1.000546
SEL COVID-19 Vaccine	0.000000	1.000000
Treatment with orthopedic aids	0.000000	1.000000
Fluid per os (Note: 'per os' means by mouth)	0.000000	1.000000
Personal care	0.000000	1.000000
Treatment and care of gastrointestinal problems	0.000000	1.000000
Nutrition	−0.000068	0.999932
RH Excretions	−0.000125	0.999875
Prepare/serve food	−0.000159	0.999841
Medication administration	−0.000623	0.999377
Excretions	−0.000810	0.999191
Mobility	−0.001389	0.998612
Supervision/care	−0.001461	0.998540
RH Prepare/serve food	−0.002781	0.997223
RH Mobility	−0.005334	0.994680
Nutritional intervention	−0.006630	0.993392
Catheterization and care	−0.009547	0.990498
Shopping	−0.027485	0.972889
0-Benefit, SEL	−0.037073	0.963606
Transfer and mobilization	−0.044897	0.956096
Care with the use of personal aids	−0.105550	0.899829
Tube feeding	−0.199580	0.819074
Respiratory treatment	−0.202534	0.816659
SEL COVID-19, Protective equipment	−0.225825	0.797858
RH Supervision/care	−0.244316	0.783240
Oxygen therapy	−0.308903	0.734252
Stoma care	−0.316058	0.729018
Rehabilitation	−0.431206	0.649725
TCNH Assistance with measurements	−0.628801	0.533231
Relief at home	−0.629875	0.532659
RH Laundry	−0.738976	0.477603
General 0-benefit	−0.866646	0.420359
RH Shopping	−1.071294	0.342565
Guidance	−1.402100	0.246080
Examinations and measurement of values	−2.258464	0.104511

to estimate hazard ratios, $\exp(\beta_i)$, with a small ridge penalty applied to ensure convergence (see Table 6). All covariates are included in the model. A β_i-value exceeding zero, or alternatively, a hazard ratio surpassing one, suggests that with an increase in the ith covariate's value, there is an associated rise in the event hazard, leading to a decrease in survival time.

Acknowledgements This work was done in connection with DigiRehab A/S, Denmark. DigiRehab A/S is a privately-held Danish company.

Author Contributions C. M. Lillelund and M. Harbo conceived the present idea. C. M. Lillelund developed the theory, gathered the data, performed the experiments, and took the lead in writing the manuscript. M. Harbo encouraged C. M. Lillelund to investigate falls and home care usage, provided valuable domain knowledge, and helped interpret the results. C. F. Pedersen provided valuable feedback on the study design and implementation, and supervised the writing process. All authors discussed the results and contributed to the final manuscript.

Funding Open access funding provided by Aarhus Universitet. This research was funded by the PRECISE project (http://www.aal-europe.eu/projects/precise/) under Grant Agreement No. AAL-2021-8-90-CP by the European AAL Association.

Data availability The data supporting the findings of this study come from a large Danish municipality. Restrictions apply to the availability of the data, which were used under license for the current study and are therefore not publicly available.

Code availability All data processing, machine learning, and evaluation were done in Python 3.8. The source code is available from the corresponding author upon request.

Declarations

Conflict of interest C. M. Lillelund and C. F. Pedersen declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. M. Harbo is the head of development at DigiRehab A/S, which have a business relationship with the Danish municipality that provided the dataset for the present study. Neither DigiRehab A/S nor the municipality in question have influenced the reported work.

Ethics approval Ethics approval is not applicable. The study did no human trials and all data were completely anonymized.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Breslow NE (1975) Analysis of survival data under the proportional hazards model. International Statistical Review pp 45–57

Chelli A, Patzold M (2019) A machine learning approach for fall detection and daily living activity recognition. IEEE Access 7:38670–38687. https://doi.org/10.1109/ACCESS.2019.2906693

Cox DR (1972) Regression models and life-tables. J Roy Stat Soc: Ser B (Methodol) 34(2):187–202

Klaperski-van Der Wal S, Bruton A, Felton L, Turner S (2023) A mixed-method exploration of the effects and feasibility of an intergenerational fall-prevention gardening programme in older adults at risk of falling: a clinical trial. J Public Health. https://doi.org/10.1007/s10389-023-02154-2

Donat H, Özcan A (2007) Comparison of the effectiveness of two programmes on older adults at risk of falling: unsupervised home exercise and supervised group exercise. Clin Rehabil 21(3):273–283. https://doi.org/10.1177/0269215506069486

Dormosh N, Schut MC, Heymans MW, Van Der Velde N, Abu-Hanna A (2022) Development and internal validation of a risk prediction model for falls among older people using primary care electronic health records. The Journals of Gerontology: Series A 77(7):1438–1445. https://doi.org/10.1093/gerona/glab311

Dos Santos HD, Silva AP, Maciel MCO, Burin HMV, Urbanetto JS, Vieira R (2019) Fall detection in EHR using word embeddings and deep learning. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, pp 265–268, https://doi.org/10.1109/BIBE.2019.00054

Dubois A, Mouthon A, Sivagnanaselvam RS, Bresciani JP (2019) Fast and automatic assessment of fall risk by coupling machine learning algorithms with a depth camera to monitor simple balance tasks. J Neuroeng Rehabil 16(1):71. https://doi.org/10.1186/s12984-019-0532-x

Franse CB, Rietjens JA, Burdorf A, van Grieken A, Korfage IJ, van der Heide A, Raso FM, van Beeck E, Raat H (2017) A prospective study on the variation in falling and fall risk among community-dwelling older citizens in 12 european countries. BMJ Open 7(6), https://doi.org/10.1136/bmjopen-2017-015827

Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, Wen A, Liu H, Rocca WA, Olson JE, Sauver JS, Sohn S (2022) A hybrid model to identify fall occurrence from electronic health records. Int J Med Informatics 162:104736. https://doi.org/10.1016/j.ijmedinf.2022.104736

Gareth J, Daniela W, Trevor H, Robert T (2021) An introduction to statistical learning: with applications in R, 2nd edn. Springer

Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18(17–18):2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

Haider H, Hoehn B, Davis S, Greiner R (2020) Effective ways to build and evaluate individual survival distributions. J Mach Learn Res 21(1):1–63

Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hoffman GJ, Ha J, Alexander NB, Langa KM, Tinetti M, Min LC (2018) Underreporting of Fall Injuries of Older Adults: Implications for Wellness Visit Fall Risk Screening: Accuracy of Self-Reported Fall Injuries. J Am Geriatr Soc 66(6):1195–1200. https://doi.org/10.1111/jgs.15360

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ (2005) Survival Ensembles. Biostatistics 7(3):355–373

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random Survival Forests. The Annals of Applied Statistics 2(3):841–860. https://doi.org/10.1214/08-AOAS169

Kaplan EL, Meier P (1958) Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc 53(282):457–481. https://doi.org/10.2307/2281868

Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ (2019) Deep learning-based survival prediction of oral cancer patients. Sci Rep 9(1):6994. https://doi.org/10.1038/s41598-019-43372-7

Kuspinar A, Hirdes JP, Berg K, McArthur C, Morris JN (2019) Development and validation of an algorithm to assess risk of first-time falling among home care clients. BMC Geriatr 19(1):264. https://doi.org/10.1186/s12877-019-1300-2

Leung KM, Elashoff RM, Afifi AA (1997) Censoring issues in survival analysis. Annu Rev Public Health 18(1):83–104. https://doi.org/10.1146/annurev.publhealth.18.1.83

Lillelund CM, Pannullo F, Jakobsen MO, Pedersen CF (2023) Predicting survival time of ball bearings in the presence of censoring. Accepted at AAAI Fall Symposium 2023 on Survival Prediction, 2309.07188

Lindberg DS, Prosperi M, Bjarnadottir RI, Thomas J, Crane M, Chen Z, Shear K, Solberg LM, Snigurska UA, Wu Y, Xia Y, Lucero RJ

(2020) Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach. Int J Med Informatics 143:104272. https://doi.org/10.1016/j.ijmedinf.2020.104272

Lo Y, Lynch SF, Urbanowicz RJ, Olson RS, Ritter AZ, O'Connor M, Keim SK, McDonald M, Moore JH, Bowles H (2019) Using machine learning on home health care assessments to predict fall risk. Studies in Health Technology and Informatics 264:684–688. https://doi.org/10.3233/SHTI190310

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, p 4768–4777

McMaughan DJ, Oloruntoba O, Smith ML (2020) Socioeconomic Status and Access to Healthcare: Interrelated Drivers for Healthy Aging. Front Public Health 8:231. https://doi.org/10.3389/fpubh.2020.00231

Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C (2020) Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study. JMIR Med Inform 8(4):e16970. https://doi.org/10.2196/16970

Patterson BW, Engstrom CJ, Sah V, Smith MA, Mendonça EA, Pulia MS, Repplinger MD, Hamedani AG, Page D, Shah MN (2019) Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits. Med Care 57(7):560–566. https://doi.org/10.1097/MLR.0000000000001140

Sa Qi, Sun W, Greiner R (2024) SurvivalEVAL: A comprehensive open-source python package for evaluating individual survival distributions. Proceedings of the AAAI Symposium Series 2:453–457. https://doi.org/10.1609/aaaiss.v2i1.27713

Salminen M, Vahlberg T, Kivelä SL (2009) The long-term effect of a multifactorial fall prevention programme on the incidence of falls requiring medical treatment. Public Health 123(12):809–813. https://doi.org/10.1016/j.puhe.2009.10.018

Santos J, D P Dos Santos H, Vieira R (2020) Fall detection in clinical notes using language models and token classifier. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, pp 283–288, https://doi.org/10.1109/CBMS49503.2020.00060

Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39(5), https://doi.org/10.18637/jss.v039.i05

Stepanova M, Thomas L (2002) Censoring issues in survival analysis. Oper Res 50(2):277–289. https://doi.org/10.1146/annurev.publhealth.18.1.83

Stepanova M, Thomas L (2002) Survival Analysis Methods for Personal Loan Data. Oper Res 50(2):277–289. https://doi.org/10.1287/opre.50.2.277.426

Stevens JA, Ballesteros MF, Mack KA, Rudd RA, DeCaro E, Adler G (2012) Gender differences in seeking care for falls in the aged medicare population. Am J Prev Med 43(1):59–62. https://doi.org/10.1016/j.amepre.2012.03.008

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30(10):1105–1117. https://doi.org/10.1002/sim.4154

Wang BY, Sherrington C, Fairhall N, Kwok WS, Michaleff ZA, Tiedemann A, Wallbank G, Pinheiro MB (2023) Exercise for fall prevention in community-dwelling people aged 60+: more effective in trials with higher fall rates in control groups. J Clin Epidemiol 159:116–127. https://doi.org/10.1016/j.jclinepi.2023.05.003

Wilmink G, Dupey K, Alkire S, Grote J, Zobel G, Fillit HM, Movva S (2020) Artificial intelligence-powered digital health platform and wearable devices improve outcomes for older adults in assisted living communities: Pilot intervention study. JMIR Aging 3(2):e19554. https://doi.org/10.2196/19554

Yang Y, Hirdes JP, Dubin JA, Lee J (2019) Fall risk classification in community-dwelling older adults using a smart wrist-worn device and the resident assessment instrument-home care: Prospective observational study. JMIR Aging 2(1):e12153. https://doi.org/10.2196/12153

Ye C, Li J, Hao S, Liu M, Jin H, Zheng L, Xia M, Jin B, Zhu C, Alfreds ST, Stearns F, Kanov L, Sylvester KG, Widen E, McElhinney D, Ling XB (2020) Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int J Med Informatics 137:104105. https://doi.org/10.1016/j.ijmedinf.2020.104105

Youssef EF, Shanb AAE (2016) Supervised versus home exercise training programs on functional balance in older subjects. Malays J Med Sci 23(6):83–93

Zhu X, Yao J, Huang J (2016) Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine, pp 544–547, https://doi.org/10.1109/BIBM.2016.7822579

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix D

An Explainable and Conformal AI Model to Detect Temporomandibular Joint Involvement in Children Suffering from Juvenile Idiopathic Arthritis

Abstract: Juvenile idiopathic arthritis (JIA) is the most common rheumatic disease during childhood and adolescence. The temporomandibular joints (TMJ) are among the most frequently affected joints in patients with JIA, and mandibular growth is especially vulnerable to arthritic changes of the TMJ in children. A clinical examination is the most cost-effective method to diagnose TMJ involvement, but clinicians find it difficult to interpret and inaccurate when used only on clinical examinations. This study implemented an explainable artificial intelligence (AI) model that can help clinicians assess TMJ involvement. The classification model was trained using Random Forest on 6154 clinical examinations of 1035 pediatric patients (67% female, 33% male) and evaluated on its ability to correctly classify TMJ involvement or not on a separate test set. Most notably, the results show that the model can classify patients within two years of their first examination as having TMJ involvement with a precision of 0.86 and a sensitivity of 0.7. The results show promise for an AI model in the assessment of TMJ involvement in children and as a decision support tool.

Authors: Lena Todnem Bach Christensen, Dikte Straadt, Stratos Vassis, Christian Marius Lillelund, Peter Bangsgaard Stoustrup, Ruben Pauwels, Thomas Klit Pedersen, and Christian Fischer Pedersen.

Status: Published in the proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–4.

An Explainable and Conformal AI Model to Detect Temporomandibular Joint Involvement in Children Suffering from Juvenile Idiopathic Arthritis

Lena Todnem Bach Christensen[*], Dikte Straadt[*], Stratos Vassis[†], Christian Marius Lillelund[*‡], Peter Bangsgaard Stoustrup[†], Ruben Pauwels[†], Thomas Klit Pedersen[†] and Christian Fischer Pedersen[*]
Email: {lena.todnem,dikte.straadt}@gmail.com {cl,cfp}@ece.au.dk
{stratos.vassis,pstoustrup,ruben.pauwels,thomas.klit}@dent.au.dk
[*]Department of Electrical and Computer Engineering, Aarhus University, Denmark
[†]Department of Dentistry and Oral Health, Aarhus University, Denmark

***Abstract*—Juvenile idiopathic arthritis (JIA) is the most common rheumatic disease during childhood and adolescence. The temporomandibular joints (TMJ) are among the most frequently affected joints in patients with JIA, and mandibular growth is especially vulnerable to arthritic changes of the TMJ in children. A clinical examination is the most cost-effective method to diagnose TMJ involvement, but clinicians find it difficult to interpret and inaccurate when used only on clinical examinations. This study implemented an explainable artificial intelligence (AI) model that can help clinicians assess TMJ involvement. The classification model was trained using Random Forest on 6154 clinical examinations of 1035 pediatric patients (67% female, 33% male) and evaluated on its ability to correctly classify TMJ involvement or not on a separate test set. Most notably, the results show that the model can classify patients within two years of their first examination as having TMJ involvement with a precision of 0.86 and a sensitivity of 0.7. The results show promise for an AI model in the assessment of TMJ involvement in children and as a decision support tool.**

***Index Terms*—Juvenile Idiopathic Arthritis, Temporomandibular Joints Involvement, Clinical Decision Support System, Machine Learning, Artificial Intelligence**

I. Introduction

It is challenging for clinicians to diagnose early involvement of the temporomandibular joints (TMJ) in children with juvenile idiopathic arthritis (JIA), as the disease is often asymptomatic [1]. Inadequate management of potential involvement of the temporomandibular joints may lead to asymmetrical jaw growth, skeletal deformities, restricted maximal incisor opening, pain or discomfort in the jaw joint area and the surrounding muscles. Furthermore, a child may experience airway constriction and sleep-disordered breathing, which can disrupt their sleeping pattern and worsen their quality of life [1]–[3]. In this context, TMJ involvement means findings believed to be caused by previous or current TMJ arthritis. There are several modalities to diagnose TMJ arthritis, e.g., clinical (orthodontic) examinations, computed tomography (CT) scans, contrast-enhanced magnetic resonance imaging (MRI), and ultrasound [4]. Clinical examinations are the most cost-effective modality to assess TMJ involvement of patients with JIA [2], however, no single parameter of the clinical examination can be suggested as a predictor of TMJ involvement in patients with JIA [2], and the sensitivity and specificity of detecting TMJ involvement using clinical examinations are low compared to CT and MRI [2], which on the other hand are costly and time-consuming [2].

Müller et al. attempted early diagnosis of TMJ involvement using clinical examinations, ultrasound and MRI [5]. Based on 33 patients with a diagnosis of JIA, they found that a clinical examination had a sensitivity of 0.47 (95% CI = 0.25-0.71) for patients and 0.66 (95% CI = 0.47-0.81) for joints. The agreement between clinical and MRI diagnoses was observed in 16 out of 30 patients (53%) and 34 out of 60 joints (57%). In particular, all 15 TMJs with condylar deformities detected by MRI were accurately diagnosed as pathological by clinical examination, but only 7 out of 18 with TMJ involvement (39%) received a correct diagnosis. This motivates the use of additional decision support if clinical examinations are to be used as an indicator of early TMJ involvement. Previous work have mainly focused on the role of clinical examination in diagnosing TMJ arthritis (i.e., the presence of active TMJ arthritis), but to the best of our knowledge, no prior work has proposed a model that can identify TMJ involvement based on clinical examinations alone. In a recent consensus article on the management of the orofacial manifestation of JIA, clinical examination was recommended as a routine screening tool [3].

In this work, we propose an explainable AI model to assess TMJ involvement in children under the age of 16 who suffer from JIA. The model is based exclusively on clinical examinations carried out in accordance with consensus-based standards for the specific patient group. The objective is to aid in early diagnosis by providing clinicians with a decision support tool that can inform them of significant clinical findings or symptoms that a patient may experience, that could indicate the presence of TMJ involvement. Early diagnosis and management can prevent some of the physical and psychological consequences that come with JIA, and lead to an anticipated better outcome for the child and their family.

[‡]Corresponding author: Christian Marius Lillelund (cl@ece.au.dk)

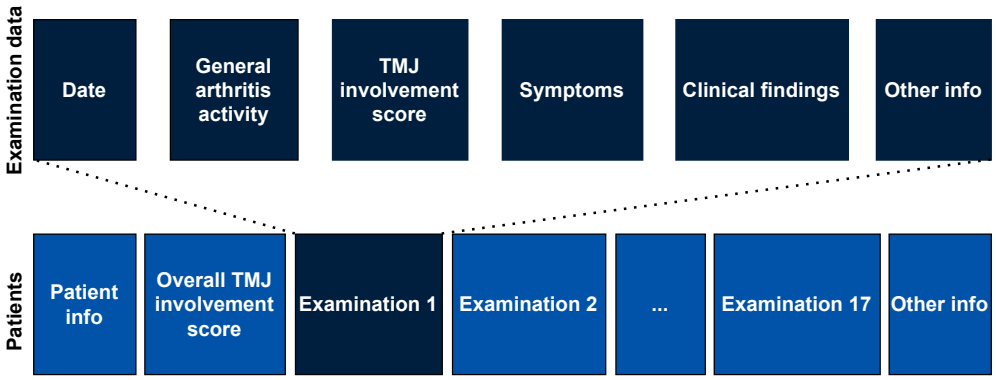


Figure 1: Each examination of a JIA patient includes a screening, where the patient's symptoms, drug usage and other clinical findings are recorded. The dataset contains historical examinations of 1035 JIA patients spanning a 25-year period with 6154 longitudinal records in total.

II. Methodology

A. Clinical data

The dataset contains 6154 longitudinal records of 1035 pediatric patients (<16 years old) over a 25-year period, recorded and provided by the Department of Dentistry and Oral Health, Aarhus University. Each patient underwent 2-17 clinical examinations, where the patient's symptoms, drug usage, general arthritis level, TMJ involvement and clinical findings were recorded (see Fig. 1). The clinical variables examined ($d = 95$) at each visit are in agreement with evidence-based standards on orofacial examination in JIA [6]. Among the cohort of patients, 690 are girls and 345 are boys, reflecting the gender distribution commonly observed in patients with JIA within an open population [7]. All patients were diagnosed with JIA prior to their first examination. However, not all patients initially showed signs of TMJ involvement; some developed it over time, while others were diagnosed with TMJ involvement after the examination was completed. The involvement of the TMJ was confirmed or refuted by an X-ray or MRI. Table I shows all recorded clinical features ($d = 95$) and a subset chosen by experts in the field ($d = 26$). We provide a complete feature description table in the Supplementary file.

B. Preprocessing and feature engineering

We encode all categorical features using label encoding and utilize entity embeddings with an embedding vector dimensionality of one to encode nominal features. The two continuous features, *openmm* and *protrusionmm*, increase with age of the patient, so we apply a transformation to indicate the deviation of the patient from the gender and age specific averages. The *drug* feature indicates what kind of medication being provided to the patient and includes 55 distinct combinations, so we split the feature into five subgroups of medication to reduce the number of features. Three general classes of medications are commonly prescribed to treat arthritis; they include non-steroidal anti-inflammatory drugs (NSAIDs), corticosteroids, and disease-modifying antirheumatic drugs (DMARDs). DMARDs can be further categorized into conventional DMARDs and biological DMARDs. By implementing this categorization, the feature space for the *drug* feature was reduced from 55 to just five. In addition, clinical features extracted from the left and right side of the face have a strong correlation, so all side-specific features were merged into a single feature taking a common value or the highest value of the left and right feature. Lastly, we apply a z-score data normalization with zero mean and one standard deviation to all features before training.

C. Classification analysis

We propose a binary classification method to predict the involvement of TMJ in children with JIA. We adopt the acclaimed Random Forest machine learning algorithm [8], since it is a flexible and easy-to-use machine learning algorithm, which has shown excellence on similar classification problems in the past. After preprocessing, we adopt an 80-10-10 data split; 80% is used for training, 10% is used for calibration, and 10% is used for testing. We report the classifier's precision, sensitivity, and macro-averaged F1 score ($\mathrm{F1_m}$) on the test set (see Tab. II). The $\mathrm{F1_m}$ is defined as:

$$F1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}, \tag{1}$$

$$F1_m = \frac{F1_{class_1} + F1_{class_2} + ... + F1_{class_i}}{N}. \tag{2}$$

D. Uncertainty estimation

To quantify the uncertainties of the predictive model, we use the Model Agnostic Prediction Interval Estimator (MAPIE) library to train a conformal classifier [9]. The classifier wraps the original classifier and produces conformal prediction sets that have a guaranteed marginal coverage rate. Thus, the model output is then a set of classes instead of a single class. We set $\alpha = 0.1$, indicating that we can expect 90% of the prediction regions to cover the true outcome. However, the guarantee is only an average for the samples from the distribution of the calibration data. In practice, we use the Rescaled Adjusted

Table I: All clinical features ($d = 95$), Expert features ($d = 26$). See the Supplementary file for a full description.

All clinical features ($d = 95$)
abrasion, aplasia, asybasis, asymenton, asyoccl, asypupilline, asyupmid, asymmetrymasseterright, asymmetrymasseterleft, backbending, bruxism, chewingfunction, clickclosingright, clickclosingleft, clicklateroleftright, clicklateroleftleft, clicklaterorightright, clicklaterorightleft, clickopeningright, clickopeningleft, clickprotrusionright, clickprotrusionleft, crepitationleft, crepitationright, deepbite, drug, dualbite, forwardbending, headache, hypermobilityleft, hypermobilityright, incisaloverjet, involvementstatus, krepitationleft, krepitationright, laterpalpleft, laterpalpright, laterotrusionleftmm, laterotrusionrightmm, lockleft, lockright, lips, lowerface, masseterleft, masseterright, micrognathism, morningstiffness, muscularpainleft, muscularpainright, neckpain, neckpalpation, neckstiffness, opening, openingfunction, openingmm, overbite, overjet, painleft, painmoveright, painmoveleft, ptext, ptint, profile, protrusion, protrusionmm, respiration, rotationleft, rotationright, sagittalrelationleft, sagittalrelationright, spacerelationship, sternoleft, sternoright, swollenjointright, swollenjointleft, swollenleft, swollenright, swollenjointright, swollenjointleft, swollenleft, swollenright, swollenright, temporalisleft, temporalisright, tempsenleft, tempsenright, tongue, tractionleft, tractionright, transversal, translationleft, translationright, dualbite, laterpalpright, laterpalpleft, postpalpright, postpalpleft

Expert-chosen clinical features ($d = 26$)
asybasis, asyoccl, asypupilline, chewingfunction, deepbite, drug, krepitationleft, krepitationright, laterotrusionleftmm, laterotrusionrightmm, laterpalpleft, laterpalpright, lowerface, openbite, opening, openingmm, overbite, overjet, painmoveleft, painmoveright, profile, protrusion, protrusionmm, retrognathism, translationleft, translationright

Partial Sums (RAPS) method with a penalty term to reduce the size of the prediction sets.

E. Explainability

Explainability helps the clinician judge whether the AI's prediction is reasonable, emphasizing clinical features of highest importance and their relationship with the predicted outcome. This can help turn AI from a black box into something that can be linked to current clinical knowledge. This also facilitates a comparative analysis of feature rankings across different feature spaces, assisting in the identification of overlapping features and their consistent significance across varied model configurations. We obtain explainability by computing Shapley Additive exPlanations (SHAP) values on the test set using the SHAP library [10].

III. Experiments and results

Table II reports the predictive performance of the classifier. We adopt three sampling strategies (IID Data, Temporal Segmentation and Lagged Features) of the clinical dataset and five evaluation metrics for each class (precision, sensitivity, macro-averaged F1 score, coverage, and set size). This section will review the sampling strategies and the classifier's results.

A. Strategy 1: IID Data

This strategy assumes that all clinical examinations are IID (independent and identically distributed), i.e., the occurrence of TMJ involvement does not change over time. This approach neglects any temporal correlation between examinations and treats them all as independent. Using all features or only the expert features yields similar predictive performance with a macro F1 score of 0.87. However, this approach has poor practical application, as it involves examinations from the early stages of the disease, where involvement is less evident, to later stages of the disease, where involvement is clearly evident.

Table II: Performance metrics on the test set. N is the number of samples in the dataset and d is the number of features. The predicted classes, TMJ_0 and TMJ_1, correspond to no and present TMJ involvement, respectively.

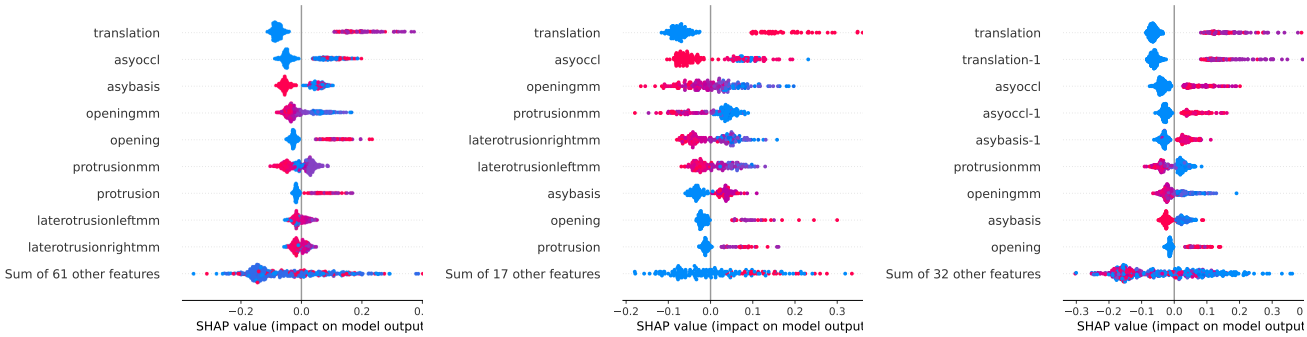
Strategy	Dimensions	Class	Precision ↑	Sensitivity ↑	F1_m ↑	Coverage ↑	Set size ↓
Sec. III-A	$N = 6006$	TMJ_0	0.88	0.93	0.8742	0.986	1.154
	$d = 95$	TMJ_1	0.88	0.81		0.899	1.207
	$N = 6006$	TMJ_0	0.88	0.93	0.8706	0.986	1.154
	$d = 26$	TMJ_1	0.88	0.80		0.899	1.207
	$N = 2682$	TMJ_0	0.89	0.95	0.8455	0.969	1.036
	$d = 26$	TMJ_1	0.86	0.70		0.740	1.078
Sec. III-B	$N = 1430$	TMJ_0	0.76	0.90	0.7451	0.988	1.233
	$d = 26$	TMJ_1	0.79	0.58		0.737	1.333
	$N = 1839$	TMJ_0	0.86	0.75	0.8144	0.967	1.565
	$d = 26$	TMJ_1	0.78	0.88		0.978	1.326
Sec. III-C	$N = 4983$	TMJ_0	0.88	0.90	0.8603	0.949	1.233
	$d = 52$	TMJ_1	0.84	0.82		0.906	1.212
	$N = 3969$	TMJ_0	0.88	0.89	0.8578	0.961	1.301
	$d = 78$	TMJ_1	0.84	0.83		0.917	1.310

B. Strategy 2: Temporal segmentation

This strategy splits examinations in three intervals: 0-2, 2-5 and 5-15 years from a patient's first examination. Thus, the first model is trained on 2682 examinations recorded in the first two years, which can be used to determine early TMJ involvement in children when they have had their first or second examination at the clinic. We see a significant drop in predictive performance of about 12.5% (from 0.8 to 0.7 sensitivity for the TMJ class), which is worse when training a model on examinations from the second to fifth year with a 27.5% drop in sensitivity (from 0.8 to 0.58). Based on the present data, it is evidently easier to detect TMJ involvement in the early stage (0-2 years) of the disease than in the middle stage (2-5). After fives years, the clinical examination can be done more thoroughly, as the clinical signs are more evident, e.g., asymmetry starts to be visible at 4-5 years; this is related to patient maturity and progression of clinical signs.

C. Strategy 3: Lagged Features

This strategy converts the timeseries dataset to a supervised-learning problem by creating columns of lag samples that contain historical feature values. The first model is trained on 4983 samples with one lag feature for each feature, and the second model is trained on 3969 samples with two lag features. This naturally increases the number of features compared to the baseline (IID Data) strategy. Adding historical information increases the sensitivity slightly by 2.5% (from 0.8 to 0.82) between the IID model using expert features only ($N = 6006$, $d = 26$) and the model with feature values from the previous observation ($N = 4983$, $d = 52$).



(a) All samples together. (b) Between 5-15 years. (c) One previous observation.

Figure 2: SHAP summary plots. The y-axis displays features, while the x-axis represents SHAP values. Features are ranked by importance, determined by the mean of their absolute SHAP values, with higher positions indicating greater significance. Each feature is represented by individual points, with a point corresponding to a single examination. The position on the x-axis signifies the impact of that feature on the model's output. The color of each point indicates whether the patient had a low (blue) or high (red) value of that specific feature. All SHAP values are computed on the test set.

IV. Discussion

The TMJ is among the most frequently affected joints in patients with JIA and mandibular growth is especially vulnerable to arthritic changes of the TMJ in children. The primary results of this study show that the proposed model with application of the "Temporal Segmentation" strategy (see Sec. III-B) can classify patients early. That is, within a period of two years (i.e., from the first visitation and two years ahead) the model can classify patients with TMJ involvement with a precision of 0.86 and a sensitivity of 0.7. This means that out of all patients predicted as having TMJ involvement, 86% had actual TMJ involvement, and 70% of patients with TMJ involvement were correctly identified. This is based on a dataset with 2682 samples and 26 clinical features. Compared to [5], who attempted to detect TMJ involvement within 1 month of the MRI for 33 patients, they report a 0.47 sensitivity (95% CI: 0.25-0.71) and a misdiagnosis rate of 61%. Therefore, using our model offers a 49% increase in sensitivity, although we note that our dataset is significantly bigger. The results show promise for an AI model in the assessment of TMJ involvement in children and as a decision support tool for clinicians. Such a tool has the potential to enhance timely diagnosis, which can improve the treatment outcome, and also assist clinicians in treatment planning.

V. Conclusion

The proposed method can be applied to current clinical examination results to help clinicians assess involvement of the temporomandibular joints in future patients, allowing for early and personalized follow-up.

References

[1] R. Rongo, A. Michelotti, T. K. Pedersen, C. Resnick, and P. Stoustrup, "Management of temporomandibular joint arthritis in children and adolescents: An introduction for orthodontists," *Orthodontics & Craniofacial Research*, vol. 26, no. S1, pp. 151–163, 2023.

[2] K. D. Kristensen, P. Stoustrup, A. Küseler, T. K. Pedersen, M. Twilt, and T. Herlin, "Clinical predictors of temporomandibular joint arthritis in juvenile idiopathic arthritis: A systematic literature review," *Seminars in Arthritis and Rheumatism*, vol. 45, no. 6, pp. 717–732, 2016.

[3] P. Stoustrup, C. M. Resnick, S. Abramowicz, T. K. Pedersen, A. Michelotti, A. Küseler, B. Koos, C. Verna, E. B. Nordal, E. J. Granquist, J. M. Halbig, K. D. Kristensen, L. B. Kaban, L. Z. Arvidsson, L. Spiegel, M. L. Stoll, M. A. Lerman, M. Glerup, P. Defabianis, P. Frid, P. Alstergren, R. Q. Cron, S. Ringold, S. E. Nørholt, T. Peltomaki, T. A. Larheim, T. Herlin, Z. S. Peacock, C. J. Kellenberger, M. Twilt, and Temporomandibular Joint Juvenile Arthritis Working Group, "Management of Orofacial Manifestations of Juvenile Idiopathic Arthritis: Interdisciplinary Consensus-Based Recommendations," *Arthritis & Rheumatology*, vol. 75, no. 1, pp. 4–14, 2023.

[4] C. Schmidt, T. Ertel, M. Arbogast, B. Hügle, T. von Kalle, and A. Neff, "The diagnosis and treatment of rheumatoid and juvenile idiopathic arthritis of the temporomandibular joint," *Deutsches Ärzteblatt International*, vol. 119, no. 4, pp. 47–54, 2022.

[5] L. Müller, C. J. Kellenberger, E. Cannizzaro, D. Ettlin, T. Schraner, I. B. Bolt, T. Peltomäki, and R. K. Saurenmann, "Early diagnosis of temporomandibular joint involvement in juvenile idiopathic arthritis: a pilot study comparing clinical examination and ultrasound to magnetic resonance imaging," *Rheumatology*, vol. 48, no. 6, pp. 680–685, 2009.

[6] P. Stoustrup, M. Twilt, L. Spiegel, K. D. Kristensen, B. Koos, T. K. Pedersen, A. Küseler, R. Q. Cron, S. Abramowicz, C. Verna, T. Peltomäki, P. Alstergren, R. E. Petty, S. Ringold, S. E. Nørholt, R. K. Saurenmann, and T. Herlin, "Clinical orofacial examination in juvenile idiopathic arthritis: International consensus-based recommendations for monitoring patients in clinical practice and research studies," *The Journal of Rheumatology*, vol. 44, no. 3, pp. 326–333, 2017.

[7] I. Cardoso, P. Frederiksen, I. O. Specht, M. N. Händel, F. Thorsteinsdottir, B. L. Heitmann, and L. E. Kristensen, "Age and Sex Specific Trends in Incidence of Juvenile Idiopathic Arthritis in Danish Birth Cohorts from 1992 to 2002: A Nationwide Register Linkage Study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, 2021.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. Cordier, V. Blot, L. Lacombe, T. Morzadec, A. Capitaine, and N. Brunel, "Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library," in *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*. PMLR, 2023, pp. 549–581.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

Appendix E

Efficient Training of Probabilistic Neural Networks for Survival Analysis

Abstract: Variational Inference (VI) is a commonly used technique for approximate Bayesian inference and uncertainty estimation in deep learning models, yet it comes at a computational cost, as it doubles the number of trainable parameters to represent uncertainty. This rapidly becomes challenging in high-dimensional settings and motivates the use of alternative techniques for inference, such as Monte Carlo Dropout (MCD) or Spectral-normalized Neural Gaussian Process (SNGP). However, such methods have seen little adoption in survival analysis, and VI remains the prevalent approach for training probabilistic neural networks. In this paper, we investigate how to train deep probabilistic survival models in large datasets without introducing additional overhead in model complexity. To achieve this, we adopt three probabilistic approaches, namely VI, MCD, and SNGP, and evaluate them in terms of their prediction performance, calibration performance, and model complexity. In the context of probabilistic survival analysis, we investigate whether non-VI techniques can offer comparable or possibly improved prediction performance and uncertainty calibration compared to VI. In the MIMIC-IV dataset, we find that MCD aligns with VI in terms of the concordance index (0.748 vs. 0.743) and mean absolute error (254.9 vs. 254.7) using hinge loss, while providing C-calibrated uncertainty estimates. Moreover, our SNGP implementation provides D-calibrated survival functions in all datasets compared to VI (4/4 vs. 2/4, respectively). Our work encourages the use of techniques alternative to VI for survival analysis in high-dimensional datasets, where computational efficiency and overhead are of concern.

Authors: Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen.

Status: Published in the IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 9, pp. 6157–6166, 2025.

Efficient Training of Probabilistic Neural Networks for Survival Analysis

Christian Marius Lillelund[1], Martin Magris[2] and Christian Fischer Pedersen[1]

***Abstract*— Variational Inference (VI) is a commonly used technique for approximate Bayesian inference and uncertainty estimation in deep learning models, yet it comes at a computational cost, as it doubles the number of trainable parameters to represent uncertainty. This rapidly becomes challenging in high-dimensional settings and motivates the use of alternative techniques for inference, such as Monte Carlo Dropout (MCD) or Spectral-normalized Neural Gaussian Process (SNGP). However, such methods have seen little adoption in survival analysis, and VI remains the prevalent approach for training probabilistic neural networks. In this paper, we investigate how to train deep probabilistic survival models in large datasets without introducing additional overhead in model complexity. To achieve this, we adopt three probabilistic approaches, namely VI, MCD, and SNGP, and evaluate them in terms of their prediction performance, calibration performance, and model complexity. In the context of probabilistic survival analysis, we investigate whether non-VI techniques can offer comparable or possibly improved prediction performance and uncertainty calibration compared to VI. In the MIMIC-IV dataset, we find that MCD aligns with VI in terms of the concordance index (0.748 vs. 0.743) and mean absolute error (254.9 vs. 254.7) using hinge loss, while providing C-calibrated uncertainty estimates. Moreover, our SNGP implementation provides D-calibrated survival functions in all datasets compared to VI (4/4 vs. 2/4, respectively). Our work encourages the use of techniques alternative to VI for survival analysis in high-dimensional datasets, where computational efficiency and overhead are of concern.**

***Index Terms*— Machine learning, Bayesian learning, uncertainty estimation, neural networks, survival analysis**

I. Introduction

Uncertainty estimation is important in risk-sensitive applications, such as healthcare, where confidence in a prognostic outcome can have life-changing implications. Machine learning algorithms often make predictions that are subject to noise or inference errors [1], which motivates the use of models that can provide uncertainty estimates in their predictions and boost the point prediction delivered by traditional approaches. In this context, we refer to *aleatoric* uncertainty, due to the inherent noise in the data, e.g., sampling noise, occlusions, or lack of quality features, and *epistemic* uncertainty, due to lack of knowledge in the model, inversely proportional to the amount of training data [2].

Survival analysis plays an important role in disease understanding and prognosis, and medical researchers use survival models to assess the importance of prognostic variables in outcomes such as death or cancer remission. Neural networks have been widely adopted for this task, e.g., DeepSurv [3], Deep Survival Machines (DSM) [4] and Deep Cox Mixtures [5], but these methods cannot provide uncertainty estimates in their predictions. Bayesian Neural Networks (BNNs), i.e., neural networks estimated with Bayesian inference, have thus been adopted in survival analysis, using pseudo-probabilities [6], the CoxPH model [7], or implemented with the Multi-Task Logistic Regression (MTLR) framework [7], [8], however, these approaches use the traditional Variational Inference (VI) technique to estimate the model parameters [9], which, among the computational drawbacks, requires doubling the number of model parameters [1]. This might not be feasible in high-dimensional settings, constituting a relevant limitation of VI. The use of alternative methods to train BNNs for survival analysis is limited, as the predominant approaches are based on VI [6]–[8], and no prior work in probabilistic survival analysis has compared VI with other approximation techniques in terms of prediction and calibration performance.

In this work, we adopt three competing approaches for training deep probabilistic survival models: VI, Monte Carlo Dropout (MCD), and Spectral-normalized Neural Gaussian Process (SNGP). We evaluate all three approaches in terms of prediction and calibration performance and compare them to established literature benchmarks. The goal is to determine whether two lightweight training techniques (MCD and SNGP) can provide the same degree of performance and uncertainty calibration as VI, but without the computational overhead. Our contributions are several[1]:

- We adopt three competing approaches for training deep probabilistic survival models: VI, MCD, and SNGP.
- We show that MCD and SNGP are suitable alternatives to VI, as they both rival the prediction and calibration performance of VI, but without needing to double the number of model parameters.
- We evaluate the approaches in four real-world datasets.

This work was supported by the PRECISE project (`http://www.aal-europe.eu/projects/precise/`) under Grant Agreement No. AAL-2021-8-90-CP by the European AAL Association.

[1]The authors are with the Department of Electrical and Computer Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark. Emails: {cl, cfp}@ece.au.dk. [2]The author is with the Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, Università degli Studi di Trieste, Via A. Valerio 4/1, 34127 Trieste, Italy. Email: martin.magris@deams.units.it

[1]The code for replicating the results is available at: `https://github.com/thecml/baysurv`.

II. Related work

Probabilistic survival analysis using BNNs has recently acquired interest [6]–[8], [10], as Bayesian models can express uncertainty about their predictions and provide implicit regularization via the prior [2]. In contrast to maximum likelihood estimation (MLE), a Bayesian model trained by maximum a posteriori averages predictions over a posterior distribution of the set of parameters: where there is no data, the confidence intervals will diverge, indicating that there are many possible extrapolations [2].

Loya et al. [8] proposed a Bayesian extension of the discrete MTLR framework that can capture patient-specific survival uncertainties, as a remedy for the assumption of temporal consistency of relative risk between two patients [11]. MTLR uses multitask regression and joint likelihood minimization to model the log-risk at a given time interval as a linear combination of the covariates. The authors adopted this structure using a BNN with an element-wise structure (the first and second layers must have the same number of neurons) with feature selection in mind. Using VI to learn the posterior distribution and drawing samples from the predictive distribution, their model improves over traditional CoxPH and MTLR in predictive accuracy while providing confidence bounds on the predicted survival probability.

Qi et al. adopted the CoxPH and MTLR framework to formulate a BNN for survival analysis using VI [7]. The authors investigated the selection of an appropriate prior distribution on the model weights and how this can enforce sparsity for later feature selection. They also proposed a novel framework for estimating individual survival distributions with credible intervals (CrIs) around the survival function, thus communicating the underlying uncertainty on the model weights. Their work is applicable as a decision support system, as predicted survival functions are specific to individual patients.

Lillelund et al. showed that MCD can provide satisfactory predictive results in survival models while being significantly faster to train than VI [10]. Using the CoxPH framework, they quantified epistemic and aleatoric uncertainty by replacing the network's weights with a distribution and sampling the network's outputs from a Gaussian. The results in Lillelund et al. [10] demonstrate that aleatoric and epistemic uncertainty can be included in a deep survival model at some computational cost, but the authors do not report the calibration performance of the models or provide any measurement of the quality of the uncertainty estimates. In addition, the evaluation is performed only on a relatively small dataset (8873 samples).

Except for [10], all previous work uses VI [9] to infer the posterior distribution over the network parameters. This method approximates the true posterior distribution with a surrogate of a tractable parametric form. However, VI requires doubling (at least) the number of model parameters and comes with a high computational cost [1], e.g., because of the extensive use of simulation and gradient computation, which is restrictive and inefficient in high-dimensional settings. This motivates adopting non-VI techniques, but using such techniques to train deep survival models is underexplored.

III. Methods

This section introduces the reader to the main concepts in survival analysis, outlines the popular CoxPH model fundamental to our work, and describes three modern and commonly used approaches for training probabilistic models. Subsection III-D discusses our proposed network architectures.

A. Survival analysis

Survival analysis models the time until an event occurs. This event can be observed or not, e.g., a censored observation due to the termination of the study. Let $\mathcal{D} = \{(y_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N}$ be the data for the ith record (individual), where y_i denotes the observed time, δ_i is a binary event indicator (the event occurred/did not occur) and $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional vector of features (covariates). Moreover, let c_i denote the censoring time and e_i denote the event time for the ith record, thus $y_i = e_i$ if $\delta_i = 1$ or $y_i = c_i$ if $\delta_i = 0$. Survival analysis then models the survival probability $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$, i.e., the probability that some event occurs at time T later than t. Let t_k denote some measurement time on some discretized event horizon of K bins, $[t_0,\ t_1,\ t_2,\ \ldots,\ t_K]$, then at t_0 the survival function $S(t_0)$ is one and monotonically decreases for $t_k > 0$, so that the probability of surviving indefinitely is zero. In the absence of censored observations, $S(t_k)$ corresponds to the fraction of survivors at time t_k. In survival analysis, the hazard function is commonly used to estimate $S(t)$:

$$h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T > t)/\Delta t, \qquad (1)$$

which provides the instantaneous failure *rate* at given time instance t, conditional on surviving t [12, Ch. 11]. The hazard function and the survival function are connected through $h(t) = f(t)/S(t)$, with $f(t)$ being the probability density of T, i.e. $f(t) := \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t)/\Delta t$, the instantaneous (unconditional) rate of failure at t. The functions $S(t)$, $h(t)$, $f(t)$, thus summarize different probabilistic aspects of the random variable T, and are related to each other.

B. The Cox Proportional Hazard model

The Cox Proportional Hazards model (CoxPH) is a standard approach for survival analysis [13]. By adopting a multiplicative form for the contribution of several covariates to each individual's survival time, the CoxPH model is a powerful yet simple tool for assessing the simultaneous effect different covariates have on survival times. The conditional individual hazard function assumed by the CoxPH model is $h(t|\boldsymbol{x}_i) = h_0(t)\exp(f(\boldsymbol{\omega}, \boldsymbol{x}_i))$, where $h_0(t)$ stands for the baseline hazard at time t, $f(\boldsymbol{\omega}, \boldsymbol{x}_i)$ is called the risk score and f is a linear function in the parameters [13]. Even if $h_0(t)$ remains unknown, it is possible to estimate $\boldsymbol{\omega}$ independently in the exponential part, with estimated hazards that satisfy the non-negative constraint. Its standard MLE of $\hat{\boldsymbol{\omega}}$ optimizes the Cox partial log-likelihood as follows:

$$\log p(\mathcal{D}|\boldsymbol{\omega}) = \sum_{i:\delta_i=1} \log \frac{h(y_i|\boldsymbol{x}_i)}{\sum_{j:T_j \geq T_i} h(y_i|\boldsymbol{x}_j)}, \\ = \sum_{i:\delta_i=1} f(\boldsymbol{\omega}, \boldsymbol{x}_i) - \sum_{j:T_j \geq T_i} f(\boldsymbol{\omega}, \boldsymbol{x}_j). \quad (2)$$

Among the available estimators for $h_0(t)$, leading to the estimation of the survival function as $\hat{S}(t) = \hat{S}_0(t)^{\exp(\boldsymbol{x}_i\hat{\boldsymbol{\omega}})}$, with $\hat{S}_0(t) = \exp(-\int_0^t \hat{h}_0(u)\mathrm{d}u)$, the Breslow estimator is the standard [14].

C. Uncertainty in Bayesian deep learning

In this subsection, we introduce three major approaches for training probabilistic machine learning models.

Variational Inference (VI): Let $\mathcal{D}$ denote the data, $p(\mathcal{D}|\boldsymbol{\omega})$ the likelihood, and let the prior distribution on the parameter be denoted by $\boldsymbol{\omega}$ as $p(\boldsymbol{\omega})$. The object of interest in Bayesian inference is the posterior distribution $p(\boldsymbol{\omega}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\omega})p(\boldsymbol{\omega})/p(\mathcal{D})$. To obtain the distribution of the response $\boldsymbol{y}^\star$ for a new data sample $\boldsymbol{x}^\star$, the so-called posterior distribution is obtained as $p(\boldsymbol{y}^\star|\boldsymbol{x}^\star) = \int_\Theta p(\boldsymbol{y}^\star|\boldsymbol{x}^\star, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{D})\mathrm{d}\boldsymbol{\omega}$.

In general, determining the posterior is not an easy task, as the model evidence $p(\mathcal{D})$ corresponds to an intractable integral. When the dimensionality of the problem is high, the use of sampling methods is prohibitive, and VI constitutes an attractive alternative. In VI, the actual unknown posterior distribution is approximated with a variational distribution $q(\boldsymbol{\omega})$ of a tractable parametric form, whose (variational) parameter is denoted by $\boldsymbol{\zeta}$. Typically, this corresponds to a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, thus $\boldsymbol{\zeta} = [\boldsymbol{\mu}; \mathrm{vec}\,\boldsymbol{\Sigma}]$. VI achieves the best variational approximation to $p(\boldsymbol{\omega}|\mathcal{D})$ by minimizing the KL divergence from $q(\boldsymbol{\omega})$ to $p(\boldsymbol{\omega}|\mathcal{D})$, i.e., determines the variational parameter $\boldsymbol{\zeta}^\star$ by minimizing:

$$\mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega}|\mathcal{D})) := \\ -\int q(\boldsymbol{\omega}) \log p(\mathcal{D}|\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} + \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})), \quad (3)$$

for which several algorithms have been developed in the last decade [2].

Monte-Carlo Dropout (MCD): [1] and [15] provide a formal connection between Monte Carlo Dropout (MCD) and BNNs. Here, we discuss the major breakthrough for the relevant non-categorical case for the $\boldsymbol{y}$s. Consider a deep Gaussian Process (GP) with a covariance function of the form:

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \int p(\boldsymbol{w})p(b)\sigma(\boldsymbol{w}'\boldsymbol{x} + b)\sigma(\boldsymbol{w}'\boldsymbol{y} + b)\mathrm{d}\boldsymbol{w}\mathrm{d}b, \quad (4)$$

where $\boldsymbol{W}_i$ is the weight matrix of dimension $K_i \times K_{i-1}$ for each layer $i = 1, \ldots, L$ and $\boldsymbol{b}_i$ is the corresponding bias vector. We use $\boldsymbol{\omega} = \{\boldsymbol{W}_i\}_{i=1}^L$, where each row of $\boldsymbol{W}_i$ is distributed according to a standard multivariate normal distribution $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, l^{-2}\boldsymbol{I})$. Let $\sigma(\cdot)$ be a general non-linear activation function and $p(b)$ some distribution on the elements of $\boldsymbol{b}_i$. Given some precision parameter $\tau > 0$, the predictive distribution of the deep GP is

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{D})\mathrm{d}\boldsymbol{\omega}, \quad (5)$$

with

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega}) = \mathcal{N}\big(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{\omega}), \tau^{-1}\boldsymbol{I}\big). \quad (6)$$

For the intractable posterior $p(\boldsymbol{\omega}|\mathcal{D})$, we adopt a variational approximation $q(\boldsymbol{\omega})$ defined as:

$$\boldsymbol{W}_i = \boldsymbol{M} \cdot \mathrm{diag}\Big([z_{i,j}]_{j=1}^{K_i}\Big), \quad (7)$$

$$z_{i,j} \sim \text{Bernoulli}(p_i),\ i = 1, \ldots, L,\ j = 1, \ldots, K_{i-1}, \quad (8)$$

where $\boldsymbol{M}_i$ are the variational parameters and p_i some given probabilities. [1] shows that minimizing the KL divergence between the variational posterior $q(\boldsymbol{\omega})$ and the actual posterior of the deep GP is equivalent to optimizing the following objective with dropout:

$$\mathcal{L}_{\mathrm{MCD}} = \frac{1}{N}\sum_{i=}^{N} E(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) + \lambda \sum_{i=1}^{N} \big(||\boldsymbol{W}_i||_2^2 + ||\boldsymbol{b}_i||_2^2\big). \quad (9)$$

L is a standard non-probabilistic optimization objective, using the L_2 regularization and tuning parameter λ and the Euclidean square loss $E(\cdot, \cdot)$. Dropout means that for every input data and every unit in each layer, a Bernoulli variable is sampled with probability p_i, and a unit is dropped if the corresponding binary variable is 0. The probability $1 - p$ of dropping a unit is generally called the dropout rate. Therefore, a DNN, with dropout applied at every layer, is equivalent to a Monte Carlo approximation to the probabilistic deep GP; i.e., it minimizes the KL divergence between the variational distribution and the posterior of the deep GP modeling the distribution over the space of distributions that can have generated the data. It is straightforward to approximate the predictive distribution as follows:

$$p(\boldsymbol{y}^\star|\boldsymbol{x}^\star) = p(\boldsymbol{y}^\star|\boldsymbol{x}^\star, \boldsymbol{\omega})q(\boldsymbol{\omega}), \quad (10)$$

$$q(\boldsymbol{\omega}) = \text{Bernoulli}(z_1) \ldots \text{Bernoulli}(z_L), \quad (11)$$

$$p(\boldsymbol{y}^\star|\boldsymbol{x}^\star, \boldsymbol{\omega}) = \mathcal{N}\big(\boldsymbol{y}^\star; \hat{\boldsymbol{y}}^\star(\boldsymbol{x}^\star, z_1, \ldots, z_L), \tau^{-1}\boldsymbol{I}\big), \quad (12)$$

and, e.g., its mean through MC sampling:

$$\mathbb{E}_{p(\boldsymbol{y}^\star|\boldsymbol{x}^\star)} \approx \frac{1}{S}\sum_{s=1}^{S} \hat{y}^\star(\boldsymbol{x}^\star, \hat{z}_{1,s}, \ldots, \hat{z}_{L,s}), \quad (13)$$

$$\hat{z}_{i,s} \sim \text{Bernoulli}(p_i). \quad (14)$$

The practical simplicity of MCD makes it an attractive and widely adopted baseline for comparison with alternative Bayesian approaches. However, within neural network applications in survival analysis, the use of MCD for uncertainty quantification is very limited [10].

Spectral-normalized Neural Gaussian Process (SNGP): SNGP [16] is an approach for improving the quality of uncertainty modeling in a deep neural network whilst preserving its overall predictive accuracy. Over a deep residual network, SNGP applies spectral normalization to the hidden layers and replaces the dense output layer with a GP layer. Compared to MCD, SNGP works for a wide range of state-of-the-art residual-based architectures. Furthermore, it does not rely on ensemble averaging and, therefore, has a similar level of latency as a deterministic network (besides requiring a covariance reinitialization of low complexity at each iteration), scaling easily to large dimensions and lengthy datasets [16].

D. Proposed architectures

We propose three separate neural network architectures for probabilistic survival analysis. For the purpose of comparability in the results, consistency in the hyperparameters, and evaluation fairness, we propose a Multi-Layer Perceptron (MLP) as the backbone network architecture. An MLP is a fully connected feedforward network; the number of hidden layers and neurons are hyperparameters tuned based on the data, see, e.g., [2]. The adoption of an MLP is aligned with the literature, as capable of providing solid prediction performance [3], [4]. The proposed architectures are depicted in Fig. 1. There are three probabilistic training approaches:

i. training with VI by optimizing Eq. (3), where the weights and biases are treated as random variables. The output nodes provide the mean and standard deviation of the output $\hat{y}_i$ as a Gaussian draw, i.e., $\hat{y}_i \sim \mathcal{N}(\mu_{\boldsymbol{x}_i}, \sigma_{\boldsymbol{x}_i})$ for a d-dimensional sample $\boldsymbol{x}_i$. The Bayesian framework captures the epistemic uncertainty in the model, and sampling the output from a Gaussian captures the aleatoric uncertainty. See Fig. 1a.
ii. training with MCD, capturing both epistemic and aleatoric uncertainty. The network architecture is equivalent to the deterministic MLP, but now MC dropout is applied at each feedforward pass. The two output nodes correspond to the mean and variance of the predictive Gaussian, from which the ys are drawn conditionally on $\boldsymbol{x}_i$ and the random network configuration at each epoch due to the random deactivation of some of the nodes. We use dropout rates $p = \{0.1, 0.2, 0.5\}$. See Fig. 1b.
iii. training with SNGP, which can estimate the uncertainty per sample without requiring random sampling. Given a batch input $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, the GP layer returns predictions $\{y_1, y_2, ..., y_N\}$ along with the predictive covariance matrix of the predictions. The model is optimized according to the MLE objective in Eq. (2). See Fig. 1c.

Table I highlights the advantages and disadvantages of adopting VI, MCD and SNGP.

IV. Experiments and Results

A. Data and setup

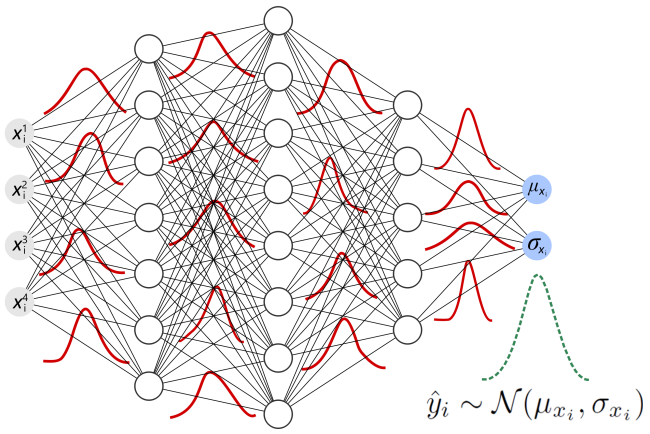
We use four datasets to evaluate the proposed architectures. (1) The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset contains gene and protein expression profiles used to identify breast cancer subgroups [18]. We obtain the dataset from [3], consisting of 1904 observations, 9 covariates, and a censoring rate of 42%. (2) The U.S. Surveillance, Epidemiology, and End Results (SEER) dataset contains survival times for cancer patients [19], specifically patients with infiltrating duct and lobular carcinoma breast cancer diagnosed between 2006 and 2010. The dataset has 4024 observations and a censoring rate of 85%. (3) The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) dataset is a large study that tracks the survival time of hospitalized adults [20]; the cleaned dataset is available from [3] and consists of 8873 observations, 14 covariates, and a censoring rate of 32%. (4) MIMIC-IV is a large electronic health record dataset with 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center in the United States between 2008 and 2019 [21]. The dataset includes patient demographics, clinical measurements, medications, and diagnoses, recorded during a patient's stay at the hospital and their survival time. See [7] for accessing the data, which has 38,520 observations, 91 covariates, and a censoring rate of 67%. After imputing missing values by sample mean for real-valued covariates or mode for categorical covariates, we apply a z-score data normalization and use one-hot encoding for categorical covariates. We split the data into train, validation and test sets by 70%, 10%, and 20% using a stratified procedure, as proposed by [7]. The stratification ensures that the event times and censoring rates are consistent across the three sets.

For the purpose of comparability in the results, we implement the traditional continuous CoxPH model using a linear estimator [13], weight regularization [22], gradient boosting [23], Random Survival Forests [24], two non-Bayesian neural networks [4], [5], and two recent BNNs based on the continuous CoxPH model and the discrete MTLR framework, respectively [7]. Our baseline MLP is equivalent to the DeepSurv model [3]. For all datasets and models, we use Bayesian optimization [25] to tune the hyperparameters over ten iterations using the validation set. These include the number of iterations, the batch size, and the network architecture, when applicable. We used the hyperparameters leading to the highest average concordance index to configure the final models. The CoxPH-based models utilize the entire dataset during training, given the nature of the partial likelihood, whereas the neural networks load stochastic subsets of the training set as mini-batches. Additionally, we use the early stopping technique, where the training process halts if the likelihood loss of the model stops improving on the validation dataset, for all neural network-based models. See App. II for more details on hyperparameter optimization.

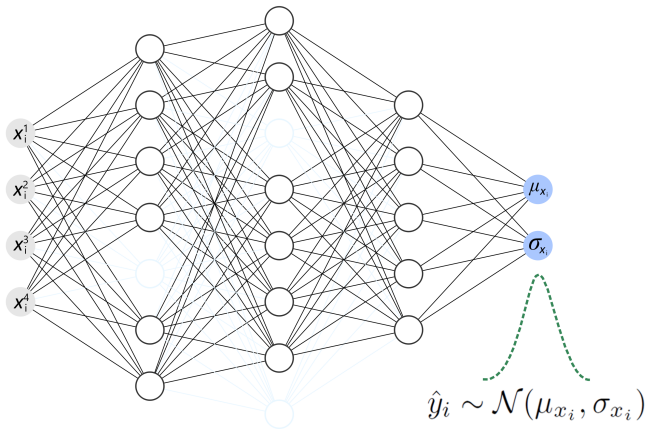
We adopt several evaluation metrics to assess the prediction and calibration performance of our baseline MLP model, its probabilistic variants, and literature benchmarks. To evaluate prediction performance, we report Antolini's Concordance Index (CI_{td}) [26], the Mean Absolute Error using hinge loss (MAE_{H}) [27], the MAE using pseudo-observations (MAE_{PO}) [27], and the Integrated Brier Score (IBS) [28]. To evaluate calibration performance, we report the Integrated Calibration Index (ICI) [29], Distribution Calibration (D-Cal) [30] and Coverage Calibration (C-Cal) [7]. See App. I for more details on the performance metrics.

B. Prediction results

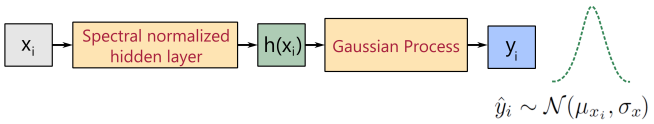
Table II shows the prediction performance of our baseline MLP, its probabilistic variants, and the literature benchmarks on the test sets. CI_{td} measures the rank correctness between pairwise observations (i, j), i.e., if individual i experiences the event before j, their predicted survival time should be shorter. MAE is the mean absolute error between the predicted and actual survival times, using either hinge loss (MAE_{H}) or pseudo-observations (MAE_{PO}) to calculate the error for censored individuals. We use the median of the survival



(a) VI.



(b) MCD.



(c) SNGP.

Fig. 1: The proposed architectures for computing $\hat{y}_i$, i.e., the predicted survival time for individual i. (a): BNN with three hidden layers trained with VI, where $\boldsymbol{x}_i \in \mathbb{R}^4$ is a four-dimensional covariate vector. (b): MLP with three hidden layers trained with MCD. Transparent objects emphasize the random activation of the nodes under MCD. (c): MLP using the SNGP technique, with spectral normalized hidden layers and the output computed from a GP layer.

TABLE I: Advantages and disadvantages of the three proposed architectures.

	Advantages:	*Disadvantages:*
VI [9]	+ Mathematically-robust & tractable + Provides a posterior distribution	- Requires training additional model parameters -Heavily relies on MC sampling - Has long training times compared to MCD and SNGP [1], [16]
MCD [1]	+ No parametric overhead compared to VI + Has shorter training times than VI on average	- Does not provide an actual posterior distribution - Only approximates the VI solution under certain assumptions [1] - Uncertainty estimation coupled to the choice of dropout rate [17]
SNGP [16]	+ Fast, deterministic approach + No parametric overhead compared to VI + Has shorter training times than VI on average	- Does not provide an actual posterior distribution - Cannot compute individual CrIs

function as the predicted survival time. IBS is the mean squared error between the binary event outcome $y \in \{0, 1\}$ and the predicted probability of the event over the entire event horizon.

For VI and MCD, the reported metrics are constructed based on predictive means by sampling 100 times from the posterior predictive distribution. We train the networks for 100 epochs using early stopping with five epochs of patience. Concerning the baseline MLP model and the probabilistic variants, in the smallest METABRIC dataset (see Table IIa), all probabilistic models report a higher CI_{td} than the baseline, and MCD in particular shows significantly lower MAE_{H} and MAE_{PO}. Thus, using VI or MCD improves the prediction performance over the MLP in this dataset. In the mid-sized SEER dataset (see Table IIb), all probabilistic variants give approximately the same CI_{td}, but have significantly lower MAE_{H} than the baseline. In the larger SUPPORT dataset (see Table IIc), only MCD outperforms both the baseline and its rival probabilistic models. In the large MIMIC-IV dataset (see Table IId), the effect of the different estimation approaches is reduced, and all measures are generally aligned, although VI and MCD still have an edge when it comes to MAE_{H}. We see that for all datasets in this experiment, using MCD with $p = 0.5$ leads to better convergence and performance than the baseline.

With respect to the existing models, MCD with $p = 0.5$ outperforms all continuous models in terms of MAE_{H} in 3/4 datasets and in terms of MAE_{PO} in 2/4 datasets. In the large MIMIC-IV dataset, it outperforms all competing solutions in terms of CI_{td}. Although BNN-MTLR does demonstrate even better MAE throughout, it is a discrete model with two

TABLE II: Prediction performance on the METABRIC, SEER, SUPPORT, and MIMIC-IV test sets. In many cases, the proposed models give notable improvements in mean absolute error over continuous state-of-the-art baselines, with similar scores between VI and MCD. Bold indicates the top 3 models including ties. N: sample size, C: pct. of censored data, d: number of covariates.

(a) METABRIC ($N = 1902$, $C = 42\%$, $d = 9$).

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH [13]	0.618	68.923	96.797	0.173
CoxNet [22]	0.611	68.288	97.631	0.173
CoxBoost [31]	0.638	69.976	102.745	0.187
RSF [24]	0.641	67.242	98.728	0.175
DSM [4]	**0.670**	70.361	100.83	0.170
DCM [5]	0.637	74.187	104.532	0.179
BNN-Cox (VI) [7]	**0.672**	65.777	94.184	**0.166**
BNN-MTLR (VI) [7]	**0.655**	**59.027**	**89.387**	**0.163**
This work:				
Baseline (MLP)	0.604	68.668	95.540	0.168
+ VI	0.634	**63.776**	**91.843**	0.168
+ MCD ($p = 0.1$)	0.621	67.235	93.981	0.167
+ MCD ($p = 0.2$)	0.628	67.371	96.184	**0.166**
+ MCD ($p = 0.5$)	0.632	**65.651**	**92.353**	**0.164**
+ SNGP	0.631	67.378	93.434	**0.166**

(b) SEER ($N = 4024$, $C = 85\%$, $d = 28$).

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH [13]	**0.723**	100.849	196.898	**0.083**
CoxNet [22]	**0.721**	95.253	197.066	0.083
CoxBoost [31]	0.689	106.635	232.204	0.092
RSF [24]	0.682	102.062	206.41	0.086
DSM [4]	0.699	153.84	413.543	0.088
DCM [5]	0.713	108.464	223.397	**0.082**
BNN-Cox (VI) [7]	**0.716**	95.184	**202.151**	**0.081**
BNN-MTLR (VI) [7]	0.702	**48.166**	**201.856**	0.085
This work:				
Baseline (MLP)	0.703	101.972	204.408	0.084
+ VI	0.699	**81.923**	**199.796**	0.084
+ MCD ($p = 0.1$)	0.702	91.202	203.457	0.084
+ MCD ($p = 0.2$)	0.691	89.103	203.323	0.085
+ MCD ($p = 0.5$)	0.692	**82.001**	207.956	0.087
+ SNGP	0.698	106.032	209.498	0.084

(c) SUPPORT ($N = 8873$, $C = 32\%$, $d = 14$).

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH [13]	0.568	409.947	754.978	0.201
CoxNet [22]	0.568	409.915	754.960	0.201
CoxBoost [31]	**0.612**	**366.928**	**721.443**	0.190
RSF [24]	0.602	377.862	736.800	0.200
DSM [4]	**0.626**	384.349	723.002	**0.185**
DCM [5]	**0.629**	434.376	759.509	**0.185**
BNN-Cox (VI) [7]	0.610	416.650	738.350	0.189
BNN-MTLR (VI) [7]	0.606	**369.060**	**707.679**	0.186
This work:				
Baseline (MLP)	0.609	391.573	722.709	0.187
+ VI	0.531	414.133	768.599	0.208
+ MCD ($p = 0.1$)	0.600	416.175	738.567	0.193
+ MCD ($p = 0.2$)	0.567	400.064	751.343	0.202
+ MCD ($p = 0.5$)	0.605	**374.046**	**717.227**	0.192
+ SNGP	0.610	397.877	726.725	**0.184**

(d) MIMIC-IV ($N = 38520$, $C = 67\%$, $d = 91$).

Model	CI_{td} ↑	MAE_H ↓	MAE_{PO} ↓	IBS ↓
CoxPH [13]	0.726	257.364	286.516	**0.174**
CoxNet [22]	0.727	255.802	283.951	**0.174**
CoxBoost [31]	0.656	263.836	297.763	0.194
RSF [24]	0.694	260.668	293.247	0.188
DSM [4]	0.727	259.110	**280.361**	0.180
DCM [5]	0.742	257.801	286.580	**0.174**
BNN-Cox (VI) [7]	0.743	256.007	288.814	**0.169**
BNN-MTLR (VI) [7]	0.744	**250.605**	**278.710**	**0.167**
This work:				
Baseline (MLP)	0.744	256.694	283.635	0.178
+ VI	0.743	254.735	282.345	0.180
+ MCD ($p = 0.1$)	**0.746**	255.815	283.672	0.177
+ MCD ($p = 0.2$)	**0.748**	**254.560**	283.156	0.175
+ MCD ($p = 0.5$)	**0.748**	**254.971**	**281.928**	0.177
+ SNGP	0.736	257.195	282.082	0.182

important caveats: (1) The MTLR framework requires training K logistic regression models, where K is the number of discrete time bins, leading to a space complexity of $O(Kd)$ instead of $O(d)$ for the continuous models [11] (see Table IV). (2) The model is trained using VI, which itself doubles the number of parameters compared to the MLP baseline and MCD. Overall, we see an advantage in combining the MLP architecture with Bayesian approximation techniques to rank individuals at risk and predict the time to event. These benefits are most apparent in small-sized datasets, as the Bayesian algorithms naturally incorporate a form of regularization (the prior) and hence are less prone to overfitting. As the size of the dataset increases, the effect of the prior diminishes, and the Bayesian solution approaches the MLE. Most importantly, the proposed architectures most often do not deteriorate prediction performance in our experiments; rather, they often improve it.

C. Calibration results

Table III shows the calibration performance of our baseline MLP and its VI, MCD, and SNGP variants and the literature benchmarks in the test sets. ICI measures the mean absolute difference between the observed and predicted probabilities of an event. D-calibration expresses to what extent we can trust the predicted survival function, $S(t)$, according to a χ^2 test. By dividing $S(t)$ into equally-sized intervals, it verifies whether the proportion of events in each interval is uniformly distributed. A p-value greater than 0.05 indicates that $S(t)$ is calibrated with respect to the event distribution. C-calibration expresses the agreement between the upper and lower bounds of the predicted CrIs and the observed intervals.

Regarding the baseline MLP model and its probabilistic variants, all models are D-calibrated on the small METABRIC dataset (see Table IIIa), with the VI and MCD variants also being C-calibrated. We assess the agreement between predicted and observed survival probabilities using the ICI metric, which is the mean absolute difference between the calibration curve and the diagonal line of perfect calibration. We see that the MCD models improve on the baseline in terms of ICI and thus provide a better-calibrated survival function. Using VI, despite being distribution and coverage calibrated, has a four-time higher ICI score than MCD in the METABRIC dataset. In the mid-sized SEER dataset (see Table IIIb), all proposed

TABLE III: Calibration performance on the METABRIC, SEER, SUPPORT, and MIMIC-IV test sets. MCD and SNGP have better calibration performances (D-Cal and C-Cal) than VI in small datasets, but similar performances in larger ones. Bold indicates the top 3 models including ties. N: sample size, C: pct. of censored data, d: number of covariates.

(a) METABRIC ($N = 1902$, $C = 42\%$, $d = 9$).

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH [13]	0.022	**0.959**	-
CoxNet [22]	**0.011**	**0.977**	-
CoxBoost [31]	0.147	0.822	-
RSF [24]	0.047	0.530	-
DSM [4]	0.025	0.017	-
DCM [5]	0.066	0.160	-
BNN-Cox (VI) [7]	**0.017**	0.101	0.812
BNN-MTLR (VI) [7]	0.044	0.438	0.763
This work:			
Baseline (MLP)	0.040	0.853	-
+ VI	0.062	0.059	0.796
+ MCD ($p = 0.1$)	**0.017**	0.735	**0.849**
+ MCD ($p = 0.2$)	0.020	0.689	**0.857**
+ MCD ($p = 0.5$)	**0.017**	0.760	**0.865**
+ SNGP	**0.017**	**0.964**	-

(b) SEER ($N = 4024$, $C = 85\%$, $d = 28$).

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH [13]	0.011	0.999	-
CoxNet [22]	0.013	**1.000**	-
CoxBoost [31]	0.051	**1.000**	-
RSF [24]	0.021	**1.000**	-
DSM [4]	**0.009**	0.931	-
DCM [5]	0.016	**1.000**	-
BNN-Cox (VI) [7]	**0.009**	**1.000**	**0.489**
BNN-MTLR (VI) [7]	0.056	0.016	0.486
This work:			
Baseline (MLP)	**0.009**	**1.000**	-
+ VI	0.027	0.678	**0.487**
+ MCD ($p = 0.1$)	0.013	0.989	**0.487**
+ MCD ($p = 0.2$)	0.014	0.983	**0.487**
+ MCD ($p = 0.5$)	0.037	0.841	0.486
+ SNGP	0.011	**1.000**	-

(c) SUPPORT ($N = 8873$, $C = 32\%$, $d = 14$).

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH [13]	0.059	**0.983**	-
CoxNet [22]	0.059	**0.983**	-
CoxBoost [31]	**0.019**	0.708	-
RSF [24]	**0.022**	0.532	-
DSM [4]	0.065	0.000	-
DCM [5]	0.090	**0.904**	-
BNN-Cox (VI) [7]	0.073	0.096	0.952
BNN-MTLR (VI) [7]	**0.042**	0.029	0.947
This work:			
Baseline (MLP)	0.069	0.229	-
+ VI	0.064	0.000	**0.977**
+ MCD ($p = 0.1$)	0.077	0.000	0.938
+ MCD ($p = 0.2$)	0.061	0.000	**0.979**
+ MCD ($p = 0.5$)	0.050	0.000	**0.956**
+ SNGP	0.075	0.689	-

(d) MIMIC-IV ($N = 38520$, $C = 67\%$, $d = 91$).

Model	ICI ↓	D-Cal ↑	C-Cal ↑
CoxPH [13]	0.037	0.549	-
CoxNet [22]	0.033	**0.553**	-
CoxBoost [31]	**0.016**	**0.997**	-
RSF [24]	0.041	**0.935**	-
DSM [4]	**0.006**	0.000	-
DCM [5]	0.041	0.011	-
BNN-Cox (VI) [7]	0.066	0.001	0.589
BNN-MTLR (VI) [7]	0.057	0.060	0.583
This work:			
Baseline (MLP)	0.026	0.068	-
+ VI	0.096	0.000	0.583
+ MCD ($p = 0.1$)	0.034	0.014	**0.600**
+ MCD ($p = 0.2$)	0.038	0.013	**0.596**
+ MCD ($p = 0.5$)	0.036	0.008	**0.597**
+ SNGP	**0.015**	0.230	-

architectures have D-calibration and C-calibration p-values greater than 0.05, indicating proper calibration; however, we observe a decrease in calibration performance in terms of ICI when using VI and MCD, also evidenced by lower p-values for D-calibration. In the SUPPORT dataset (see Table IIIc), only the baseline MLP and SNGP variants are statistically D-calibrated, and we see very poor D-calibration scores using VI or MCD. Indeed, increasing the complexity of the model (the number of parameters) and making the model more flexible can worsen D-calibration [30].

We observe a similar trend in the large MIMIC-IV dataset (see Table IIId), as none of the probabilistic methods are D-calibrated except for the SNGP. Regarding the tight coupling between the quality of the MCD uncertainty estimates and the parameter choices [17], e.g., the choice of dropout rate, we conduct empirical experiments with various dropout rates to measure its effect. In terms of calibration performance, we do not see significant differences using a low or high dropout rate; however, in the SEER, SUPPORT, and MIMIC-IV datasets, setting $p = 0.1$ provides higher p-values for D-calibration.

With respect to the existing models, our SNGP and MCD variants and the BNN-Cox model are among those that give the lowest ICI scores on average but do not necessarily provide D-calibrated survival functions. In the large MIMIC-IV dataset, neither the existing neural network models nor our proposed VI and MCD variants are D-calibrated, except for the BNN-MTLR model. In contrast, the SNGP approach, comprising a deterministic MLP with the output computed from a GP layer, is properly D-calibrated across all experimental datasets. It is important to rigorously assess multiple metrics to see how calibrated a model truly is because each calibration metric assesses a different kind of calibration, and it may be application-specific which type of calibration is needed. We note that the proposed models have been optimized according to the MLE objective in Eq. (2), which optimizes discriminative performance, not calibration performance.

D. Model complexity

Table IV compares two literature benchmark models and our proposed models in terms of space complexity and number of trainable parameters on the MIMIC-IV dataset. For a fair comparison, we configure all models with a single hidden

layer with 32 nodes. Notice the doubling of the number of parameters when using VI as opposed to the baseline MLP and MCD, resulting in increased overhead. Despite the fact that both are of $\mathcal{O}(d)$ complexity, doubling the number of trainable parameters can make a significant difference in high-dimensional models, so that VI is no longer an attractive solution.

TABLE IV: Space complexity of the models and number of trainable parameters (for the MIMIC-IV data, 91 covariates).

Model	Space complex.	#Params.
BNN-Cox (VI) [7]	$\mathcal{O}(d)$	5570
BNN-MTLR (VI) [7]	$\mathcal{O}(Kd)$	10718
This work:		
Baseline (MLP)	$\mathcal{O}(d)$	2849
+ VI	$\mathcal{O}(d)$	5700
+ MCD	$\mathcal{O}(d)$	2882
+ SNGP	$\mathcal{O}(d)$	3776

E. Uncertainty estimation

We adopt the concept of Credible Intervals (CrIs) from [7]. CrIs belong to the Bayesian framework and offer a way to quantify the uncertainty associated with a prediction. A credible interval is defined by two values, called the lower and upper bounds, which enclose a certain percentage of the predictive distribution. The main difference between CrIs and confidence intervals is that CrIs treat the parameter of interest as a random variable and the bounds as fixed values, while confidence intervals treat the parameter of interest as a fixed value and the bounds as random variables [32, Ch. 4]. Fig. 2 shows two ways in which uncertainty can be expressed along with a survival prediction for decision support purposes. For this, we selected a random sample from the METABRIC test set and drew 1000 samples from the predictive distribution to estimate its survival function. Although MCD and VI predict similar median survival times (in Fig. 2, 122 and 128, respectively, where the actual time is 135), the latter predicts a greater amount of variance around the survival function (left column) and also the histogram of the predicted survival time displays higher variance (right column). The width of the CrIs expresses the uncertainty of the model, which is especially useful when making predictions about novel individuals (new data samples). This example underlines the importance of communicating the uncertainty associated with a prediction. Adopting a model that can express uncertainty about its predictions can help medical professionals make custom treatment plans with an additional level of risk assessment. This, in turn, can justify the level of trust and confidence that physicians and patients have in the prediction. Non-probabilistic models cannot offer these advantages.

V. Conclusion

In this work, we have compared three approaches for training probabilistic survival models for the first time in terms of prediction and calibration performance. The main finding of our work is that, evidently, MCD and SNGP constitute two attractive alternatives to VI for this task, as they rival VI's performance, but without the computational overhead. In many cases, our MCD model improved the mean absolute error over continuous state-of-the-art baselines while giving C-calibration scores similar to VI, and our SNGP model improved D-calibration over VI in all datasets. Limitations of our work include a lack of high-dimensional datasets (e.g., imaging or video data), where efficient computation is even more critical than in tabular data. Moreover, it is difficult for our models to simultaneously optimize prediction and calibration performance, as the results show. Future work should investigate how to improve the proposed model's predictions without degrading their calibration.

Conflict of Interests

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Appendix I
Performance Metrics

$\mathbf{CI_{td}}$: The time-dependent Concordance Index [26] is the probability that $\{i, j\}$ are concordant given they are comparable, where concordance is defined as such: given a comparable pair $\{i, j\}$, with event times $y_i < y_j$ and $\delta_i = 1$, the predicted survival probability for j is greater than the predicted survival probability for i, at the time where i experienced the event, and j is still free from the event. This can be written as:

$$\pi_{\text{comp}} = \Pr\big(y_i < y_j \;\&\; \delta_i = 1\big), \tag{15}$$

$$\pi_{\text{conc}} = \Pr\big(S(y_i|\boldsymbol{X}_i(t)) < S(y_i|\boldsymbol{X}_j(t)) \;\&\; y_i < y_j; \delta_i = 1\big), \tag{16}$$

$$CI_{\text{td}} = \frac{\pi_{\text{conc}}}{\pi_{\text{comp}}} = \Pr\big(S(y_i|\boldsymbol{X}_i(t)) < S(y_i|\boldsymbol{X}_j(t)) \;\&\; y_i < y_j; \delta_i = 1\big), \tag{17}$$

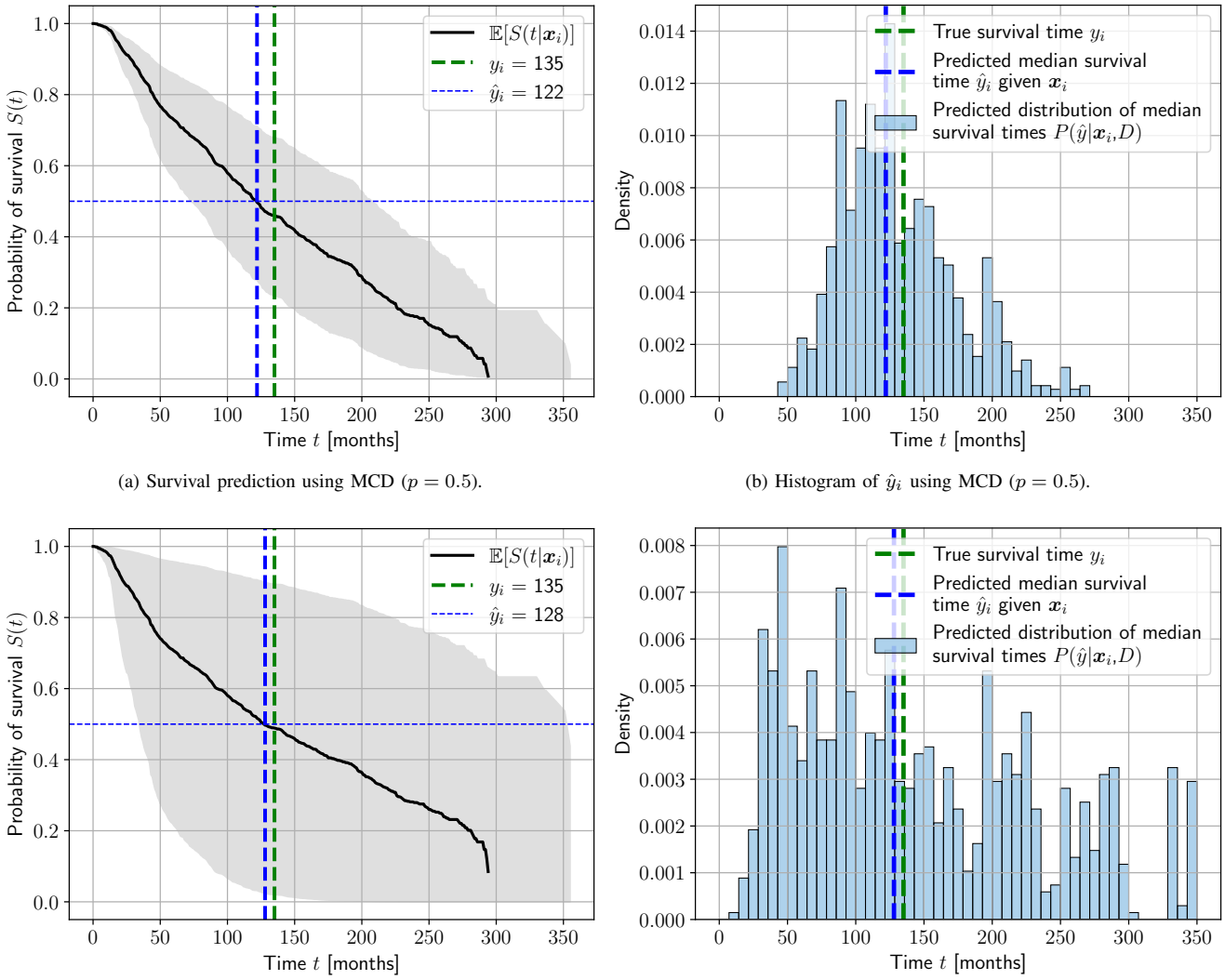
where π_{comp} is the probability that $\{i, j\}$ are comparable, and π_{conc} is the probability that $\{i, j\}$ are concordant.

IBS: The Integrated Brier Score [28] is the expectation of the single-time Brier Score (BS) over some interval $[t_k, t_K]$, where t_k denotes some measurement time on some discretized event horizon of K bins, i.e., $[t_0, t_1, t_2, \ldots, t_K]$. The BS is the mean squared error between the observed binary outcome $y_i(t_k)$ and the predicted probability of the outcome $\hat{y}_i(t_k)$ at t_k. If we consider the entire event horizon $[t_0, t_K]$, then the IBS is defined as [28]:

$$IBS = \frac{1}{K}\sum_{t=0}^{K}\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i(t_k) - y_i(t_k))^2. \tag{18}$$

MAE: The mean absolute error is the absolute difference between the predicted and actual survival times. We use the median of the survival function as the predicted survival time. Given an individual survival distribution, $S(t \mid \boldsymbol{x}_i) = \Pr(T > t \mid \boldsymbol{x}_i)$, we calculate the predicted survival time $\hat{y}_i$ as the median survival time [27]:

$$\hat{y}_i = \text{median}\,(S(t \mid \boldsymbol{x}_i)) = S^{-1}(0.5 \mid \boldsymbol{x}_i). \tag{19}$$



(a) Survival prediction using MCD ($p = 0.5$).

(b) Histogram of $\hat{y}_i$ using MCD ($p = 0.5$).

(c) Survival prediction using VI.

(d) Histogram of $\hat{y}_i$ using VI.

Fig. 2: Model prediction on the METABRIC dataset using MCD and VI. For an individual i, we use the median of the predicted survival function as the predicted survival time $\hat{y}_i$. Although MCD and VI predict similar median survival times, VI has more variance around this particular prediction than MCD. Left column: mean individual survival function and its corresponding 90% CrI. Right column: histogram (as an approximation of the underlying density) of the predicted survival time.

In this work, we report MAE_{H} and MAE_{PO}. The MAE_{H} [33] uses hinge loss to calculate the mean absolute error for censored individuals: if the predicted survival time is smaller than the censoring time, the loss is the censoring time minus the predicted time. If the predicted survival time is equal to or greater than the censoring time, the loss is zero. MAE_{PO} [33] employs pseudo-observation to estimate the actual time of survival for censored individuals.

ICI: The Integrated Calibration Index [29] is the absolute difference on average between the observed and predicted probabilities of the event. Let $f(x) = |x - x_c|$ denote the absolute difference between the calibration curve and the diagonal line of perfect calibration, where x is a predicted probability and x_c is the value of the calibration curve at x. Let $\phi(x)$ denote the density function of the predicted probabilities distribution. The ICI is then a weighted average of the absolute differences between the calibration curve and the diagonal line of perfect calibration, i.e., $ICI = \int_0^1 f(x)\phi(x)\,dx$.

D-calibration: Distribution calibration [30] measures the calibration performance of $S(t)$, expressing to what extent the predicted probabilities can be trusted. For any probability interval $[a, b] \in [0, 1]$, we define $D_m(a, b)$ as the subset of individuals in the dataset D whose predicted probability of event is in the interval $[a, b]$ [33]. A model is D-calibrated if the proportion of individuals $|D_m(a, b)|/|D|$ is statistically similar to the proportion $b-a$, which we assess using a χ^2 test. For example, compare the time-of-death empirical quantiles against the computed ones: e.g., with a D-calibrated curve, approximately 50% of these individuals should die prior to their common predicted median survival time.

C-calibration: Coverage calibration [7] is a statistical test to measure the agreement between the upper and lower bounds of the predicted Credible Intervals (CrIs) and the observed probability intervals. For example, a 90% CrI should cover an individual's likelihood of an event with a probability of 90%. In practice, we calculate the observed coverage rates using CrIs with different percentages (ranging from 10% to 90%) and then use a χ^2 test to evaluate the calibration of expected and observed coverage rates.

APPENDIX II
MODEL HYPERPARAMETERS

Table V reports the selected hyperparameters. All proposed models use the same hyperparameters unless otherwise stated. We use Bayesian optimization [25] to tune hyperparameters over ten iterations on the validation set, adopting the hyperparameters leading to the highest concordance index (CI_{td}).

TABLE V: Selected hyperparameters for the different datasets.

Hyperparameter	METABRIC	SEER	SUPPORT	MIMIC-IV
Optimizer	Adam	Adam	Adam	Adam
Decay	None	None	0.001	0.0001
Activation fn	ReLU	ReLU	ReLU	ReLU
Batch size	32	32	32	128
# Nodes per layer	[64, 128]	[64]	[128]	[32]
Learning rate	0.005	0.001	0.001	0.005
L2 regularization	0.001	0.001	0.001	0.001
Dropout rate	0.25	0.25	0.25	0.5

REFERENCES

[1] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 1050–1059.

[2] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: an algorithmic survey," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 11 773–11 823, 2023.

[3] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–12, 2018.

[4] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3163–3175, 2021.

[5] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller, "Deep Cox Mixtures for survival regression," in *Machine Learning for Healthcare Conference*, 2021, pp. 674–708.

[6] D. Feng and L. Zhao, "BDNNSurv: Bayesian deep neural networks for survival analysis using pseudo values," *Journal of Data Science*, vol. 19, no. 4, pp. 542–554, 2021.

[7] S.-a. Qi, N. Kumar, R. Verma, J.-Y. Xu, G. Shen-Tu, and R. Greiner, "Using Bayesian neural networks to select features and compute credible intervals for personalized survival prediction," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 12, pp. 3389–3400, 2023.

[8] H. Loya, P. Poduval, D. Anand, N. Kumar, and A. Sethi, "Uncertainty estimation in cancer survival prediction," *arXiv preprint arXiv:2003.08573*, 2020, accepted at AI4AH Workshop at ICLR 2020.

[9] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 2348—2356.

[10] C. M. Lillelund, M. Magris, and C. F. Pedersen, "Uncertainty estimation in deep Bayesian survival models," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2023, pp. 1–4.

[11] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 1845–1853.

[12] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*, 2nd ed. Springer, 2021.

[13] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[14] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review*, pp. 45–57, 1975.

[15] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.

[16] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7498–7512.

[17] I. Osband, C. Blundell, A. Pritzel, and B. V. Roy, "Deep exploration via bootstrapped DQN," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 4033–4041.

[18] METABRIC Group, C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[19] L. A. Gloeckler Ries, M. E. Reichman, D. R. Lewis, B. F. Hankey, and B. K. Edwards, "Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) program," *Oncologist*, vol. 8, no. 6, pp. 541–552, 2003.

[20] W. A. Knaus, "The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults," *Annals of Internal Medicine*, vol. 122, no. 3, pp. 191–203, 1995.

[21] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, pp. 1–9, 2023.

[22] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[24] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[25] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 2951—2959.

[26] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statistics in Medicine*, vol. 24, no. 24, pp. 3927–3944, 2005.

[27] S.-a. Qi, W. Sun, and R. Greiner, "SurvivalEVAL: A comprehensive open-source Python package for evaluating individual survival distributions," in *Proceedings of the 2023 AAAI Fall Symposia*, vol. 2, no. 1, 2023, pp. 453–457.

[28] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[29] P. C. Austin and E. W. Steyerberg, "The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models," *Statistics in Medicine*, vol. 38, no. 21, pp. 4051–4065, 2019.

[30] H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective ways to build and evaluate individual survival distributions," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 1–63, 2020.

[31] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2005.

[32] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[33] S.-a. Qi, N. Kumar, M. Farrokh, W. Sun, L. Kuan, R. Ranganath, R. Henao, and R. Greiner, "An effective meaningful way to evaluate survival models," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, 2023, pp. 28 244–28 276.

Appendix F

A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis

Abstract: Amyotrophic lateral sclerosis (ALS) is a degenerative disorder of the motor neurons that causes progressive paralysis in patients. Current treatment options aim to prolong survival and improve quality of life. However, due to the heterogeneity of the disease, it is often difficult to determine the optimal time for potential therapies or medical interventions. In this study, we propose a novel method to predict the time until a patient with ALS experiences significant functional impairment (ALSFRS-R$\leq$2) for each of five common functions: speaking, swallowing, handwriting, walking, and breathing. We formulate this task as a multi-event survival problem and validate our approach in the PRO-ACT dataset ($N = 3{,}220$) by training five covariate-based survival models to estimate the probability of each event over the 500 days following the baseline visit. We then predict five event-specific individual survival distributions (ISDs) for a patient, each providing an interpretable estimate of when that event is likely to occur. The results show that covariate-based models are superior to the Kaplan-Meier estimator at predicting time-to-event outcomes in the PRO-ACT dataset. Additionally, our method enables practitioners to make individual counterfactual predictions – where certain covariates can be changed – to estimate their effect on the predicted outcome. In this regard, we find that Riluzole has little or no impact on predicted functional decline. However, for patients with bulbar-onset ALS, our model predicts significantly shorter time-to-event estimates for loss of speech and swallowing function compared to patients with limb-onset ALS (log-rank $p < 0.001$, Bonferroni-adjusted $\alpha = 0.01$). The proposed method can be applied to current clinical examination data to assess the risk of functional decline and thus allow more personalized treatment planning.

Authors: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT).

Status: Published in PLOS ONE, vol. 20, no. 11, pp. 1–16, 2025.

PLOS One

RESEARCH ARTICLE

A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis

Christian Marius Lillelund[1,2]*, **Sanjay Kalra**[1,3,4], **Russell Greiner**[1,5], **The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT)**¶

1 Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, **2** Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark, **3** Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta, Canada, **4** Division of Neurology, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada, **5** Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada

¶ A complete list of members of the Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT) can be found in the Acknowledgments.
* clillelund@ualberta.ca

Check for updates

OPEN ACCESS

Citation: Lillelund CM, Kalra S, Greiner R (2025) A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis. PLoS One 20(11): e0336476. https://doi.org/10.1371/journal.pone.0336476

Editor: Guanghui Liu, State University of New York at Oswego, UNITED STATES OF AMERICA

Received: December 28, 2024

Accepted: October 27, 2025

Published: November 18, 2025

Copyright: © 2025 Lillelund et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The PRO-ACT dataset can be downloaded from the Neurological Clinical Research Institute's website (https://ncri1.partners.org/ProACT). Users

Abstract

Amyotrophic lateral sclerosis (ALS) is a degenerative disorder of the motor neurons that causes progressive paralysis in patients. Current treatment options aim to prolong survival and improve quality of life. However, due to the heterogeneity of the disease, it is often difficult to determine the optimal time for potential therapies or medical interventions. In this study, we propose a novel method to predict the time until a patient with ALS experiences significant functional impairment (ALSFRS-R ≤ 2) for each of five common functions: speaking, swallowing, handwriting, walking, and breathing. We formulate this task as a multi-event survival problem and validate our approach in the PRO-ACT dataset ($N = 3{,}220$) by training five covariate-based survival models to estimate the probability of each event over the 500 days following the baseline visit. We then predict five event-specific individual survival distributions (ISDs) for a patient, each providing an interpretable estimate of when that event is likely to occur. The results show that covariate-based models are superior to the Kaplan-Meier estimator at predicting time-to-event outcomes in the PRO-ACT dataset. Additionally, our method enables practitioners to make individual counterfactual predictions—where certain covariates can be changed—to estimate their effect on the predicted outcome. In this regard, we find that Riluzole has little or no impact on predicted functional decline. However, for patients with bulbar-onset ALS, our model predicts significantly shorter time-to-event estimates for loss of speech and swallowing function compared to patients with limb-onset ALS (log-rank $p < 0.001$, Bonferroni-adjusted $\alpha = 0.01$). The proposed method can be applied to current clinical examination data to assess the risk of functional decline and thus allow more personalized treatment planning.

must create an account to access the data, but account creation is free of charge.

Funding: Funding received from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Alberta Machine Intelligence Institute (AMII), and the Canadian Institute for Advanced Research (CIFAR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no competing interests.

Introduction

Amyotrophic lateral sclerosis (ALS) is a neurological disease that causes a gradual loss of upper and lower motor neurons in the central nervous system. Unfortunately, most patients will die within 2-5 years after onset [1], primarily due to complications from respiratory failure [2]. The underlying causes of the disease are not yet understood, and as there are no known cures, current treatments aim only to prolong survival and improve quality of life. Moreover, due to its heterogeneous and unpredictable nature, and the high variability in its progression rate and clinical phenotype [5], it is often difficult to determine the appropriate treatment. This also complicates decisions about the optimal timing of medical interventions, e.g., non-invasive ventilation, or if a specific treatment can effectively slow the progression of the disease. There is therefore a significant need for tools that can predict disease progression to facilitate personalized treatment plans and improve patient care.

The ALS Functional Rating Scale-Revised (ALSFRS-R) [6] is the most widely used questionnaire to evaluate the progression of ALS. It contains 12 questions to assess functional capacity in patients, with a focus on bulbar, motor, and respiratory functions—e.g., swallowing, walking, and breathing, respectively. Each response is scored between 0 and 4, where 0 represents complete inability with regard to that function, and 4 represents normal function. As this questionnaire is administered to patients at multiple time points over the course of the disease, it provides a longitudinal view of functional decline across several areas. Based on this questionnaire, an interesting task is to predict the degree of functional decline a patient experiences during the course of the disease, rather than simply predicting when a patient is likely going to die [7,8]. Many previous projects have therefore attempted to predict the future ALSFRS-R score based on patient covariates [3,9–12]. Ong et al. [12] proposed a binary model to predict the functional decline class (i.e., whether ALS progresses slowly or rapidly in a specific patient) after baseline (trial entry), based on several patient characteristics. Their model was able to predict slow or rapid decline, but did not predict when functional decline would occur over time. Similarly, Jabbar et al. [11] proposed a method to predict disease progression as fast or slow, but as in [12], their model could only give a yes/no answer to whether a patient would experience functional decline over time. Amaral et al. [10] applied several clustering algorithms to clinical patient records, which revealed four separate groups regarding disease progression. Their method could assign newly-diagnosed patients to a progression group, but did not predict when a patient would experience functional decline based on their covariates. Finally, Vieira et al. [9] trained two machine learning models to predict a patient's future ALSFRS-R score: a voice model based on speech recordings and a movement model based on accelerometer measurements. These models showed strong predictive performance, but similar to Amaral et al. [10] and others, they predicted future ALSFRS-R scores rather than estimating the time until functional decline would occur. Moreover, disease progression in ALS is neither constant nor linear, as those studies implicitly assumed, but can vary substantially over time and between individuals [4].

Based on these shortcomings, we propose a novel method that can predict when an ALS patient will experience each of five common types of functional decline, based on the ALSFRS-R protocol (see Table 1): significant difficulty (≤ 2) in each of *Speaking*, *Swallowing*, *Handwriting*, *Walking*, and *Dyspnea*. These five events are distinct but not mutually exclusive and likely share some information in their covariates; it may be that a patient (e.g., Mr. Smith), with limb-onset ALS, will experience difficulty walking and writing earlier than another patient (e.g., Mr. Johnson), who has bulbar-onset ALS. On the other hand, Mr. Johnson may have difficulty speaking and swallowing solid foods before Mr. Smith. As a newly-diagnosed patient with ALS, our approach will help answer important questions, such as:

- How much longer will I be able to speak clearly enough to be understood?
- How much longer will I be able to eat most solid foods without difficulty?
- How much longer will I be able to write a shopping list that my wife can read?
- How much longer will I be able to walk at a normal pace and keep my balance?
- How much longer will I be able to breathe normally during daily activities?

In practice, we formulate the problem as a multi-event survival problem: given an instance (e.g., a description of Mr. Smith at a ”baseline” point in time), our method predicts the time until said Mr. Smith will experience each of the events listed above. We present each prediction as an individual survival distribution (ISD) [13], which provide the probability of each event at each future time point, after the baseline visit. We describe each patient by the values of a set of covariates such as age, site of onset, forced vital capacity (FVC), and others, which are recorded in clinical records obtained at or before the baseline visit. To our knowledge, no existing method can explicitly predict the timing of distinct individual events in ALS. Fig 1 shows an outline of the proposed method. The source code is publicly available at: https://github.com/thecml/FunctionalALS.

Materials and methods

Clinical data and event annotation

For our empirical analyses, we use the publicly-available Pooled Resources Open-Access Clinical Trials (PRO-ACT) (Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials

Table 1. The definition of the five included functional assessments from the ALS Functional Rating Scale-Revised (ALSFRS-R) [6] and their scores. Each score ranges from 0 to 4, where 0 represents complete inability with regard to the function, and 4 represents normal function. For each patient, our method predicts the time until that patient will experience significant difficulty (ALSFRS-R ≤ 2) for that specific assessment.

Domain	Assessment	Description	Score
Bulbar	Speech	Ability to speak clearly and be understood.	4 = Speaks clearly; 2 = Intelligible with repeating; 0 = Loss of useful speech
Bulbar	Swallowing	Ability to swallow without choking.	4 = Normal eating habits; 2 = Difficulty with some foods; 0 = Needs feeding tube
Fine motor	Handwriting	Ability to write or type.	4 = Normal writing; 2 = Not all words legible; 0 = Unable to grip pen
Gross motor	Walking	Ability to walk without assistance.	4 = Walks normally; 2 = Walks with assistance; 0 = Cannot walk
Respiratory	Dyspnea	Shortness of breath.	4 = None; 2 = Moderate shortness of breath; 0 = Severe shortness of breath

https://doi.org/10.1371/journal.pone.0336476.t001

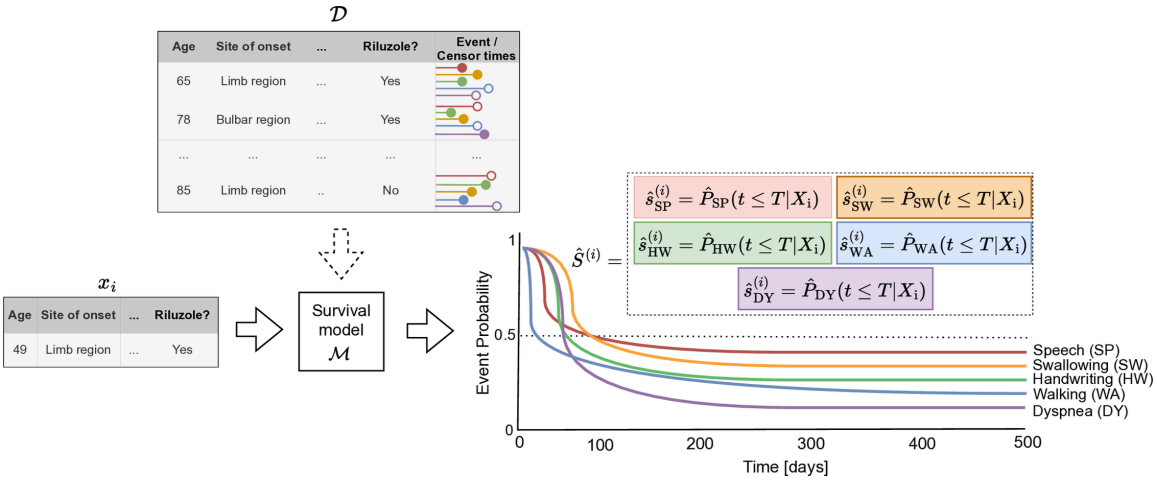


Fig 1. Outline of the proposed method. Historical data about ALS patients form a patient dataset $\mathcal{D}$ with *N* training instances and *d* covariates, and the time to event (filled dot) or censoring (hollow dot) from a patient's first visit—the "baseline time". Censoring indicates that the event of interest was not observed for a patient within the study period, so their exact event time remains unknown. We consider five separate but related events, i.e., *Speech*, *Swallowing*, *Handwriting*, *Walking*, and *Dyspnea*. We use the recorded covariates (taken at the baseline time) and event information to train a survival model $\mathcal{M}$ that can accurately estimate the individual survival distribution (ISD) of each of these five events, for a novel patient $\boldsymbol{x}_i$, denoted as $\hat{S}^{(i)}$. These ISDs give the probability of each of these five events occurring after *t* days after the baseline visit, for all *t*>0. They can also be used to estimate the time to event for this $\boldsymbol{x}_i$ patient, for example, when the survival curve intersects the dashed horizontal line at 50%, which is called the median survival time, for each of the five events.

https://doi.org/10.1371/journal.pone.0336476.g001

(PRO-ACT) Database. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members. The dataset was obtained on October 1st, 2024.) [14] dataset, which is the largest ALS dataset in the world. It includes patient demographics, lab and medical records, as well as family histories, of over 11,600 ALS patients from 23 clinical trials (see Table 2). In each clinical trial, patients follow a schedule of visits until the end of the study or until they become too ill to participate. Covariates recorded at each visit include the total ALSFRS-R score, age, sex, region of onset (e.g., limb, bulbar), the mean FVC score, time since diagnosis, disease progression rate (Defined as (48 - the total ALSFRS-R score)/(days since diagnosis/30).), and whether or not the patient is taking Riluzole. In this study, we only use covariates recorded at the baseline visit for each individual patient.

To define our events of interest, each patient's disease state is assessed by a clinician at every visit using the ALSFRS-R scale. We focus on five specific events: *Speech*, *Swallowing*, *Handwriting*, *Walking*, and *Dyspnea*. To

Table 2. Patient demographics in the PRO-ACT [14] dataset.

Demographic/Data	**PRO-ACT (N = 3,220)** Pct. or mean (± SD)
Age (years)	55.8 (11.6)
Height (cm)	172 (9.5)
Weight (kg)	76.3 (14.6)
Body mass index (BMI)	25.7 (4)
% Female	32.2
% Caucasian	94.6
Site of Onset; % Limb, % Bulbar	43.3, 13.4
Baseline ALSFRS-R Score	39.4 (5)
Time in study (days)	271.4 (122.7)
% Riluzole	75.7

https://doi.org/10.1371/journal.pone.0336476.t002

Table 3. Overview of the dataset. "SP", "SW", "HA", "WA" and "DY", are the *Speech*, *Swallowing*, *Handwriting*, *Walking* and *Dyspnea* events, respectively. Event distribution indicates the percentage of uncensored instances of each type in the dataset. *N*: sample size, *d*: number of covariates, *K*: number of event types, *T*: observed time (of the uncensored instances) for any event in days.

Dataset	N	d	K	T_{min}	T_{max}	T_{mean}	Event distribution (%):
PRO-ACT	3,220	8	5	1	498	130	SP: 37.8, SW: 31.9 HA: 50.3, WA: 60.6 DY: 27.0

https://doi.org/10.1371/journal.pone.0336476.t003

annotate our dataset, we define an event as having occurred the first time a patient scores 2 or fewer in that specific assessment during any follow-up visit following their baseline visit. These five events can occur in any order and are not mutually exclusive. There are several types of censoring, including death, being unable to perform the assessment or leaving the study prematurely. Censoring thus represents incomplete event information, where the true event time is only known to exceed the observed follow-up time. We use a maximum follow-up time of 500 days to focus on the early stages of the disease and because data become sparse with high censoring afterward. We consider patients who score 2 or fewer on any of the five assessments at baseline as having already had the event, indicating that the loss of function occurred before the start of the trial, and exclude these patients from the study. We also exclude patients with no recorded ALSFRS-R history. After applying these criteria and accounting for censoring, the final dataset comprises 3,220 patients with observed or right-censored event times. The code for loading and annotating the data is available in the source code repository. Table 3 provides an overview of the dataset. Fig 2 shows the distribution of event and censoring times. A list of covariates is available in the Supplement (S1 Table).

Survival analysis and notation

Survival analysis models the time until some event of interest occurs. This event can either be observed or censored—the latter typically happens if the patient is still alive at the end of the study or decides to drop out [15, Ch. 11]. Survival analysis has become an important tool for understanding diseases and predicting outcomes, which helps medical researchers evaluate the significance of prognostic factors in applications like comatose after cardiac arrest [16], liver transplantation [17] or fall risk [18]. In ALS, survival analysis has been used mainly to predict risk scores with respect to death [4,7,19].

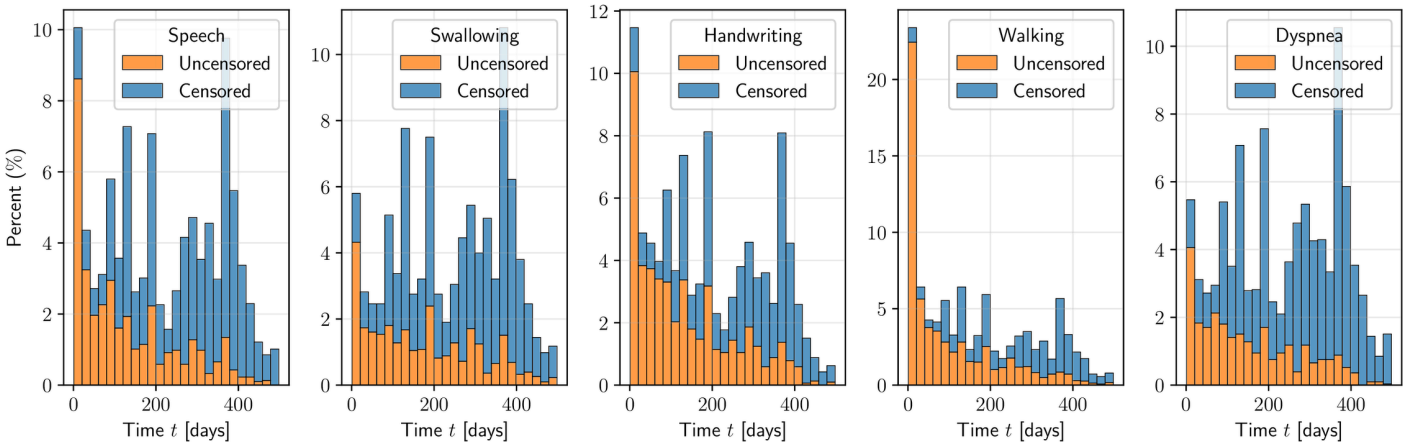


Fig 2. Distribution of uncensored and censored times in the PRO-ACT dataset for the five events.

https://doi.org/10.1371/journal.pone.0336476.g002

We briefly introduce the notation: Let $\mathcal{D} = \left\{(\boldsymbol{x}^{(i)}, \boldsymbol{t}^{(i)}, \delta^{(i)})\right\}_{i=1}^{N}$ be the dataset, where for each i, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ is a d-dimensional vector of covariates. We consider discrete times, where $\boldsymbol{t}^{(i)} \in \{1, 2, \ldots, T_{\max}\}^K$ is a vector of observed times for the i-th patient for K different events, each at one of these $T_{\max}$ possible times, and $\delta^{(i)} \in \{0, 1\}^K$ is a vector of event indicators for each event. Moreover, let $e_k^{(i)} \in \{1, 2, \ldots, T_{\max}\}$ denote the event time and let $c_k^{(i)} \in \{1, 2, \ldots, T_{\max}\}$ denote the censoring time for the i-th instance for event k over the event horizon, thus we have $t_k^{(i)} = e_k^{(i)}$ if $\delta_k^{(i)} = 1$, otherwise, $t_k^{(i)} = c_k^{(i)}$ if $\delta_k^{(i)} = 0$. A survival model predicts the probability that some event k occurs at time T later than t, i.e., the so-called survival or event probability, which is denoted $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$. In multi-event survival analysis, we want to predict the probability of different events occurring over time. For each patient i, this results in a matrix of survival functions, $\boldsymbol{S}^{(i)} = \{\boldsymbol{s}_1^{(i)}, \boldsymbol{s}_2^{(i)}, \ldots, \boldsymbol{s}_K^{(i)}\}$. Here, $\boldsymbol{s}_k^{(i)}$ represents the probabilities that patient i does not experience event k up to each time point t, i.e., $\boldsymbol{s}_k^{(i)} = (P(e_k^{(i)} > 1), P(e_k^{(i)} > 2), \ldots, P(e_k^{(i)} > T_{\max}))$. See Fig 1 for a visual overview.

Experiments and results

Setup and preprocessing

We follow Sechidis et al. [20] and first split the dataset into a training, validation and test set by 70%, 10%, and 20% using a stratified procedure. This ensures that the event times and censoring rates are consistent across the three sets and the K events. The stratification procedure uses a random seed (0-9). After splitting the data, we impute missing values using the sample mean for real-valued covariates or the mode for categorical covariates based on the training set only. A list of missing values is available in the Supplement (S2 Table). After imputing missing values, we encode categorical covariates using a one-hot encoding strategy and apply a z-score data normalization to speed up the training process and improve stability. No outlier detection was performed.

Our evaluation procedure consists of training five survival models on the training set and evaluating them on the holdout test set. For evaluation fairness, we do not perform hyperparameter optimization for specific models, but instead configure all models with sensible default parameters. A list of hyperparameters is available in the Supplement (S3 Table). For deep learning models, we use a single hidden layer with 32 nodes. For methods that support it, we use early stopping during training to terminate the process if the validation loss does not improve for 10 consecutive epochs. All data preprocessing and experiments were implemented in Python 3.9. A list of software packages used is available in the source code repository.

Survival models

To evaluate both single-event and multi-event survival modeling approaches, we distinguish between models that predict the time to a single event and those that jointly model multiple event types. As single-event models, we implement the traditional Cox Proportional Hazards (CoxPH) model using a linear estimator [21], Random Survival Forests (RSF) [22], DeepSurv [23] and MTLR [24]. As a multi-event model, we implement the Multi-Event Network for Survival Analysis (MENSA) model [25]. To explain these models: **CoxPH** is a popular semiparametric method to fit a regression model to survival data. By adopting a multiplicative form for the contribution of several covariates to each individual's event time, the CoxPH model is a powerful yet simple tool for assessing the simultaneous effect that different covariates have on the event times. CoxPH assumes proportional hazards, which means that the effect of covariates on the hazard function is constant over time. **RSF** extends decision trees for survival analysis and recursively splits training data based on some survival-specific criterion, e.g., log-rank splitting, with the goal of maximizing separation between event times between nodes. RSF can model nonlinear relationships between the covariates and the event, and does not assume proportional hazards. **DeepSurv** is a multilayer perceptron (MLP) based on the CoxPH model, where the risk score is a nonlinear function of the covariates. DeepSurv assumes proportional hazards. **MTLR** is a discrete survival model that estimates the individual survival function as a multinomial logistic regression model and does not assume proportional hazards. Lastly,

MENSA is an MLP that learns the individual survival distribution for K events as a mixture of Weibull distributions with a shared covariate layer to model dependencies between events. MENSA does not assume proportional hazards.

Evaluation metrics

To comprehensively assess the survival models, we select a range of evaluation metrics that capture three key aspects of model performance: discrimination, absolute/squared error, and calibration. Discrimination, usually measured by the concordance index (CI), is relevant when we want to identify which patients are likely to experience functional decline earlier versus later at the group level. Squared or absolute error metrics, like the mean absolute error (MAE), are used to assess how accurately the model predicts the time of functional impairment for individual patients. Finally, calibration measures how well the predicted probabilities match the observed frequencies of events at specific time points. For example, if a model predicts that a patient has an 80% chance of not experiencing significant functional decline before 100 days, then among 100 similar patients, about 80 should not experience significant functional decline by that time. We use the following metrics to assess discrimination, absolute/squared error, and calibration:

CI: Harrell's CI [26] measures discrimination performance by calculating the proportion of concordant pairs among all comparable pairs given predicted risk scores. A pair is considered comparable if the event order can be determined. Specifically, for a comparable pair $\{i,j\}$ with event times $e_i < e_j$ and $\delta_i = 1$ (indicating that i experienced the event), we calculate risk scores $\hat{\eta}_i$ and $\hat{\eta}_j$ as the negative median survival times. The median survival time is the time point at which the estimated survival probability drops to 0.5. If the predicted risk score for j is greater than that for i at the time i experienced the event (while j remains event-free), the pair is concordant. A CI of 0.5 indicates chance-level ranking. We also report Uno's CI [27], which is a CI that uses a technique to correct for the bias introduced by censoring.

BS/IBS: The Brier Score (BS) calculates the squared error between the predicted probability of the event and the Heaviside step function of the observed event. The integrated Brier score [28] (IBS) aggregates Brier scores over multiple time points to provide a single measure of the squared error.

MAE: The MAE calculates the average absolute difference between the true event time and the predicted event time. Given an ISD, $S(t \mid \boldsymbol{x}_i) = \Pr(T > t \mid \boldsymbol{x}_i)$, we calculate the predicted event time $\hat{t}_i$ as the median survival time [29]. Since our dataset contains censoring, we calculate the margin MAE (mMAE) as proposed by [13], which assigns a "best-guess" estimate to each censored patients using the Kaplan-Meier (KM) [30] estimator.

D-calibration: Distribution calibration [13], also known as D-Cal, measures how well the predicted survival function, $\hat{S}(t)$, is calibrated for each event. D-calibration assesses this using a Pearson's chi-squared test with $\alpha = 0.05$. For any probability interval $[a, b] \in [0, 1]$, we define $D_m(a, b)$ as the group of patients in the dataset D whose predicted probability of an event is in the interval $[a,b]$ [31]. A model is D-calibrated if the proportion of patients $|D_m(a, b)|/|D|$ is statistically similar to the amount $b–a$. Intuitively, if a model is well-calibrated, the predicted probabilities match observed event frequencies—for instance, about 20% of patients predicted to have a 0.2 event probability should actually experience the event. In our experiments, we report the number of times each model is D-calibrated (at p>0.05) for each event across 10 experiments.

Model results

We present our empirical results (Table 4 and Fig 3) in terms of the aforementioned aspects with respect to model performance: discrimination, absolute/squared error, and calibration. Below, we discuss the results and reflect on their implications.

Discrimination: Good discrimination performance is important if we need to rank patients by their disease severity to prioritize access to limited clinical resources, for example, portable ventilators. Across all models, we see the highest Harrell's and Uno's C-indices for the *Speech* (approx. 74–75% and 70–71%, respectively) and *Swallowing* (approx. 75–76% and 70–71%, respectively) events. On the other hand, functional decline with respect to limb functions is harder

Table 4. Mean (± SD.) prediction performance, averaged over 10 independent experiments. D-calibration counts the number of times the respective model was D-calibrated. Harrell's CI, Uno's CI and IBS results are multiplied by 100.

Model	Event	Harrell's CI ↑	Uno's CI ↑	IBS ↓	mMAE ↓	D-Cal ↑
CoxPH [21]	Speech	74.52±1.39	70.44±3.34	14.41±0.43	206.85±9.88	(10/10)
	Swallowing	75.75±1.53	70.96±4.58	12.58±0.72	146.15±25.50	(10/10)
	Handwriting	67.57±0.92	63.75±2.29	16.73±0.70	141.85±7.40	(10/10)
	Walking	68.67±1.20	67.15±1.40	16.60±0.50	145.04±4.50	(7/10)
	Dyspnea	68.54±1.71	66.99±6.30	13.79±0.41	235.46±27.57	(10/10)
RSF [22]	Speech	75.05±1.13	70.08±2.75	14.69±0.44	164.51±19.12	(10/10)
	Swallowing	75.53±1.46	69.65±2.98	13.02±0.74	111.71±11.46	(10/10)
	Handwriting	67.38±1.39	63.70±2.05	16.87±0.57	146.97±6.88	(10/10)
	Walking	68.44±1.47	66.68±1.54	16.78±0.56	145.98±4.63	(10/10)
	Dyspnea	78.23±1.07	78.95±4.01	11.89±0.33	850.45±471.39	(10/10)
DeepSurv [23]	Speech	74.18±3.05	70.77±3.57	14.36±0.89	232.15±28.48	(10/10)
	Swallowing	75.61±1.37	70.67±5.52	12.48±0.65	178.77±39.68	(10/10)
	Handwriting	67.12±1.08	63.72±2.15	16.81±0.84	151.83±18.59	(9/10)
	Walking	69.53±0.93	68.44±1.11	16.23±0.51	148.17±6.60	(8/10)
	Dyspnea	71.72±3.99	70.61±7.60	13.11±0.61	384.96±284.17	(10/10)
MTLR [24]	Speech	74.22±1.67	70.17±3.23	13.95±0.44	179.35±5.34	(10/10)
	Swallowing	75.13±1.43	70.63±5.41	12.34±0.59	122.93±19.52	(10/10)
	Handwriting	67.32±1.11	63.71±2.42	16.72±0.66	140.94±6.29	(9/10)
	Walking	68.97±1.16	67.71±1.06	16.30±0.51	149.93±10.99	(9/10)
	Dyspnea	66.54±2.22	65.90±6.85	13.50±0.50	203.36±24.57	(10/10)
MENSA [25]	Speech	74.21±1.11	70.56±3.47	14.53±0.41	181.81±7.92	(9/10)
	Swallowing	74.94±1.36	70.85±5.20	12.61±0.56	119.16±7.12	(10/10)
	Handwriting	66.46±0.90	63.80±2.65	17.10±0.66	141.97±5.70	(9/10)
	Walking	68.63±1.29	66.94±1.50	17.04±0.61	149.50±6.49	(6/10)
	Dyspnea	66.56±1.77	64.40±7.07	13.78±0.30	189.43±24.28	(10/10)

https://doi.org/10.1371/journal.pone.0336476.t004

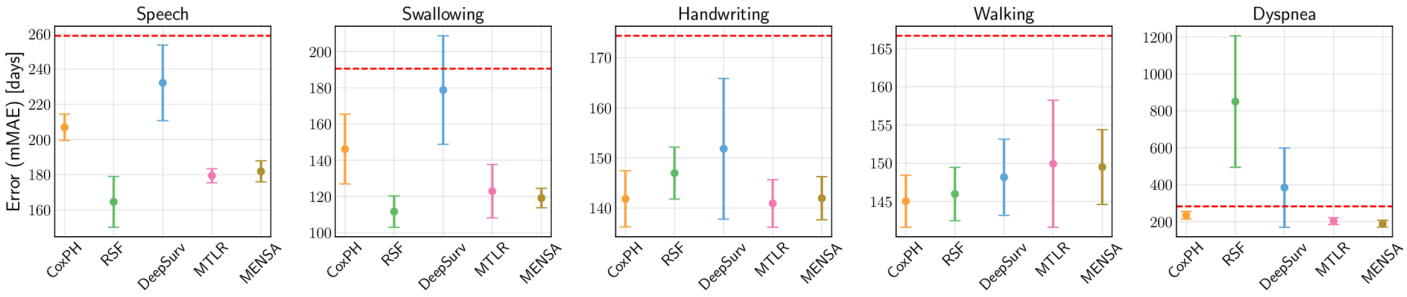


Fig 3. The mMAE (in days) as a function of covariate-based models (x-axis) in the PRO-ACT test set, with error bars representing empirical 95% confidence intervals. The horizontal dashed line is the KM estimator. Lower is better.

https://doi.org/10.1371/journal.pone.0336476.g003

to discriminate, with lower C-indices (approx. 67–69%) for the *Handwriting* and *Walking* events. Although the *Dyspnea* event typically has moderate CI values (approx. 66–68%), RSF achieves a Harrell's CI of 78.23% and a Uno's CI of 78.95% for this event, making it the best model in this case. DeepSurv and MTLR also demonstrate consistently good discrimination performance across events, and MENSA matches or slightly exceeds these for the *Speech* and *Swallowing* events, though its performance dips slightly for the *Walking* and *Dyspnea* events.

Absolute/squared error: Accurate time-to-event estimates or survival probabilities are important when a decision or intervention depends on the precise timing of the event. Across all models, the *Swallowing* event consistently exhibits some of the lowest IBS (approx. 12-13) and mMAE (approx. 110-150) numbers, meaning that this event is the easiest

to predict the timing of. For the *Dyspnea* event, the IBS is also moderate (11.9-13.8), but the RSF model is notable with a low IBS but a relatively high mMAE (approx. 850±471 days). In contrast, other models produce much more moderate mMAE numbers for this event (approx. 200-400 days), as seen with the MENSA model. This highlights an important point: simply calculating the squared error of the predicted survival probabilities does not fully capture a model's ability to predict the timing of an event. MENSA achieves reasonable mMAE numbers (approx. 120-190 days) in all cases, except for the *Speech* event (181±8 days), however, and tends to perform particularly well for the *Swallowing* event, similar to MTLR.

Fig 3 shows the prediction error in days between the population-level KM estimator [30] and the covariate-based models. The KM estimate is calculated per event on observed and censored event times, without considering individual differences (i.e., covariates). In all five events, the covariate-based models generally give superior time-to-event estimates compared to the KM method, with some variation between models. This suggests that incorporating patient-specific information can lead to more accurate time-to-event predictions when predicting functional decline, rather than simply relying on a population-level estimate.

Calibration: Good calibration performance is important when we have to use the actual predicted probabilities to guide our decision-making. For example, underestimating the risk of respiratory decline or bulbar involvement may result in mistimed interventions or suboptimal treatment choices. Across all models, the *Walking* event consistently poses a calibration challenge, as predicted event probabilities generally deviate from the observed event frequencies here (i.e., the models are sometimes not calibrated). However, RSF achieves perfect D-calibration (10/10) across all events and experiments, demonstrating robust and consistent calibration performance. DeepSurv and MTLR also perform well, with most experiments achieving D-calibration. MENSA achieves perfect calibration for the *Swallowing* and *Dyspnea* (10/10) events, but shows slightly poorer calibration for the *Speech* (9/10) and *Handwriting* (9/10) events, and especially the *Walking* event (only 6/10).

Individual disease predictions

Our method can predict a patient's disease trajectory as K individual survival distributions, where K is the number of events (here, $K = 5$). Fig 4 shows such ISDs, from baseline ($t = 0$) to 500 days later, for the i-th patient. As an example, let us return to our patient, Mr. Smith. He is a 72-year-old male, who had been diagnosed with limb-onset ALS prior to study entry and had experienced difficulty speaking and weakness in his legs, but all five of his scores were above 2, so that he could participate in the study. Mr. Smith's covariates are available in the Supplement (S4 Table). The neurologist predicted that talking and walking difficulties would be among Mr. Smith's earliest functional impairments, and our method

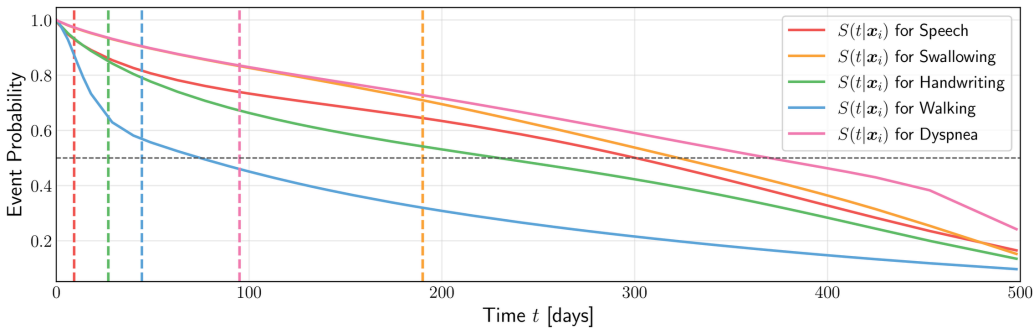


Fig 4. Predicted ISDs for Mr. Smith (*i*). The point where the survival curve intersects the dashed horizontal line at 50% indicates the predicted time of event—so this predicts, for example, *Walking* at approximately 80 days, *Handwriting* at 240 days, and *Speech* at 300 days. The dashed vertical lines are Mr. Smith's actual time of event for the respective events—so, for example, *Speech* at approximately 7 days, *Handwriting* at 35 days, and *Walking* at 45 days.

https://doi.org/10.1371/journal.pone.0336476.g004

confirmed her suspicion. From the outset, Mr. Smith was provided with a walking aid and a personalized long-term care plan that included therapy, support systems, and home modifications. Shortly after his hospital visit, he had to repeat many of his words to be understood and relied on assistance from a cane or his wife to walk. Meanwhile, handwriting also became challenging for him, but he did not experience shortness of breath or difficulty swallowing whole foods until much later.

Individual counterfactual predictions

Our method can also make individual counterfactual predictions [32], that is, alternative scenarios in which certain patient covariates are changed to see their (predicted) effect on the probability of the event. Such predictions can give novel insights into disease mechanisms and assess treatment options on an individual level. In the following figures, we have used MENSA [25] to make individual counterfactual predictions based on our PRO-ACT dataset. These provide insights into the relationship between common disease predictors and functional decline, and ask how changing certain covariates affects disease outcomes. Again, we use our patient, Mr. Smith, as the instance *i*. Additionally, we separate all patients in the test set into groups by the same covariate we used for the individual patient and employ the log-rank test [33] to measure the statistical difference between the groups' event times. We use a Bonferroni correction to correct for multiple comparisons by dividing the significance level ($\alpha = 0.05$) by the number of comparisons ($n = 5$), i.e., the significance level is then $\alpha_{adjusted} = 0.05/5 = 0.01$. For the remainder of this section, we will discuss our counterfactual predictions.

As a starting point, many newly diagnosed patients with ALS need to decide whether to take certain pharmaceutical drugs. Currently, there are three approved drugs by the US Food and Drug Administration (FDA) to treat ALS, the most common being Riluzole. Riluzole works by blocking the release of glutamate, which delays the onset of ventilator dependence or tracheostomy in some people and may increase overall survival by two to three months [34]. However, the drug can cause significant side effects, including nausea, anorexia, and diarrhea [4]. Therefore, it is important to determine whether the drug will be effective—i.e., significantly delay functional decline in this regard—for each individual, in order to decide whether that person should receive the treatment. Upon diagnosis, Mr. Smith decided not to take Riluzole, but what if he instead had decided to take it. Fig 5 shows the predicted ISDs for each of the five events using Mr. Smith's covariates ($\boldsymbol{x}_i$), with the Riluzole covariate being yes (resp., no) at the baseline visit. The two survival curves have nearly the same slopes, with Riluzole only being marginally above non-Riluzole, and nearly identical estimated event times. There are no statistical differences in event times between the groups either (p>0.01). As an example, for the *Speech* event, this model predicts that taking Riluzole would delay the onset of speaking difficulties by approximately 20 days.

In ALS, the site of onset refers to the specific part of the body where the initial symptoms began. ALS can affect different regions, and the site of onset can influence the progression of the disease. The two most common sites of onset are bulbar-onset and limb-onset, the former being associated with faster disease progression in general [11]. Mr. Smith

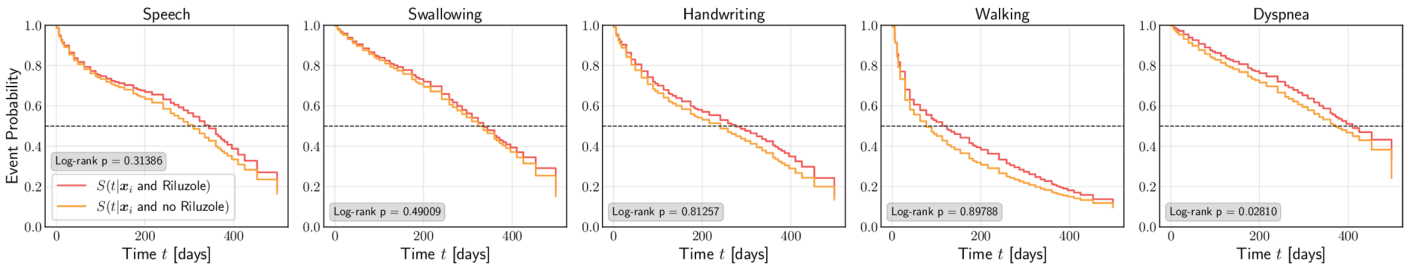


Fig 5. Predicted ISDs for Mr. Smith based on his use of Riluzole.

https://doi.org/10.1371/journal.pone.0336476.g005

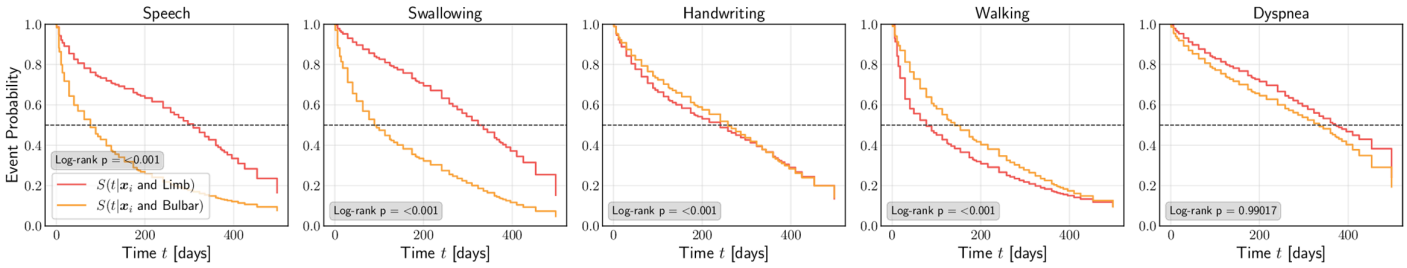


Fig 6. Predicted ISDs for Mr. Smith based on whether he has limb-onset or bulbar-onset.

https://doi.org/10.1371/journal.pone.0336476.g006

was diagnosed with limb-onset ALS, but what if instead he had been diagnosed with bulbar-onset ALS. Fig 6 shows the predicted ISDs for Mr. Smith based on whether he has limb-onset or bulbar-onset. We see significant differences in the predictions for the *Speech* and *Swallowing* events, indicating a worse prognosis for Mr. Smith if he has bulbar-onset, but more agreement between the curves for the *Handwriting*, *Walking* and *Dyspnea* events.

Next, forced Vital Capacity (FVC) is a measure of lung function and represents the total amount of air a person can forcibly exhale after taking a deep breath. It is commonly used to assess lung respiratory function and is measured in liters. In ALS, FVC is a key clinical assessment for evaluating respiratory muscle strength, which progressively weakens as the disease progresses. Mr. Smith's total FVC score at baseline was 2.69. Again, we ask the counterfactual question of what if instead he had a higher FVC or a lower FVC. Fig 7 shows the predicted ISDs for Mr. Smith based on whether he has a high or low FVC. As expected, the *Handwriting* event shows only a weak relationship with FVC, as evidenced by the high *p*-value and the alignment of the two survival curves. In contrast, the *Swallowing* event exhibits a significant relationship with FVC, and the model provides substantially different time-to-event predictions based solely on this covariate.

Lastly, we consider the total ALSFRS-R score at baseline and its effect on predicted functional decline. This score provides a snapshot of Mr. Smith's initial physical capabilities when he is relatively healthy before the disease has advanced notably. Mr. Smith's total ALSFRS-R score at baseline was 37. Again, we make a counterfactual prediction if instead Mr. Smith had a significantly higher or lower ALSFRS-R score. Fig 8 shows the predicted ISDs for Mr. Smith based on whether he has a high or low ALSFRS-R score. Consistent with the literature [4], if Mr. Smith had a high ALSFRS-R score to begin with, he would (presumably) experience functional decline much later than if he had a low ALSFRS-R score. This is indicated by the discrepancy between the survival curves and the low *p*-values across all the events. Obviously, initial physical ability has a strong relationship with loss of function after a certain time, and patients may be at different stages of the disease and have different levels of strength when they enter the trial. In addition, research in ALS has suggested the presence of various disease phenotypes with different progression rates [35].

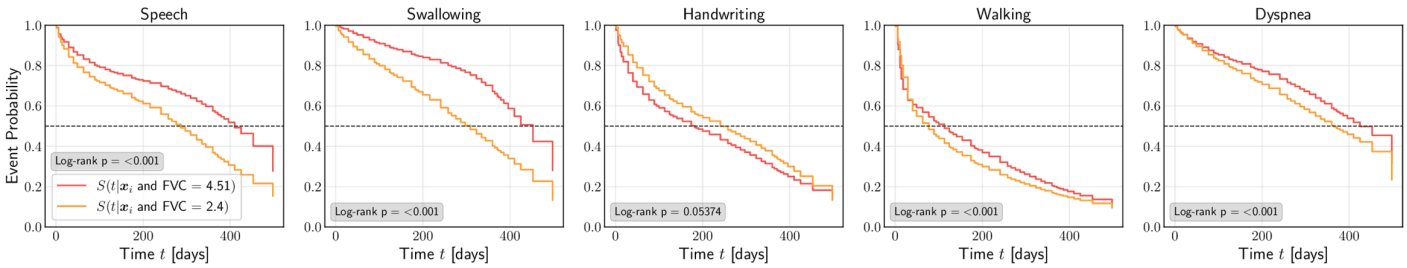


Fig 7. Predicted ISDs for Mr. Smith based on his FVC score being high or low.

https://doi.org/10.1371/journal.pone.0336476.g007

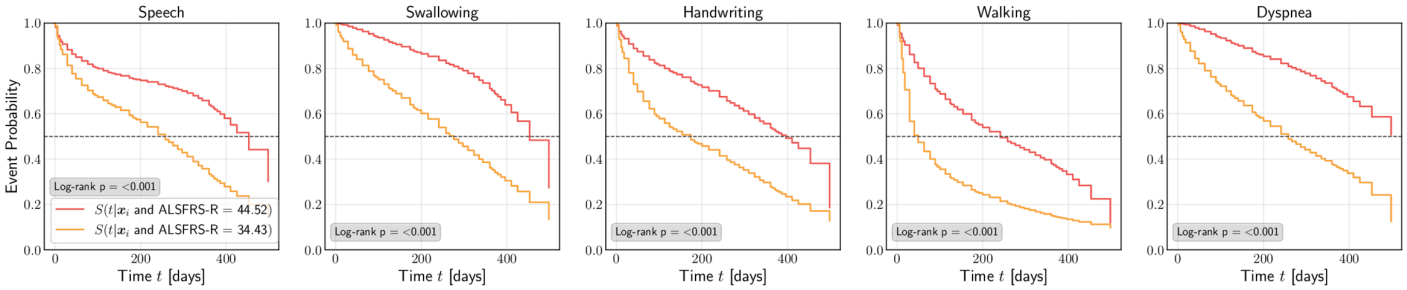


Fig 8. Predicted ISDs for Mr. Smith based on whether his initial ALSFRS-R score is high or low.

https://doi.org/10.1371/journal.pone.0336476.g008

Discussion

Contribution

This research has three goals: (1) determine whether covariate-based survival models can predict the time to individual functional decline (ALSFRS-R $\leq$ 2) in ALS patients more accurately than population estimators, (2) precisely estimate risk scores and event times, and (3) explore how counterfactual scenarios affect the probability of functional decline over the course of the disease.

For the first goal, individualized predictions are crucial for tailoring healthcare and therapeutic decisions to the individual patient. Models that estimate an individual's risk rather than a group average enable us to answer important questions from the patient's point of view such as: how much longer will I be able to speak clearly enough to be understood? We make a novel contribution in this regard, as we empirically show that most covariate-based models, even simple linear models, outperform the KM estimator (which does not use covariates) in predicting the time to functional decline. Given that ALS affects multiple functional systems (bulbar, motor, respiratory) at varying rates, our analysis also shows clear advantages of adopting non-proportional hazards models, such as Random Survival Forests and MENSA. Unlike other methods for predicting functional decline in ALS [3,9–12], our method uniquely provides patients and clinicians with an actual time (in days) when functional decline is likely to occur.

For the second goal, we rigorously evaluate five survival models across three aspects of model performance: discrimination, absolute/squared error, and calibration. The main takeaway is that some types of functional decline in ALS patients are easier to predict than others. For example, the *Speech* event has an average Harrell's CI of approximately 74% across the five models—the highest among all events—but also an average mMAE (a variant of the MAE that supports censoring) of 192.4 days, the second-highest. This means that even if a model is good at ranking patients by their relative risk (independent of time) for some event, it cannot necessarily predict accurately when the event will occur. In contrast, the *Walking* event has a lower average CI of approximately 67%, but an average mMAE of only 147.7 days. This suggests that, although it is harder to rank patients by risk for this event compared to *Speech*, it is easier, by contrast, to predict when the event will occur. This leads to the realization that strong ranking performance does not necessarily translate into accurate time-to-event predictions.

Finally, for the third goal, we use our method to make counterfactual predictions, investigating how modifying specific covariates affects predicted disease outcomes for individual patients. To our knowledge, no prior work has explored this use case in ALS. However, such predictions are highly relevant for medical professionals who must make important treatment decisions after first meeting a patient. For example, consider our patient, Mr. Smith: based on his covariates, there were no significant differences in the predicted survival probabilities (i.e., the risk of functional decline) should he decide to take Riluzole, a drug used to treat ALS. In other words, whether he decided to take Riluzole or not, this decision would not significantly change his predicted risk of functional decline over the next 500 days from the first time he saw the

neurologist. We believe that these individualized predictions can be very beneficial for treatment planning and clinical decision-making.

Application

Our method has several important applications. First, it can help medical doctors make informed decisions by tailoring treatments and interventions to the individual's predicted disease trajectory. This is crucial in ALS, where timely interventions, such as assistive devices and nutritional or respiratory support, can significantly impact a patient's quality of life. Second, patients and their families are often faced with difficult decisions regarding treatment options, lifestyle adjustments, and end-of-life planning. By providing a clear, individualized prognosis, clinicians can help patients make more informed decisions about their care, improve their quality of life, and give them a sense of control over their treatment. As part of the evaluation, we show how our method can make counterfactual predictions, answering the what-if questions: Mr. Smith was diagnosed with limb-onset ALS and the model predicts a specific path of functional decline for him, but what would have happened if he had been diagnosed with bulbar-onset ALS instead. Third, personalized predictions can also help design and evaluate clinical trials. Understanding the likely timing of functional decline in patients with ALS can help researchers identify suitable candidates for clinical studies, ensuring that the trials are conducted with participants at the appropriate stage of disease. This can lead to more effective testing of potential therapies and interventions that could slow or stop disease progression.

Concretely, if practitioners wish to implement our method in everyday clinical practice, we recommend the following steps:

1. Integrate model training with existing electronic health record (EHR) systems or clinical decision support tools to automatically retrieve patient data (covariates) and past ALSFRS-R scores (event information). After obtaining the relevant data, they should train the model, evaluate it, deploy it, and expose an application programming interface (API) that allows clinicians to make predictions on new patients.
2. Develop a user-friendly software interface or web-based application that allows clinicians to easily input or access patient information and obtain individual predictions of functional decline over time. When a new patient is diagnosed with ALS and has their first visit at the hospital, clinicians can use the application to enter the patient's covariates and obtain a prediction over the next 500 days.
3. Update, retrain, and deploy the model regularly with new patient data to maintain accuracy and incorporate new biomarkers or additional clinical variables, if such become available. This can be done using Machine Learning Operations (MLOps), which is a set of practices that automate and simplify machine learning workflows and deployments.

Limitations

All survival models evaluated in this work assume conditional independent censoring, i.e., that the event time is independent of the censoring time, given the patient's covariates. For example, patients who leave the study early do so for reasons unrelated to the event of interest or because of poor lung function captured by the forced vital capacity covariate. However, ALS is a heterogeneous disease that can quickly deteriorate and no study can capture all intrinsic disease mechanisms. Generally, clinical trials in ALS experience high dropout rates [36], which is also evident in our study through the high censoring rates shown in Table 3 and Fig 2. It is likely that this phenomenon is related to disease progression and therefore functional decline, but since we do not observe functional decline in patients who leave the study, this kind of censoring may violate the assumption of conditional independent censoring and bias the results. According to Atassi et al. [36], the three most common reasons for attrition (i.e., dropout) in ALS trials were withdrawal of consent for no specific reason (37%), death (28%), and withdrawal secondary to an adverse event (17%).

Future work

Our method has several avenues for future research. First, addressing the aforementioned dependency between the event and censoring times is relevant, given the relatively few covariates provided in the PRO-ACT dataset and because all evaluated survival models assume that censoring is conditionally independent. To address this issue, when estimating model parameters, we can learn a so-called copula function [37]. This is a link function that describes the dependency structure between the event and censoring distributions. Copulas have recently gained attention in survival analysis due to their flexibility in capturing complex dependency structures [38]. Second, it is highly relevant to include covariates captured using magnetic resonance imaging (MRI) into the predictive models, since such features can explain early signs of neurodegeneration, for example, cerebral atrophy in the motor cortex and corticospinal tract [39]. At present, such features have only been used to predict mortality or disease aggressiveness in patients with ALS [19,39,40], but not to predict functional decline. Thus, this presents an exciting avenue for future research.

Conclusion

This article proposes a novel and interpretable method for estimating the time to functional decline in ALS patients. We formulate this task as a multi-event survival problem and validate our method in the PRO-ACT dataset by training five covariate-based survival models to predict significant functional impairment with respect to five common functions. The proposed method has the following advantages over state-of-the-art approaches: (1) It explicitly predicts the time until functional decline is likely to occur for an individual patient. (2) It gives superior time-to-event estimates over commonly used population estimators, such as the KM estimator. (3) It allows for individual counterfactual predictions, where certain covariates can be changed to see their effect on the predicted outcome. In conclusion, we recommend using the proposed method on patient records in clinical settings, subject to additional rigorous validation, to assess the risk of imminent functional decline and enable more personalized treatment planning.

Supporting information

S1 Table. List of covariates.
(DOCX)

S2 Table. List of missing data rows.
(DOCX)

S3 Table. List of model hyperparameters.
(DOCX)

S4 Table. Mr. Smith's covariates.
(DOCX)

Acknowledgments

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. The following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: ALS Therapy Alliance, Cytokinetics, Inc., Amylyx Pharmaceuticals, Inc., Knopp Biosciences, Neuraltus Pharmaceuticals, Inc., Neurological Clinical Research Institute (MGH), Northeast ALS Consortium, Novartis, Orion Corporation, Prize4Life Israel, Regeneron Pharmaceuticals, Inc., Sanofi, Teva Pharmaceutical Industries, Ltd., and The ALS Association.

Author contributions

Conceptualization: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner.

Data curation: Christian Marius Lillelund.

Formal analysis: Christian Marius Lillelund.

Funding acquisition: Russell Greiner.

Investigation: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner.

Methodology: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner.

Project administration: Russell Greiner.

Resources: Christian Marius Lillelund, Sanjay Kalra.

Software: Christian Marius Lillelund.

Supervision: Sanjay Kalra, Russell Greiner.

Validation: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner.

Visualization: Christian Marius Lillelund.

Writing – original draft: Christian Marius Lillelund.

Writing – review & editing: Christian Marius Lillelund, Sanjay Kalra, Russell Greiner.

References

1. Morris J. Amyotrophic Lateral Sclerosis (ALS) and related motor neuron diseases: an overview. Neurodiagn J. 2015;55(3):180–94. https://doi.org/10.1080/21646821.2015.1075181 PMID: 26630810
2. Talbot K. Motor neuron disease: the bare essentials. Pract Neurol. 2009;9(5):303–9. https://doi.org/10.1136/jnnp.2009.188151 PMID: 19762894
3. Turabieh H, Afshar AS, Statland J, Song X. Towards a machine learning empowered prognostic model for predicting disease progression for amyotrophic lateral sclerosis. AMIA Annu Symp Proc. 2024;2023:718–25.
4. Kjældgaard A-L, Pilely K, Olsen KS, Jessen AH, Lauritsen AØ, Pedersen SW, et al. Prediction of survival in amyotrophic lateral sclerosis: a nationwide, Danish cohort study. BMC Neurol. 2021;21(1):164. https://doi.org/10.1186/s12883-021-02187-8 PMID: 33865343
5. Feldman EL, Goutman SA, Petri S, Mazzini L, Savelieff MG, Shaw PJ, et al. Amyotrophic lateral sclerosis. Lancet. 2022;400(10360):1363–80. https://doi.org/10.1016/S0140-6736(22)01272-7 PMID: 36116464
6. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci. 1999;169(1–2):13–21. https://doi.org/10.1016/s0022-510x(99)00210-5 PMID: 10540002
7. Kuan L-H, Parnianpour P, Kushol R, Kumar N, Anand T, Kalra S, et al. Accurate personalized survival prediction for amyotrophic lateral sclerosis patients. Sci Rep. 2023;13(1):20713. https://doi.org/10.1038/s41598-023-47935-7 PMID: 38001260
8. van der Burgh HK, Schmidt R, Westeneng H-J, de Reus MA, van den Berg LH, van den Heuvel MP. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. Neuroimage Clin. 2016;13:361–9. https://doi.org/10.1016/j.nicl.2016.10.008 PMID: 28070484
9. Vieira FG, Venugopalan S, Premasiri AS, McNally M, Jansen A, McCloskey K, et al. A machine-learning based objective measure for ALS disease severity. NPJ Digit Med. 2022;5(1):45. https://doi.org/10.1038/s41746-022-00588-8 PMID: 35396385
10. M Amaral D, Soares DF, Gromicho M, de Carvalho M, Madeira SC, Tomás P, et al. Temporal stratification of amyotrophic lateral sclerosis patients using disease progression patterns. Nat Commun. 2024;15(1):5717. https://doi.org/10.1038/s41467-024-49954-y PMID: 38977678
11. Din Abdul Jabbar MA, Guo L, Nag S, Guo Y, Simmons Z, Pioro EP, et al. Predicting amyotrophic lateral sclerosis (ALS) progression with machine learning. Amyotroph Lateral Scler Frontotemporal Degener. 2024;25(3–4):242–55. https://doi.org/10.1080/21678421.2023.2285443 PMID: 38052485
12. Ong M-L, Tan PF, Holbrook JD. Predicting functional decline and survival in amyotrophic lateral sclerosis. PLoS One. 2017;12(4):e0174925. https://doi.org/10.1371/journal.pone.0174925 PMID: 28406915
13. Haider H, Hoehn B, Davis S, Greiner R. Effective ways to build and evaluate individual survival distributions. J Mach Learn Res. 2020;21(1):3289–351.

14. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database: design, initial analyses, and predictive features. Neurology. 2014;83(19):1719–25. https://doi.org/10.1212/WNL.0000000000000951 PMID: 25298304
15. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. 2nd ed. New York: Springer; 2021.
16. Shen X, Elmer J, Chen GH. Neurological prognostication of post-cardiac-arrest coma patients using EEG data: a dynamic survival analysis framework with competing risks. In: Proceedings of the 8th Machine Learning for Healthcare Conference. 2023. p. 667–90.
17. Andres A, Montano-Loza A, Greiner R, Uhlich M, Jin P, Hoehn B, et al. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. PLoS One. 2018;13(3):e0193523. https://doi.org/10.1371/journal.pone.0193523 PMID: 29543895
18. Lillelund CM, Harbo M, Pedersen CF. Prognosis of fall risk in home care clients: a noninvasive approach using survival analysis. J Public Health (Berl). 2024. https://doi.org/10.1007/s10389-024-02317-9
19. Lajoie I, Canadian ALS Neuroimaging Consortium (CALSNIC), Kalra S, Dadar M. Regional cerebral atrophy contributes to personalized survival prediction in amyotrophic lateral sclerosis: a multicentre, machine learning, deformation-based morphometry study. Ann Neurol. 2025;97(6):1144–57. https://doi.org/10.1002/ana.27196 PMID: 39985309
20. Sechidis K, Tsoumakas G, Vlahavas I. On the stratification of multi-label data. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2011. p. 145–58. https://doi.org/10.1007/978-3-642-23808-6_10
21. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1972;34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3). https://doi.org/10.1214/08-aoas169
23. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. https://doi.org/10.1186/s12874-018-0482-1 PMID: 29482517
24. Yu CN, Greiner R, Lin HC, Baracos V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 2011. p. 1845–53.
25. Lillelund CM, Gharari Foomani AH, Sun W, Qi S, Greiner R. MENSA: a multi-event network for survival analysis with trajectory-based likelihood estimation. arXiv preprint 2023. https://arxiv.org/abs/2409.06525
26. Harrell FE. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543. https://doi.org/10.1001/jama.1982.03320430047030
27. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105–17. https://doi.org/10.1002/sim.4154 PMID: 21484848
28. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statist Med. 1999;18(17–18):2529–45. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5
29. Qi S, Sun W, Greiner R. SurvivalEVAL: a comprehensive open-source python package for evaluating individual survival distributions. AAAI-SS. 2024;2(1):453–7. https://doi.org/10.1609/aaaiss.v2i1.27713
30. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457–81. https://doi.org/10.1080/01621459.1958.10501452
31. Qi S-A, Kumar N, Farrokh M, Sun W, Kuan L-H, Ranganath R, et al. An effective meaningful way to evaluate survival models. Proc Mach Learn Res. 2023;202:28244–76. PMID: 40895293
32. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369–75. https://doi.org/10.1038/s42256-020-0197-y
33. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966;50(3):163–70. PMID: 5910392
34. Jaiswal MK. Riluzole and edaravone: a tale of two amyotrophic lateral sclerosis drugs. Med Res Rev. 2019;39(2):733–48. https://doi.org/10.1002/med.21528 PMID: 30101496
35. Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis. Nat Rev Neurol. 2014;10(11):661–70. https://doi.org/10.1038/nrneurol.2014.184 PMID: 25311585
36. Atassi N, Yerramilli-Rao P, Szymonifka J, Yu H, Kearney M, Grasso D, et al. Analysis of start-up, retention, and adherence in ALS clinical trials. Neurology. 2013;81(15):1350–5. https://doi.org/10.1212/WNL.0b013e3182a823e0 PMID: 24005339
37. Emura T, Chen Y-H. Analysis of survival data with dependent censoring. Singapore: Springer; 2018. https://doi.org/10.1007/978-981-10-7164-5
38. Foomani AHG, Cooper M, Greiner R, Krishnan RG. Copula-based deep survival models for dependent censoring. In: Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence. 2023. p. 669–80.
39. Dadar M, Manera AL, Zinman L, Korngut L, Genge A, Graham SJ, et al. Cerebral atrophy in amyotrophic lateral sclerosis parallels the pathological distribution of TDP43. Brain Commun. 2020;2(2):fcaa061. https://doi.org/10.1093/braincomms/fcaa061 PMID: 33543125
40. Dieckmann N, Roediger A, Prell T, Schuster S, Herdick M, Mayer TE, et al. Cortical and subcortical grey matter atrophy in amyotrophic lateral sclerosis correlates with measures of disease accumulation independent of disease aggressiveness. Neuroimage Clin. 2022;36:103162. https://doi.org/10.1016/j.nicl.2022.103162 PMID: 36067613

Appendix G

MENSA: A Multi-Event Network for Survival Analysis with Trajectory-based Likelihood Estimation

Abstract: Most existing time-to-event methods focus on either single-event or competing-risks settings, leaving multi-event scenarios relatively underexplored. In many healthcare applications, for example, a patient may experience multiple clinical events, that can be non-exclusive and semi-competing. A common workaround is to train independent single-event models for such multi-event problems, but this approach fails to exploit dependencies and shared structures across events. To overcome these limitations, we propose MENSA (Multi-Event Network for Survival Analysis), a deep learning model that jointly learns flexible time-to-event distributions for multiple events, whether competing or co-occurring. In addition, we introduce a novel trajectory-based likelihood term that captures the temporal ordering between events. Across four multi-event datasets, MENSA improves predictive performance over many state-of-the-art baselines. Source code is available at `https://github.com/thecml/mensa`.

Authors: Christian Marius Lillelund, Ali Hossein Gharari Foomani, Weijie Sun, Shiang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT).

Status: Published in the proceedings of ML4H 2025, Proceedings of Machine Learning Research, vol. 297, pp. 543–571, 2026.

Proceedings of Machine Learning Research 297, 2025 Machine Learning for Health (ML4H) 2025

MENSA: A Multi-Event Network for Survival Analysis with Trajectory-based Likelihood Estimation

Christian Marius Lillelund[1,2*] CLILLELUND@UALBERTA.CA

Ali Hossein Gharari Foomani[1] HOSSEING@UALBERTA.CA

Weijie Sun[1] WEIJIE2@UALBERTA.CA

Shi-ang Qi[1] SHIANG@UALBERTA.CA

Russell Greiner[1,3] RGREINER@UALBERTA.CA

The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT)[4]

[1] *Department of Computing Science, University of Alberta, Edmonton, Canada*
[2] *Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark*
[3] *Alberta Machine Intelligence Institute, Edmonton, Canada*
[4] *PRO-ACT Consortium*†

Abstract

Most existing time-to-event methods focus on either single-event or competing-risks settings, leaving multi-event scenarios relatively underexplored. In many healthcare applications, for example, a patient may experience multiple clinical events, that can be non-exclusive and semi-competing. A common workaround is to train independent single-event models for such multi-event problems, but this approach fails to exploit dependencies and shared structures across events. To overcome these limitations, we propose MENSA (Multi-Event Network for Survival Analysis), a deep learning model that jointly learns flexible time-to-event distributions for multiple events, whether competing or co-occurring. In addition, we introduce a novel trajectory-based likelihood term that captures the temporal ordering between events. Across four multi-event datasets, MENSA improves predictive performance over many state-of-the-art baselines. Source code is available at https://github.com/thecml/mensa.

Keywords: Survival analysis, time-to-event prediction, density estimation, multiple events, competing risks, precision medicine

Data and Code Availability All datasets used in this study are publicly available. Appendix E.1 describes how to obtain these datasets and preprocess them accordingly.

Institutional Review Board (IRB) This study uses only publicly available datasets and does not involve any interaction with human subjects or access to identifiable private information. Therefore, IRB approval was not required.

1. Introduction

Many survival models estimate the probability that an event of interest (typically death) will occur for an individual at some future time. This can involve: a *single event* of interest, where we want to model the time until one specific event (Gareth et al., 2021, Ch. 11), *competing risks*, where the occurrence of one event excludes other events from occurring (Fine and Gray, 1999), or *multiple events*, where several events may occur in some order (Andersen and Keiding, 2002). As an example of the latter, a patient with a chronic condition may experience several medical events over the course of their disease.

In current practice, multi-event problems are often approached by training multiple single-event models (Armstrong et al., 2014; Solomon et al., 2017). However, this approach (1) fails to exploit shared covariate patterns, (2) ignores temporal or causal dependencies between events, and (3) becomes computationally inefficient as the number of events increases. Other methods mitigate these issues through shared frailty terms (Jiang and Haneuse, 2017), multi-task regularization (Li et al., 2016; Wang et al.,

* Corresponding author.

© 2025 C.M. Lillelund *et al.*

2017), or joint survival modeling (Hsieh and Wang, 2018), but these rely on restrictive assumptions about frailty, proportional hazards, or a fixed survival distribution. Recent work has explored structured transition models (Groha et al., 2020), hierarchical timescale modeling (Tjandra et al., 2021), latent ordinary differential equations (ODEs) for estimating event-specific hazard curves (Moon et al., 2022), and transition models based on Monte Carlo sampling (Rossman et al., 2022). These offer partial solutions but typically assume predefined state transitions, scale poorly to long time horizons, cannot handle multiple events, or depend on sampling to predict an individual survival distributions (ISD) (Haider et al., 2020). To date, most research remains confined to single-event or competing-risks scenarios. Motivated by these limitations, we introduce a new and scalable deep learning model for multi-event survival analysis, called MENSA (Multi-Event Network for Survival Analysis). Our main contributions are:

- We propose a multi-state survival model based on the framework by Andersen and Keiding (2002), where each event corresponds to a transition between two states (*e.g.*, from healthy to sick). This enables prediction of how long a patient will remain in a particular state, or when the patient will experience one of multiple events. Each event-specific marginal distribution is modeled as a mixture of Weibull distributions, relaxing the proportional hazards assumption. Also, this architecture naturally extends to single-event and competing-risks settings.

- We introduce a training objective that combines the standard negative log-likelihood (NLL) loss with an additional trajectory-based term. This new term encodes known temporal relationships between events, guiding the model to predict ISDs that align with prior knowledge.

- We show empirically that MENSA offers solid discrimination and calibration performance, improving over many baseline models. MENSA is also more computationally efficient than several baselines in the multi-event setting. Two ablation studies show that training MENSA jointly across all events leads to a notable increase in the local concordance index (C-index), reflecting better within-patient event ordering, and that using the proposed trajectory-based likelihood term further improves this metric.

2. Related work

Andersen and Keiding (2002) proposed a model for multi-state processes (for example, disease progression leading to relapse and eventually death), capturing sequences of events. This framework has since become a fundamental tool for multi-state analysis (Gorfine et al., 2025). Each transition between states (events) is modeled with its own hazard function, allowing state-specific effects, and naturally incorporates competing risks, recurrent events, and absorbing states (*e.g.*, death). However, individual event probabilities must be derived from transition probabilities across states, so the event time cannot be computed directly. Moreover, the model assumes a linear relationship between covariates and the transition-specific hazard functions.

Groha et al. (2020) proposed an ODE-based neural network for multi-state survival analysis that incorporates a predefined state-transition structure. This structure is encoded using a binary transition matrix, indicating which transitions are allowed. The model then learns time-varying transition intensities by solving Kolmogorov forward equations, but only for those transitions permitted by the predefined graph. A key limitation of this approach is that it enforces a fixed sequence of events – for example, once cancer occurs, the model does not allow subsequent heart failure and vice versa. This framework requires practitioners to specify the full state-transition graph in advance, which can be restrictive when the event order is uncertain or variable.

Tjandra et al. (2021) proposed a hierarchical multi-event model that predicts event occurrences across multiple time scales, using coarse estimates (*e.g.*, monthly) to guide finer predictions (*e.g.*, daily). This approach decomposes long-horizon tasks into simpler intermediate steps, progressively refining predictions from broad to fine temporal resolutions. However, it has two main limitations. First, the hierarchical structure incurs significant parameter overhead and scales poorly with longer durations. Second, the method is optimized solely for discriminative loss. Strong discrimination performance does not necessarily translate into accurate time-to-event predictions, and vice versa (Qi et al., 2023a).

Moon et al. (2022) proposed an ODE-based model for multi-event analysis. This model maps temporal features into continuous latent trajectories and estimates cause-specific hazard functions at discrete time intervals without relying on parametric hazard

assumptions. However, the model is only designed for competing risks, supports at most two events, and assumes access to time-varying features. It cannot be directly applied on datasets with only static (baseline) covariates, which remain the most common in survival analysis. Moreover, while latent ODE-based models offer high flexibility, they face practical limitations. As noted by Chen (2024, p. 66), these models are typically slower to train and more prone to numerical instability than parametric alternatives.

Rossman et al. (2022) introduced a multi-state survival model, *PyMSM*, which estimates transition-specific hazard functions using time-dependent covariates and Monte Carlo sampling. PyMSM is well-suited for settings where individuals move through a sequence of well-defined states (*e.g.*, from illness to recovery to death), and covariates can be updated at each transition. However, its reliance on Monte Carlo sampling introduces two key limitations. First, to estimate an ISD for a patient, the model must sample many possible trajectories, which can be computationally expensive. Second, since the model only approximates the ISDs, it lacks closed-form survival or density functions, which prevents analytical comparisons (*e.g.*, hazard ratios) or mathematical guarantees (*e.g.*, proportional hazards).

3. The proposed method

3.1. Problem formulation

We consider a survival dataset with multiple events and right-censoring, consisting of N time-to-event triplets, denoted as $\mathcal{D} = \left\{ \left(\boldsymbol{x}^{(i)}, \boldsymbol{t}^{(i)}, \boldsymbol{\delta}^{(i)} \right) \right\}_{i=1}^{N}$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ denotes a d-dimensional vector of features (covariates), $\boldsymbol{t}^{(i)} \in \{1, 2, \ldots, t_{\max}\}^K$ is a vector of observed times for the i-th instance across K events, and $\boldsymbol{\delta}^{(i)} \in \{0, 1\}^K$ is a vector of event indicators for each event[1]. We consider each instance to have (potential) event times, $e_k^{(i)} \in \{1, 2, \ldots, t_{\max}\}$, and censoring times, $c_k^{(i)} \in \{1, 2, \ldots, t_{\max}\}$. We then have $t_k^{(i)} = e_k^{(i)}$ if $\delta_k^{(i)} = 1$, meaning that the event was observed, otherwise, $t_k^{(i)} = c_k^{(i)}$ if $\delta_k^{(i)} = 0$, meaning that the event was right-censored.

Now, as a practical example, consider a drug trial over $t \in [0, 100]$ days (see Figure 1), where a patient i with features $\boldsymbol{x}^{(i)}$ can experience four side effects: nausea (A), fatigue (B), fever (C), or headache (D). Based on $\boldsymbol{x}^{(i)}$, we want to estimate the so-called survival function $S_E(t) = \Pr(t < T_E)$[2] for each of these possible events (side effects), which is the probability that the patient has not yet experienced the event by $t \geq 0$. By estimating this function for every event, we obtain a matrix of survival functions, $\hat{\boldsymbol{S}}^{(i)} = [\boldsymbol{s}_1^{(i)}, \boldsymbol{s}_2^{(i)}, \ldots, \boldsymbol{s}_K^{(i)}]$, for the i-th patient, where K is the number of events. Finally, we can estimate the patient's likely time to each event as the point where each curve crosses the 50% (median) line.

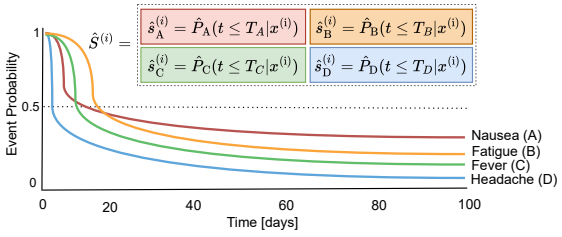


Figure 1: Example of a multi-event prediction.

3.2. Mixture model architecture

We propose a multi-state architecture (see Figure 2) in which events are modeled as transitions between discrete states – *e.g.*, from *healthy* to *sick*, or from *alive* to *dead* – based on the framework by Andersen and Keiding (2002), and further described in Appendix B. Let $\mathcal{P} = \{0, 1, \ldots, P-1\}$ denote the finite set of states, where state 0 is the initial (event-free) state. Each event $k \in \{1, \ldots, K\}$ corresponds to a transition into a state $p_k \in \mathcal{P} \setminus \{0\}$. For example, if event k represents a *relapse*, then p_k denotes the absorbing state for *relapse*. This mapping naturally supports scenarios with sequential or co-occurring events – *e.g.*, first becoming ill and then later dying.

For each instance i, we estimate a matrix of survival distributions $\boldsymbol{S}^{(i)}$: one distribution for each state $p \in \mathcal{P}$, including the initial event-free state ($p = 0$). The marginal density function $f_p(t)$ for state p, which represents the probability of entering state

1. In our terminology, we refer to *competing-risks* problems as those in which only one event can be observed per instance, and the event time is effectively a scalar: $t^{(i)}$. In contrast, *multi-event* problems involve multiple events per instance, and the event time is instead represented as a vector: $\boldsymbol{t}^{(i)}$. Cases involving *semi-competing risks* – where one event precludes another but not vice versa – fall under our definition of multi-event problems.

2. We use T_E to denote the time of event, and T_C to denote the time of censoring. The observed time is then $T = \min(T_E, T_C)$.

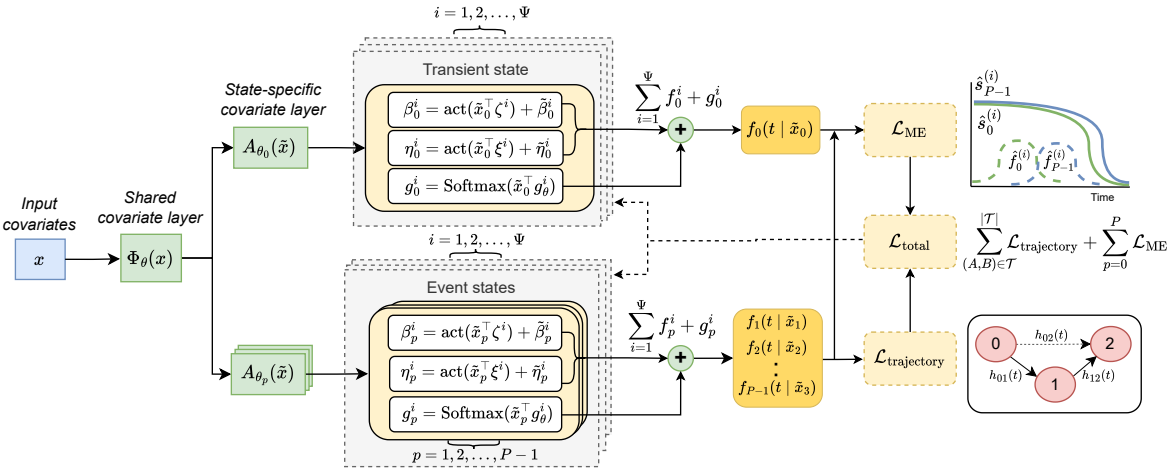


Figure 2: Overview of the MENSA architecture. The model first maps input covariates through a shared multilayer perceptron (MLP) layer, producing a common representation $\tilde{\boldsymbol{x}}$. This is passed through event-specific covariate layers, each producing $\tilde{\boldsymbol{x}}_p$ for state $p \in \mathcal{P}$. Each adapted representation then parameterizes a mixture of Weibull distributions to predict a time-to-event density function, $f_p(t)$, capturing the time of transition from the transient state (state 0) to each event state. The model is trained using a combined loss consisting of a multi-event likelihood, $\mathcal{L}_{\text{ME}}$, which captures how well the model fits event times across all events, and a trajectory likelihood, $\mathcal{L}_{\text{trajectory}}$, which enforces known temporal ordering constraints between pairs of events.

p at time point t, is modeled as a convex combination of $\Psi \in \mathbb{N}_+$ Weibull distributions. This formulation offers two key advantages. First, it relaxes the proportional hazards assumption on the underlying marginal survival distribution (proof in Appendix I), allowing the hazard function to vary over time. Second, it leverages the closed-form probability density function (PDF) and cumulative distribution function (CDF) of the Weibull distribution, respectively:

$$f(t;\beta,\eta) = \frac{\eta}{\beta}\left(\frac{t}{\beta}\right)^{\eta-1} \exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \quad (1)$$

$$F(t;\beta,\eta) = 1 - \exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \quad (2)$$

where $\beta > 0$ is the scale and $\eta > 0$ the shape parameter. We parameterize each such Weibull distribution by two covariate-dependent transformations, that is, $h_p(\boldsymbol{x}^{(i)}) = A_{\theta,p}\big(\Phi_\theta(\boldsymbol{x}^{(i)})\big)$, such that:

$$\log \beta_{\psi,p}(\boldsymbol{x}^{(i)}) = \tilde{\beta}_{\psi,p} + \text{act}\big(h_p(\boldsymbol{x}^{(i)})^\top \zeta_{\psi,p}\big), \quad (3)$$

$$\log \eta_{\psi,p}(\boldsymbol{x}^{(i)}) = \tilde{\eta}_{\psi,p} + \text{act}\big(h_p(\boldsymbol{x}^{(i)})^\top \xi_{\psi,p}\big), \quad (4)$$

where $\Phi_\theta(\cdot)$ is a shared covariate transformation, $A_{\theta,p}(\cdot)$ is a state-specific adapter applied to its output, $\text{act}(\cdot)$ denotes a non-linear activation function, $\psi \in \{1, 2, \ldots, \Psi\}$ indexes the Weibull mixture distributions, and $\zeta_{\psi,p}, \xi_{\psi,p} \in \mathbb{R}^d$ are learnable weight vectors associated with the scale and shape parameters, respectively. The terms $\tilde{\beta}_{\psi,p}$ and $\tilde{\eta}_{\psi,p}$ are learnable biases that initialize the log-scale and log-shape, respectively, for each state p. We define $\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)}) = \big(\beta_{\psi,p}(\boldsymbol{x}^{(i)}), \eta_{\psi,p}(\boldsymbol{x}^{(i)})\big)$, and the corresponding distribution PDF and CDF for transitioning into state p at time t given covariates $\boldsymbol{x}^{(i)}$ are then given by:

$$f_p(t;\boldsymbol{x}^{(i)}) = \sum_{\psi=1}^{\Psi} g_{\psi,p}(\boldsymbol{x}^{(i)})\, f_{\psi,p}\big(t;\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)})\big), \quad (5)$$

$$F_p(t;\boldsymbol{x}^{(i)}) = \sum_{\psi=1}^{\Psi} g_{\psi,p}(\boldsymbol{x}^{(i)})\, F_{\psi,p}\big(t;\boldsymbol{\theta}_{\psi,p}(\boldsymbol{x}^{(i)})\big), \quad (6)$$

where $g_{\psi,p}(\boldsymbol{x}^{(i)})$ are normalized mixture weights produced by a state-specific gating network, *i.e.*, a softmax over linear outputs of $h_p(\boldsymbol{x}^{(i)})$. The corresponding survival function is $S_p(t;\boldsymbol{x}^{(i)}) = 1 - F_p(t;\boldsymbol{x}^{(i)})$.

3.3. Trajectory-based likelihood

To estimate (5) and (6), we extend the standard survival likelihood (defined in Appendix C) to our multi-

Table 1: Possible scenarios for two events, A and B, with corresponding transition probabilities and log-likelihood contributions. δ_A and δ_B indicate whether each event is observed (1) or censored (0). $P_{0\to A}(t_A)$ and $P_{0\to B}(t_B)$ denote the probabilities of experiencing events A and B at times t_A and t_B, respectively; $S_A(t)$ and $S_B(t)$ are survival functions; and $f_A(t)$ and $f_B(t)$ are probability density functions. If prior knowledge dictates that event A must precede event B when both occur, the joint-event likelihood ($\delta_A = 1$, $\delta_B = 1$) incorporates the probability of surviving event B at the time of event A, as highlighted in bold.

#	δ_A	δ_B	Transition probability	Log-likelihood
1	0	0	$1 - (P_{0\to A}(t_A) + P_{0\to B}(t_B))$	$S_A(t_A) + S_B(t_B)$
2	1	0	$P_{0\to A}(t_A)$	$f_A(t_A) + S_B(t_B)$
3	0	1	$P_{0\to B}(t_B)$	$S_A(t_A) + f_B(t_B)$
4	1	1	$P_{0\to A}(t_A) * P_{0\to B}(t_B)$	$f_A(t_A) + f_B(t_B) + \boldsymbol{S_B(t_A)}$

state architecture – where each instance i may experience multiple events. However, rather than indexing events directly, we express the likelihood in terms of state transitions. Specifically, we define one indicator and time pair $(\delta_p^{(i)}, t_p^{(i)})$ per state $p \in \mathcal{P}$, where each event k corresponds to a transition into state p. The indicator $\delta_p^{(i)}$ denotes whether a transition into state p was observed, and $t_p^{(i)}$ is the corresponding observed or censored transition time. The likelihood includes the state's density $f_p(t_p^{(i)} \mid \boldsymbol{x}^{(i)})$, if observed ($\delta_p^{(i)} = 1$), and the survival function $S_p(t_p^{(i)} \mid \boldsymbol{x}^{(i)})$, if censored ($\delta_p^{(i)} = 0$).

To account for the imbalance between state transitions, we introduce non-negative weights, w_p, computed from the inverse frequency of transitions into each state using the training set, normalized so that $\sum_{p\in\mathcal{P}} w_p = |\mathcal{P}|$. Assuming conditional independence between state transitions[3], we can then define our multi-event likelihood as:

$$\mathcal{L}_{\text{ME}}(\mathcal{D}) = \sum_{i=1}^{N} \sum_{p\in\mathcal{P}} w_p \Big[\delta_p^{(i)} \log f_p\Big(t_p^{(i)} \mid \boldsymbol{x}^{(i)}\Big) + (1-\delta_p^{(i)}) \log S_p\Big(t_p^{(i)} \mid \boldsymbol{x}^{(i)}\Big)\Big], \tag{7}$$

where $f_p(\cdot)$ and $S_p(\cdot)$ denote the conditional density and survival function for state p, respectively.

To guide the model, we incorporate prior knowledge of temporal event ordering when available. As an example, consider the illness-death model (Fix and Neyman, 1951), where patients start out healthy (state 0), become ill (event A), and may eventually die (event B). This means there are four possible scenarios (see Table 1): (1) neither event occurred (both censored), (2) only A occurred (B censored), (3) only B occurred (A censored), and (4) both events occurred, with A before B. We assume that if such semi-competing events are observed, one must come before the other – *e.g.*, death cannot occur before illness. To incorporate this constraint into the optimization, we define an auxiliary likelihood term that encourages the model to assign a high survival probability for event B at the time event A occurs, *i.e.*, maximize $S_B(t_A)$. This ensures the model assigns a low probability to B when A has occurred, effectively handling semi-competing risks.

Let $\mathcal{T} = \{(A, B) \mid A, B \in \mathcal{K}, \text{A occurs before B}\}$ denote the set of known event orderings, and for each instance i, we consider all uncensored event pairs $(A, B) \in \mathcal{T}$. The trajectory likelihood is computed by summing the survival probability of event B at the time of event A, thus $S_B(T_A \mid \boldsymbol{x}^{(i)})$, weighted by the event indicators $\delta_A^{(i)}$ and $\delta_B^{(i)}$, which equal 1 if events A and B, respectively, have occurred for instance i, and 0 otherwise. The likelihood is then:

$$\mathcal{L}_{\text{trajectory}}(\mathcal{D}) = \sum_{i=1}^{N} \sum_{(A,B)\in\mathcal{T}} \delta_A^{(i)} \delta_B^{(i)} \log S_B\big(T_A \mid \boldsymbol{x}^{(i)}\big), \tag{8}$$

where $\mathcal{T}$ is the set of trajectories. Finally, we have:

$$\mathcal{L}_{\text{total}}(\mathcal{D}) = (1-\lambda)\frac{1}{N}\mathcal{L}_{\text{ME}}(\mathcal{D}) + \lambda \cdot \frac{1}{N}\mathcal{L}_{\text{trajectory}}(\mathcal{D}), \tag{9}$$

where λ is a weighting hyperparameter that controls the trade-off between the two likelihoods. The training algorithm can be found in Appendix E.4.

3. Conditional independence means that, given the covariates, the transition to one state is assumed to be independent of transitions to other states. Therefore, any relationship between states is assumed explained by the covariates.

4. Experiments

4.1. Experimental setup

Datasets. We use five datasets to test the effectiveness of our method. Table 7 in Appendix E summarizes the dataset statistics, and Appendix E.1 contains details of preprocessing steps, and flow charts of event trajectories. These datasets cover single-event, competing-risks and multi-event settings.

Baselines. We compare 11 baseline survival models: Cox proportional hazards (CoxPH) (Cox, 1972), CoxNet (Simon et al., 2011), Weibull AFT (Klein and Moeschberger, 2006), GBSA (Ridgeway, 1999), RSF (Ishwaran et al., 2008), NFG[4] (Jeanselme et al., 2023), DeepSurv (Katzman et al., 2018), DeepHit (Lee et al., 2018), Hierarchical (Tjandra et al., 2021), MTLR (Yu et al., 2011; Kim et al., 2021), and DSM (Nagpal et al., 2021). Appendix E.2 provides implementation details.

Hyperparameters. To ensure a fair comparison, all models – including baselines – were tuned using Bayesian optimization (Snoek et al., 2012) over ten iterations. For each model and dataset, we selected the hyperparameters that achieved the highest Harrell's C-index (Harrell Jr et al., 1996) on the validation set. Appendix E.3 lists the chosen parameters.

Evaluation metrics. To evaluate the discrimination performance, we report the global C-index (CI_G) (Tjandra et al., 2021), which measures how well the model can discriminate instances within an event, and the local C-index (CI_L) (Tjandra et al., 2021), which measures how well the model can discriminate events within an instance, and the time-dependent AUC, which measures discrimination at a fixed time point. We calculate the AUC by computing it at the 25th, 50th, and 75th percentiles of observed event times in the test set and then averaging these values. To evaluate calibration performance, we report the mean absolute error (MAE) using a margin loss (mMAE) (Haider et al., 2020), the integrated Brier score (IBS) (Graf et al., 1999) and the distribution calibration (D-calibration) score (Haider et al., 2020). Appendix D gives more details on the metrics.

4.2. Implementation details

Following recent literature (Katzman et al., 2018; Nagpal et al., 2021; Lillelund et al., 2025), we use a multilayer perceptron (MLP) for the backbone network architecture. This transforms the raw input features into a learned representation, which is then used to parameterize the Weibull distributions. On top of this shared representation, we add small state-specific MLPs to fine-tune representations for each prediction task. We use a single hidden layer, the ReLU6 activation function for the MLP, and a mixture of Ψ Weibull distributions. Adam is utilized for optimization (Kingma and Ba, 2017). Our model is implemented in Pytorch (Paszke et al., 2019). See Appendix E.4 for further details.

In each of 10 experiments, we follow Sechidis et al. (2011) and split each dataset into stratified training (70%), validation (10%), and test (20%) sets with random seeds 0–9. After splitting, we impute missing values using the sample mean for continuous covariates and the mode for categorical covariates, based on the training set. We apply a z-score normalization to numerical features and encode categorical features using one-hot encoding. We train MENSA with early stopping on the validation set with a patience of 20 epochs. Concerning reporting, results are first computed separately for each random seed and event. Metrics are then averaged across events within each seed, producing one value per seed per model–dataset combination. We report the mean and standard deviation of these per-seed averages.

4.3. Results and takeaways

Here, we report results on the multi-event datasets. Concerning the single-event and competing-risks settings, these results are presented in Appendix G.

Overall performance. Table 2 presents the discrimination results of our proposed model and existing baselines on four multi-event datasets. Models marked with † in the tables are significantly and practically worse than MENSA (details of the statistical testing and minimum effect of interest (MEI) thresholds are provided in Appendix F). Overall, MENSA outperforms many deep learning baselines in discrimination tasks across multiple datasets and performs comparably to strong tree-based methods such as GBSA and RSF. Unlike DeepHit or Hierarchical, MENSA avoids unstable behavior (large variance or metric collapse). For these multi-event datasets, we also assess local discrimination performance using the local C-index, which captures the accuracy of within-patient event orderings and reflects the added value of a multi-event model. In cases where MENSA out-

4. The Neural Fine Gray (NFG) model is used exclusively for the competing-risks experiments, see Appendix G.2.

Table 2: Discrimination performance (mean ± SD.) between different models on multi-event benchmark datasets, averaged over 10 experiments. For all metrics, higher (↑) is better. Model types are abbreviated as SE = single event, CR = competing risks, and ME = multiple events. Bold indicates the top-3 models. † indicates models that are significantly and practically worse than MENSA, according to paired t-tests with Holm correction ($\alpha = 0.05$) and a minimum effect of interest (MEI) threshold of 1.0.

Model	Type	MIMIC-IV ($K=3$)			Rotterdam ($K=3$)			PRO-ACT ($K=4$)			EBMT ($K=5$)		
		CI_G	CI_L	AUC	CI_G	CI_L	AUC	CI_G	CI_L	AUC	CI_G	CI_L	AUC
CoxPH	SE	70.9$_{\pm 0.52}$	90.3$_{\pm 0.48}$	74.3$_{\pm 0.71}$	69.7$_{\pm 1.16}$	83.6$_{\pm 2.09}$†	75.1$_{\pm 1.48}$	70.6$_{\pm 0.47}$	**76.5**$_{\pm 0.55}$	76.7$_{\pm 0.91}$	54.9$_{\pm 1.04}$	66.3$_{\pm 1.16}$†	54.6$_{\pm 1.10}$
CoxNet	SE	71.0$_{\pm 0.53}$	90.4$_{\pm 0.45}$	74.4$_{\pm 0.71}$	69.8$_{\pm 1.16}$	83.8$_{\pm 2.05}$†	75.1$_{\pm 1.49}$	70.6$_{\pm 0.46}$	76.5$_{\pm 0.57}$	76.7$_{\pm 0.89}$	**55.1**$_{\pm 1.12}$	67.2$_{\pm 1.13}$†	54.5$_{\pm 1.17}$
Weibull	SE	71.2$_{\pm 0.54}$	**90.8**$_{\pm 0.43}$	74.8$_{\pm 0.70}$	69.7$_{\pm 1.18}$	87.9$_{\pm 1.99}$†	75.1$_{\pm 1.45}$	70.7$_{\pm 0.49}$	**77.1**$_{\pm 0.57}$	76.7$_{\pm 0.90}$	54.9$_{\pm 1.36}$	**68.0**$_{\pm 0.97}$	54.2$_{\pm 1.35}$
GBSA	SE	**71.9**$_{\pm 0.43}$	**91.4**$_{\pm 0.51}$	**75.2**$_{\pm 0.53}$	**75.3**$_{\pm 1.62}$	**90.5**$_{\pm 1.05}$	71.2$_{\pm 2.02}$†	**72.0**$_{\pm 0.31}$	75.6$_{\pm 0.66}$	**77.6**$_{\pm 0.44}$	**55.3**$_{\pm 0.93}$	67.2$_{\pm 0.82}$†	**55.4**$_{\pm 1.26}$
RSF	SE	69.1$_{\pm 0.48}$†	**90.8**$_{\pm 0.40}$	74.6$_{\pm 0.56}$	**70.7**$_{\pm 1.15}$	84.9$_{\pm 2.19}$†	**77.3**$_{\pm 1.74}$	**71.1**$_{\pm 0.51}$	75.6$_{\pm 0.83}$	**77.5**$_{\pm 0.60}$	55.0$_{\pm 0.75}$	64.6$_{\pm 0.92}$†	**55.6**$_{\pm 1.07}$
MTLR	SE	70.2$_{\pm 0.50}$†	88.3$_{\pm 0.95}$†	74.4$_{\pm 0.71}$	69.8$_{\pm 1.16}$	82.8$_{\pm 2.37}$†	**75.7**$_{\pm 1.85}$	70.7$_{\pm 0.53}$	**76.5**$_{\pm 0.53}$	77.4$_{\pm 0.83}$	53.6$_{\pm 1.04}$	61.6$_{\pm 2.38}$†	55.4$_{\pm 1.19}$
DeepSurv	SE	**71.8**$_{\pm 0.55}$	89.9$_{\pm 0.49}$	**75.8**$_{\pm 0.68}$	70.0$_{\pm 1.24}$	84.6$_{\pm 2.93}$†	75.4$_{\pm 1.75}$	71.0$_{\pm 0.59}$	76.0$_{\pm 0.57}$	76.7$_{\pm 1.00}$	**55.1**$_{\pm 1.02}$	67.6$_{\pm 0.82}$	54.9$_{\pm 1.22}$
DeepHit	CR	70.9$_{\pm 0.75}$	63.1$_{\pm 9.74}$†	75.1$_{\pm 0.94}$	66.5$_{\pm 1.09}$†	**90.2**$_{\pm 1.03}$	74.2$_{\pm 1.50}$	64.9$_{\pm 1.15}$†	65.3$_{\pm 3.35}$†	73.4$_{\pm 0.73}$†	52.4$_{\pm 1.29}$†	60.4$_{\pm 2.76}$†	55.2$_{\pm 1.46}$
DSM	CR	70.5$_{\pm 0.60}$	88.1$_{\pm 1.37}$†	74.5$_{\pm 0.83}$	**70.6**$_{\pm 1.26}$	86.0$_{\pm 2.81}$†	**76.4**$_{\pm 1.68}$	70.8$_{\pm 0.68}$	76.0$_{\pm 0.68}$	**77.4**$_{\pm 0.76}$	54.8$_{\pm 1.00}$	**68.0**$_{\pm 0.99}$	55.3$_{\pm 1.36}$
Hierarch.	ME	65.2$_{\pm 1.67}$†	85.3$_{\pm 1.73}$†	48.7$_{\pm 0.51}$†	67.8$_{\pm 2.06}$	55.7$_{\pm 8.11}$†	50.0$_{\pm 0.05}$†	70.8$_{\pm 0.60}$	73.6$_{\pm 1.42}$†	57.3$_{\pm 0.89}$†	54.7$_{\pm 1.33}$	44.5$_{\pm 1.12}$†	53.5$_{\pm 1.03}$†
MENSA (Ours)	ME	**71.4**$_{\pm 0.29}$	90.3$_{\pm 0.54}$	**75.2**$_{\pm 0.41}$	69.4$_{\pm 1.20}$	**90.1**$_{\pm 1.55}$	73.8$_{\pm 2.31}$	**70.8**$_{\pm 0.45}$	76.4$_{\pm 0.61}$	77.2$_{\pm 0.60}$	55.1$_{\pm 1.37}$	**68.8**$_{\pm 0.66}$	**55.7**$_{\pm 1.62}$

Table 3: Calibration performance (mean ± SD.) between different models on four multi-event benchmark datasets, averaged over 10 experiments. For IBS and mMAE, lower (↓) is better. For D-calibration, higher (↑) is better. D-calibration counts the number of experiments in which the respective model was D-calibrated. Bold indicates the top-3 models. † indicates models that are significantly and practically worse than MENSA, based on paired t-tests with Holm correction ($\alpha = 0.05$) and a predefined MEI of 1.0 for IBS and mMAE.

Model	Type	MIMIC-IV ($K=3$)			Rotterdam ($K=3$)			PRO-ACT ($K=4$)			EBMT ($K=5$)		
		IBS	mMAE	D-Cal	IBS	mMAE	D-Cal	IBS	mMAE	D-Cal	IBS	mMAE	D-Cal
CoxPH	SE	15.3$_{\pm 0.51}$	10.8$_{\pm 5.77}$	30/30	19.6$_{\pm 1.53}$	23.2$_{\pm 1.15}$†	20/20	16.6$_{\pm 0.39}$	129.0$_{\pm 3.31}$†	36/40	20.4$_{\pm 0.68}$	48.6$_{\pm 0.71}$	49/50
CoxNet	SE	15.3$_{\pm 0.52}$	8.6$_{\pm 0.16}$	30/30	19.6$_{\pm 1.54}$	23.0$_{\pm 1.11}$	20/20	16.6$_{\pm 0.40}$	128.6$_{\pm 3.23}$†	36/40	20.3$_{\pm 0.68}$	**48.3**$_{\pm 0.70}$	49/50
Weibull	SE	15.6$_{\pm 0.42}$	11.3$_{\pm 0.14}$†	30/30	19.2$_{\pm 1.64}$	23.1$_{\pm 0.78}$	15/20	16.7$_{\pm 0.35}$	220.7$_{\pm 11.24}$†	36/40	**20.1**$_{\pm 0.98}$	59.4$_{\pm 0.86}$	18/50
GBSA	SE	16.7$_{\pm 0.63}$†	**6.9**$_{\pm 0.07}$	29/30	19.3$_{\pm 1.44}$	24.1$_{\pm 0.74}$†	20/20	16.7$_{\pm 0.42}$	**120.8**$_{\pm 2.14}$	35/40	20.4$_{\pm 0.75}$	**44.7**$_{\pm 0.44}$	50/50
RSF	SE	16.6$_{\pm 0.49}$†	**7.5**$_{\pm 0.08}$	27/30	**18.6**$_{\pm 1.29}$	**22.6**$_{\pm 1.05}$	18/20	**16.3**$_{\pm 0.41}$	170.1$_{\pm 22.18}$†	39/40	**18.9**$_{\pm 1.25}$	68.9$_{\pm 1.00}$	50/50
MTLR	SE	**14.9**$_{\pm 0.52}$	9.7$_{\pm 0.15}$†	30/30	**18.1**$_{\pm 1.65}$	23.0$_{\pm 0.72}$	19/20	**16.4**$_{\pm 0.42}$	**126.3**$_{\pm 4.56}$	40/40	**18.9**$_{\pm 1.25}$	68.9$_{\pm 1.00}$†	50/50
DeepSurv	SE	**14.9**$_{\pm 0.51}$	9.1$_{\pm 0.22}$	30/30	19.4$_{\pm 1.47}$	**22.7**$_{\pm 1.02}$	18/20	**16.3**$_{\pm 0.42}$	139.1$_{\pm 5.27}$†	37/40	20.5$_{\pm 0.73}$	**45.8**$_{\pm 0.63}$	49/50
DeepHit	CR	32.7$_{\pm 0.98}$†	20.0$_{\pm 0.31}$†	0/30	21.4$_{\pm 2.64}$†	27.3$_{\pm 0.79}$†	18/20	25.6$_{\pm 1.21}$†	196.5$_{\pm 4.00}$†	0/40	21.1$_{\pm 3.99}$	74.3$_{\pm 0.84}$†	50/50
DSM	CR	16.0$_{\pm 0.47}$†	17.0$_{\pm 1.92}$†	22/30	**18.3**$_{\pm 1.54}$	23.2$_{\pm 0.76}$†	15/20	16.5$_{\pm 0.35}$	231.5$_{\pm 15.51}$†	38/40	20.7$_{\pm 1.46}$	59.6$_{\pm 0.82}$	17/50
Hierarch.	ME	51.0$_{\pm 1.16}$†	101.9$_{\pm 233.61}$†	0/30	56.0$_{\pm 5.09}$†	43.5$_{\pm 0.76}$†	0/20	53.9$_{\pm 1.86}$†	280.5$_{\pm 5.47}$†	0/40	71.8$_{\pm 7.84}$†	92.2$_{\pm 0.80}$†	0/50
MENSA (Ours)	ME	**15.1**$_{\pm 0.59}$	**8.3**$_{\pm 0.09}$	21/30	18.8$_{\pm 1.59}$	**22.6**$_{\pm 0.68}$	18/20	16.6$_{\pm 0.41}$	**124.1**$_{\pm 2.14}$	39/40	22.0$_{\pm 2.33}$	63.1$_{\pm 2.88}$	10/50

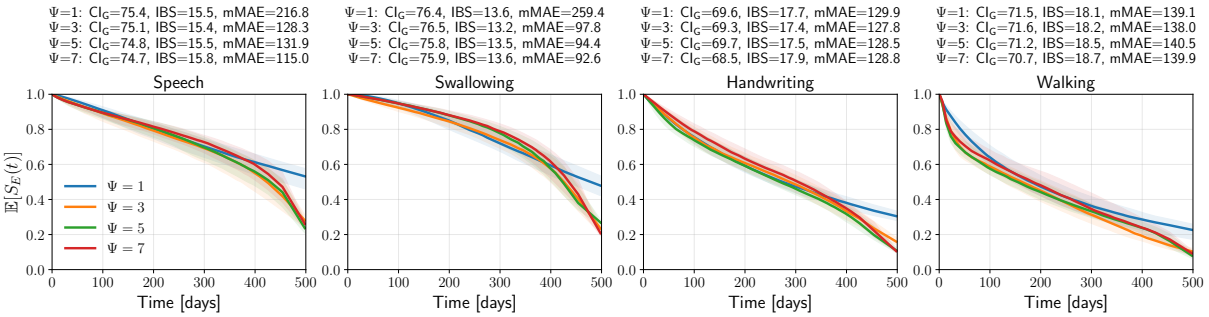


Figure 3: Population-level summaries of predicted survival functions across the PRO-ACT test set for each event. Each panel shows the median expected survival $\mathbb{E}[S_E(t)]$ with narrow (40–60th percentile) variability bands. Different colors correspond to MENSA models trained with varying numbers of Weibull distributions $\Psi \in \{1, 3, 5, 7\}$. The inset boxes report event-specific performance metrics (Global C-index, IBS, mMAE).

performs the baselines in local C-index, we attribute the improvement to its multi-state architecture and trajectory-based likelihood term. Table 3 summarizes calibration performance. MENSA maintains low IBS

Table 4: Ablation study comparing joint vs. independent training of MENSA on two multi-event datasets. Results are averaged over 10 experiments, with changes shown relative (%) to the baseline.

Dataset	Training	CI_G	CI_L	AUC	IBS	mMAE	D-Cal
Rotterdam	Joint	69.45±1.25	90.24±1.36	73.88±2.33	18.78±1.59	22.62±0.71	18/20
	Indep.	69.82±1.31 (+0.53)	87.81±1.74 (-2.69)	74.57±2.10 (+0.93)	18.84±1.67 (+0.32)	22.48±0.67 (-0.63)	20/20
PRO-ACT	Joint	70.87±0.50	76.07±0.51	77.12±1.20	16.60±0.52	124.37±3.44	36/40
	Indep.	70.98±0.64 (+0.15)	75.92±0.88 (-0.19)	77.29±0.75 (+0.23)	16.54±0.42 (-0.34)	122.78±2.33 (-1.27)	38/40

Table 5: Ablation study comparing training MENSA with/without the trajectory likelihood on two multi-event datasets. Results are averaged over 10 experiments, with changes shown relative (%) to the baseline.

Dataset	$\mathcal{L}_{trajectory}$	CI_G	CI_L	AUC	IBS	mMAE	D-Cal
Rotterdam	With	69.45±1.25	90.24±1.36	73.88±2.33	18.78±1.59	22.62±0.71	18/20
	Without	69.54±1.22 (+0.13)	89.94±1.45 (-0.34)	74.16±2.09 (+0.38)	18.79±1.72 (+0.09)	22.58±0.71 (-0.18)	18/20
EBMT	With	55.26±1.24	68.73±0.56	55.66±1.47	22.13±2.30	62.98±2.97	9/50
	Without	54.93±1.15 (-0.6)	68.39±0.56 (-0.5)	55.53±1.43 (-0.24)	22.36±2.59 (+1.04)	63.76±3.90 (+1.25)	14/50

and mMAE – often close to or better than classical survival methods (*e.g.*, CoxPH, RSF) – and exhibits comparable D-calibration to other deep learning variants (*e.g.*, DeepSurv and DSM). Compared to the Hierarchical model, our approach demonstrates consistently superior performance across all datasets.

Impact of number of distributions. Using functional decline prediction in patients with amyotrophic lateral sclerosis (ALS) as a test case (PRO-ACT dataset; four events: *Speaking*, *Swallowing*, *Handwriting*, and *Walking*; see Appendix E.1), we train MENSA to estimate event-specific population survival curves. Each curve represents the median survival (*i.e.*, event) probability $\mathbb{E}[S_E(t)]$ across all $N = 670$ patients under a given mixture configuration Ψ. Setting $\Psi = 1$ means that a single distribution models the time-to-event distribution for each event. Increasing Ψ introduces additional distributions, allowing the model to represent more complex and non-proportional risk structures. As shown in Figure 3, higher mixture complexity shifts the population survival downward across all events by the end of follow-up (about 500 days), reflecting a more pessimistic risk estimate. Quantitatively, this added flexibility lowers prediction error (mMAE: 183→118 days) and slightly improves calibration (IBS: 0.158→0.155) while maintaining stable discrimination (Global CI: ≈0.73). These results suggest that moderate mixture complexity ($\Psi \approx 3-5$) better captures disease progression in ALS without overfitting, improving time-to-event accuracy at the population level.

Computational analysis. We analyze the computational complexity of our method and several deep baselines[5]. For fairness, we configure all models with a default architecture (see Appendix E.3). The full results are presented in Appendix H. In the multi-event setting, all baselines (except the Hierarchical model) must be trained K times, whereas MENSA handles all events in a single forward pass. On a (hypothetical) dataset with 100 features and four events, MENSA requires 1.4× fewer parameters than DeepSurv, 2.5× fewer than DeepHit, 3.2× fewer than MTLR, 1.5× fewer than DSM, and 13× fewer than the Hierarchical model. Reductions in FLOP follow a similar trend, ranging from 1.5× (DeepSurv, DSM) and 2.5× (DeepHit) to 5.7× (MTLR) and 13× (Hierarchical).

Ablation studies. We perform two ablation studies evaluating (1) joint versus independent training, and (2) the trajectory-based likelihood term. Table 4 shows the results for training events jointly. Here, *jointly* refers to learning all event-specific parameters within a single shared architecture. We see in both datasets that the global metrics (*e.g.*, global C-index, AUC and mMAE) increase slightly when training independently, but the local C-index (within-patient event ordering) drops significantly in Rotterdam (-2.69% decrease) and PRO-ACT (-0.19% decrease). Table 5 presents the results for the trajectory term. In Rotterdam, it improves the local C-index with only small shifts in the global C-index, AUC and mMAE. In EBMT, the trajectory term consistently improves performance.

5. We report computational metrics only for the deep learning models; FLOPs and parameter counts are not well-defined or directly comparable for classical methods.

5. Discussion

5.1. Limitations

This work has limitations. First, MENSA assumes conditionally independent censoring, which may be violated if relevant covariates that influence both censoring and event times are unobserved. Second, as a parametric model, MENSA relies on predefined event-time distributions. While this choice enhances interpretability and stability, it can reduce flexibility and require larger datasets for reliable parameter estimation. In contrast, semiparametric approaches such as CoxPH and DeepSurv make fewer distributional assumptions and can sometimes generalize better in small-sample settings. Third, the current implementation of MENSA is limited to static baseline covariates and does not yet incorporate time-varying covariates or longitudinal measurements, which are common in many clinical applications.

5.2. Application

MENSA provides particular value in clinical contexts where the *local ordering* of events is most relevant. For example, in the PRO-ACT dataset, the goal is to predict the time to different types of functional decline in ALS patients as measured by the ALSFRS-R scale (see Appendix E.1). A model with a high local C-index can accurately rank event occurrences within a single patient, enabling clinically meaningful questions such as: *Which physical functions am I likely to lose first, and in what order?* This prioritization of local ordering is especially useful in neurodegenerative diseases like ALS, where functional decline follows distinct trajectories. Prior studies have shown, for instance, that bulbar-onset ALS typically affects speech functionality before gait and walking ability (Feldman et al., 2022). Beyond ALS, multi-event modeling is crucial in many diseases where multiple outcomes unfold over time. In chronic heart failure, for example, patients may experience repeated hospitalizations, arrhythmic events, and eventual cardiac death, each with complex temporal dependencies. In oncology (*e.g.*, the SEER cohort), treatment response, relapse, and metastasis form interdependent event pathways that conventional single-event models cannot adequately capture. Similarly, in hematologic and transplant settings (*e.g.*, the EBMT cohort), graft-versus-host disease, relapse, and non-relapse mortality compete and co-occur, requiring models that jointly estimate their risks.

Regarding interpretability, practitioners can extract feature importances for each event-specific prediction. Model-agnostic methods tailored to survival analysis, such as SurvLIME (Kovalev et al., 2020) and SurvSHAP (Krzyziński et al., 2023), can be applied to MENSA's event-specific outputs to quantify the contribution of individual covariates to each predicted event time.

5.3. Conclusion

We have presented MENSA, a new method for multi-event survival analysis. MENSA jointly learns flexible time-to-event distributions for multiple events, which may be competing or co-occurring. Across four multi-event survival datasets, MENSA demonstrated solid discrimination and calibration performance, outperforming many deep learning baselines while remaining computationally efficient in the multi-event setting. Compared to the Hierarchical survival model – the main existing deep learning approach for multi-event survival analysis – MENSA achieved superior results across all evaluation metrics. Ablation studies further showed that training all events within a shared architecture improves within-patient event ordering, as reflected by higher local C-index values, and that the proposed trajectory-based likelihood term further enhances this metric.

Acknowledgments

†Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: ALS Therapy Alliance, Cytokinetics, Inc., Amylyx Pharmaceuticals, Inc., Knopp Biosciences, Neuraltus Pharmaceuticals, Inc., Neurological Clinical Research Institute, MGH, Northeast ALS Consortium, Novartis, Orion Corporation, Prize4Life Israel, Regeneron Pharmaceuticals, Inc., Sanofi, Teva Pharmaceutical Industries, Ltd., The ALS Association.

This research received support from the Natural Science and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), and the Alberta Machine Intelligence Institute (Amii).

References

Per K. Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115, 2002.

Gregory T. Armstrong, Toana Kawashima, Wendy Leisenring, Kayla Stratton, Marilyn Stovall, Melissa M. Hudson, Charles A. Sklar, Leslie L. Robison, and Kevin C. Oeffinger. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. *Journal of Clinical Oncology*, 32(12):1218, 2014.

Nazem Atassi, James Berry, Amy Shui, Neta Zach, Alexander Sherman, Ervin Sinani, Jason Walker, Igor Katsovskiy, David Schoenfeld, Merit Cudkowicz, and Melanie Leitner. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi. The ALSFRS-R: A Revised ALS Functional Rating Scale That Incorporates Assessments of Respiratory Function. *Journal of Neurological Sciences*, 169(1-2):13–21, 1999.

George H. Chen. An Introduction to Deep Survival Analysis Models for Predicting Time-to-Event Outcomes. *Foundations and Trends in Machine Learning*, 17(6):921–1100, 2024.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40): 1317, 2019.

Takeshi Emura and Yi-Hau Chen. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.

Eva L. Feldman, Stephen A. Goutman, Stefan Petri, Letizia Mazzini, Michael G. Savelieff, Pamela J. Shaw, and Gen Sobue. Amyotrophic Lateral Sclerosis. *The Lancet*, 400(10360):1363–1380, 2022.

Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

Evelyn Fix and Jerzy Neyman. A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients. *Human Biology*, 23(3):205–241, Sep 1951.

James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Springer, 2 edition, 2021.

Malka Gorfine, Richard J. Cook, Per Kragh Andersen, Terry M. Therneau, Pierre Joly, Hein Putter, Maja Pohar Perme, and Michal Abrahamowicz. An Overview and Recent Developments in the Analysis of Multistate Processes. *arXiv preprint arXiv:2502.06492*, 2025.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Stefan Groha, Sebastian M. Schmon, and Alexander Gusev. A general framework for survival analysis and multi-state modelling. *arXiv preprint arXiv:2006.04893*, 2020.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An Extensive Data Processing Pipeline for MIMIC-IV. In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193, pages 311–325, 2022.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, 21(1):1–63, 2020.

Frank E. Harrell Jr, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Jin-Jian Hsieh and Jian-Lin Wang. Quantile residual life regression based on semi-competing risks data. *Journal of Applied Statistics*, 45(10):1770–1780, 2018.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random Survival Forests. *The Annals of Applied Statistics*, 2 (3):841–860, 2008.

Vincent Jeanselme, Chang Ho Yoon, Brian Tom, and Jessica Barrett. Neural Fine-Gray: Monotonic neural networks for competing risks. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209, pages 379–392, 2023.

Fei Jiang and Sebastien Haneuse. A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics*, 44(1):112–129, 2017.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1): 1–9, 2023.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

Sejin Kim, Michal Kazmierski, and Benjamin Haibe-Kains. Deep-CR MTLR: a Multi-Modal Approach for Cancer Survival Prediction with Competing Risks. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications*, volume 146, pages 223–231, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2017.

John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. SurvLIME: A Method for Explaining Machine Learning Survival Models. *Knowledge-Based Systems*, 203:106164, 2020.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. SurvSHAP(t): Time-Dependent Explanations of Machine Learning Survival Models. *Knowledge-Based Systems*, 262:110234, 2023.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1145–1152, Apr. 2018.

Yan Li, Jie Wang, Jieping Ye, and Chandan K. Reddy. A Multi-Task Learning Formulation for Survival Analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1715–1724, 2016.

Christian M. Lillelund, Martin Magris, and Christian F. Pedersen. Efficient Training of Probabilistic Neural Networks for Survival Analysis. *IEEE Journal of Biomedical and Health Informatics*, 29 (9):6157–6166, 2025.

Intae Moon, Stefan Groha, and Alexander Gusev. SurvLatent ODE: A Neural ODE based time-to-event model with competing risks for longitudinal data improves cancer-associated Venous Thromboembolism (VTE) prediction. In *Machine Learning for Healthcare Conference*, pages 800–827. PMLR, 2022.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data with Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021.

Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data. *arXiv preprint arXiv:2204.07276*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037, 2019.

Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-

learn. *Journal of Machine Learning Research*, 21 (212):1–6, 2020.

Shi-ang Qi, Neeraj Kumar, Jian-Yi Xu, Jaykumar Patel, Sambasivarao Damaraju, Grace Shen-Tu, and Russell Greiner. Personalized breast cancer onset prediction from lifestyle and health history information. *PLOS ONE*, 17(12):e0279174, 2022.

Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28244–28276, 2023a.

Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. In *Proceedings of the 2023 AAAI Fall Symposia*, volume 2, pages 453–457, 2023b.

Greg Ridgeway. The State of Boosting. *Computing Science and Statistics*, 31:172–181, 1999.

Lynn A. Gloeckler Ries, Marsha E. Reichman, Denise Riedel Lewis, Benjamin F. Hankey, and Brenda K. Edwards. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist*, 8(6):541–552, 2003.

Hagai Rossman, Ayya Keshet, and Malka Gorfine. PyMSM: Python package for Competing Risks and Multi-State models for Survival Data. *Journal of Open Source Software*, 7(78):4566, 2022.

Patrick Royston and Douglas G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13:33, 2013.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the Stratification of Multi-label Data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158, 2011.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5), 2011.

Richard M. Simon, Edward L. Korn, Lisa M. McShane, Michael D. Radmacher, George W. Wright, and Yingdong Zhao. *Design and analysis of DNA microarray investigations*, volume 209. Springer, 2003.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pages 2951—2959, 2012.

Scott D. Solomon, Adel R. Rizkala, Jianjian Gong, Wenyan Wang, Inder S. Anand, Junbo Ge, Carolyn S. P. Lam, Aldo P. Maggioni, Felipe Martinez, Milton Packer, Marc A. Pfeffer, Burkert Pieske, Margaret M. Redfield, Jean L. Rouleau, Dirk J. Van Veldhuisen, Faiez Zannad, Michael R. Zile, Akshay S. Desai, Victor C. Shi, Martin P. Lefkowitz, and John J. V. McMurray. Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the PARAGON-HF trial. *JACC: Heart Failure*, 5(7): 471–482, 2017.

Donna Tjandra, Yifei He, and Jenna Wiens. A Hierarchical Approach to Multi-Event Survival Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):591–599, 2021.

Lu Wang, Yan Li, Jiayu Zhou, Dongxiao Zhu, and Jieping Ye. Multi-task Survival Analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 485–494, 2017.

Zifeng Wang and Jimeng Sun. SurvTRACE: Transformers for Survival Analysis with Competing Events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2022.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24, pages 1845—1853, 2011.

Appendix A. Notation

Table 6 provides a list of notations used in this work.

Table 6: Table of notation.

Symbol	Definition
$c_p^{(i)}$	Censoring time for instance i in state p
$e_p^{(i)}$	True event/transition time for instance i in state p
$t_p^{(i)}$	Observed time for instance i in state p: $\min(e_p^{(i)}, c_p^{(i)})$
$\delta_p^{(i)}$	Event indicator: $\delta_p^{(i)} = \mathbb{1}[e_p^{(i)} \leq c_p^{(i)}]$
$\mathcal{D}$	A time-to-event dataset
N	Number of individuals/instances in a dataset
d	Number of features/covariates in a dataset
$\boldsymbol{x}^{(i)}$	Covariate vector for instance i, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$
$h_0(t)$	Baseline hazard function
$h(t \mid \boldsymbol{x}^{(i)})$	Hazard function given covariates $\boldsymbol{x}^{(i)}$
$K \in \mathbb{N}_+$	Number of distinct events
$P \in \mathbb{N}_+$	Number of distinct states
$\mathcal{K}$	Set of possible events, $\lvert\mathcal{K}\rvert = K$
$\mathcal{P}$	Set of possible states, $\lvert\mathcal{P}\rvert = P$
$\mathcal{T}$	Set of event trajectories over $\mathcal{K}$
$\mathcal{L}$	Loss/objective function used during training
L	Number of training epochs
κ	Learning rate for optimization
λ	Weighting hyperparameter for the trajectory loss
$\Psi \in \mathbb{N}_+$	Number of mixture distributions
$G(t) \in \mathcal{P}$	Latent state process (*e.g.*, Markov process) over state space $\mathcal{P}$
$f(\boldsymbol{\theta}, \boldsymbol{x}^{(i)})$	Risk or output function given parameters $\boldsymbol{\theta}$ and input $\boldsymbol{x}^{(i)}$
$S_E(t)$	Event survival function: $S_E(t) = \Pr(T_E > t)$
$f_E(t)$	Event density function: $f_E(t) = \frac{d}{dt}(1 - S_E(t))$
$F_E(t)$	Event cumulative distribution function: $F_E(t) = 1 - S_E(t)$

Appendix B. Multi-state analysis

Single event. Survival analysis typically models the time until a single event takes place (Gareth et al., 2021, Ch. 11). We can represent such problems using a two-state model, as in Figure 4.

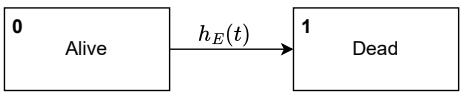


Figure 4: A single-event model with a single transition function.

In this simplest case, there is one transient state (0, *e.g.*, "alive") and one absorbing state (1, *e.g.*, "dead"). Once an instance enters an absorbing state, no further transitions are possible. The observation for a given instance will in the simplest form consist of a random variable, T_E, representing the time from a given origin ($t = 0$) to the occurrence of the event. The distribution of T_E can be characterized by the probability distribution function $F_E(t) = \Pr(T_E \leq t)$ or, equivalently, by the survival function $S_E(t) = 1 - F_E(t) = \Pr(T_E > t)$. Note that $S_E(t)$ and $F_E(t)$, respectively, correspond to the probabilities

of being in state 0 or 1 at time t. In continuous time, the distribution of T_E is commonly characterized by the *hazard rate function*:

$$h_E(t) = -d \log S_E(t)/dt = \lim_{\Delta t \to 0} \frac{\Pr(t < T_E \leq t + \Delta t | T_E > t)}{\Delta t}, \tag{10}$$

where $S_E(t)$ is the *survival function*:

$$S_E(t) = \exp\left(-\int_0^t h_E(u)\, du\right). \tag{11}$$

Here, $h_E(t)$ is the transition intensity from state 0 to 1, *i.e.*, the instantaneous risk of going from state 0 to state 1 at time t. We can extend this framework to multiple state transitions using *marked point processes* (Andersen and Keiding, 2002). Let $G(t)$, $t \in \mathbb{T}$, $\mathbb{T} = [0, t_{\max}]$ denote a stochastic process with a finite state space $\mathcal{P} = \{0, ..., P-1\}$ and with right-continuous sample paths: $G(t+) = G(t)$. The process has an initial distribution $\pi_p(0) = [\Pr(G(0) = p) \mid p \in \mathcal{P}]$. A multi-state process $G(\cdot)$ generates a history $\mathcal{G}_t$, which consists of observations of the process in the interval $[0, t_{\max}]$ – *e.g.*, the patient is healthy at the start of the study ($\mathcal{G}_0$), but becomes ill 10 days in ($\mathcal{G}_{10}$), before the study terminates after 100 days ($t_{\max} = 100$). Given this history, we can define the so-called transition probabilities between two distinct events by:

$$\Pr_{p \to q}(s, t) = \Pr(G(t) = q \mid G(s) = p, \mathcal{G}_{s^-}), \tag{12}$$

for $p, q \in \mathcal{P}, s, t \in \mathbb{T}, s \leq t$ and transition intensities by the derivatives, assuming such exist:

$$h_{p \to q}(t) = \lim_{\Delta t \to 0} \frac{\Pr_{p \to q}(t, t + \Delta t)}{\Delta t}. \tag{13}$$

Some transition intensities may be zero for all t between two states, *i.e.*, $h_{p \to q}(t) = 0 \; \forall t$ and $\forall p, q \in \mathcal{P}$. A state $p \in \mathcal{P}$ is absorbing if, for all $t \in \mathbb{T}, p \in \mathcal{P}, p \neq q, h_{p \to q}(t) = 0$; otherwise p is transient. The state probabilities $\pi_p(t) = \Pr(G(t) = p)$ are given by:

$$\pi_p(t) = \sum_{q \in \mathcal{P}} \pi^q(0)\, \Pr_{q \to p}(0, t). \tag{14}$$

Note that $\Pr_{p \to q}(\cdot, \cdot)$ and $h_{p \to q}(\cdot)$ depend on both the probability measure $\Pr(\cdot)$, and on the history.

Competing risks. In competing risks, multiple events (*i.e.*, absorbing states) exist, but only one event of $k = 1, \ldots, K$ different types can occur for each instance, illustrated in Figure 5.

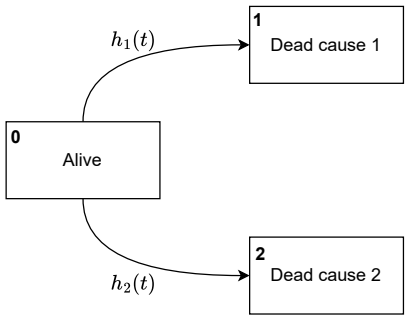


Figure 5: A competing-risks model for mortality from two causes.

In our multi-state framework, a competing-risks model has one transient state, aka state 0, and a number of absorbing states. Entering an absorbing state corresponds to experiencing event k, for $k \in \{1, \ldots, K\}$.

Thus, we have P total states and $K = P - 1$ events. In this model, if an instance is censored before experiencing any events, it simply remained in the transient state and is thus event-free. Using our state notation, the transition intensities $h_p(t)$ for $p \in \{1, \ldots, P-1\}$ are given by the cause-specific hazard function (omitting state 0):

$$h_p(t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{Transition to state } p \text{ at } t + \Delta t \mid T_p \geq t)}{\Delta t}, \tag{15}$$

where T_p is the random variable corresponding to the time of the transition to state p. There are only edges originating from state 0, the only transient state of the model, *i.e.*, $h_{p \to q}(t) = 0$ for all $p \neq 0$ and all q. The transition probabilities are given by the survival function:

$$\Pr_{0 \to 0}(0, t) = S_0(t) = \Pr(T_0 > t) = \exp\left(- \int_0^t \sum_{p=1}^{P-1} h_p(u)\, du \right), \tag{16}$$

and the cumulative incidence function (CIF) for transitioning from state 0 to any state $p \neq 0$:

$$\Pr_{0 \to p}(0, t) = \int_0^t S_0(u-)\, h_p(u)\, du, \quad p = 1, \ldots, P-1. \tag{17}$$

Multiple events. In multi-event scenarios, we can observe multiple transitions (events) for each instance i, thus the transitions (or event occurrences) are not necessarily mutually exclusive, as in Figure 6.

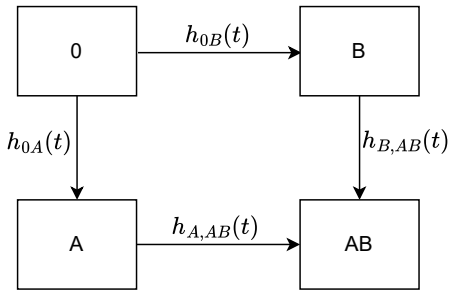


Figure 6: A multi-event model describing the joint behavior of A and B.

Such a model can, for example, describe the joint behavior of two events A and B; if $h_{0B} = h_{A,AB}$ but $h_{0A} \neq h_{B,AB}$, A is called *locally dependent* on B but B is not locally dependent on A. The temporal ordering of events enables an asymmetric notion of dependence, providing more information for causal inference than the symmetric association concepts used in cross-sectional studies.

Appendix C. Survival likelihood

In the standard survival analysis setting, considering a single event of interest, we define the survival time T_E and censoring time T_C as follows:

$$\begin{aligned} T_E = t^{(i)} \text{ and } T_C \geq t^{(i)} \quad &\text{if } \delta^{(i)} = 1, \\ T_E > t^{(i)} \text{ and } T_C = t^{(i)} \quad &\text{if } \delta^{(i)} = 0. \end{aligned}$$

Combining these cases, the likelihood for the i-th observation is:

$$\mathcal{L}^{(i)} = \Pr(T_E = t^{(i)}, T_C \geq t^{(i)} \mid x^{(i)})^{\delta^{(i)}} \Pr(T_E \geq t^{(i)}, T_C = t^{(i)} \mid x^{(i)})^{1-\delta^{(i)}} \tag{18}$$

Under the assumption of independent censoring (Emura and Chen, 2018), we have:

$$\begin{aligned}\mathcal{L}^{(i)} &= \Big[\Pr(T_E = t^{(i)} \mid x^{(i)}) \Pr(T_C \geq t^{(i)} \mid x^{(i)})\Big]^{\delta^{(i)}} \Big[\Pr(T_E \geq t^{(i)} \mid x^{(i)}) \Pr(T_C = t^{(i)} \mid x^{(i)})\Big]^{1-\delta^{(i)}} \\ &= \Big[f_E(t^{(i)} \mid x^{(i)})\, S_C(t^{(i)} \mid x^{(i)})\Big]^{\delta^{(i)}} \Big[S_E(t^{(i)} \mid x^{(i)})\, f_C(t^{(i)} \mid x^{(i)})\Big]^{1-\delta^{(i)}} \\ &= \Big[f_E(t^{(i)} \mid x^{(i)})^{\delta^{(i)}} S_E(t^{(i)} \mid x^{(i)})^{1-\delta^{(i)}}\Big] \Big[f_C(t^{(i)} \mid x^{(i)})^{1-\delta^{(i)}} S_C(t^{(i)} \mid x^{(i)})^{\delta^{(i)}}\Big] \end{aligned} \tag{19}$$

where $S_E(t \mid x^{(i)}) = \Pr(T_E \geq t \mid x^{(i)})$, $f_E(t \mid x^{(i)}) = -dS_E(t \mid x^{(i)})/dt$, $S_C(t \mid x^{(i)}) = \Pr(T_C \geq t \mid x^{(i)})$, and $f_C(t \mid x^{(i)}) = -dS_C(t \mid x^{(i)})/dt$. Under the non-informative censoring assumption, the term $f_C(t^{(i)} \mid x^{(i)})^{1-\delta^{(i)}} S_C(t^{(i)} \mid x^{(i)})^{\delta^{(i)}}$ is unrelated to the likelihood for the event times and can simply be ignored (Emura and Chen, 2018). Therefore, the likelihood function can be written as:

$$\mathcal{L} = \prod_{i=1}^{N} f_E(t^{(i)} \mid \boldsymbol{x}^{(i)})^{\delta^{(i)}} S_E(t^{(i)} \mid \boldsymbol{x}^{(i)})^{1-\delta^{(i)}}, \tag{20}$$

where $f_E(t^{(i)} \mid x^{(i)})$ is the density function of the event time and $S_E(t^{(i)} \mid x^{(i)})$ is the survival function of the event time. Note how the censoring terms $f_C(t^{(i)})$ and $S_C(t^{(i)})$ are absent in (20). If the event is observed, the likelihood contribution is the probability density of the event occurring at that time. If the observation is censored, the likelihood contribution is the probability of surviving beyond that time. The overall likelihood is the product of these contributions across all instances.

Appendix D. Evaluation metrics

Harrell's C-index: The concordance index (C-index or CI) measures the discriminative performance of a survival model by calculating the proportion of concordant pairs among all comparable pairs. A pair is considered comparable if we can determine who has the event first. It is defined as (Harrell Jr et al., 1996):

$$\text{C-index} = \frac{\sum_{i,j \in \mathcal{D}} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \mathbb{1}[\eta^{(i)} > \eta^{(j)}] \cdot \delta^{(i)}}{\sum_{i,j \in \mathcal{D}} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \delta^{(i)}} \tag{21}$$

where $\eta^{(i)}$ and $\eta^{(j)}$ represent risk scores for individuals i and j, respectively.

Global C-index: The global C-index is computed by averaging C-index scores across multiple instances for each event, and was proposed by Tjandra et al. (2021). It thus provides an overall measure of a model's discriminative performance in multiple events scenarios. This metric evaluates global consistency, *i.e.*, how well does the model discriminate instances *within* an event. It is defined as:

$$\text{Global C-index} = \frac{\sum_{k \in \mathcal{K}} \sum_{i,j \in \mathcal{D}_k} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \mathbb{1}[\eta^{(i)} > \eta^{(j)}] \cdot \delta^{(i)}}{\sum_{k \in \mathcal{K}} \sum_{i,j \in \mathcal{D}_k} \mathbb{1}[t^{(i)} < t^{(j)}] \cdot \delta^{(i)}}, \tag{22}$$

where $\mathcal{K}$ represents the set of all distinct events, and the index k corresponds to the k-th specific event.

Local C-index: The local C-index is computed by averaging C-index scores across multiple events for each instance, and was proposed by Tjandra et al. (2021). This metric evaluates local consistency, *i.e.*, how well does the model discriminate events *within* an individual. It is defined as:

$$\text{Local C-index} = \frac{\sum_{i \in N} \sum_{k_1,k_2 \in \mathcal{K}^{(i)}} \mathbb{1}[t_{k_1} < t_{k_2}] \cdot \mathbb{1}[\eta_{k_1} > \eta_{k_2}] \cdot \delta_{k_1}}{\sum_{i \in N} \sum_{k_1,k_2 \in \mathcal{K}^{(i)}} \mathbb{1}[t_{k_1} < t_{k_2}] \cdot \delta_{k_1}}, \tag{23}$$

where $\mathcal{K}^{(i)}$ represents the set of all K different events for instance i, and η_{k1} and η_{k2} denote the risk scores associated with the two events, respectively.

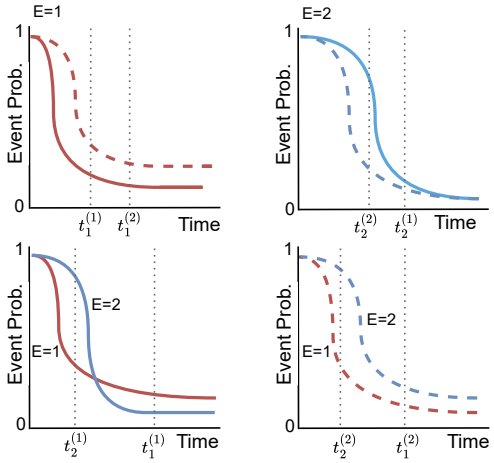


Figure 7: Global and local consistency. Red/blue lines represent survival functions $S_1(t)$ and $S_2(t)$ for event 1 and 2, respectively. Solid/dashed lines represent individuals $i = 1, 2$, respectively. Top-left: globally consistent at $t_1^{(1)}$ and $t_1^{(2)}$ since $S_1^{(1)}(t) < S_2^{(2)}(t)$ both times. Top-right: globally consistent at $t_2^{(2)}$ and $t_2^{(1)}$ since $S_2^{(2)}(t) < S_2^{(1)}(t)$ both times. Bottom-left: locally inconsistent at $t_2^{(1)}$ since $S_2^{(1)}(t) > S_1^{(1)}(t)$ and at $t_1^{(1)}$ since $S_1^{(1)}(t) > S_2^{(1)}(t)$. Bottom-right: locally inconsistent at $t_2^{(2)}$, since $S_2^{(2)}(t) > S_1^{(2)}(t)$.

Figure 7 shows a visual example of global and local consistency with respect to the survival function. Global consistency is satisfied between instances if the instance-specific curves align with the observed events. Local consistency is satisfied within an instance if the event-specific curves align with the observed events.

BS/IBS: The Brier Score (BS) is defined as the mean squared difference between the predicted survival curve and the Heaviside step function of the observed event. The integrated Brier Score (Graf et al., 1999) (IBS) aggregates the Brier Scores across multiple time points to provide a single measure of model performance. We use inverse probability weighting (IPCW) to handle censored events. The BS is defined as:

$$\mathrm{BS}(t^*) = \frac{1}{N}\sum_{i\in\mathcal{D}}\left[\frac{S(t^* \mid \mathbf{x}^{(i)})^2 \cdot \mathbb{1}[t^{(i)} \le t^*, \delta^{(i)} = 1]}{S_C(t^{(i)})} + \frac{(1 - S(t^* \mid \mathbf{x}^{(i)}))^2 \cdot \mathbb{1}[t^{(i)} > t^*]}{S_C(t^*)}\right], \tag{24}$$

which is the mean square error between observed survival status and survival probability at time t^*, and $S_C(t^*)$ is the non-censoring probability at time t^*. The IBS is defined as:

$$\mathrm{IBS} = \frac{1}{N}\sum_{i\in\mathcal{D}}\frac{1}{t_{\max}} \cdot \int_0^{t_{\max}} \mathrm{BS}(t)\, dt, \tag{25}$$

where $t_{\max}$ is the maximum observed time.

MAE: The mean absolute error (MAE) is the absolute difference between the predicted and actual survival times. Given an individual survival distribution, $S(t \mid \boldsymbol{x}_i) = \Pr(T > t \mid \boldsymbol{x}_i)$, we calculate the predicted survival time $\hat{t}^{(i)}$ using the median time of the predicted survival distribution (Qi et al., 2023b):

$$\hat{t}^{(i)} \;=\; \text{median}\,(S(t \mid \boldsymbol{x}_i)) = S^{-1}(\tau = 0.5 \mid \boldsymbol{x}_i), \tag{26}$$

$$\text{MAE}\,(\hat{t}_i, t_i, \delta_i = 1) \;=\; \mid t_i - \hat{t}_i \mid. \tag{27}$$

For censored individuals, we calculate the marginal MAE (mMAE), as proposed by Haider et al. (2020):

$$\text{mMAE} = \frac{1}{\sum_{i=1}^{N} \omega^{(i)}} \sum_{i=1}^{N} \omega^{(i)} \left| [(1-\delta^{(i)}) \cdot e_{\text{m}}(t^{(i)}) + \delta^{(i)} \cdot t^{(i)}] - \hat{t}^{(i)} \right|,$$
$$\text{where} \quad e_{\text{m}}(t^{(i)}) = \begin{cases} t^{(i)} + \frac{\int_{t^{(i)}}^{\infty} S_{\text{KM}}(t)dt}{S_{\text{KM}}(t)} & \text{if} \quad \delta^{(i)} = 0 \\ t^{(i)} & \text{if} \quad \delta^{(i)} = 1 \end{cases},$$
$$\text{and} \quad \omega^{(i)} = \begin{cases} 1 - S_{\text{KM}}(t^{(i)}) & \text{if} \quad \delta^{(i)} = 0 \\ 1 & \text{if} \quad \delta^{(i)} = 1 \end{cases}. \tag{28}$$

D-calibration: Distribution calibration (D-calibration) (Haider et al., 2020) measures the calibration performance of $S(t)$, expressing to what extent the predicted probabilities can be trusted. We assess this using a Pearson's chi-squared goodness-of-fit test. For any probability interval $[a, b] \in [0, 1]$, we define $D_m(a, b)$ as the group of individuals in the dataset D whose predicted probability of event is in the interval $[a, b]$ (Qi et al., 2023a). A model is D-calibrated if the proportion of individuals $|D_m(a, b)|/|D|$ is statistically similar to the amount $b - a$. In the results, we count the number of times a model is D-calibrated (*i.e.*, $p > 0.05$) for each event over 10 experiments.

Appendix E. Experimental details

E.1. Datasets and preprocessing

Table 7 shows an overview of the survival datasets used in this work.

Table 7: Overview of the raw datasets. Symbols indicate single-event ($*$), competing-risks (†) or multi-event (‡) task. Number of events (K) is denoted for the respective task. In SEER, "D", "BC" and "HF" denote the death event, death by breast cancer and death by heart failure. In MIMIC-IV, "A", "S", and "D" denote the ARF, shock and death events. In Rotterdam, "R" denotes relapse and "D" denotes death. In PRO-ACT, "SP", "SW", "HA" and "WA" denote the speech, swallowing, handwriting and walking events. In EBMT, "R", "AE", "RAE", "REL" and "D" denote the recovery, adverse events, recovery and adverse events, relapse, and death events.

Dataset	#Instances (N)	#Features (d)	#Events (K)	Max t	Event distribution (%)
SEER$*$†	19,246	17	1/2	121	D: 45.5 BC: 12.4, HF: 45.5
MIMIC-IV$*$‡	24,516	100	1/3	4,686	D: 37.2 A: 20.7, S: 19.2, D: 37.2
Rotterdam†‡	2,982	10	2/3	7,043	R: 6.5, D: 50.9 R: 50.9, D: 42.7
PRO-ACT‡	3,220	8	4	498	SP: 37.8, SW: 31.9, HA: 50.3, WA: 60.6
EBMT‡	2,279	6	5	6229	R: 34.4, AE: 39.8, RAE: 29, REL: 16.2, D: 23.4

SEER: The U.S. Surveillance, Epidemiology, and End Results (SEER) dataset (Ries et al., 2003) is a comprehensive collection of cancer patient data from approximately 49% of the U.S. population. It provides information about survival times for cancer patients post-diagnosis. For the single-event case, we consider the outcome of death due to breast cancer, and for the competing-risks case, we consider the outcomes of death due to breast cancer and death due to heart failure, and consider other outcomes as censored. Concerning preprocessing, we select a cohort of 19,246 newly-diagnosed patients (first year of disease) with 17 features,

which include demographics and tumor characteristics. We use the raw feature "COD to site recode" to extract the event indicator and apply some special value processing for some of the features, based on the guideline provided by Wang and Sun (2022). For categorical features, missing values are imputed using the mode, followed by one-hot encoding. The data is then standardized to ensure consistency in our analyses. The SEER cohort is available for download at www.seer.cancer.gov.

MIMIC-IV: The Medical Information Mart for Intensive Care (MIMIC-IV) dataset by Johnson et al. (2023) comprises electronic health records from patients admitted to intensive care units at the Beth Israel Deaconess Medical Center in the U.S., covering the years from 2008 to 2019. For the single-event case, we consider death since hospital admission and for the multi-event case, we consider the outcomes of acute respiratory failure (ARF), shock and death. Concerning preprocessing, we use the method described by Gupta et al. (2022) to extract 1,672 static features from the dataset, for 26,236 patients whose ages are between 60 and 65 years. To refine our feature set, we apply the unicox feature selection method (Qi et al., 2022; Simon et al., 2003), reducing the number of features to the 100 that appear to be the most relevant. In the training set, we assessed the statistical significance of each feature by calculating the p-values using lifelines's (Davidson-Pilon, 2019) CoxPHFitter summary. The top 100 features with the smallest p-values were selected. The final set of features is included in the source code. The MIMIC-IV v2.2 is available for download at https://physionet.org/content/mimiciv/2.2/.

Rotterdam: The Rotterdam dataset (Royston and Altman, 2013) includes 2,982 primary breast cancer patients from the Rotterdam tumor bank, with 1,546 having node-positive disease. The follow-up time ranged from 1 to 231 months (median 107 months). The dataset has 10 features, which include demographics, tumor characteristics, and treatment information. In the competing-risks case, we consider the outcome of recurrence-free survival time (RFS), which is defined as the time from primary surgery to the earlier of disease recurrence or death from any cause. In the multi-event case, we consider the time to death or disease recurrence, where death censors disease recurrence, but disease recurrence does not censor death. Figure 8 shows the event trajectories in the Rotterdam dataset. Concerning preprocessing, we converted the categorical size groups ('$\leq$ 20', '20-50', '$>$ 50') to their approximate median values (10, 35, and 75, respectively). The Rotterdam dataset is available at https://rdrr.io/cran/survival/man/rotterdam.html.

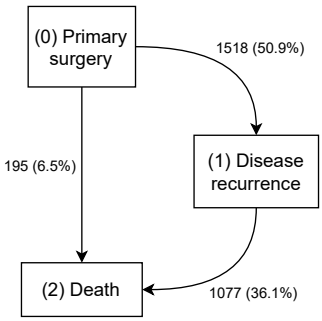


Figure 8: Event trajectories in the Rotterdam dataset (N=2,982).

PRO-ACT: The Pooled Resources Open-Access Clinical Trials (PRO-ACT)[6] database by Atassi et al. (2014) is the largest ALS dataset in the world, which includes patient demographics, lab and medical records and family history of over 11,600 ALS patients, where ALS disease state is evaluated by the ALS functional rating scale revised (ALSFRS-R) (Cedarbaum et al., 1999). The assessment contains 12 questions and a maximum score of 48, each representing physical functionality, *e.g.*, the ability to walk. Each ALSFRS-R score is between 0 and 4, where 0 represents complete inability with regard to the function, and 4

6. Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members.

represents normal function. We use the ALSFRS-R scores to formulate four events: "Speech", "Swallowing", "Handwriting" and "Walking". We consider an event to have occurred if a patient scores 2 or lower on the ALSFRS-R scale during any subsequent follow-up visitation following their baseline visitation. The events can occur in any order and are not mutually exclusive. ALSFRS-R scores are generally monotonic decreasing, but due to the subjectivity of the assessment, scores can occasionally increase post-baseline. We use a maximum follow-up time of 500 days and exclude patients with no recorded history. The PRO-ACT dataset is available for download at https://ncri1.partners.org/ProACT.

EBMT: The European Society for Blood and Marrow Transplantation (EBMT) dataset (Rossman et al., 2022) consists of 2,279 patient records who underwent blood or marrow transplantation between 1985 and 1998. The dataset records survival times after transplantation for patients suffering from blood cancer. We consider five events in our analysis: "Recovery" (Rec), "Adverse Event" (AE), a combination of both (AE and Rec), "Relapse", and "Death". These events allowed us to evaluate their individual and combined effects on patient prognosis. The dataset includes six features belonging to four baseline covariates: donor-recipient match (where a gender mismatch is defined as a female donor and a male recipient), prophylaxis type, year of transplant, and age at transplant. Concerning preprocessing, we encode the year of transplantation as boolean variables indicating whether the transplant took place during 1990-1994 and 1995-1998, and we classify the age at the transplantation into boolean variables indicating whether the age was less than 20, between 20 and 40, or more than 40 years. The EBMT dataset is available in the PyMSM package at https://github.com/hrossman/pymsm.

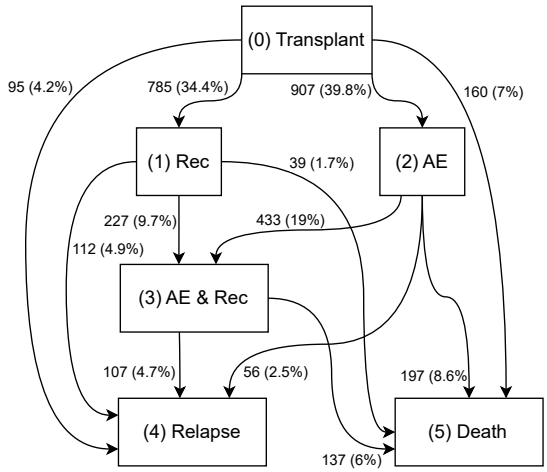


Figure 9: Event trajectories in the EBMT dataset (N=2,279).

E.2. Baselines models

CoxPH: The Cox Proportional Hazards model (CoxPH) is a widely used semiparametric model for survival analysis (Cox, 1972). It assumes that the hazard function for an individual i with covariates $\boldsymbol{x}^{(i)}$ can be expressed as $h\big(t|\boldsymbol{x}^{(i)}\big) = h_0(t)\exp\big(f\big(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}\big)\big)$, where $h_0(t)$ is an unspecified baseline hazard function, and f is typically a linear function $f(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}) = \boldsymbol{x}^{(i)\top}\boldsymbol{\theta}$ producing a risk score. The key assumption of CoxPH is that covariates have a multiplicative effect on the hazard and that this effect is constant over time (*i.e.*, proportional hazards). Model parameters $\boldsymbol{\theta}$ are estimated by maximizing the Cox partial log-likelihood, which does not require modeling $h_0(t)$ explicitly. The model is implemented from the *scikit-survival* package (Pölsterl, 2020).

CoxNet: CoxNet is an extension of the Cox Proportional Hazards model that incorporates Elastic Net regularization into the partial log-likelihood optimization (Simon et al., 2011). The hazard function retains

the same form as in CoxPH, $h\big(t|\boldsymbol{x}^{(i)}\big) = h_0(t)\exp\big(\boldsymbol{x}^{(i)\top}\boldsymbol{\theta}\big)$, but the parameter estimation adds a penalty term $\omega\left[\alpha\|\boldsymbol{\theta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\theta}\|_2^2\right]$, where $\omega > 0$ controls the overall strength of regularization, and $\alpha \in [0,1]$ controls the balance between L1 (lasso) and L2 (ridge) penalties. This approach enables feature selection (via L1) while maintaining stability in the presence of multicollinearity (via L2), making it particularly effective in high-dimensional settings. The model is implemented from the *scikit-survival* package (Pölsterl, 2020).

Weibull AFT: The Weibull Accelerated Failure Time (AFT) model is a parametric survival model in which survival times are assumed to follow a Weibull distribution whose parameters depend on covariates. In the absence of covariates, the scale parameter can be interpreted as the time at which 37% of the population has experienced the event, while the shape parameter determines whether the cumulative hazard is convex or concave, corresponding to accelerating or decelerating hazards. The cumulative hazard function has the closed form $H(t) = (t/\beta)^{\eta}$, where $\beta > 0$ is the scale parameter and $\eta > 0$ is the shape parameter. When covariates $\boldsymbol{x}^{(i)}$ are present, the scale parameter is modeled as $\beta^{(i)} = \exp(-\boldsymbol{x}^{(i)\top}\boldsymbol{\theta})$ and, optionally, the shape parameter may also depend on $\boldsymbol{x}^{(i)}$. Model parameters are estimated by maximizing the full likelihood. The model is implemented from the *lifelines* package (Davidson-Pilon, 2019).

GBSA: Gradient Boosted Survival Analysis (GBSA) (Ridgeway, 1999) is a shallow ensemble model that extends gradient boosting methods to survival settings by optimizing a penalized version of the Cox partial log-likelihood. Specifically, it employs an ℓ_2-norm regularization to prevent overfitting and enhance generalization. Unlike classical gradient boosting, which typically updates a single weak learner or fits all features to the negative gradient, GBSA uses an offset-based approach to flexibly update multiple candidate features at each boosting iteration. This leads to more efficient gradient approximation and model convergence. The model is implemented from the *scikit-survival* package (Pölsterl, 2020).

RSF: Random Survival Forests (RSF) (Ishwaran et al., 2008) adapt the random forest framework to survival analysis by building an ensemble of survival trees, each trained on a bootstrap sample of the data. Each tree recursively partitions the feature space using a survival-specific splitting criterion—typically the log-rank test—to maximize differences in survival outcomes between child nodes. RSF naturally handles nonlinear covariate effects and complex interactions, and it is well-suited for high-dimensional settings without requiring strong parametric assumptions. Censoring is handled natively by the splitting criterion and ensemble aggregation. The model is implemented from the *scikit-survival* package (Pölsterl, 2020).

NFG: Neural Fine-Gray (NFG) (Jeanselme et al., 2023) extends the classical sub-distribution hazard approach of Fine and Gray (1999) to a flexible neural-network framework for competing risks. The architecture employs a shared embedding network for covariates, followed by separate monotonic neural networks that directly model each instance's CIF, ensuring non-decreasing behavior over time through weight constraints. The CIFs are combined using a risk-balancing mechanism so that their sum remains less than or equal to one, preserving a valid joint probability structure. Training maximizes the exact log-likelihood derived analytically from the CIFs via automatic differentiation, avoiding numerical integration. This formulation allows the model to capture non-linear covariate effects and complex, risk-specific dynamics without assuming a parametric hazard form. The model is implemented from `https://github.com/Jeanselme/NeuralFineGray`.

DeepSurv: DeepSurv (Katzman et al., 2018) is a deep learning extension of the Cox Proportional Hazards model that replaces the linear risk function with a nonlinear neural network. The individual hazard function takes the form $h\big(t|\boldsymbol{x}^{(i)}\big) = h_0(t)\exp\big(f\big(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}\big)\big)$, where $f(\boldsymbol{\theta}, \boldsymbol{x}^{(i)})$ is a feedforward neural network that maps covariates to a continuous risk score. The model is trained by numerically maximizing the Cox partial log-likelihood with respect to the network parameters $\boldsymbol{\theta}$. By capturing nonlinear interactions among covariates, DeepSurv can improve predictive performance over traditional Cox models in complex data settings. The model is implemented from `https://github.com/shi-ang/BNN-ISD/`.

DeepHit: DeepHit (Lee et al., 2018) is a deep discrete-time model designed to handle competing risks, where the occurrence of one event prevents the occurrence of another, as it assumes $\sum_{t=1}^{t_{\max}}\sum_{k=1}^{K} p_k[t] = 1$, where $t_{\max} \in \mathbb{N}_+$ is the number of discrete time bins (the maximum discretized time index) in the prediction horizon and $K \in \mathbb{N}_+$ is number of competing events (causes). DeepHit contains a shared covariate layer among all events and cause-specific subnetworks for each competing event. The model outputs a probability distribution for each discrete time bin and each competing event via a softmax layer. This architecture allows DeepHit to estimate the joint distribution of survival times and events directly without making

assumptions about the underlying stochastic process. It accounts for dependencies among events in $\boldsymbol{\theta}$ but not for individuals who experience no event by time T. The model is implemented from https://github.com/chl8856/DeepHit.

Hierarchical: The hierarchical model (Tjandra et al., 2021) is a deep discrete-time survival model, designed to handle multiple events by predicting event probabilities across multiple temporal granularities. It begins with coarse-scale predictions (*e.g.*, monthly) and progressively refines them into finer ones (*e.g.*, daily), effectively decomposing long-horizon tasks into simpler intermediate steps. Formally, the model is a multi-task network that takes input x and produces $\hat{P}$, a set of event-specific probability distributions $\hat{p}_k$, computed hierarchically by subnetworks ϕ_k. At granularity m, the probability $\hat{p}_{km}$ is derived from both $\psi_{km}(\Phi(x))$ and the preceding level $\hat{p}_{k,m-1}$, enforcing top-down consistency. During training, a composite likelihood with ranking penalties provides supervision across the event horizon, optimized only for discriminative performance using Harrell's C-index (Harrell Jr et al., 1996). The model is implemented from https://github.com/MLD3/Hierarchical_Survival_Analysis.

MTLR: The Multi-Task Logistic Regression (MTLR) model (Yu et al., 2011) is a discrete-time survival model that estimates the full survival distribution by training a sequence of dependent logistic regression models—one for each discrete time interval. Unlike models that rely on hazard functions, MTLR directly estimates the probability of survival at each time step, making it flexible and naturally suited for handling censored data. The model captures temporal dependencies between intervals by sharing parameters across the logistic regressors. MTLR has been extended to the competing-risk setting as Deep-CR MTLR (Kim et al., 2021), which replaces the logistic regressors with neural networks to model nonlinear effects and inter-event dependencies. MTLR is implemented from https://github.com/shi-ang/BNN-ISD/ and Deep-CR MTLR is implemented from https://github.com/bhklab/aaai21_survival_prediction.

DSM: Deep Survival Machines (DSM) (Nagpal et al., 2021) is a deep continuous-time survival model that estimates the survival function as a weighted mixture of k parametric survival distributions (*e.g.*, Weibull or LogNormal). Both the parameters of each distribution and their associated mixture weights are modeled as nonlinear functions of the input covariates, using a shared multilayer perceptron (MLP). This mixture-based formulation allows DSM to capture heterogeneity in the event-time distributions. The model is trained by maximizing the likelihood of observed and censored data. The model is implemented from the *auton-survival* package (Nagpal et al., 2022).

E.3. Hyperparameter settings

Tables 8-19 list the hyperparameter settings used for each model across all datasets. These values were selected via Bayesian optimization (Snoek et al., 2012) over ten iterations, adopting the hyperparameters leading to the highest Harrell's C-index in the validation set (seed = 0). When applicable, early stopping was used based on the validation loss. Note that the default configuration is used to evaluate computational complexity only.

Table 8: Hyperparameters for CoxPH.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Ties method	breslow	breslow	breslow	breslow	breslow	breslow
Regularization	0	10	10	0.01	0	0.01
Iterations	100	200	100	500	500	100
Tolerance	1e-9	1e-5	1e-5	1e-7	1e-9	1e-5

Table 9: Hyperparameters for CoxNet.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Number of alphas	100	150	150	150	50	150
Alpha min ratio	auto	0.0001	0.0001	0.0001	0.001	0.0001
L1 ratio	0.5	0.1	1	0.75	1	0.25
Tolerance	1e-7	1e-05	1e-7	1e-06	1e-7	1e-05
Max iterations	100000	100000	200000	100000	200000	200000

Table 10: Hyperparameters for Weibull AFT.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Penalizer	0	0.001	0.0001	0.1	0.001	0
L1 Ratio	0	1	1	0	0.5	0

Table 11: Hyperparameters for GBSA.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Number of estimators	100	50	800	800	800	50
Learning rate	0.1	1	1	1	1	1
Max depth	1	10	1	10	10	3
Min samples split	10	2	5	10	10	2
Min samples leaf	5	10	5	1	10	1
Max features	`sqrt`	`sqrt`	None	`sqrt`	None	None
Dropout rate	0.0	0.5	0.5	0.5	0.5	0.25
Subsample	0.8	1.0	1.0	0.25	0.25	0.25

Table 12: Hyperparameters for RSF.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Number of estimators	100	500	200	200	500	1000
Max depth	1	10	5	5	5	10
Min samples split	10	2	5	10	10	2
Min samples leaf	5	10	5	1	10	1
Max features	`sqrt`	`sqrt`	`log2`	`log2`	None	`log2`

Table 13: Hyperparameters for NFG.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Layers	[32]	[64]	[128]	–	–	–
Activation	ReLU	ReLU	ReLU	–	–	–
Survival layers	[32]	[32]	[32]	–	–	–
Dropout	0.0	0.25	0.5	–	–	–
Optimizer	Adam	Adam	Adam	–	–	–
Multihead	True	True	True	–	–	–

Table 14: Hyperparameters for DeepSurv.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Hidden size	32	128	64	32	32	64
Learning rate	0.001	0.005	0.001	0.0005	0.005	0.005
Weight decay	0.01	0.01	0.1	0.01	0	0.1
Dropout rate	0.25	0	0.5	0	0	0.5
Epochs	1000	2000	2000	1000	2000	500
Early stopping	True	True	True	True	True	True
Patience	10	10	20	10	20	20

Table 15: Hyperparameters for DeepHit.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Shared layer nodes	32	16	16	32	64	64
Individual layer nodes	32	64	64	64	128	128
Dropout rate	0.25	0.25	0.25	0.5	0.1	0.1
Learning rate	0.001	0.0001	0.005	0.0001	0.0005	0.0005
Weight decay	0.01	0.1	0	0.1	0	0.01
Alpha	0.2	0.4	0.4	0	0	0.8
Sigma	0.1	0.5	0.05	0.1	0.1	0.1
Eta multiplier	0.8	1.0	1.0	1.0	0.5	1.0

Table 16: Hyperparameters for Hierarchical.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Theta layer size	[32]	[50]	[200]	[200]	[200]	[50]
Fine bin layer size	[32, 5]	[50, 5]	[100, 10]	[150, 10]	[150, 10]	[100, 10]
Learning rate	0.001	0.0005	0.001	0.001	0.001	0.001
Regularization constant	0.01	0.01	0.01	0.01	0.01	0.01
Alpha	0.0001	1e-5	1e-5	0	0.0001	0.001
Sigma	10	10	5	20	5	5

Table 17: Hyperparameters for MTLR.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Learning rate	0.001	0.0001	0.001	0.0001	0.0001	0.0005
Weight decay	0.01	0.001	0.001	0.1	0	0.1
Dropout rate	0.25	0	0	0.1	0.1	0.25

Table 18: Hyperparameters for DSM.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Hidden layers	32	32	32	64	64	64
Learning rate	0.001	0.005	0.005	0.005	0.005	0.001
Iterations	10000	20000	10000	10000	10000	10000
Batch size	32	32	64	32	64	64
Number of dists.	3	5	1	5	5	1

Table 19: Hyperparameters for MENSA.

Hyperparameter	Default	SEER	Rotterdam	MIMIC-IV	PRO-ACT	EBMT
Hidden nodes	32	128	32	32	16	128
Learning rate	0.001	0.0005	0.001	0.001	0.0001	0.001
Weight decay	0.01	0.005	0.001	0.0001	0.001	0.0001
Dropout rate	0.25	0	0.25	0.25	0.1	0.1
Batch size	32	128	32	64	32	32
Number of distributions (Ψ)	1	3	3	1	3	3
Weighting parameter (λ)	-	-	0.25	-	-	0.5

E.4. Implementation details

To train MENSA, we optimize the joint likelihood of event times and event trajectories using stochastic gradient descent (SGD). The algorithm operates over mini-batches of survival data, computing both the multi-event likelihood and the trajectory likelihood at each iteration, and updating the shared and event-specific parameters accordingly. Algorithm 1 summarizes the training procedure, including initialization, loss computation, and parameter updates over the training epochs. All experiments were run in Python 3.9 with PyTorch 1.13.1, NumPy 1.24.3, and Pandas 1.5.3 on a workstation with an Intel Core i9-10980XE (3.00 GHz), 64 GB RAM, and an NVIDIA RTX 3090 GPU (CUDA 11.7). All datasets are publicly available, and the code for loading and preprocessing the datasets is provided in the source code repository.

Appendix F. Statistical and practical significance

To assess whether performance differences between MENSA and baseline models were statistically and practically meaningful, we conducted paired tests across all datasets and metrics. For each dataset-metric pair, we compared MENSA to every baseline model using per-seed results averaged across events.

Statistical testing. We performed two-sided paired t-tests using the per-seed metric values, treating runs with the same random seed as paired observations. The resulting p-values were adjusted for multiple comparisons using the Holm correction method ($\alpha = 0.05$). Models whose adjusted p-values fell below the

Algorithm 1 Training algorithm for MENSA.

Input: A survival dataset $\mathcal{D}$ of the form $\{(X^{(i)}, T_{\text{obs}}^{(i)}, \delta^{(i)})\}_{i=1}^{N}$; $\mathcal{M}$: the model with initialized parameters $\{\Phi_\theta, \tilde{\beta}, \tilde{\eta}, \zeta, \xi, g_\theta\}$; $\mathcal{T}$: list of trajectories; κ: learning rate; L: number of training epochs; λ: weighting hyperparameter.
Result: $\mathcal{M} = \{\hat{\Phi}_\theta, \hat{\tilde{\beta}}, \hat{\tilde{\eta}}, \hat{\zeta}, \hat{\xi}, \hat{g}_\theta\}$: trained model parameters.
Initialization: $\mathcal{M} \leftarrow \texttt{Instantiate}(\mathcal{M})$
for $i = 1, \ldots, L$ **do**
  $[\log f(T_{\text{obs}}), \log S(T_{\text{obs}})] \leftarrow \mathcal{M}(\mathcal{D})$ // $\log f$ and $\log S$ are $N \times P$
  $\mathcal{L}_{\text{ME}} \leftarrow -\sum_{p \in \mathcal{P}} w_p \big[\delta_p \log f_p(T_{\text{obs}}) + (1 - \delta_p) \log S_p(T_{\text{obs}})\big]$ // Multi-event loss
  $\mathcal{L}_{\text{trajectory}} \leftarrow 0$
  for each pair $(A, B) \in \mathcal{T}$ **do**
    $\mathcal{L}_{\text{trajectory}} \mathrel{+}= \sum_i \delta_A^{(i)} \delta_B^{(i)} \big[- \log S_B(T_A^{(i)})\big]$ // Accumulate trajectory loss
  end for
  $\mathcal{L}_{\text{total}} \leftarrow (1 - \lambda) \frac{1}{N} \mathcal{L}_{\text{ME}} + \lambda \frac{1}{N} \mathcal{L}_{\text{trajectory}}$ // Weighted total loss
  $\{\hat{\Phi}_\theta, \hat{\tilde{\beta}}, \hat{\tilde{\eta}}, \hat{\xi}, \hat{\zeta}, \hat{g}_\theta\}^{(i)} \leftarrow \text{SGD}\big(\{\cdot\}^{(i-1)}, \nabla \mathcal{L}_{\text{total}}, \kappa\big)$ // Update step
end for
return $\mathcal{M}$

threshold were considered significantly different from MENSA. The correction was applied independently for each dataset–metric combination.

Practical significance. In addition to statistical testing, we required a minimum effect size of interest (MEI) to ensure that reported differences were not only statistically significant but also practically meaningful. For all discrimination (CI_G, CI_L, AUC) and calibration (IBS, mMAE) metrics, the MEI threshold was set to 1.0 point on the 0-100 scale used for evaluation. For metrics where higher is better (C-indices and AUC), a baseline was considered practically worse if its mean was at least 1.0 point lower than MENSA's mean. For metrics where lower is better (IBS, mMAE), the baseline had to exceed MENSA's mean by at least 1.0 point to be considered practically worse.

Reporting. A dagger symbol (†) in the tables indicates models that were both statistically significant (Holm-corrected $p < 0.05$) and practically worse than MENSA according to the MEI criterion. D-calibration was excluded from dagger marking since it is reported as a count of calibrated experiments rather than a continuous measure. The full analysis pipeline – including scaling conventions, dataset ordering, and Holm-adjusted significance checks – is implemented in a source code repository.

Appendix G. Additional results

G.1. Single-event results

Table 20 shows the discrimination and calibration results on two single-event datasets. In the SEER dataset, MENSA demonstrates solid discriminative ability, ranking third in Harrell's C-index and second in AUC, while also achieving the second-best IBS and mMAE scores, which underscores its accuracy in survival prediction. In the MIMIC-IV dataset, MENSA shows similarly strong performance, obtaining the second-best Harrell's C-index, the best overall AUC and IBS, and the lowest mMAE among the deep learning baselines. Taken together, these findings indicate that MENSA is not only competitive with shallow single-event survival methods but also surpasses existing deep learning approaches on several key metrics. This highlights its effectiveness for both risk discrimination and time-to-event prediction in large, single-event datasets.

Table 20: Discrimination and calibration performance (mean ± SD.) between different models on single-event benchmark datasets, averaged over 10 experiments. For Harrell's CI, AUC, and D-calibration, higher (↑) is better. For IBS and mMAE, lower (↓) is better. D-calibration counts the number of experiments in which the respective model was D-calibrated.

Dataset	Model	Harrell's CI	AUC	IBS	mMAE	D-Cal
SEER ($K=1$)	CoxPH	69.8±0.61	74.3±0.78	16.8±0.30	29.9±0.39	4/10
	CoxNet	69.8±0.62	74.3±0.79	16.8±0.30	29.8±0.37	4/10
	Weibull	69.8±0.62	74.3±0.78	16.9±0.30	34.3±0.82	10/10
	GBSA	77.1±0.53	81.7±0.62	15.1±0.28	25.6±0.33	1/10
	RSF	77.1±0.55	82.1±0.69	14.2±0.29	29.1±0.62	8/10
	MTLR	70.5±0.58	75.7±0.81	16.4±0.30	29.9±0.70	10/10
	DeepSurv	76.0±0.70	80.7±0.78	14.8±0.35	29.6±0.84	0/10
	DeepHit	75.8±0.79	80.8±0.93	14.8±0.38	26.7±0.46	2/10
	DSM	75.9±0.93	80.8±1.03	14.9±0.47	42.3±2.31	7/10
	MENSA (Ours)	76.0±0.65	81.0±0.77	14.8±0.36	27.1±0.71	8/10
MIMIC-IV ($K=1$)	CoxPH	74.2±0.48	78.7±0.60	12.9±0.38	5.9±0.03	10/10
	CoxNet	74.3±0.48	78.8±0.61	12.8±0.38	5.9±0.04	10/10
	Weibull	74.5±0.49	79.1±0.63	13.0±0.38	5.8±0.05	10/10
	GBSA	74.9±0.43	79.5±0.56	13.9±0.39	5.9±0.05	8/10
	RSF	72.5±0.47	78.9±0.56	14.2±0.40	6.0±0.06	8/10
	MTLR	74.3±0.52	79.5±0.77	12.5±0.36	7.1±0.24	9/10
	DeepSurv	75.1±0.43	80.0±0.60	12.4±0.39	6.8±0.10	10/10
	DeepHit	74.5±0.49	79.7±0.75	20.5±0.57	11.7±0.14	0/10
	DSM	74.4±0.68	79.2±0.93	13.2±0.44	6.4±0.56	5/10
	MENSA (Ours)	75.4±0.46	80.1±0.58	12.4±0.34	6.1±0.05	4/10

G.2. Competing-risks results

Table 21 shows the discrimination and calibration results on two competing-risks datasets. In SEER, MENSA performs competitively across all metrics, combining strong discrimination with low prediction error and good calibration. In Rotterdam, MENSA also provides solid discrimination and competitive calibration, though with somewhat higher prediction error compared to the best-performing single-event baselines. However, it has the third-lowest mMAE in this dataset as well. MENSA also achieves good D-calibration results across both datasets. These findings suggest that MENSA extends robustly to competing-risk settings, offering a balanced trade-off between discrimination and calibration performance.

Appendix H. Computational analysis

Table 22 compares selected deep learning baselines and MENSA in terms of space complexity. This analysis is conducted in single-event, competing-risks and multi-event scenarios. For a fair comparison, we configure all models with a single hidden layer with 32 nodes. By nature of design, models that cannot learn events jointly have to be trained K times in the multi-event scenario. In this case, the DeepSurv model, as an example, has a $K\mathcal{O}(d)$ space complexity, where d is the number of features, assuming training is not done in parallel. This can lead to significant increases in the number of trainable parameters in high-dimensional datasets with many events, making it a less-attractive solution. MENSA does not suffer from this problem, as it can train on multiple events at once with a space complexity of $\mathcal{O}(P(\Psi + d) + d)$.

Figure 10 shows the number of trainable parameters and FLOPs (computed using the `fvcore` library[7]) as a function of the number of features. In the single-event case, MENSA uses fewer parameters and FLOPs than Hierarchical, MTLR, and DeepHit, and only slightly more than DeepSurv and DSM. In competing-risks, MENSA remains more efficient than most models except DSM, whose overhead is marginally lower. For

7. Meta Research's fvcore library, available at https://github.com/facebookresearch/fvcore (September 1, 2025).

Table 21: Discrimination and calibration performance (mean ± SD.) between different models on competing-risks benchmark datasets, averaged over 10 experiments. For global CI, local CI, and AUC, higher (↑) is better. For IBS and mMAE, lower (↓) is better. D-calibration counts the number of experiments in which the respective model was D-calibrated.

Dataset	Model	Global CI	Local CI	AUC	IBS	mMAE	D-Cal
SEER ($K = 2$)	CoxPH	67.9±0.61	78.4±0.71	66.9±0.54	13.3±0.18	29.0±0.60	14/20
	CoxNet	68.8±0.61	78.2±0.60	64.8±0.62	13.4±0.18	26.0±0.71	14/20
	Weibull	67.9±0.61	78.5±0.71	66.8±0.56	13.5±0.18	54.7±0.94	20/20
	GBSA	74.8±0.54	78.5±0.70	72.7±0.67	12.4±0.15	23.8±0.68	11/20
	RSF	74.4±0.56	81.0±0.49	73.7±0.66	12.0±0.16	38.8±1.78	18/20
	NFG	72.8±0.55	80.0±0.46	68.6±0.61	12.6±0.16	185.8±11.84	7/20
	MTLR-CR	67.4±0.70	73.8±1.06	66.1±0.82	16.0±0.39	40.6±0.73	0/20
	DeepSurv	73.4±0.61	79.5±0.56	70.3±0.47	12.3±0.18	31.7±2.54	11/20
	DeepHit	71.5±0.42	78.5±0.66	68.5±0.58	12.7±0.14	135.2±15.01	8/20
	DSM	73.0±0.73	79.2±1.00	70.1±0.65	12.5±0.24	74.7±9.09	15/20
	Hierarch.	71.4±1.00	69.8±1.33	68.2±1.10	15.4±0.40	38.0±1.48	0/20
	MENSA (Ours)	72.9±0.66	78.7±0.80	69.9±0.66	12.4±0.19	26.0±0.89	17/20
Rotterdam ($K = 2$)	CoxPH	69.0±1.49	86.9±1.85	78.8±1.57	13.8±2.71	252.2±84.82	20/20
	CoxNet	69.0±1.50	86.9±1.79	78.9±1.49	13.8±2.67	307.6±125.96	20/20
	Weibull	69.1±1.56	88.4±1.49	78.9±1.61	13.9±3.09	420.0±72.95	15/20
	GBSA	75.2±1.85	88.9±1.53	71.8±2.47	14.7±2.58	21.0±1.30	20/20
	RSF	69.7±1.49	88.6±1.29	78.6±1.77	13.4±2.35	224.4±58.86	20/20
	NFG	66.2±3.86	89.0±1.57	75.3±3.80	15.1±3.05	810.1±129.17	13/20
	MTLR-CR	68.4±1.81	86.2±1.69	78.3±1.73	16.0±3.89	54.3±7.57	3/20
	DeepSurv	69.1±1.62	88.3±1.93	78.7±1.78	13.9±2.53	103.4±30.50	20/20
	DeepHit	62.6±2.52	88.8±1.50	77.0±2.32	15.1±2.74	33.3±5.26	15/20
	DSM	69.9±1.40	88.8±1.37	78.2±2.01	13.5±2.47	139.3±4.09	17/20
	Hierarch.	65.2±1.72	73.4±4.57	50.0±0.06	76.1±13.31	83.4±7.46	0/20
	MENSA (Ours)	67.3±1.71	88.9±1.54	77.1±1.84	15.2±5.23	48.9±7.46	18/20

Table 22: Comparison between the proposed model and selected state-of-the-art baselines in terms of space complexity. d: number of features, K: number of events, B: number of discrete time bins, M: levels of granularity, Ψ: number of mixture distributions, P: number of states.

Model	Single-event	Competing risks	Multi-event
DeepSurv	$\mathcal{O}(d)$	$K\mathcal{O}(d)$	$K\mathcal{O}(d)$
DeepHit	$\mathcal{O}(Bd)$	$\mathcal{O}(KBd)$	$K\mathcal{O}(Bd)$
MTLR	$\mathcal{O}(Bd)$	$\mathcal{O}(KBd)$	$K\mathcal{O}(Bd)$
DSM	$\mathcal{O}(\Psi + d)$	$\mathcal{O}(K\Psi + d)$	$K\mathcal{O}(\Psi + d)$
Hierarch.	$\mathcal{O}(MB + d)$	$\mathcal{O}(MKB + d)$	$\mathcal{O}(MKB + d)$
MENSA (Ours)	$\mathcal{O}(\Psi + d)$	$\mathcal{O}(P(\Psi + d) + d)$	$\mathcal{O}(P(\Psi + d) + d)$

datasets with over 100 features, differences are negligible. In the multi-event scenario, baseline models trained separately per event require K trainings, while MENSA trains all events jointly, achieving a notably lower parameter count and FLOPs than all baselines, including the Hierarchical model (see Table 23). Overall, MENSA is the most computationally efficient among deep learning models in multi-event settings.

Appendix I. Proof of non-proportional hazards

This section addresses how the proposed method overcomes the limitations associated with the proportional hazards assumption (Cox, 1972). We begin by reviewing how a traditional Weibull distribution adheres to the proportional hazards property and then show, intuitively, why MENSA, which uses a mixture of Weibull distributions, does not meet this assumption.

Table 23: Relative reduction in parameters and FLOPs for MENSA compared to deep learning baselines in the multi-event setting ($K = 4$) with 100 features.

Model	#Params. reduction	#FLOPs reduction
DeepSurv	1.4×	1.5×
Deephit	2.5×	2.5×
MTLR	3.2×	5.7×
DSM	1.5×	1.5×
Hierarch.	13×	13×

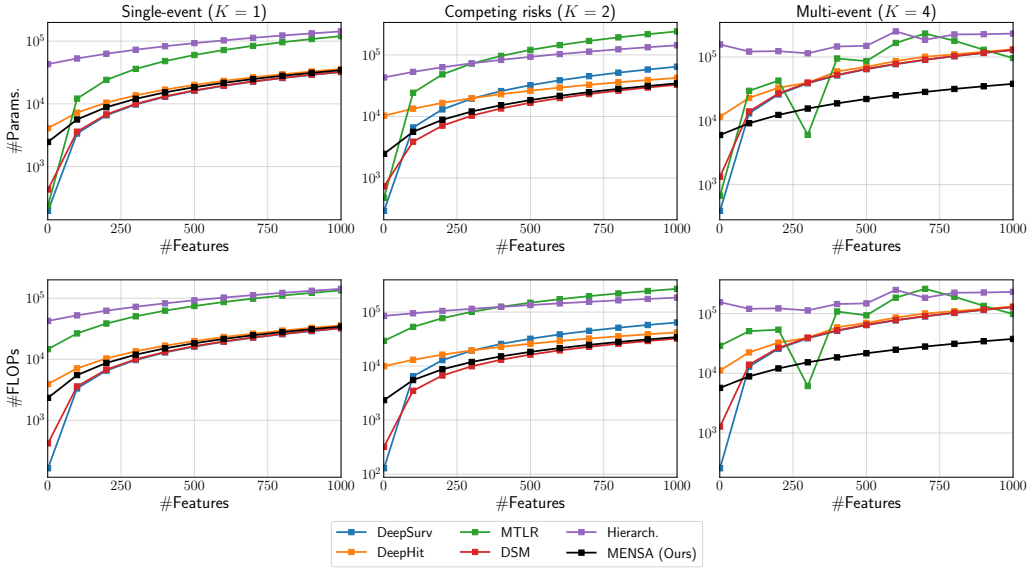


Figure 10: Number of features vs. number of trainable parameters (upper row) and number of floating-point operations (lower row) for single-event ($K = 1$), competing-risks ($K = 2$) and multi-event ($K = 4$) scenarios.

It is well-established that the Weibull distribution satisfies the proportional hazards assumption when its shape parameter, η, is independent of the covariates $\boldsymbol{x}^{(i)}$ (Klein and Moeschberger, 2006). The PDF, CDF, survival function, and hazard function for a Weibull distribution are given by:

$$f(t;\beta,\eta) = \frac{\eta}{\beta}\left(\frac{t}{\beta}\right)^{\eta-1}\exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \tag{29}$$

$$F(t;\beta,\eta) = \int_{\tau=0}^{t} f(\tau;\beta,\eta)\,d\tau = 1-\exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \tag{30}$$

$$S(t;\beta,\eta) = 1-F(t;\beta,\eta) = \exp\left(-\left(\frac{t}{\beta}\right)^{\eta}\right), \tag{31}$$

$$h(t;\beta,\eta) = \frac{f(t;\beta,\eta)}{S(t;\beta,\eta)} = \frac{\eta}{\beta}\left(\frac{t}{\beta}\right)^{\eta-1}. \tag{32}$$

To demonstrate that the Weibull distribution satisfies the proportional hazards assumption, we parameterize the scale parameter, β, as a function of the covariates $\boldsymbol{x}^{(i)}$:

$$\beta(\boldsymbol{x}^{(i)}) \;=\; \beta_0\exp(-\gamma^{\top}\boldsymbol{x}^{(i)}), \tag{33}$$

where β_0 is the baseline scale parameter, and γ is a vector of regression coefficients. Substituting $\beta(\boldsymbol{x}^{(i)})$ into the hazard function yields:

$$\begin{aligned} h(t;\boldsymbol{x}^{(i)}) &= \frac{\eta}{\beta(\boldsymbol{x}^{(i)})}\left(\frac{t}{\beta(\boldsymbol{x}^{(i)})}\right)^{\eta-1} \\ &= \frac{\eta}{\beta_0\exp\left(-\gamma^\top\boldsymbol{x}^{(i)}\right)}\left(\frac{t}{\beta_0\exp\left(-\gamma^\top\boldsymbol{x}^{(i)}\right)}\right)^{\eta-1} \\ &= \frac{\eta\exp\left(\gamma^\top\boldsymbol{x}^{(i)}\right)}{\beta_0}\left(\frac{t\cdot\exp\left(\gamma^\top\boldsymbol{x}^{(i)}\right)}{\beta_0}\right)^{\eta-1} \\ &= \frac{\eta}{\beta_0^\eta}t^{\eta-1}\exp\left(\eta\gamma^\top\boldsymbol{x}^{(i)}\right). \end{aligned} \tag{34}$$

We define the baseline hazard function as:

$$h_0(t) = \frac{\eta}{\beta_0^\eta}\cdot t^{\eta-1}. \tag{35}$$

Therefore, the hazard function can be expressed as:

$$h(t;\,\boldsymbol{x}^{(i)}) = h_0(t)\exp\left(\eta\gamma^\top\boldsymbol{x}^{(i)}\right). \tag{36}$$

This formulation shows that the Weibull hazard function is the product of a time-dependent baseline hazard, $h_0(t)$, and a time-independent relative hazard, $\exp\left(\eta\gamma^\top\boldsymbol{x}^{(i)}\right)$. Hence, when the shape parameter η is constant, and only the scale parameter β varies with the covariates, the Weibull distribution satisfies the proportional hazards assumption. In contrast, both the Deep Survival Machines (DSM) (Nagpal et al., 2021) and the proposed MENSA model, which utilize a mixture of Weibull distributions, do not satisfy this assumption. This can be attributed to two reasons:

1. In DSM and MENSA, both the shape parameter η and the scale parameter β depend on the covariates $\boldsymbol{x}^{(i)}$. Substituting an $\boldsymbol{x}^{(i)}$-dependent $\eta(\boldsymbol{x}^{(i)})$ into the baseline hazard definition prevents the hazard function from being expressed in a proportional hazards form, as shown in (36), thus violating the assumption.

2. Since both DSM and MENSA model the survival distribution as a mixture of Weibulls, their hazard functions can be expressed through a weighted combination of individual Weibull distributions. Thus, given mixture weights, $g_\psi(\boldsymbol{x}^{(i)})$, the PDF (5) and CDF (6) functions, their hazard function can be described as:

$$h\left(t;\,\beta,\eta,\boldsymbol{x}^{(i)}\right) = \frac{f\left(t;\,\beta,\eta,\boldsymbol{x}^{(i)}\right)}{S\left(t;\,\beta,\eta,\boldsymbol{x}^{(i)}\right)} = \frac{\sum_{\psi=1}^{\Psi} g_\psi(\boldsymbol{x}^{(i)})f_\psi\left(t;\,\beta_\psi(\boldsymbol{x}^{(i)}),\eta_\psi(\boldsymbol{x}^{(i)})\right)}{1-\sum_{\psi=1}^{\Psi} g_\psi(\boldsymbol{x}^{(i)})F_\psi\left(t;\,\beta_\psi(\boldsymbol{x}^{(i)}),\eta_\psi(\boldsymbol{x}^{(i)})\right)}. \tag{37}$$

Because both the shape and scale parameters, as well as the mixture weights, depend on the covariates $\boldsymbol{x}^{(i)}$, and the marginal density is modeled as a mixture of Weibull distributions, the resulting hazard function is not proportional across covariates or have a simple linear relationship in t or $\log(t)$. Unlike the traditional Weibull model, mixture-based approaches, such as DSM and MENSA, deviate from the proportional hazards assumption, offering greater flexibility and potentially more accurate representations of patient risk over time.

Appendix H

Overcoming Dependent Censoring in the Evaluation of Survival Models

Abstract: Dependent censoring occurs when the event time and censoring time are not conditionally independent given the observed covariates. This complicates survival model evaluation because widely used metrics, such as the Brier score, typically handle right-censoring using inverse probability of censoring weighting (IPCW). Unfortunately, IPCW is valid only when the estimated censoring distribution is independent of the event time. We propose a dependent Brier score based on an Archimedean copula and the Copula-Graphic estimator, and establish consistency and asymptotic normality of its margin-time estimator. To evaluate the metric, we introduce a semi-synthetic framework that creates realistic dependent censoring while preserving the original covariate structure and known event times. Across 12 datasets, the proposed metric reduces estimation error by 12-16% on average relative to IPCW. Source code is available at `https://github.com/thecml/DependentEVAL`.

Authors: Christian Marius Lillelund, Shi-ang Qi, and Russell Greiner.

Status: Published in the proceedings of UAI 2026, Proceedings of Machine Learning Research, vol. 337, pp. 3841–3866, 2026.

Overcoming Dependent Censoring in the Evaluation of Survival Models

Christian Marius Lillelund[1,2] **Shi-ang Qi**[1,3,4] **Russell Greiner**[1,3]

[1]Department of Computing Science, University of Alberta, Edmonton, Canada
[2]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark
[3]Alberta Machine Intelligence Institute, Edmonton, Canada
[4]Vector Institute, Toronto, Canada

Abstract

Dependent censoring occurs when the event time and censoring time are not conditionally independent given the observed covariates. This complicates survival model evaluation because widely used metrics, such as the Brier score, typically handle right-censoring using inverse probability of censoring weighting (IPCW). Unfortunately, IPCW is valid only when the estimated censoring distribution is independent of the event time. We propose a dependent Brier score based on an Archimedean copula and the Copula-Graphic estimator, and establish consistency and asymptotic normality of its margin-time estimator. To evaluate the metric, we introduce a semi-synthetic framework that creates realistic dependent censoring while preserving the original covariate structure and known event times. Across 12 datasets, the proposed metric reduces estimation error by 12-16% on average relative to IPCW. Source code is available at `https://github.com/thecml/DependentEVAL`.

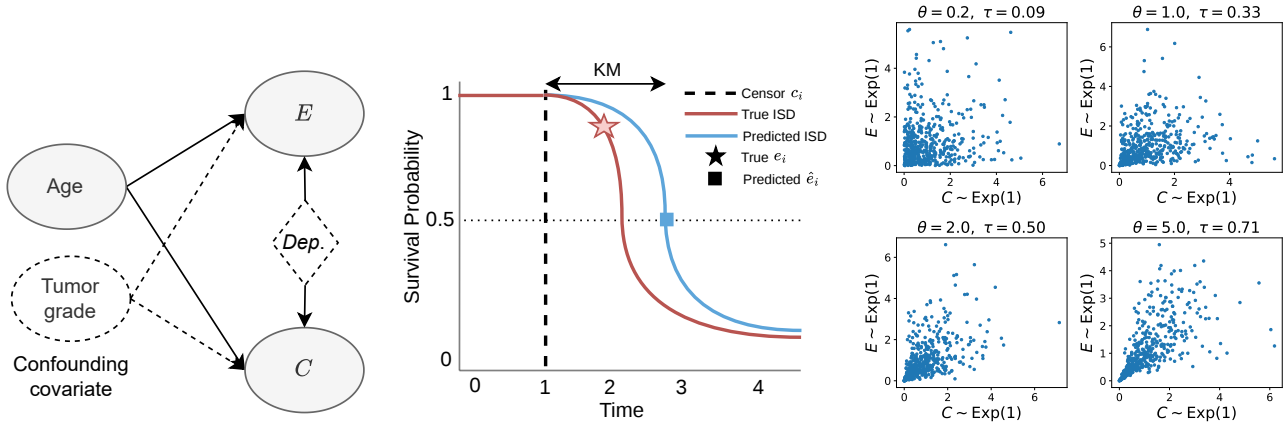
1 INTRODUCTION

Survival analysis differs from standard regression because outcomes may be *censored*: for some individuals, we only know that the event time exceeds a certain specified value. A fundamental assumption underlying most methods and evaluation metrics is *independent censoring*, meaning that finding someone censored at time c does not convey information about their true survival time e, beyond $e > c$ [Kleinbaum and Klein, 2012]. Censoring is considered *dependent* whenever the censoring time C and the event time E are not conditionally independent given the observed covariates X – *i.e.*, $E \not\perp\!\!\!\perp C \mid X$ (see Appendix A for details). While dependent censoring has been relatively well studied in the context of *model estimation*, with several methods proposed to address it [Foomani et al., 2023, Zhang et al., 2024, Liu et al., 2025], its impact on *model evaluation* remains largely unaddressed. In particular, the integrated Brier score (IBS) [Graf et al., 1999] is typically implemented via inverse probability of censoring weighting (IPCW) based on the Kaplan-Meier (KM) [Kaplan and Meier, 1958] estimator or a Cox proportional hazards (CoxPH) model [Cox, 1972]. These approaches assume independence between event and censoring times. When this assumption is violated, the IPCW weights are misspecified, potentially biasing performance estimates and distorting comparisons between survival models.

Figure 1a illustrates how the independent censoring assumption fails when a confounding covariate – such as "Tumor grade" – is not observed. Because tumor grade affects both relapse (E) and dropout (C), advanced tumor grades lead to shorter event and censoring times, inducing dependence. Standard KM-based evaluation relies on independent censoring, assuming that censored individuals are like those remaining at risk. As shown in Figure 1b, when a patient i is censored before the event for reasons related to the event, the KM estimator may overestimate survival beyond c_i. Because this estimated survival distribution is used to construct the IPCW weights, the distortion produces misspecified weights and, potentially, biased performance estimates.

Contribution. We introduce a dependent Brier score and its integrated version, IBS-Dep, for evaluating survival models under dependent censoring. Our approach models the joint distribution of event and censoring times and uses the Copula-Graphic (CG) estimator to impute (marginal) event times for censored instances. Our theory establishes consistency and asymptotic normality of the CG-based margin-time estimator; consequently, IBS-Dep is consistent for the corresponding margin-time surrogate Brier risk. We evaluate how well this surrogate approximates the oracle IBS based on uncensored event times in a semi-synthetic framework, and find that IBS-Dep reduces estimation error compared to IBS using IPCW. To our knowledge, this is the first metric for survival analysis that supports dependent censoring.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3841–3866.



(a) (b) (c)

Figure 1: (a) Residual dependence between event time E (cancer relapse) and censoring time C (dropout) denoted by the diamond "Dep.", due to an unobserved confounding covariate – here "Tumor grade". (b) Evaluation under dependent censoring, where a patient i is censored at time c_i just prior to the event, e_i, due to reasons related to the event (but unobserved). This induces bias in the predicted individual survival distribution and the predicted event time $\widehat{e}_i$ produced by the KM estimator, because the censored patient is assumed to be similar to those still at risk. Note that the true and KM-estimated survival curves coincide on the interval $[0, c_i]$ – here $c_i = 1$ – but diverge beyond due to dependent censoring. (c) Scatter plots of (e_i, c_i), $i = 1, \ldots, 500$, simulated from a Clayton copula [Clayton, 1978] with parameter θ (reported via the corresponding Kendall's τ) and standard exponential marginals. Samples are generated by drawing $(u_{1i}, u_{2i}) \sim C_\theta$ and applying the inverse exponential CDF. The dependence structure is induced by the Archimedean generator φ_θ, where $C_\theta(u_1, u_2) = \varphi_\theta^{-1}\big(\varphi_\theta(u_1) + \varphi_\theta(u_2)\big)$. This is the same generator function that appears in the CG estimator in Definition 2.2.

2 BACKGROUND AND RELATED WORK

2.1 SURVIVAL DATA

We consider a survival dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, where for each instance i, $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes the covariates, $t_i \in \mathbb{R}_+$ denotes the observed time (either event or censoring), and $\delta_i \in \{0, 1\}$ denotes the event indicator. Each instance has latent event and censoring times, e_i and c_i, with $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbf{1}\{e_i \leq c_i\}$. At the population-level, let E and C denote non-negative random variables for the true event and censoring times, with cumulative distribution functions $F_E(t) = \Pr(E \leq t)$ and $F_C(t) = \Pr(C \leq t)$ and survival functions $S_E(t) = 1 - F_E(t)$ and $S_C(t) = 1 - F_C(t)$. A common nonparametric estimator of $S_E(t)$ is the KM estimator.

Definition 2.1. Let $u_1 < \cdots < u_J$ denote the distinct observed event times, and let $t_{\max} = \max_{1 \leq i \leq N} t_i$. For each event time u_j, define $d_j = \sum_{i=1}^N \mathbb{1}[t_i = u_j, \delta_i = 1]$ as the number of events at time u_j, and $n_j = \sum_{i=1}^N \mathbb{1}[t_i \geq u_j]$ as the number of instances at risk immediately before u_j. The Kaplan-Meier (KM) estimator is

$$\widehat{S}_E(t) = \prod_{u_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \qquad 0 \leq t \leq t_{\max}. \quad (2.1)$$

The empty product is defined as one, and $\widehat{S}_E(t)$ is undefined for $t > t_{\max}$. See Appendix D.1 for the derivation.

2.2 COPULA-BASED ESTIMATION

Zheng and Klein [1995] generalized the KM estimator to the CG estimator, which relaxes its independent censoring assumption by modeling dependence through a copula. A copula links two marginal distributions into a joint distribution and can thereby capture dependence between E and C. Rivest and Wells [2001] derived the first explicit form of the CG estimator under an *assumed copula* from the Archimedean family (*e.g.*, the Clayton [Clayton, 1978] and Frank [Frank, 1979] copulas), which specifies dependence through a single parameter θ via a generator function φ_θ [Emura and Chen, 2018] (see details in Appendix B). Figure 1c shows sampling from a Clayton copula, which can model *lower-tail dependence* – that is, a tendency for early events (small values of E) to occur alongside early censoring (small values of C). Now, assuming an Archimedean copula, we can estimate $\widehat{S}_E(t)$ using the CG estimator.

Definition 2.2. Let $u_1 < \cdots < u_J$ denote the distinct observed event times, and let $t_{\max} = \max_{1 \leq i \leq N} t_i$. The Copula-Graphic (CG) estimator is

$$\widehat{S}_E(t) = \varphi_\theta^{-1}\left(\sum_{u_j \leq t}\left[\varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\left(\frac{n_j}{N}\right)\right]\right). \quad (2.2)$$

For $0 \leq t \leq t_{\max}$, where φ_θ is an Archimedean copula generator, $d_j = \sum_{i=1}^N \mathbb{1}[t_i = u_j, \delta_i = 1]$ is the number of

events at u_j, and $n_j = \sum_{i=1}^{N} \mathbb{1}[t_i \geq u_j]$ is the number at risk immediately before u_j. The empty sum is zero, and $\widehat{S}_E(t)$ is undefined for $t > t_{\max}$ [Emura and Chen, 2018]. For independent copula $\varphi_{\text{ind}}(u) = -\log u$, the CG estimator reduces to the KM estimator. See Appendix D.2 for the derivation.

2.3 IPCW-WEIGHTED BRIER SCORE

The Brier score (BS) measures the squared error between the predicted survival probability and the observed binary outcome at a time t^* [Brier, 1950], which for uncensored instances, is:

$$\text{BS}\big(\mathcal{D}; \widehat{S}_E(t^* \mid \cdot)\big) = \frac{1}{N} \sum_i \big(\mathbb{1}[t_i > t^*] - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\big)^2. \tag{2.3}$$

To incorporate censoring, Graf et al. [1999] used inverse probability of censoring weights (IPCW), where $\widehat{S}_C(t)$ is typically estimated using the marginal KM estimator in standard implementations [Qi et al., 2023b][1]:

$$w_i(t^*) \;=\; \frac{\delta_i \, \mathbb{1}[t_i \leq t^*]}{\widehat{S}_C(t_i)} + \frac{\mathbb{1}[t_i > t^*]}{\widehat{S}_C(t^*)}, \tag{2.4}$$

where the first term applies to events before t^* and the second to those surviving beyond. Individuals censored before t^* receive weight zero and therefore do not contribute to the sum. Then, the BS-IPCW metric can be expressed as:

$$\begin{aligned} &\text{BS}_{\text{IPCW}}\big(\mathcal{D}; \widehat{S}_E(t^* \mid \cdot)\big) \\ &= \frac{1}{N} \sum_{i=1}^{N} w_i(t^*) \Big(\mathbb{1}[t_i > t^*] - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\Big)^2. \end{aligned} \tag{2.5}$$

Often, practitioners want to summarize Equation 2.5 over a time interval. Therefore, Graf et al. [1999] proposed the integrated Brier score (IBS), defined as the uniform time average of the Brier score over a prespecified interval $[t_1, t_2]$:

$$\text{IBS}\big(\mathcal{D}; \widehat{S}_E, t_1, t_2\big) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}_{\text{IPCW}}\big(\widehat{S}_E(\tau \mid \cdot)\big) \, \mathrm{d}\tau. \tag{2.6}$$

A common choice is $[t_1, t_2] = [0, t_{\max}]$, where $t_{\max}$ is the maximum observed time in the dataset.

The IPCW-weighted Brier score depends on estimating the censoring distribution. When $\widehat{S}_C(t)$ is obtained via the KM, its consistency relies on the independent censoring assumption. Under dependent censoring, KM can be biased and inconsistent [Emura and Chen, 2018, Campigotto and Weller, 2014]. Conditional IPCW estimators [Gerds and Schumacher, 2006] attempt to address this by estimating $\widehat{S}_C(t \mid X)$, but two limitations remain: (1) *model misspecification*, which may give unreliable weights, and (2) *residual dependence*, where unobserved factors induce dependence not captured by the conditional model.

[1] More generally, it may use a conditional estimator $\widehat{S}_C(t \mid X)$.

3 THE DEPENDENT BRIER SCORE

3.1 DEFINITION

We propose a Brier score that relaxes independent censoring by modeling dependence via an Archimedean copula. Our method consists of two parts: (1) the CG estimator to impute *margin time* estimates for censored instances, and (2) an *uncertainty weighting* mechanism.

The margin time is the conditional expectation of the event time given that the event time is greater than the censoring time. Given that an instance is censored at time t_i, we can calculate its margin time using [Haider et al., 2020]:

$$e^{\text{margin}}(S_E, t_i) \;=\; \mathbb{E}[E \mid E > t_i] \;=\; t_i + \frac{\int_{t_i}^{\infty} S_E(t)\, dt}{S_E(t_i)}. \tag{3.1}$$

Because S_E is unknown, we estimate the margin time using $e^{\text{margin}}\Big(\widehat{S}_E^{\text{CG}}, t_i\Big)$,[2] where $\widehat{S}_E^{\text{CG}}$ is obtained from the CG estimator (Definition 2.2) under an Archimedean copula with parameter θ. The resulting margin time serves as an imputed event time for each censored instance. This construction requires specification of a copula family and its parameter and accounts for marginal dependence between E and C through their joint copula model.

To account for uncertainty in these imputed event times, we adaptively down-weight censored instances according to how early censoring occurs. For an instance censored at t_i, we use its estimated progress through the event-time distribution, $\widehat{F}_E^{\text{CG}}(t_i) = 1 - \widehat{S}_E^{\text{CG}}(t_i)$, as an uncertainty weight. We therefore define $w_i = 1$ for uncensored instances and $w_i = 1 - \widehat{S}_E^{\text{CG}}(t_i)$ for censored instances. Early censoring yields a smaller w_i, whereas later censoring generally yields a larger weight. This leads to the proposed Brier score, which is free from IPCW and targets a margin-time surrogate of the Brier risk:

$$\begin{aligned} &\text{BS}_{\text{Dep}}\Big(\mathcal{D}; \widehat{S}_E(t^* \mid \cdot)\Big) \\ &= \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \Big(\mathbb{1}[\tilde{e}_i > t^*] - \widehat{S}_E(t^* \mid \boldsymbol{x}_i)\Big)^2, \end{aligned} \tag{3.2}$$

where

$$\tilde{e}_i \;=\; \begin{cases} e_i, & \delta_i = 1, \\ e^{\text{margin}}\Big(\widehat{S}_E^{\text{CG}}, t_i\Big), & \delta_i = 0. \end{cases}$$

Similarly, we define the integrated dependent Brier score by averaging Equation 3.2 over $[t_1, t_2]$:

$$\text{IBS}_{\text{Dep}}\big(\mathcal{D}; \widehat{S}_E, t_1, t_2\big) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}_{\text{Dep}}\big(\widehat{S}_E(\tau \mid \cdot)\big) \, \mathrm{d}\tau. \tag{3.3}$$

[2] For brevity, we write $e^{\text{margin}}(t_i)$ when the survival estimator is fixed.

Section 3.2 establishes consistency of the CG-based margin-time estimator for its copula-induced target, which implies consistency of IBS-Dep for the associated surrogate Brier risk.

3.2 THEORETICAL ANALYSIS

We now develop a concise large-sample theory for the CG-based margin-time functional under dependent censoring [Rivest and Wells, 2001]. The theory concerns the imputed margin times used by IBS-Dep, rather than the oracle Brier score based on unobserved event times. For ease of understanding, we sketch the main ideas and implications here, and defer the formal definitions, regularity conditions, and detailed proofs to Appendix D.

First, we show that the margin time admits a direct probabilistic interpretation: it equals the conditional mean remaining event time beyond a landmark,

$$e^{\mathrm{margin}}(c) \;=\; \mathbb{E}[E \mid E > c],$$

provided $S_E(c) > 0$ and the corresponding tail integral is finite (see Appendix D.3.1). This justifies margin time as a clinically interpretable summary of the post-c prognosis.

Second, we establish *consistency* of the plug-in estimator $\widehat{e}^{\mathrm{margin}}(c)$ for a large-sample target $e^{*,\mathrm{margin}}(c)$ induced by the (possibly misspecified) Archimedean copula model considered by Rivest and Wells [2001]. Under uniform consistency of $\widehat{S}_E$ for S_E^* on a finite horizon $[0, t_{\max})$ and mild regularity ($S_E^*(c) > 0$ and negligible tail contribution beyond $t_{\max}$), a deterministic perturbation bound shows that small uniform errors in $\widehat{S}_E$ translate into small error in $\widehat{e}^{\mathrm{margin}}(c)$. When the copula model is correctly specified, $S_E^* = S_E$ and hence $e^{*,\mathrm{margin}}(c) = e^{\mathrm{margin}}(c)$ (see details in Appendix D.3.2).

Finally, we derive a *first-order (delta-method) expansion* and an asymptotic normal limit. Writing the margin time as a functional $\mathcal{T}_c(S)$, we obtain an exact linearization

$$\mathcal{T}_c(S + h) - \mathcal{T}_c(S) \;=\; \dot{\mathcal{T}}_{c;S}(h) \;+\; R_{c;S}(h),$$

with an explicit remainder $R_{c;S}(h)$ that is second order in the perturbation (see Appendix D.3.3). Combining this expansion with the functional central limit theorem (CLT) of Rivest and Wells [2001, Theorem 2] for $\sqrt{n}(\widehat{S}_E - S_E^*)$ yields the asymptotic linear representation (converges in distribution)

$$\sqrt{n}\big(\widehat{e}^{\mathrm{margin}}(c) - e^{*,\mathrm{margin}}(c)\big) \implies \frac{1}{S_E^*(c)} \int_c^{t_{\max}} \mathbb{G}_E(u)\, du \;-\; \frac{\int_c^{t_{\max}} S_E^*(u)\, du}{(S_E^*(c))^2}\, \mathbb{G}_E(c),$$

where $\mathbb{G}_E(c)$ stands for the Gaussian process. If the right-hand side is non-degenerate, this implies an $N^{-1/2}$ convergence rate and enables principled uncertainty quantification for margin time via the induced influence form.

4 SYNTHETIC EXPERIMENTS

4.1 EXPERIMENTAL SETUP

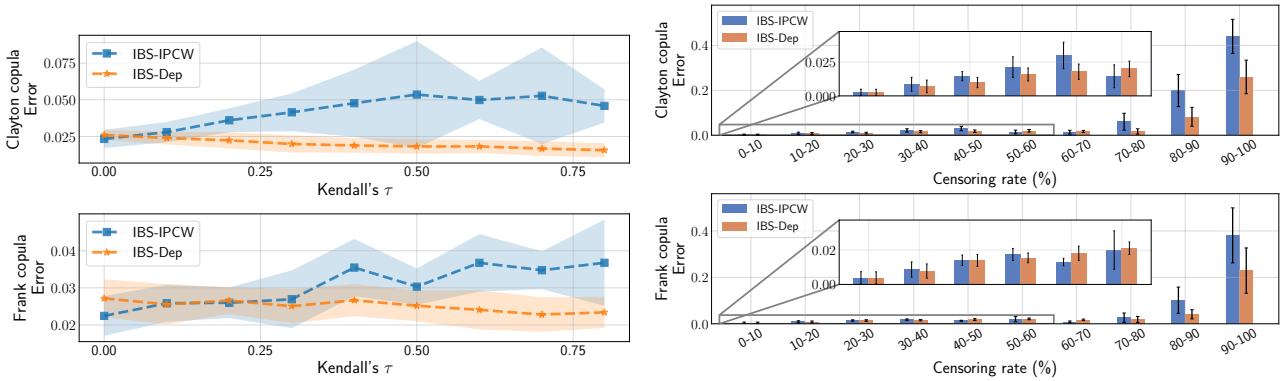
To validate our dependent Brier score, we simulate synthetic survival data with known ground-truth event times. Covariates $x_i \sim \mathrm{Unif}([0,1]^d)$ ($d = 10$) are fixed per seed. Event and censoring times follow covariate-dependent Weibull proportional hazards models with linear risk functions $g(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ and fixed shape/scale parameters (see Appendix E.1). Dependence is parameterized by Kendall's τ: when $\tau = 0$, event and censoring times are independent; for $\tau > 0$, dependence is introduced via a copula, with family and parameter assumed known to the dependent metric. For each of 10 random seeds, we generate $N = 10{,}000$ and use fixed training (70%) and test (30%) splits. The event times and fitted CoxPH survival model are held fixed across censoring settings, so changes in evaluation error reflect the censoring mechanism rather than changes in the prediction model. Oracle IBS is computed on the test set using the fully observed event times prior to censoring.

4.2 RESULTS AND TAKEAWAYS

Increasing dependence. Figure 2 (a) shows the estimation error as a function of Kendall's τ at fixed censoring ($\approx 50\%$). IBS-IPCW becomes increasingly unstable as τ grows, while our IBS-Dep remains comparatively stable and achieves lower (or comparable) error. This reflects the breakdown of independent censoring: as dependence increases, IPCW becomes biased and exhibits higher variance because marginal inverse weights no longer correct the induced selection. The effect is stronger under the Clayton copula due to its lower-tail dependence, which increases the probability that early (high-risk) events coincide with early censoring and thereby more severely distorts the observable risk set. As a result, the empirical IBS-IPCW deviates increasingly from the oracle IBS computed from the fully observed data, whereas IBS-Dep tracks the oracle much more closely.

Increasing censoring. Figure 2 (b) shows error as a function of the censoring rate with $\tau = 0.5$. Both metrics deteriorate under heavy censoring, but IBS-IPCW degrades much more rapidly, particularly beyond $\sim 80\%$ censoring. As censoring increases, IPCW relies on fewer uncensored observations with larger inverse weights, inflating variance and bias under dependence. In contrast, IBS-Dep shows slower error growth, since it adjusts contributions using the joint (E, C) structure rather than relying solely on extreme reweighting.

Misspecification of C_θ. Figure 3 evaluates IBS-Dep when the assumed copula differs from the data-generating model. We consider three settings: (1) *Wrong family*, where Clayton data are evaluated with Frank, or vice versa, at the same Kendall's τ; (2) *Wrong dependence*, where the family is correct but data generated with $\tau = 0.5$ are evaluated using



(a) Increasing dependence (Censoring = 50%).

(b) Increasing censoring (Kendall's $\tau = 0.5$).

Figure 2: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and the proposed IBS-Dep under a *known* copula, averaged over 10 experiments. (a) Error as a function of the dependence between E and C with censoring held constant. (b) Error as a function of the censoring rate with Kendall's τ held constant.

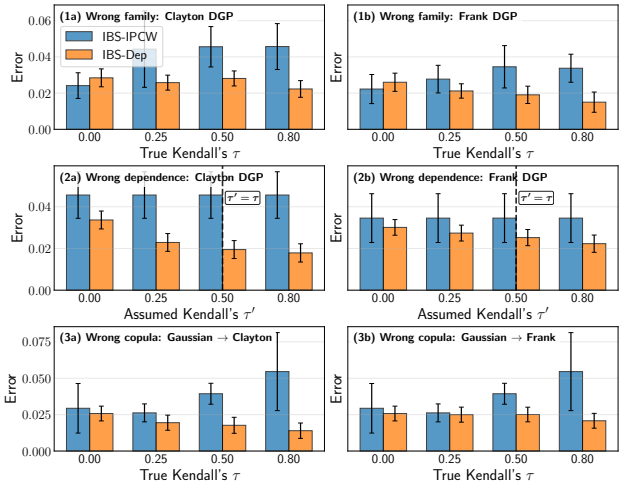


Figure 3: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and IBS-Dep under copula misspecification, averaged over 10 experiments. The rows show (respectively) wrong family, wrong dependence strength, and wrong copula experiments. In the first two rows, columns separate Clayton and Frank data-generating copulas; in the final row, the Gaussian data-generating copula is evaluated using Clayton (resp., Frank) copulas. Censoring is held fixed at 50%.

varying τ'; and (3) *Wrong copula*, where non-Archimedean Gaussian-copula data are evaluated using Clayton or Frank at the same Kendall's τ. Across settings, IBS-Dep remains less sensitive to dependent censoring than IBS-IPCW, even with an imperfect assumed copula. However, the smallest error does not necessarily occur at the true τ: IBS-Dep is slightly upward biased, and stronger assumed dependence can reduce this bias. We therefore interpret these results as a robustness analysis rather than a parameter-recovery diagnostic.

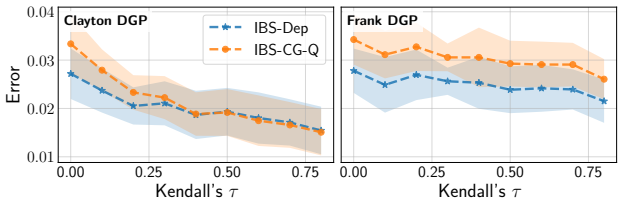


Figure 4: Mean ($\pm$ SD) absolute estimation error of IBS-Dep and IBS-CG-Q under correctly specified copulas, shown separately for Clayton and Frank and averaged over 10 experiments. Censoring is held fixed at 50%.

Integrating over censored event-time uncertainty. IBS-Dep imputes each censored individual with a single CG-based margin time. To test whether this point estimate discards useful uncertainty, we also evaluate IBS-CG-Q, an ablation that integrates over the conditional event-time distribution after censoring. For evaluation time t, define

$$\widehat{q}_i^{\mathrm{CG}}(t) = \begin{cases} \mathbb{1}[t_i > t], & \delta_i = 1, \\ \Pr_{\mathrm{CG}}(E > t \mid E > t_i), & \delta_i = 0. \end{cases}$$

where $\Pr_{\mathrm{CG}}$ denotes the conditional survival probability estimated by the CG estimator. The CG-Q Brier score replaces the binary survival status by $\widehat{q}_i^{\mathrm{CG}}(t)$:

$$\mathrm{BS}_{\text{CG-Q}}(\mathcal{D}; \widehat{S}_E, t) = \frac{1}{N} \sum_{i=1}^{N} \Big[\widehat{q}_i^{\mathrm{CG}}(t) \left(1 - \widehat{S}_E(t \mid \boldsymbol{x}_i)\right)^2 + \left(1 - \widehat{q}_i^{\mathrm{CG}}(t)\right) \widehat{S}_E(t \mid \boldsymbol{x}_i)^2 \Big]. \tag{4.1}$$

The integrated version, IBS-CG-Q, is obtained by averaging Equation 4.1 over a prespecified interval $[t_1, t_2]$, analogously to IBS-Dep. Figure 4 compares IBS-Dep and IBS-CG-Q under correctly specified copulas. Overall, IBS-Dep gives slightly lower error than its conditional counterpart, al-

though the two become nearly indistinguishable for stronger Clayton dependence. Both variants are upward biased across all dependence levels, with the largest bias at weak dependence and particularly under the Frank copula for IBS-CG-Q. As Kendall's τ increases, the upward bias decreases, so the apparent improvement at stronger dependence reflects reduced bias rather than an intrinsically easier censoring problem.

5 SEMI-SYNTHETIC EXPERIMENTS

5.1 EXPERIMENTAL SETUP

To evaluate IBS-Dep on real data, we require ground-truth event times, which are unavailable due to censoring. Following Qi et al. [2023a], we construct semi-synthetic datasets (see Figure 5) in which event and censoring mechanisms are learned from real data and used to resample outcomes while preserving the covariate structure. Dependent censoring is induced via targeted feature selection, creating settings where censoring depends on omitted information. The resulting dataset $\mathcal{D}''$ provides known event times for oracle evaluation while keeping realistic marginal behavior.

Datasets (Step 1 in Figure 5): We apply the pipeline to 12 datasets: WHAS, METABRIC, Churn, GBSG, NACD, FLCHAIN, SUPPORT, Employee, MIMIC-IV, and 3 from SEER (death from brain, liver and stomach cancer). Table 1 summarizes the datasets; details are in Appendix E.2.

Table 1: Overview of the raw datasets. †Number of features before (after) one-hot encoding.

Dataset	#Instances	#Features†	%Censored
WHAS	500	14 (14)	57.0%
GBSG	686	8 (8)	56.4%
METABRIC	1,902	9 (9)	42.1%
Churn	1,958	12 (25)	52.4%
NACD	2,396	48 (48)	36.4%
FLCHAIN	7,871	9 (31)	72.5%
SUPPORT	8,873	14 (27)	32.0%
Employee	14,999	8 (17)	76.2%
MIMIC-IV	38,520	91 (103)	66.7%
SEER-brain	73,703	10 (10)	40.1%
SEER-liver	82,841	14 (14)	37.6%
SEER-stomach	100,360	14 (14)	43.4%

Feature selection (Step 2 in Figure 5): To induce dependent censoring, we use three strategies: (1) *Original*, keeping all features; (2) *Top-k*, selecting the most predictive features based on permutation importance [Breiman, 2001] using Harrell's C-index [Harrell Jr. et al., 1996]; and (3) *Random-%*, selecting a random subset of the original features.

Copula model (Step 6 in Figure 5): To estimate dependence under censoring, we train Clayton and Frank copulas separately with linear Weibull proportional hazards models for the event-time and censoring-time marginals. In this semi-synthetic setting, where the true copula parameter is unknown, the dependence parameter θ is thus estimated from the data. Model parameters are learned by minimizing the penalized negative log-likelihood (NLL) under dependent censoring on the training set:

$$\begin{aligned}&\ell_{\text{pen}}(\theta;\mathcal{D})\\&= -\frac{1}{N}\sum_{i=1}^{N}\Big\{\delta_i\Big[\log f_E(t_i \mid \mathbf{x}_i) + \log \partial_{u_1} C_\theta(u_1,u_2)\Big]\\&\qquad + (1-\delta_i)\Big[\log f_C(t_i \mid \mathbf{x}_i) + \log \partial_{u_2} C_\theta(u_1,u_2)\Big]\Big\}\\&\qquad + \lambda\theta^2.\end{aligned} \tag{5.1}$$

Here, $u_1 = S_E(t_i \mid \mathbf{x}_i)$ and $u_2 = S_C(t_i \mid \mathbf{x}_i)$, where C_θ is an Archimedean copula and $\lambda = 0.01$ is the regularization coefficient. The full derivation of the unpenalized likelihood is in Appendix C. We select the candidate model $\widehat{C}_\theta$ with the lowest validation penalized NLL. We optimize Equation 5.1 using Adam [Kingma and Ba, 2017], with separate learning rates for the Weibull marginal parameters and the copula model. The validation penalized NLL is also used for early stopping. Algorithm 1 summarizes the procedure. Only the selected copula family and parameter are used when computing IBS-Dep. See Appendix E.4 for further implementation details.

Survival learners (Step 7 in Figure 5): We compare predictions from five survival learners: CoxPH [Cox, 1972], Gradient Boosting Survival Analysis (GBSA) [Ridgeway, 1999], Random Survival Forests (RSF) Ishwaran et al. [2008], DeepSurv Katzman et al. [2018] and Multi-Task Logistic Regression (MTLR) [Yu et al., 2011]. These learners are trained to predict survival curves $\widehat{S}_E(t \mid \boldsymbol{x}_i)$ for each instance i in the test set. See Appendix E.3 for implementation details and chosen hyperparameters.

We repeat the evaluation pipeline in Figure 5 over 10 independent experiments. In each experiment, the semi-synthetic dataset $\mathcal{D}''$ is split into stratified training (70%), validation (10%), and test (20%) sets (seeds 0–9). Missing values are imputed using training-set statistics (mean for continuous, mode for categorical variables). Numerical features are z-score normalized and categorical features one-hot encoded. To limit computational cost, large datasets ($N \geq 10{,}000$) are subsampled after preprocessing (Appendix E.4). As a sanity check, we plot empirical KM curves and event/censoring time distributions (see Figure E.1).

5.2 RESULTS AND TAKEAWAYS

Estimation error. Table 2 reports the mean estimation error on the semi-synthetic datasets, measured as the absolute deviation from the oracle IBS computed using the true event times. On average, the CG-based estimator reduces error

by 4-8%, while the uncertainty-weighted variant achieves the largest reductions, at 12-16%. The relative improvement becomes more pronounced when important features are omitted (*e.g.*, Random 25%). This does not imply that

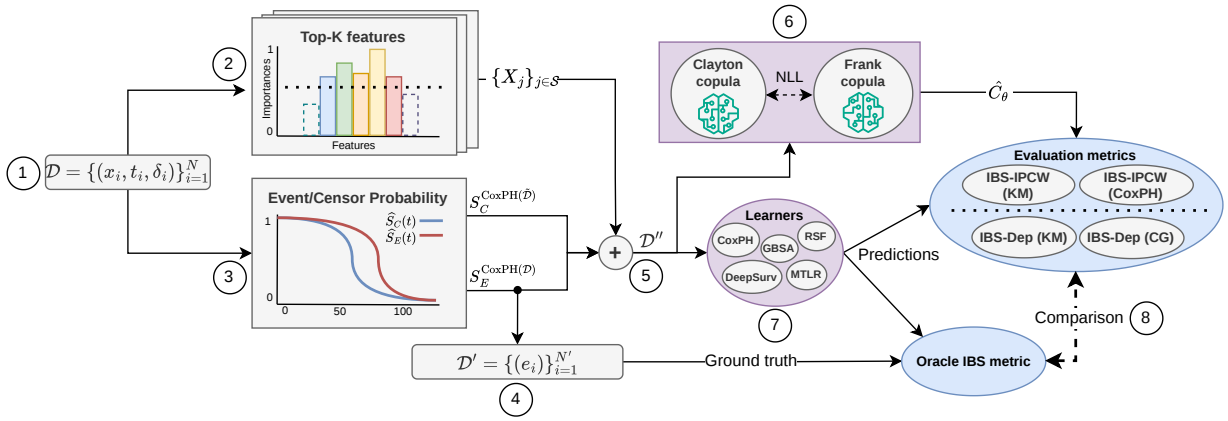


Figure 5: Evaluation pipeline. (1) We begin with a real-world dataset $\mathcal{D}$. (2) A feature selection procedure (*e.g.*, top-k selection) yields a reduced feature set. In parallel, (3) two CoxPH models are fitted to estimate the event- and censoring-time distributions, and (4) uncensored event times are sampled from the event distribution and stored as $\mathcal{D}'$ for oracle evaluation. (5) We construct a semi-synthetic dataset $\mathcal{D}''$ by combining the selected features with event and censoring times sampled from the fitted distributions. Next, (6) two Archimedean copula models are trained to estimate $\widehat{C}_\theta$, and (7) five survival learners are trained on $\mathcal{D}''$. Finally, (8) test-set predictions are evaluated using IBS-IPCW with KM or CoxPH censoring estimates and IBS-Dep with KM- or CG-based dependence estimates using $\widehat{C}_\theta$, and compared with the oracle IBS computed from $\mathcal{D}'$.

Table 2: Mean absolute estimation error of independent and proposed dependent IBS-based metrics across 12 datasets and 4 feature selection strategies. Results are averaged over 5 learners and 10 experiments. Relative (%) change is shown against the KM-based IBS-IPCW baseline. *IBS-IPCW (CoxPH)* denotes IPCW with a CoxPH censoring model, corresponding to conditional independent censoring. *Dep (KM)* denotes the proposed metric using KM instead of CG as an ablation, and *UW* denotes uncertainty weighting (see Section 3.1). Full results, including variances, are provided in Appendix F.2.

	Metric	WHAS	METABRIC	Churn	GBSG	NACD	FLCHAIN	SUPPORT	Employee	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stom.)	Average
Original	IPCW (KM)	0.073	0.061	0.117	**0.125**	0.092	0.156	0.108	0.152	0.026	0.077	0.104	0.119	0.101
	IPCW (CoxPH)	0.106	0.051	**0.115**	0.129	0.148	0.162	0.109	**0.100**	0.039	**0.055**	0.097	**0.095**	0.100 (-0.4%)
	Dep (KM)	**0.055**	0.047	0.120	0.130	0.091	0.135	0.111	0.148	**0.023**	0.089	0.118	0.125	0.099 (-1.4%)
	Dep (CG)	**0.055**	**0.047**	0.120	0.126	0.091	0.135	0.109	0.154	0.024	0.084	0.106	0.112	0.097 (-4.0%)
	Dep (CG)^{UW}	0.063	0.049	0.124	0.134	**0.082**	**0.090**	**0.093**	0.152	0.024	0.067	**0.092**	0.097	**0.089** (-11.8%)
Top-5	IPCW (KM)	0.071	0.062	0.121	**0.124**	0.103	0.133	0.109	0.155	0.027	0.076	0.109	0.124	0.101
	IPCW (CoxPH)	0.065	0.061	**0.114**	0.126	0.098	0.138	0.109	**0.103**	0.027	0.083	0.109	0.128	0.097 (-4.4%)
	Dep (KM)	**0.055**	0.048	0.120	0.130	0.098	0.108	0.112	0.149	0.026	0.090	0.120	0.126	0.098 (-2.8%)
	Dep (CG)	**0.055**	**0.048**	0.119	0.126	0.097	0.108	0.109	0.156	0.027	0.085	0.111	0.116	0.096 (-4.9%)
	Dep (CG)^{UW}	0.065	0.049	0.125	0.134	**0.089**	**0.066**	**0.094**	0.154	**0.025**	**0.068**	**0.096**	**0.101**	**0.089** (-12.2%)
Top-10	IPCW (KM)	0.071	0.061	0.119	**0.125**	0.098	0.143	0.109	0.152	0.026	0.077	0.103	0.119	0.100
	IPCW (CoxPH)	0.082	0.051	**0.111**	0.130	0.093	0.150	0.109	**0.101**	0.031	**0.055**	0.097	0.099	0.092 (-7.9%)
	Dep (KM)	**0.054**	0.047	0.121	0.131	0.095	0.123	0.111	0.147	**0.023**	0.090	0.118	0.124	0.099 (-1.4%)
	Dep (CG)	**0.054**	**0.047**	0.121	0.126	0.095	0.123	0.109	0.153	0.024	0.085	0.107	0.111	0.096 (-4.0%)
	Dep (CG)^{UW}	0.063	0.049	0.125	0.134	**0.086**	**0.078**	**0.093**	0.152	0.025	0.068	**0.093**	**0.097**	**0.088** (-11.7%)
Rand. 25%	IPCW (KM)	0.090	0.061	0.129	0.153	0.109	0.140	0.142	0.161	**0.016**	0.111	0.128	0.132	0.114
	IPCW (CoxPH)	0.091	0.056	0.122	0.157	0.111	0.147	0.142	**0.110**	0.017	0.111	0.197	0.133	0.116 (+1.6%)
	Dep (KM)	**0.081**	0.049	0.126	0.153	0.101	0.140	0.145	0.137	0.031	0.113	0.127	0.135	0.111 (-2.6%)
	Dep (CG)	**0.081**	**0.049**	**0.117**	**0.147**	0.099	0.140	0.141	0.120	0.032	0.102	0.109	0.120	0.105 (-8.4%)
	Dep (CG)^{UW}	0.095	0.050	0.131	0.161	**0.091**	**0.032**	**0.125**	0.147	0.021	**0.089**	**0.098**	**0.106**	**0.096** (-16.4%)

Algorithm 1 Copula estimation: PyTorch-like Pseudocode

```
# model_e, model_c: Weibull models for E/C
# copula: Copula model with parameter theta
def find_copula(model_e, model_c, copula,
    d_train, d_val):

   opt = Adam(params=[model_e, model_c,
       copula])

   for epoch in range(n_epochs):
      for batch in DataLoader(d_train):
         train_nll = loss(model_e, model_c,
              batch, copula)
         opt.zero_grad()
         train_nll.backward()
         opt.step()

      # validate and track best loss
      val_losses = []
      for batch in DataLoader(d_val):
         val_losses.append(loss(model_e,
             model_c, batch, copula))
      val_nll = mean(val_losses)

      # check early stopping
      if theta_stabilized(copula) and
          no_val_improve(val_nll): break

   return copula

# Computes dependent NLL given a copula
def loss(model_e, model_c, batch, copula):
   T, E, X = batch.T, batch.E, batch.X
   s1 = model_e.survival(T, X) # S_E(T|X)
   s2 = model_c.survival(T, X) # S_C(T|X)
   f1 = model_e.pdf(T, X) # f_E(T|X)
   f2 = model_c.pdf(T, X) # f_C(T|X)

   # dependent likelihood via copula terms
   S = cat([s1, s2])
   p1 = log(f1) + log(copula.ccdf("u", S))
   p2 = log(f2) + log(copula.ccdf("v", S))

   reg = 0.01 * sum(copula.theta ** 2)
   ll = p1 * E + (1 - E) * p2
   return -mean(ll) + reg
```

Notes: `cat` concatenates matrices. `log` calculates the natural logarithm. `ccdf` computes a conditional CDF via differentiation.

the KM estimate of the censoring distribution becomes more biased; however, the resulting Brier-score bias can still depend on the learner's prediction errors. When informative features are removed, prediction errors increase; under dependent censoring, IPCW applies misspecified inverse weights to these larger squared errors, amplifying the resulting bias. In addition, with weaker (random) covariates, the marginal model explains less variation in E, leaving more unexplained (residual) dependence between E and C. The CoxPH-based IPCW estimator also depends strongly on the selected feature set: it *reduces* error by up to 8% for the Original, Top-5, and Top-10 strategies, but *increases* error by 2% under the Random 25% strategy. This highlights a key limitation of conditional IPCW, since an uninformative feature set can yield a misspecified censoring model and invalidate conditional independent censoring. The ablation study highlights the importance of the CG estimator. Replacing censored event times with margin times derived from KM yields only modest gains of 1-3% over IBS-IPCW. In contrast, the CG estimator estimates the marginal event-time distribution while accounting for dependence between event and censoring times through the specified copula model, and uses this distribution to compute the conditional margin time for censored instances. This increases the reduction to 4-8% across feature selection strategies, indicating that accounting for dependent censoring when estimating the event-time margin provides substantial improvements over KM-based margin imputation.

Table 3: Ranking performance of the independent and proposed dependent metrics across 12 datasets and 4 feature selection strategies. The results show how often each metric recovered at least two of the top three survival learners according to the oracle IBS, across 10 experiments per feature selection strategy. Each entry is therefore reported out of 40. Higher is better.

Dataset	IPCW (KM)	IPCW (CoxPH)	Dep (KM)	Dep (CG)	Dep $(CG)^{UW}$
WHAS	38/40	39/40	**40/40**	**40/40**	39/40
METABRIC	**40/40**	**40/40**	39/40	39/40	**40/40**
Churn	**37/40**	**37/40**	34/40	29/40	32/40
GBSG	**40/40**	**40/40**	38/40	37/40	35/40
NACD	39/40	**40/40**	38/40	36/40	38/40
FLCHAIN	**39/40**	**39/40**	18/40	18/40	35/40
SUPPORT	**37/40**	36/40	35/40	34/40	31/40
Employee	33/40	**35/40**	12/40	11/40	13/40
MIMIC (IV)	**33/40**	29/40	1/40	1/40	13/40
SEER (brain)	37/40	**38/40**	**38/40**	36/40	36/40
SEER (liver)	**39/40**	**39/40**	38/40	30/40	35/40
SEER (stom.)	36/40	38/40	**40/40**	35/40	35/40

Ranking. Table 3 reports the ranking experiment, which evaluates whether a censored metric identifies the same top-performing models as the oracle IBS. We first compute the oracle IBS for each model (*i.e.*, without censoring) and identify the top-3 models, then assess whether each censored metric recovers at least two of them. IBS-IPCW preserves the oracle top-3 most consistently overall, while the dependent metrics perform comparably on several datasets but worse on FLCHAIN, Employee, and MIMIC-IV. Thus, although the copula-based IBS substantially reduces absolute estimation error under dependent censoring, this does not necessarily translate into better separation between closely

performing models. IBS-IPCW may therefore remain reasonable for within-dataset ranking when its bias affects competing models similarly. However, it does not estimate the true IBS under dependent censoring and may give unfair rankings when models differ in how they handle censoring.

Computational analysis. The primary computational cost lies in copula fitting, which jointly optimizes the marginal survival models and copula parameter with model selection over candidate families (Clayton and Frank). Empirically, this required 5-75 minutes, depending on dataset size (Appendix F.1). Once estimated, IBS-Dep adds negligible cost relative to IPCW. The incremental peak PyTorch GPU memory allocation is modest ($\leq$ 11 MiB), indicating that the method requires little additional GPU memory.

6 DISCUSSION

6.1 MAIN FINDINGS AND IMPLICATIONS

Our results show that accounting for marginal dependent censoring at evaluation time can meaningfully change how we judge model performance. Across synthetic and semi-synthetic experiments, IBS-Dep consistently reduced bias relative to IBS-IPCW when event and censoring times were dependent. The copula-based estimator achieved average reductions of 4-8%, while the uncertainty-weighted variant yielded improvements of 12-16%. These effects were observed across 12 generated survival datasets with sample sizes from 500 to 10,000, feature dimensions from 14 to 91, and censoring rates between 32% and 76%. Overall, this suggests that dependent censoring is not a rare edge case, but a relevant trait of many real-world datasets. Although IBS-Dep reduced estimation error, it did not consistently improve the ranking of models, because bias reduction does not necessarily improve separation between models.

Most prior work concerning dependent censoring has concentrated on *model estimation*. Copula-based modeling approaches [Foomani et al., 2023, Zhang et al., 2024, Liu et al., 2025] explicitly model the joint distribution of event and censoring times to capture their dependence. However, these models are typically evaluated using metrics such as Harrell's C-index or Brier score with IPCW, both of which assume independent censoring. This creates a mismatch: models are designed for dependence, but evaluation still assumes independence. As a result, reported performance in dependent censoring settings may itself be biased. Our work shifts the focus to *model evaluation*, introducing a flexible copula-based extension of the Brier score.

6.2 LIMITATIONS AND FUTURE WORK

Our work also points to several directions for future research. More broadly, the framework introduced here opens the door to dependence-aware versions of other common evaluation metrics. Widely used measures – such as Harrell's C-index, Uno's C-index [Uno et al., 2011], and D-calibration [Haider et al., 2020] – rely on independence assumptions. Extending copula-based corrections to concordance and calibration measures is a natural next step toward having metrics that all support dependent censoring.

At the same time, several limitations remain. First, we assume a stable dependence structure; under temporal distribution shift, the copula parameter must be re-estimated, suggesting the need for adaptive or time-varying dependence models. Second, the true copula is not identifiable in real-world survival datasets because for each individual only one of the competing times is observed [Tsiatis, 1975]. Some misspecification is therefore unavoidable. Third, likelihood-based estimation can be unstable in small samples, particularly under dependent censoring, because weak identification of the copula parameter can make its estimate highly sensitive to sampling variation. These limitations motivate reporting sensitivity analyses across plausible copula families and dependence strengths. Agreement across specifications would strengthen confidence in the evaluation, whereas disagreement would reveal that model comparisons are sensitive to assumptions about the censoring mechanism.

6.3 USER GUIDE

We conclude with a brief guide on how to use IBS-Dep. The central choice is the dependence model used to approximate the relationship between event and censoring times. In standard right-censored data, we observe $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}(e_i \leq c_i)$, but only the smaller of e_i and c_i. Thus, the joint distribution of the population-level variables E and C, and hence their true dependence structure, is generally not identifiable without additional assumptions. We therefore propose not attempting to recover the true latent copula, but instead using a quantitative model comparison to determine whether an Archimedean copula fits the data better than independence.

In practice, one specifies candidate Archimedean copula families and parametric marginal models for E and C, whose parameters are estimated by minimizing the dependent NLL in Equation 5.1. As a baseline, one should fit the independence copula using the same marginal specification. The model with the best validation likelihood is selected; if none of the candidate dependent copulas improves upon the independence copula, one should default to independence. Algorithm 1 summarizes the copula estimation procedure. The selected copula is treated as a working model for the censoring mechanism, and the CG estimate of the event-time distribution is used to impute margin times for censored instances. The selected family is not claimed to be the true latent copula, but a likelihood-based approximation of the dependence structure between event and censoring times.

Acknowledgements

This research received support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), and the Alberta Machine Intelligence Institute (Amii).

References

Leo Breiman. Random Forests. *Machine Learning*, 45(1): 5–32, 2001.

Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pages 45–57, 1975.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Federico Campigotto and Edie Weller. Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials. *Journal of Clinical Oncology*, 32(27):3068–3074, 2014.

David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, and METABRIC Group. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

Angela Dispenzieri, Jerry A. Katzmann, Robert A. Kyle, Dirk R. Larson, Terry M. Therneau, Colin L. Colby, Raynell J. Clark, Graham P. Mead, Shaji Kumar, L. Joseph Melton, and S. Vincent Rajkumar. Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population. *Mayo Clinic Proceedings*, 87(6):517–523, 2012.

Takeshi Emura and Yi-Hau Chen. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.

Ali H. G. Foomani, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 669–680, 2023.

Stephane Fotso et al. *PySurvival: Open Source Package for Survival Analysis Modeling*, 2019. Available from `https://www.pysurvival.io/`.

Maurice J. Frank. On the simultaneous associativity of (x, y) and $x + y - f(x, y)$. *Aequationes Mathematicae*, 19(1):194–226, 1979.

Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, 21(1):1–63, 2020.

Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis*. John Wiley & Sons, Inc., 2008.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random Survival Forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9, 2023.

Edwin L Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2017.

David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, 1 edition, 2012.

William A. Knaus. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.

Christian Marius Lillelund, Ali Hossein Gharari Foomani, Weijie Sun, Shi-ang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). MENSA: A Multi-Event Network for Survival Analysis with Trajectory-Based Likelihood Estimation. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pages 543–571, 2026a.

Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Position: Stop Chasing the C-index when Evaluating Survival Analysis Models. *arXiv preprint arXiv:2506.02075*, 2026b.

Xin Liu, Weijia Zhang, and Min-Ling Zhang. HACSurv: A Hierarchical Copula-Based Approach for Survival Analysis with Dependent Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3079–3087, 2025.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, 2 edition, 2006.

Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li Hao Kuan, et al. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28244–28276, 2023a.

Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. In *Proceedings of the 2023 AAAI Fall Symposia*, volume 2, pages 453–457, 2023b.

Shi-ang Qi, Yakun Yu, and Russell Greiner. Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 41303–41339, 2024.

Greg Ridgeway. The State of Boosting. *Computing Science and Statistics*, 31:172–181, 1999.

Lynn A. Gloeckler Ries, Marsha E. Reichman, Denise Riedel Lewis, Benjamin F. Hankey, and Brenda K. Edwards. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist*, 8(6):541–552, 2003.

Louis-Paul Rivest and Martin T. Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79:138–155, 2001.

Patrick Royston and Douglas G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13:1–15, 2013.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

Terry M. Therneau. *A Package for Survival Analysis in R*, 2024. URL `https://CRAN.R-project.org/package=survival`. R package version 3.6-4.

Anastasios Tsiatis. A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72 (1):20–22, 1975.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10): 1105–1117, 2011.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24, pages 1845—1853, 2011.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep Copula-Based Survival Analysis for Dependent Censoring with Identifiability Guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20613–20621, 2024.

Ming Zheng and John P. Klein. Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula. *Biometrika*, 82(1):127–138, 1995.

Overcoming Dependent Censoring in the Evaluation of Survival Models (Supplementary Material)

Christian Marius Lillelund[1,2] **Shi-ang Qi**[1,3,4] **Russell Greiner**[1,3]

[1]Department of Computing Science, University of Alberta, Edmonton, Canada
[2]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark
[3]Alberta Machine Intelligence Institute, Edmonton, Canada
[4]Vector Institute, Toronto, Canada

A CENSORING ASSUMPTIONS

Random censoring. Instances censored at time t are representative of all instances still at risk at t, in terms of their survival experience [Kleinbaum and Klein, 2012, Ch. 1]. In a medical study, we would assume that the death rate of censored individuals is the same as that of uncensored individuals who remain in the risk set. Formally, $E \perp\!\!\!\perp C$, where E is the event time and C is the censoring time, meaning $\Pr(E > t \mid C = t) = \Pr(E > t)$ for all t.

Independent Censoring. In the absence of competing risks (*i.e.*, when no alternative event precludes the main event), within any subgroup of interest (*e.g.*, a smoker and a non-smoker group), instances censored at time t are representative of those who remain at risk at time t with respect to their survival experience [Kleinbaum and Klein, 2012]. This is also known as conditional independent censoring, since we assume that the observed covariates account for the relationship between event and censoring. Thus, once we condition on the covariates, the censoring time is assumed to be independent of the event time. Formally, $E \perp\!\!\!\perp C \mid X$, where X are the covariates, meaning $\Pr(E > t \mid C = t, X) = \Pr(E > t \mid X)$ for all t. Independent censoring is a less restrictive assumption than random censoring on the data-generating process.

Dependent Censoring. In the absence of competing risks, within any subgroup of interest, instances censored at time t are *not* representative of those who remain at risk at time t with respect to their survival experience [Foomani et al., 2023, Lillelund et al., 2026b]. Equivalently, censoring is said to be dependent when, even after conditioning on the observed covariates, the censoring mechanism remains related to the event process. Thus, the observed covariates do not fully account for the relationship between E and C, and the censoring time is not independent of the event time given the covariates.

B COPULAS FOR SURVIVAL ANALYSIS

Bivariate survival analysis can be viewed as a *competing risks* problem, where the occurrence of an event (*e.g.*, death) and censoring (*e.g.*, dropout) are mutually exclusive. Tsiatis [1975] referred to this as the *identifiability issue*; competing risks data do not allow one to observe two event times simultaneously (one event censors the other), and hence, the data may not identify the dependence structure between event times. Zheng and Klein [1995] gave a partial solution to the identifiability problem of dependent censoring by an assumed *copula* between two event times. A copula links two or more random variables by specifying their dependence structure [Emura and Chen, 2018]. A mathematician, Abe Sklar, first used the word "copula" in his study of probabilistic metric spaces [Sklar, 1959]. In his work, he gave a mathematical definition of copulas and established the most fundamental theorem about copulas, known as Sklar's theorem [Sklar, 1959]. It demonstrated that any joint cumulative distribution function can be written in terms of a copula over the quantiles of its marginal cumulative distribution functions. Among the various types of copulas, the Archimedean copula is the most widely used, and enables the modeling of dependencies in arbitrarily high dimensions using a single parameter. This family of copulas is defined by a convex generator function, $\varphi_\theta : [0, 1] \to [0, \infty]$, which is continuous, strictly decreasing, and satisfies the boundary conditions $\varphi_\theta(0) = \infty$ and $\varphi_\theta(1) = 0$ [Emura and Chen, 2018]. Given event time E and censoring time C, their joint survival function under an Archimedean copula can be written as:

$$\Pr(E > t, C > c) = \varphi_\theta^{-1}\left[\varphi_\theta(S_E(t)) + \varphi_\theta(S_C(c))\right], \tag{B.1}$$

where φ_θ is the generator function, and $S_E(t) = \Pr(E > t)$ and $S_C(c) = \Pr(C > c)$ are the marginal survival functions. In the bivariate case, φ_θ describes the dependence structure (*i.e.*, correlation) between two random variables, *e.g.*, E and C, and the (rank) correlation can be described with Kendall's tau (τ):

$$\tau = \Pr\left(t_2 - t_1)(c_2 - c_1) \geq 0 - \Pr\left(t_2 - t_1)(c_2 - c_1) < 0, \right. \tag{B.2}$$

where (t_1, c_1) and (t_2, c_2) are sampled from a (survival) copula. The Clayton [Clayton, 1978] and Frank [Frank, 1979] copulas are two common Archimedean copulas in survival literature. Within these families, the copula C_θ is parameterized by θ, interpreted as the degree of dependence between E and C in the bivariate case. When $\theta \approx 0$, it yields the independence copula for Clayton and Frank copulas; thus, a larger value of θ implies greater dependence between the marginal distributions. Formally, we write $C(u_1, \ldots, u_d) : [0, 1]^d \to [0, 1]$, which is a d-dimensional copula (*i.e.*, where d is the number of variables) with $u_1, \ldots, u_d$ uniform marginal probabilities and the following properties [Nelsen, 2006]:

1. **Groundedness**: If there exists an $i \in \{1, \ldots, d\}$ such that $u_i = 0$, then

$$C(u_1, \ldots, u_d) = 0.$$

2. **Uniform Margins**: For all $i \in \{1, \ldots, d\}$, if $\forall j \neq i \Rightarrow u_j = 1$, then

$$C(u_1, \ldots, u_d) = u_i.$$

3. d**-Increasingness**: For all $u = (u_1, \ldots, u_d)$ and $v = (v_1, \ldots, v_d)$ where $u_i < v_i$ for all $i = 1, \ldots, d$, the following holds:

$$\sum_{\ell \in \{0,1\}^d} (-1)^{\ell_1 + \cdots + \ell_d} C\left(u_1^{\ell_1} v_1^{1-\ell_1}, \ldots, u_d^{\ell_d} v_d^{1-\ell_d}\right) \geq 0.$$

Let E be the event distribution, C the censoring distribution, $\boldsymbol{x}$ a vector of covariates, and $S_E(t \mid \boldsymbol{x}) = \Pr(E > t \mid \boldsymbol{x})$ and $S_C(c \mid \boldsymbol{x}) = \Pr(C > c \mid \boldsymbol{x})$ the marginal survival functions given $\boldsymbol{x}$. We can then define a bivariate (survival) copula C_θ [Nelsen, 2006] that describes the degree of dependence between E and C as follows:

$$\Pr(E > t, C > c \mid \boldsymbol{x}) = C_\theta(S_E(t \mid \boldsymbol{x}), S_C(c \mid \boldsymbol{x}))\,. \tag{B.3}$$

Kendall's tau (τ) can be defined as:

$$\tau = \Pr\left(t_2 - t_1)(c_2 - c_1) \geq 0 \mid \boldsymbol{x} \right. - \Pr\left(t_2 - t_1)(c_2 - c_1) < 0 \mid \boldsymbol{x}\right., \tag{B.4}$$

where (t_1, c_1) and (t_2, c_2) are sampled from the model (B.3). Kendall's τ can be solely expressed as a function of the copula C_θ by [Emura and Chen, 2018]:

$$\tau_\theta = 4 \int_0^1 \int_0^1 C_\theta(u, v), C_\theta(du, dv), \tag{B.5}$$

which implies that Kendall's τ does not depend on the marginal survival functions. An Archimedean copula is a type of copula that is defined using a generator function, which allows for a flexible way to model dependence structures between random variables, *e.g.*, E and C. Such a copula is defined as [Emura and Chen, 2018]:

$$C_\theta(u, v) = \varphi_\theta^{-1}\left(\varphi_\theta(u) + \varphi_\theta(v)\right). \tag{B.6}$$

where the function $\varphi_\theta : [0, 1] \to [0, \infty]$ is called a generator of the copula, which is continuous and strictly decreasing from $\varphi_\theta(0) = \infty$ to $\varphi_\theta(1) = 0$. The following Archimedean copulas are considered in this work:

- **The independence copula:**

$$C(u, v) = uv. \tag{B.7}$$

- **The Clayton copula [Clayton, 1978]:**

$$C_\theta(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}, \quad \theta \geq 0. \tag{B.8}$$

- **The Frank copula [Frank, 1979]:**

$$C_\theta(u, v) = -\frac{1}{\theta} \log\left[1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{e^{-\theta} - 1}\right], \quad \theta \neq 0. \tag{B.9}$$

C SURVIVAL LIKELIHOOD

In the standard survival analysis framework, we define the survival time E and censoring time C such that:

$$\begin{aligned} E \;=\; t_i \text{ and } C \;\geq\; t_i \qquad & \text{if } \delta_i \;=\; 1, \\ E \;>\; t_i \text{ and } C \;=\; t_i \qquad & \text{if } \delta_i \;=\; 0. \end{aligned}$$

Combining these cases, the likelihood for the i-th observation is:

$$\mathcal{L}_i = \Pr\left(E = t_i, C \geq t_i \,|\, x_i\right)^{\delta_i} \; \Pr\left(E \geq t_i, C = t_i \,|\, x_i\right)^{1-\delta_i} \tag{C.1}$$

Under the assumption of independent censoring [Emura and Chen, 2018], we have:

$$\begin{aligned} \mathcal{L}_i &= \left[\Pr\left(E = t_i \,|\, x_i\right) \Pr\left(C \geq t_i \,|\, x_i\right)\right]^{\delta_i} \left[\Pr\left(E \geq t_i \,|\, x_i\right) \Pr\left(C = t_i \,|\, x_i\right)\right]^{1-\delta_i} \\ &= \left[f_E\left(t_i \,|\, x_i\right) \; S_C\left(t_i \,|\, x_i\right)\right]^{\delta_i} \left[S_E\left(t_i \,|\, x_i\right) \; f_C\left(t_i \,|\, x_i\right)\right]^{1-\delta_i} \\ &= \left[f_E\left(t_i \,|\, x_i\right)^{\delta_i} \, S_E\left(t_i \,|\, x_i\right)^{1-\delta_i}\right] \left[f_C\left(t_i \,|\, x_i\right)^{1-\delta_i} \, S_C\left(t_i \,|\, x_i\right)^{\delta_i}\right] \end{aligned} \tag{C.2}$$

where $S_E(t \mid x_i) = \Pr(E \geq t \mid x_i)$, $f_E(t \mid x_i) = -dS_E(t \mid x_i)/dt$, $S_C(t \mid x_i) = \Pr(C \geq t \mid x_i)$, and $f_C(t \mid x_i) = -dS_C(t \mid x_i)/dt$. Under the non-informative censoring assumption, the term $f_C(t_i \mid x_i)^{1-\delta_i} S_C(t_i \mid x_i)^{\delta_i}$ is unrelated to the likelihood for the event times and can simply be ignored [Emura and Chen, 2018]. Therefore, the likelihood function can be written as:

$$\mathcal{L} = \prod_{i=1}^{N} f_E(t_i \mid \boldsymbol{x}_i)^{\delta_i} S_E(t_i \mid \boldsymbol{x}_i)^{1-\delta_i}, \tag{C.3}$$

where $f_E(t_i \mid x_i)$ is the density function of the event time and $S_E(t_i \mid x_i)$ is the survival function of the event time. Note how the censoring terms $f_C(t_i)$ and $S_C(t_i)$ are absent in (C.3). If the event is observed, the likelihood contribution is the probability density of the event occurring at that time. If the observation is censored, the likelihood contribution is the probability of surviving beyond that time. The overall likelihood is the product of these contributions across all instances.

However, if censoring is dependent, we can no longer rely on this clean decomposition of the likelihood as in Equation C.2. Instead, given some Archimedean copula (*e.g.*, Clayton or Frank), we can rewrite Equation C.2 as the log-likelihood under dependent censoring [Foomani et al., 2023]:

$$\begin{aligned} \ell(\theta; \mathcal{D}) = \sum_{i=1}^{N} \delta_i \; \log\left[f_E(t_i|\boldsymbol{x}_i)\right] \; + \; \delta_i \log\left[\frac{\partial}{\partial u_1} C(u_1, u_2)\Big|_{\substack{u_1 = S_E(t_i|\boldsymbol{x}_i) \\ u_2 = S_C(t_i|\boldsymbol{x}_i)}}^{u_1 = S_E(t_i|\boldsymbol{x}_i)}\right] \\ + \; (1-\delta_i) \log\left[f_C(t_i|\boldsymbol{x}_i)\right] + (1-\delta_i) \; \log\left[\frac{\partial}{\partial u_2} C(u_1, u_2)\Big|_{\substack{u_1 = S_E(t_i|\boldsymbol{x}_i) \\ u_2 = S_C(t_i|\boldsymbol{x}_i)}}^{u_1 = S_E(t_i|\boldsymbol{x}_i)}\right] \end{aligned} \tag{C.4}$$

In this expression, the first term represents the log-likelihood of observing the event at time t_i. The second term captures the conditional probability of the censoring time occurring after the event time, given that the event occurred at t_i. By symmetry, the third and fourth terms mirror these quantities but with respect to the censoring time. Although Equation C.4 may appear complex, the partial derivatives of any Archimedean copula have closed-form expressions, which allows us to maximize the log-likelihood using gradient descent.

D THEORETICAL PROPERTIES AND DERIVATIONS

This appendix develops the theoretical foundations of the proposed CG margin-time estimator under dependent censoring. We first derive the KM and CG estimators to establish the survival estimation framework under independent and dependent censoring, respectively, which forms the basis for the subsequent margin-time analysis. We then recall the asymptotic target of the CG survival estimator under possible copula misspecification and introduce the margin-time functional and its plug-in CG estimator.

D.1 THE KAPLAN-MEIER ESTIMATOR

The following derivation follows the presentation of Emura and Chen [2018], generalized to accommodate tied event times.

Recall that (t_i, δ_i), $i = 1, \ldots, N$, denotes survival data without covariates, where $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}[e_i \leq c_i]$. Let $u_1 < \cdots < u_J$ denote the distinct observed event times. For each u_j, define

$$d_j = \sum_{i=1}^{N} \mathbb{1}[t_i = u_j, \delta_i = 1]$$

as the number of events at u_j, and

$$n_j = \sum_{i=1}^{N} \mathbb{1}[t_i \geq u_j]$$

as the number of instances at risk immediately before u_j.

We represent $S_E(t)$ as a decreasing step function whose jumps occur at the observed event times. Its value at time t can therefore be written as

$$S_E(t) = \prod_{u_j \leq t} \frac{S_E(u_j)}{S_E(u_j^-)} \tag{D.1}$$

$$= \prod_{u_j \leq t} \left(1 - \frac{\Pr(E = u_j)}{\Pr(E \geq u_j)}\right), \tag{D.2}$$

where $S_E(u_j^-)$ denotes the survival probability immediately before the jump at u_j.

Suppose that E and C are independent. Then,

$$S_E(t) = \prod_{u_j \leq t} \left(1 - \frac{\Pr(E = u_j, C \geq u_j)}{\Pr(E \geq u_j, C \geq u_j)}\right) \tag{D.3}$$

$$= \prod_{u_j \leq t} \left(1 - \frac{\Pr(\min(E, C) = u_j, E \leq C)}{\Pr(\min(E, C) \geq u_j)}\right). \tag{D.4}$$

Replacing each probability ratio by its empirical estimate gives

$$\widehat{S}_E(t) = \prod_{u_j \leq t} \left(1 - \frac{\sum_{i=1}^{N} \mathbb{1}[t_i = u_j, \delta_i = 1]}{\sum_{i=1}^{N} \mathbb{1}[t_i \geq u_j]}\right) \tag{D.5}$$

$$= \prod_{u_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \qquad 0 \leq t \leq t_{\max}. \tag{D.6}$$

Thus, all events tied at u_j are incorporated through the common factor $1 - d_j/n_j$. The empty product is defined as one, and the estimator is undefined for $t > t_{\max}$.

D.2 THE COPULA-GRAPHIC ESTIMATOR

The following derivation is based on the Copula-Graphic estimator introduced by Zheng and Klein [1995] and its explicit Archimedean representation derived by Rivest and Wells [2001], generalized here to accommodate tied event times.

Consider the Archimedean copula model

$$\Pr(E > t, C > c) = \varphi_\theta^{-1} \left[\varphi_\theta(S_E(t)) + \varphi_\theta(S_C(c))\right], \tag{D.7}$$

where $\varphi_\theta : [0, 1] \to [0, \infty]$ is a continuous, strictly decreasing generator satisfying $\varphi_\theta(0) = \infty$ and $\varphi_\theta(1) = 0$ [Emura and Chen, 2018]. The marginal survival functions are $S_E(t) = \Pr(E > t)$ and $S_C(t) = \Pr(C > t)$.

Let $u_1 < \cdots < u_J$ denote the distinct observed event times, and let d_j and n_j be defined as in Section 2. We treat all events tied at u_j as a single jump of the event survival function. When events and censorings occur at the same observed time, we use the event-before-censoring convention: the event jump is applied before any censoring removals at that time.

Immediately before the event jump at u_j, the empirical joint survival probability is

$$\widehat{\Pr}(E \geq u_j, C \geq u_j) = \frac{n_j}{N}.$$

Immediately after removing the d_j events at u_j, but before removing any censorings tied at that time, it is

$$\frac{n_j - d_j}{N}.$$

Because only the event survival function changes across this event jump, the copula representation gives

$$\varphi_\theta\left(\frac{n_j - d_j}{N}\right) = \varphi_\theta\Big(\widehat{S}_E(u_j)\Big) + \varphi_\theta\Big(\widehat{S}_C(u_j^-)\Big), \tag{D.8}$$

whereas immediately before the jump,

$$\varphi_\theta\Big(\frac{n_j}{N}\Big) = \varphi_\theta\Big(\widehat{S}_E(u_j^-)\Big) + \varphi_\theta\Big(\widehat{S}_C(u_j^-)\Big). \tag{D.9}$$

Subtracting Equation D.9 from Equation D.8 yields the jump recursion

$$\begin{aligned}&\varphi_\theta\Big(\widehat{S}_E(u_j)\Big) - \varphi_\theta\Big(\widehat{S}_E(u_j^-)\Big)\\ &\quad= \varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\Big(\frac{n_j}{N}\Big).\end{aligned} \tag{D.10}$$

Since $\widehat{S}_E(t) = 1$ before the first observed event and $\varphi_\theta(1) = 0$, summing the jumps up to time t gives

$$\varphi_\theta\Big(\widehat{S}_E(t)\Big) = \sum_{u_j \leq t}\Big[\varphi_\theta\Big(\widehat{S}_E(u_j)\Big) - \varphi_\theta\Big(\widehat{S}_E(u_j^-)\Big)\Big] \tag{D.11}$$

$$= \sum_{u_j \leq t}\left[\varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\Big(\frac{n_j}{N}\Big)\right]. \tag{D.12}$$

Applying φ_θ^{-1} therefore gives the CG estimator

$$\widehat{S}_E(t) = \varphi_\theta^{-1}\left(\sum_{u_j \leq t}\left[\varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\Big(\frac{n_j}{N}\Big)\right]\right), \qquad 0 \leq t \leq t_{\max}. \tag{D.13}$$

The empty sum is zero, and the estimator is undefined for $t > t_{\max}$.

For the independence generator $\varphi_{\mathrm{ind}}(u) = -\log u$, Equation D.13 becomes

$$\widehat{S}_E(t) = \exp\left(-\sum_{u_j \leq t}\left[-\log\left(\frac{n_j - d_j}{N}\right) + \log\Big(\frac{n_j}{N}\Big)\right]\right) \tag{D.14}$$

$$= \prod_{u_j \leq t}\frac{n_j - d_j}{n_j} = \prod_{u_j \leq t}\left(1 - \frac{d_j}{n_j}\right), \tag{D.15}$$

which is the Kaplan-Meier estimator.

For the Clayton generator $\varphi_\theta(u) = u^{-\theta} - 1$ [Clayton, 1978], the CG estimator becomes

$$\widehat{S}_E(t) = \left[1 + \sum_{u_j \leq t}\left\{\left(\frac{n_j - d_j}{N}\right)^{-\theta} - \Big(\frac{n_j}{N}\Big)^{-\theta}\right\}\right]^{-1/\theta}. \tag{D.16}$$

D.3 THEORETICAL ANALYSIS

Rivest and Wells [2001] study the CG estimator for $\widehat{S}_E(t)$ under dependent censoring. Without assuming the fitted Archimedean copula is the true copula, their estimator targets a marginal survival function $S_E^*(t)$ defined via their equation (8) (at page 6). Moreover, when the dependence is truly Archimedean with the same generator used to construct the estimator, $\widehat{S}_E(t)$ is uniformly consistent for $S_E(t)$ [Rivest and Wells, 2001, Theorem 1].

In this section, we provide theoretical properties of this margin time. We start the analysis by formally (re-)stating the calculation.

For any $c \geq 0$ with $S_E(c) > 0$ and $\int_c^\infty S_E(u)\,du < \infty$, define

$$e^{\text{margin}}(c; S_E) \;:=\; c + \frac{\displaystyle\int_c^\infty S_E(u)\,du}{S_E(c)}. \tag{D.17}$$

We estimate it by the plug-in CG estimator

$$\widehat{e}^{\,\text{margin}}(c) \;:=\; e^{\text{margin}}(c; \widehat{S}_E) \;=\; c + \frac{\displaystyle\int_c^\infty \widehat{S}_E(u)\,du}{\widehat{S}_E(c)}. \tag{D.18}$$

Similarly, define the large-sample target

$$e^{*,\text{margin}}(c) \;:=\; e^{\text{margin}}(c; S_E^*) \;=\; c + \frac{\displaystyle\int_c^\infty S_E^*(u)\,du}{S_E^*(c)}. \tag{D.19}$$

When the copula model is correctly specified, $S_E^* = S_E$ and hence $e^{*,\text{margin}}(c) = e^{\text{margin}}(c; S_E)$.

To connect (D.17)–(D.18) to Rivest and Wells [2001] results (proved on $[0, t_0)$), without loss of generality, we assume the right tail beyond some finite horizon (*i.e.*, t_0) does not contribute. Under such a setting, we can rewrite the integrals in (D.18)-(D.19) using

$$\begin{aligned}\int_c^\infty \widehat{S}_E(u)\,du \;&=\; \int_c^{t_0} \widehat{S}_E(u)\,du \\ \int_c^\infty S_E^*(u)\,du \;&=\; \int_c^{t_0} S_E^*(u)\,du,\end{aligned}$$

so the theory reduces to controlling $\widehat{S}_E$ on the finite interval $[0, t_0)$ without any tail extrapolation.

D.3.1 Margin time as conditional expectation

Now we prove our method in (D.17) corresponds to the conditional expectation. This following theorem and proof are adapted from Haider et al. [2020, Theorem B.1]. Here, we rephrase the proof for the completeness.

Lemma D.1. *Let $E \geq 0$ have survival function $S_E(t) = \Pr(E > t)$. Fix $c \geq 0$ with $S_E(c) > 0$ and assume $\int_c^\infty S_E(u)\,du < \infty$. Then*

$$\mathbb{E}[E \mid E > c] \;=\; c + \frac{\int_c^\infty S_E(u)\,du}{S_E(c)}. \tag{D.20}$$

Proof. Define $(E - c)_+ := \max\{E - c, 0\}$. For any nonnegative Y, $\mathbb{E}[Y] = \int_0^\infty \Pr(Y > v)\,dv$. Apply this conditionally

on $\{E > c\}$:

$$\begin{aligned}\mathbb{E}[(E-c)_+ \mid E > c] &= \int_0^\infty \Pr((E-c)_+ > v \mid E > c)\, dv \\ &= \int_0^\infty \Pr(E > c+v \mid E > c)\, dv \\ &= \int_0^\infty \frac{\Pr(E > c+v)}{\Pr(E > c)}\, dv \\ &= \frac{1}{S_E(c)} \int_0^\infty S_E(c+v)\, dv \\ &= \frac{1}{S_E(c)} \int_c^\infty S_E(u)\, du,\end{aligned}$$

where we substituted $u = c + v$. Finally, on $\{E > c\}$, $E = c + (E - c)$, hence

$$\begin{aligned}\mathbb{E}[E \mid E > c] &= c + \mathbb{E}[E - c \mid E > c] \\ &= c + \mathbb{E}[(E-c)_+ \mid E > c] \\ &= c + \frac{\int_c^\infty S_E(u)\, du}{S_E(c)}.\end{aligned}$$

□

D.3.2 Consistency of the margin time

To simplify the notation, we denote the numerator term in (D.17) as

$$A(c; S) := \int_c^\infty S(u)\, du.$$

Lemma D.2. *Let $s_{\min}$ be a small positive constant that satisfies $S(c) \geq s_{\min} > 0$, $A(c; S) < \infty$, and $A(c; \widehat{S}) < \infty$. Let*

$$\Delta A := A(c; \widehat{S}) - A(c; S) = \int_c^\infty \big(\widehat{S}(u) - S(u)\big)\, du, \qquad \Delta S(c) := \widehat{S}(c) - S(c).$$

Assume $|\Delta S(c)| \leq s_{\min}/2$ (so that $\widehat{S}(c) \geq s_{\min}/2$). Then

$$\left|e^{\text{margin}}(c; \widehat{S}) - e^{\text{margin}}(c; S)\right| \leq \frac{1}{s_{\min}} |\Delta A| + \frac{2A(c; S)}{s_{\min}^2} |\Delta S(c)| + \frac{2}{s_{\min}^2} |\Delta A|\, |\Delta S(c)|. \tag{D.21}$$

In particular, for small perturbations, the last term is second order in the errors.

Proof. By definition,

$$e^{\text{margin}}(c; \widehat{S}) - e^{\text{margin}}(c; S) = \frac{A(c; \widehat{S})}{\widehat{S}(c)} - \frac{A(c; S)}{S(c)}.$$

Write $A(c; \widehat{S}) = A(c; S) + \Delta A$ and $\widehat{S}(c) = S(c) + \Delta S(c)$:

$$\frac{A(c; \widehat{S})}{\widehat{S}(c)} - \frac{A(c; S)}{S(c)} = \frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S)}{S(c)}.$$

Subtract and re-add $\frac{A(c;S)+\Delta A}{S(c)}$:

$$\frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S)}{S(c)} = \underbrace{\left(\frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S) + \Delta A}{S(c)}\right)}_{I} + \underbrace{\left(\frac{A(c; S) + \Delta A}{S(c)} - \frac{A(c; S)}{S(c)}\right)}_{II}. \tag{D.22}$$

Term II equals $\Delta A/S(c)$, hence $|(II)| \le |\Delta A|/s_{\min}$.

For term I,

$$(I) = (A(c;S)+\Delta A)\left(\frac{1}{S(c)+\Delta S(c)} - \frac{1}{S(c)}\right) = (A(c;S)+\Delta A)\frac{-\Delta S(c)}{S(c)\,(S(c)+\Delta S(c))}.$$

Taking absolute values and using $S(c) \ge s_{\min}$ and $S(c)+\Delta S(c) \ge s_{\min}/2$ gives

$$|(I)| \le |A(c;S)+\Delta A| \cdot \frac{2|\Delta S(c)|}{s_{\min}^2} \le \frac{2A(c;S)}{s_{\min}^2}|\Delta S(c)| + \frac{2}{s_{\min}^2}|\Delta A|\,|\Delta S(c)|.$$

Combining the bounds for both terms yields (D.21). □

Corollary D.3 (Consistency)**.** *Assume:*

(a) Theorem 1 in Rivest and Wells [2001] holds at some $t_0 > c$, i.e.*, $\widehat{S}_E(t)$ is uniformly consistent for $S_E^*(t)$ on $[0,t_0)$;*

(b) no tail contribution beyond t_0;

(c) $S_E^*(c) \ge s_{\min} > 0$.

Then $\widehat{e}^{\,\mathrm{margin}}(c) \xrightarrow{p} e^{*,\mathrm{margin}}(c)$. If the copula model is correctly specified so that $S_E^* = S_E$, then $\widehat{e}^{\,\mathrm{margin}}(c) \xrightarrow{p} e^{\mathrm{margin}}(c;S_E)$.

Proof. Because (b) assumes no tail contribution beyond t_0, we have

$$\Delta A \;=\; \int_c^{\infty} (\widehat{S}_E(u) - S_E^*(u))\,du \;=\; \int_c^{t_0} (\widehat{S}_E(u) - S_E^*(u))\,du.$$

Hence

$$\begin{aligned} |\Delta A| &\le (t_0 - c) \\ \sup_{u\in[c,t_0]} |\widehat{S}_E(u) - S_E^*(u)| &\le (t_0 - c) \\ \sup_{u\in[0,t_0]} |\widehat{S}_E(u) - S_E^*(u)| &\xrightarrow{p} 0, \end{aligned}$$

by uniform consistency on $[0,t_0)$. Also $|\Delta S(c)| = |\widehat{S}_E(c) - S_E^*(c)| \le \sup_{u\in[0,t_0]} |\widehat{S}_E(u) - S_E^*(u)| \xrightarrow{p} 0$.

Since $\sup_{u\in[0,t_0)} \left|\widehat{S}_E(u) - S_E^*(u)\right| \xrightarrow{p} 0$ and $s_{\min} > 0$, we have $\Pr\left(|\Delta S(c)| < \frac{s_{\min}}{2}\right) \xrightarrow{p} 1$, hence $\widehat{S}_E(c) \ge \frac{s_{\min}}{2}$ with probability close to 1.

Then we can apply Lemma D.2 (because we have $|\Delta S(c)| \le s_{\min}/2$) with $S = S_E^*$ and $\widehat{S} = \widehat{S}_E$ to conclude $\widehat{e}^{\,\mathrm{margin}}(c) - e^{*,\mathrm{margin}}(c) \xrightarrow{p} 0$. The correctly specified case follows from $S_E^* = S_E$. □

D.3.3 First-order expansion and delta method

In the preceding section, we prove the "law of large numbers property" of the margin time – in the limit, we have asymptotic consistency. Now, we want to demonstrate the convergence rate.

Let us first define some notation for convenience. Fix $c \ge 0$. Define

$$\mathcal{T}_c(S) \;:=\; e^{\mathrm{margin}}(c;S) \;=\; c + \frac{A(c;S)}{S(c)}.$$

Lemma D.4. *Let S be a survival function with $S(c) > 0$ and $A(c;S) < \infty$. Let h be any bounded function on $[c,\infty)$ such that $S(c)+h(c) > 0$ and $A(c;S+h) < \infty$. Write $S_h := S + h$ and define $\Delta A := A(c;S_h) - A(c;S) = \int_c^{\infty} h(u)\,du$. Then*

$$\mathcal{T}_c(S_h) - \mathcal{T}_c(S) \;=\; \underbrace{\frac{1}{S(c)}\int_c^{\infty} h(u)\,du - \frac{A(c;S)}{S(c)^2}\,h(c)}_{\dot{\mathcal{T}}_{c;S}(h)} + R_{c;S}(h), \tag{D.23}$$

where the remainder is

$$R_{c;S}(h) \;=\; -\frac{\Delta A\; h(c)}{S(c)^2} + \frac{A(c;S_h)\,h(c)^2}{S(c)^2\,(S(c)+h(c))}. \tag{D.24}$$

In particular, if $|h(c)| \le S(c)/2$, *then* $R_{c;S}(h) \;=\; O\big(|h(c)|^2 + |\Delta A|\,|h(c)|\big)$.

Proof. The proof is identical to the finite-horizon case, replacing $\int_c^{t_0}$ by $\int_c^{\infty}$. Start from

$$\mathcal{T}_c(S_h) - \mathcal{T}_c(S) \;=\; \frac{A(c;S_h)}{S_h(c)} - \frac{A(c;S)}{S(c)}.$$

Add and subtract $A(c;S_h)/S(c)$ to obtain

$$\begin{aligned}\frac{A(c;S_h)}{S_h(c)} - \frac{A(c;S)}{S(c)} \;&=\; A(c;S_h)\left(\frac{1}{S_h(c)} - \frac{1}{S(c)}\right) + \frac{A(c;S_h)-A(c;S)}{S(c)} \\ &=\; A(c;S_h)\left(\frac{1}{S_h(c)} - \frac{1}{S(c)}\right) + \frac{\Delta A}{S(c)}.\end{aligned}$$

Use the identity $\frac{1}{a+b} - \frac{1}{a} \;=\; -\frac{b}{a^2} + \frac{b^2}{a^2(a+b)}$ with $a = S(c)$ and $b = h(c)$, and then decompose $A(c;S_h) = A(c;S) + \Delta A$. Collecting terms yields (D.23)–(D.24). □

Theorem 2 in Rivest and Wells [2001] states that, for any t_0 satisfying $\pi(t_0) > 0$ and suitable bounded-derivative conditions, $\sqrt{n}\big(\widehat{S}_E(t) - S_E^*(t)\big)$ converges weakly on $D[0,t_0)$ to a mean-zero Gaussian process, where $\pi(t) = \Pr(\min(E,C) \ge t)$. Because the margin-time functional depends only on $\widehat{S}_E$ over $[0,t_0]$, this functional CLT is sufficient.

Theorem D.5. *Assume:*

(a) Rivest–Wells Theorem 2 holds at some $t_0 > c$, i.e., $\sqrt{n}(\widehat{S}_E - S_E^*) \Rightarrow \mathbb{G}_E$ *on* $D[0,t_0)$, *where* $\mathbb{G}_E$ *has almost surely continuous sample paths;*

(b) no tail contribution beyond t_0*;*

(c) $S_E^*(c) > 0$.

Then

$$\sqrt{n}\Big(\widehat{e}^{\,\mathrm{margin}}(c) - e^{*,\mathrm{margin}}(c)\Big) \quad\Rightarrow\quad \dot{\mathcal{T}}_{c;S_E^*}(\mathbb{G}_E) \;=\; \frac{1}{S_E^*(c)}\int_c^{\infty} \mathbb{G}_E(u)\,du - \frac{\int_c^{\infty} S_E^*(u)\,du}{(S_E^*(c))^2}\,\mathbb{G}_E(c). \tag{D.25}$$

Moreover, under no tail contribution assumption, the integrals in (D.25) *equal* $\int_c^{t_0}(\cdot)\,du$.

Proof. Let $h_n := \widehat{S}_E - S_E^*$. By Lemma D.4 applied with $S = S_E^*$ and $h = h_n$,

$$\widehat{e}^{\,\mathrm{margin}}(c) - e^{*,\mathrm{margin}}(c) \;=\; \dot{\mathcal{T}}_{c;S_E^*}(h_n) + R_{c;S_E^*}(h_n).$$

Multiply by $\sqrt{n}$:

$$\sqrt{n}\Big(\widehat{e}^{\,\mathrm{margin}}(c) - e^{*,\mathrm{margin}}(c)\Big) \;=\; \dot{\mathcal{T}}_{c;S_E^*}\big(\sqrt{n}h_n\big) + \sqrt{n}\,R_{c;S_E^*}(h_n).$$

Under no tail contribution, $\widehat{S}_E(u) = S_E^*(u) = 0$ for $u \ge t_0$ (w.p. $\to 1$), so $\int_c^{\infty} h_n(u)\,du = \int_c^{t_0} h_n(u)\,du$ and likewise for $A(c;S_E^*)$. Thus $\dot{\mathcal{T}}_{c;S_E^*}$ is linear and is continuous at every path that is continuous at c.

By Theorem 1 in Rivest and Wells [2001], $\widehat{S}_E$ is uniformly consistent for S_E^* on $[0,t_0)$, so $\|h_n\|_{\infty,[0,t_0]} \xrightarrow{p} 0$. The remainder in Lemma D.4 is second order in $h_n(c)$ and $\Delta A = \int_c^{t_0} h_n$, hence $\sqrt{n}R_{c;S_E^*}(h_n) = o_p(1)$ under the same tightness argument implied by the functional CLT in (a).

Since $\sqrt{n}h_n \Rightarrow \mathbb{G}_E$ on $D[0,t_0)$ and $\mathbb{G}_E$ has almost surely continuous sample paths, $\dot{\mathcal{T}}_{c;S_E^*}$ is continuous at $\mathbb{G}_E$ almost surely. The continuous mapping theorem therefore yields

$$\dot{\mathcal{T}}_{c;S_E^*}\big(\sqrt{n}h_n\big) \Rightarrow \dot{\mathcal{T}}_{c;S_E^*}(\mathbb{G}_E).$$

Combining this with $\sqrt{n}R_{c;S_E^*}(h_n) = o_p(1)$ and Slutsky's theorem gives (D.25). □

E EXPERIMENTAL DETAILS

E.1 SYNTHETIC EXPERIMENT DETAILS

We evaluate the behavior of survival evaluation metrics under controlled censoring mechanisms using a fully synthetic data-generating process (DGP) with known ground-truth event times. The primary goal of these experiments is to isolate the effect of dependent censoring on metric bias while keeping the predictive model fixed.

All experiments use Weibull marginal distributions for both event and censoring times and induce dependence through a copula function. Synthetic datasets are generated using a linear Weibull DGP, and all experimental factors other than the evaluation metric are held fixed.

Covariates. For each experiment, we generate N independent covariate vectors

$$\boldsymbol{x}_i \sim \mathrm{Unif}([0,1]^d),$$

with $d = 10$ and $N = 10{,}000$. Covariates are fixed across different censoring dependence levels for each random seed to isolate the effect of the censoring mechanism.

Event and censoring models. Both the event time E and censoring time C follow Weibull distributions with covariate-dependent hazards. Conditional on $\boldsymbol{x}_i$, the hazard and survival functions take the form

$$h(t \mid \boldsymbol{x}_i) = \left(\frac{v}{\rho}\right)\left(\frac{t}{\rho}\right)^{v-1} \exp\big(g(\boldsymbol{x}_i)\big), \tag{E.1}$$

$$S(t \mid \boldsymbol{x}_i) = \exp\left(-\left(\frac{t}{\rho}\right)^{v} \exp\big(g(\boldsymbol{x}_i)\big)\right). \tag{E.2}$$

We use the following parameters throughout:

$$(v_E, \rho_E) = (4, 17), \qquad (v_C, \rho_C) = (3, 12).$$

The risk functions are linear:

$$g_E(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_E, \qquad g_C(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_C,$$

where $\boldsymbol{\beta}_E, \boldsymbol{\beta}_C \in \mathbb{R}^d$ are sampled independently for each random seed from $\mathrm{Unif}([-1,1]^d)$.

Copula-based dependence. Dependence between event and censoring times is introduced using copula-based constructions. For each instance i, a pair of uniform random variables

$$(u_i, v_i) \in [0,1]^2$$

is sampled from a copula with a specified Kendall's τ.

For Archimedean copulas (Clayton and Frank), (u_i, v_i) is sampled directly using the standard Archimedean construction, with the copula parameter chosen to match the target Kendall's τ via the corresponding analytical relationship. For the Gaussian copula, dependence is induced by first sampling a bivariate normal vector (z_{1i}, z_{2i}) with zero mean, unit variance, and correlation coefficient ρ. The correlation ρ is chosen such that the resulting Kendall's τ matches the target value. The uniform variables are then obtained via the probability integral transform:

$$u_i = \Phi(z_{1i}), \qquad v_i = \Phi(z_{2i}).$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

When $\tau = 0$, u_i and v_i are sampled independently, yielding conditional independence between E and C given $\boldsymbol{X}$.

Event and censoring times are subsequently generated via inverse transform sampling:

$$e_i = S_E^{-1}(v_i \mid \boldsymbol{x}_i), \qquad c_i = S_C^{-1}(u_i \mid \boldsymbol{x}_i).$$

Censoring rate calibration. To ensure that differences in metric behavior are not confounded by changes in censoring severity, the censoring distribution is calibrated to achieve approximately matched marginal censoring rates across copula families and dependence strengths.

This is accomplished by introducing a multiplicative scaling factor on the censoring hazard. For each copula family, a grid of scaling factors is evaluated in pilot simulations, and the resulting empirical censoring rates are recorded. Scaling factors are then selected to span a fixed set of target censoring-rate intervals, ensuring that each copula family contributes comparable censoring levels despite differences in dependence structure.

The selected scaling factors are reused across random seeds, yielding stable censoring rates with low variability while preserving the intended dependence between event and censoring times.

Observed data. The observed time and event indicator are given by

$$t_i = \min(e_i, c_i), \qquad \delta_i = \mathbb{I}\{e_i \leq c_i\}.$$

For oracle evaluation, the true event times e_i are retained for all instances.

Model fitting. For each simulated dataset, we fit a CoxPH model using the observed data $(t_i, \delta_i, \boldsymbol{x}_i)$. Survival curves $\widehat{S}(t \mid \boldsymbol{x}_i)$ are obtained using the Breslow estimator. The CoxPH model specification and hyperparameters are held fixed across experiments, while the model is refitted for each simulated dataset.

Metric evaluation. We compare two evaluation approaches:

- **IPCW Brier Score**, computed using inverse probability of censoring weights as implemented in the `SurvivalEVAL` package [Qi et al., 2023b].
- **Dependent Brier Score**, computed using the proposed dependent Brier score based on CG margin-time imputation with uncertainty weighting and with the copula family and dependence strength assumed known.

Metric error is defined as the absolute deviation from the oracle Brier score computed using the true event times. All results are averaged over 10 random seeds. Empirical censoring rates are additionally reported to verify calibration stability across dependence levels.

E.2 DATASETS AND PREPROCESSING

WHAS: Worcester Heart Attack Study dataset (WHAS) [Hosmer et al., 2008] contains 500 patients with acute myocardial infarction, focusing on the time to death post-hospital admission. The data was already post-processed and can be downloaded from the `scikit-survival` package [Pölsterl, 2020].

METABRIC: The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset [Curtis et al., 2012] contains survival information for breast cancer patients. This dataset includes a diverse range of feature sets that encompass clinical traits, expression profiles, copy number variation (CNV) profiles, and single-nucleotide polymorphism (SNP) genotypes. All these features are derived from breast tumor samples collected during the METABRIC trial. The dataset can be downloaded from `https://www.cbioportal.org/study/summary?id=brca_metabric`.

Churn: Customer churn prediction dataset (Churn) focuses on predicting customer attrition. Each record in the file corresponds to a customer and contains the same set of features, such as age, gender, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. We apply one-hot encoding on the US region feature and exclude subjects who are censored at time 0. The dataset can be downloaded from the PySurvival package [Fotso et al., 2019].

GBSG: The German Breast Cancer Study Group (GBSG) dataset [Royston and Altman, 2013] includes 686 patients with node-positive breast cancer and complete prognostic information. It is publicly available in R's `survival` package [Therneau, 2024]. While the original GBSG dataset features a higher censoring rate and a broader set of variables, we use a modified version that combines the dataset with uncensored cases from the Rotterdam dataset. This results in fewer variables, a lower censoring rate, and a larger sample size than the original. This combined dataset, which was already post-processed in the DeepSurv study [Katzman et al., 2018], is available for download at `https://github.com/jaredleekatzman`.

NACD: The Northern Alberta Cancer Dataset (NACD) [Haider et al., 2020] contains 2402 patients and 53 features. It is a conglomerate of many different cancer patients, including lung, colorectal, head and neck, esophageal, stomach, and other cancers. The event of interest in this dataset is failure time. We drop patients with negative or zero survival time leaving 2,396 patients. There are no missing values in this dataset. The dataset can be downloaded from `http://pssp.srv.ualberta.ca` under "Public Predictors".

FLCHAIN: The Serum Free Light Chain (FLCHAIN) dataset is a stratified random sample comprising half of the participants from a study investigating the association between serum free light chain (FLC) levels and mortality [Dispenzieri et al., 2012]. It is available in R's `survival` package [Therneau, 2024]. After downloading, we perform several preprocessing steps: we remove three subjects with events occurring at time zero, impute missing values in the `creatinine` feature using its median, and exclude the `chapter` feature—which encodes the cause of death according to ICD chapter headings—since it is only defined for deceased (uncensored) subjects and would therefore leak the event indicator to the model.

SUPPORT: The Study to Understand Prognoses, Preferences, Outcomes, and Risks of Treatment (SUPPORT) dataset [Knaus, 1995] includes 9,105 participants and was designed to investigate survival outcomes and clinical decision-making among seriously ill hospitalized patients. The dataset contains a substantial proportion of missing values across many features. We obtained the dataset from Qi et al. [2024] and follow their preprocessing steps.

Employee: Employee dataset contains employee activity information that can be used to predict when an employee will quit. The dataset can be downloaded from the PySurvival package [Fotso et al., 2019]. It contains duplicate entries; after dropping these duplicates, the number of subjects in the dataset is reduced from 14,999 to 11,991. We also apply one-hot encoding to the department information.

MIMIC-IV: The Medical Information Mart for Intensive Care IV (MIMIC-IV) [Johnson et al., 2023] contains critical care data from patients admitted to hospitals and intensive care units (ICU). We create a version using the MIMIC-IV database, which contains patients who are alive at least 24 hours after being admitted to ICU. Their date of death is derived from hospital records or state records, which means the cause of mortality is not limited to the reason for ICU admission. We follow the instructions in [Qi et al., 2023a] to process the dataset, and the code is available in their GitHub repository: `https://github.com/shi-ang/CensoredMAE`.

SEER: The SEER Program dataset [Ries et al., 2003] is a comprehensive collection of cancer patient data from approximately 49% of the U.S. population. Our study focuses on three subsets: SEER-brain, SEER-liver, and SEER-stomach, containing data on patients diagnosed with brain, liver, and stomach cancers, respectively. The goal is to model the time from diagnosis to failure events, such as death or disease progression. We preprocess the dataset as in [Qi et al., 2024]. Features analyzed include age, sex, cancer stage, grade, and various clinical factors. We excluded features with over 70% missing data, patients without follow-up times, duplicates, and those with a survival time of zero. The SEER cohort is available for download at `www.seer.cancer.gov`.

Figure E.1 shows the empirical KM survival curves and the distribution of event and censoring times for the generated semi-synthetic datasets used for training and evaluation. Reported censoring rates are computed from the observed indicators.

E.3 SURVIVAL LEARNERS

CoxPH: The Cox Proportional Hazards model (CoxPH) is a semi-parametric survival analysis model [Cox, 1972], consisting of a non-parametric baseline hazard function $h_0(t)$ and a parametric partial hazard function $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$, typically estimated using the Breslow method [Breslow, 1975]. The hazard function is $h(t \mid \boldsymbol{x}_i) = h_0(t) \exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$, where f is linear. The MLE of $\widehat{\boldsymbol{\theta}}$ maximizes the Cox partial log-likelihood [Cox, 1972]. CoxPH is implemented from the `scikit-survival` package [Pölsterl, 2020].

GBSA: GBSA [Ridgeway, 1999] is a likelihood-based boosting model that optimizes an ℓ_2-norm penalized Cox partial log-likelihood [Cox, 1972]. It updates candidate variables flexibly at each boosting step using an offset-based gradient boosting method, unlike traditional gradient boosting, which updates either a single component or all features for fitting the gradient. GBSA is implemented from the `scikit-survival` package [Pölsterl, 2020].

RSF: Random Survival Forests (RSF) [Ishwaran et al., 2008] extend decision trees for survival analysis by constructing multiple trees on bootstrapped samples. Each tree predicts an individual's survival function, with splits based on criteria like the log-rank test. RSF handles censored data as right-censored and accounts for it during tree construction, with final predictions as the average of all trees' survival functions. RSF is implemented from the `scikit-survival` package [Pölsterl, 2020].

DeepSurv: DeepSurv [Katzman et al., 2018] is a neural network implementation of the CoxPH model of the form $h(t \mid \boldsymbol{x}_i) = h_0(t) \exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$, where $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ denotes a risk score as a nonlinear function of the features, *i.e.*, $f(\boldsymbol{\theta}, \boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i \boldsymbol{\theta})$, where σ is a nonlinear activation function. Popular functions include the hyperbolic tangent (Tanh) function or the Rectified Linear Unit (ReLU). The MLE for $\widehat{\boldsymbol{\theta}}$ is derived by numerically maximizing the partial Cox log-likelihood. DeepSurv is implemented from `https://github.com/shi-ang/BNN-ISD/`.

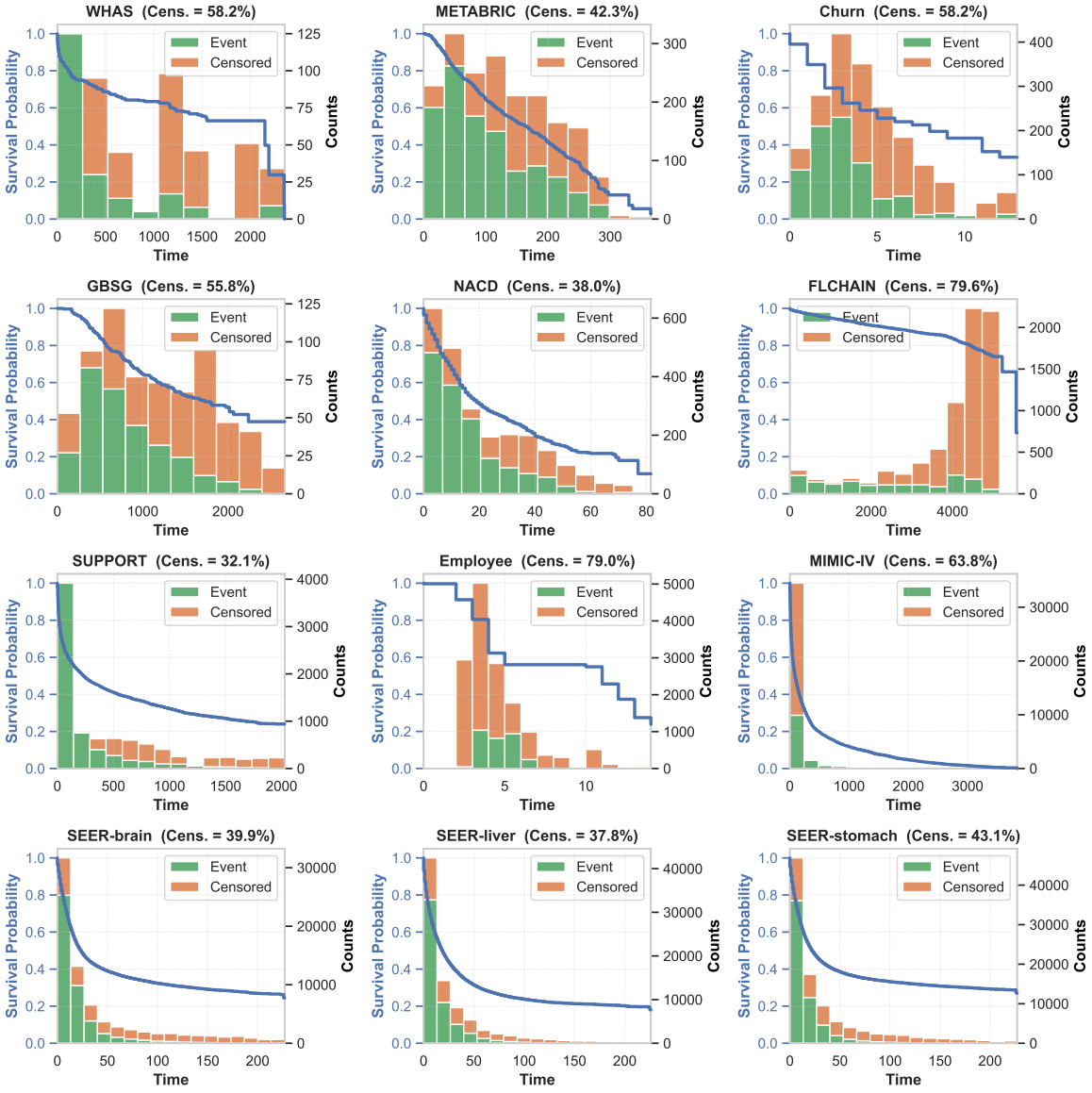


Figure E.1: Empirical KM survival curve and event/censoring time distribution for the generated semi-synthetic datasets used in training and evaluation.

MTLR: Multi-Task Logistic Regression (MTLR) [Yu et al., 2011] is a neural network implementation of a discrete-time model that predicts the survival distribution using a sequence of dependent logistic regressors. The number of discrete times is determined by the square root of the number of uncensored patients, and we use quantiles to divide those uncensored instances evenly into each time interval. MTLR is implemented from `https://github.com/shi-ang/BNN-ISD/`.

Table E.1 reports the selected hyperparameters for the survival learners.

E.4 IMPLEMENTATION DETAILS

Subsampling. For computational efficiency, we subsample the largest semi-synthetic datasets while preserving their event-time distributions. For MIMIC-IV, SEER, and Employee, we draw $N = 10{,}000$ rows stratified by event status and discretized time intervals.

Reproducibility. Experiments used Python 3.9, PyTorch 1.13.1, NumPy 1.24.3, and Pandas 1.5.3, with tensor computations in double precision (fp64). They were run on a workstation with an AMD Ryzen 5 5600X 3.70 GHz CPU, 32 GB RAM,

Table E.1: Hyperparameters for the survival learners.

Model	Parameter	Value
CoxPH	alpha	0.01
	ties	'Breslow'
	n_iter	100
	tol	1e-9
GBSA	n_estimators	100
	max_depth	1
	min_samples_split	2
	min_samples_leaf	1
	max_features	'sqrt'
	sub_sample	0.8
	random_state	0
RSF	n_estimators	100
	max_depth	1
	min_samples_split	2
	min_samples_leaf	1
	max_features	'sqrt'
	random_state	0
DeepSurv	hidden_size	100
	verbose	False
	lr	0.001
	c1	0.01
	num_epochs	1000
	batch_size	32
	dropout	0.25
	early_stop	True
	patience	10
MTLR	verbose	False
	lr	0.001
	c1	0.01
	num_epochs	1000
	batch_size	32
	early_stop	True
	patience	10

and an NVIDIA GeForce RTX 2060 (6 GB) GPU using CUDA 11.7. Parts of the pipeline and preprocessing utilities build on the *MENSA* codebase [Lillelund et al., 2026a].

Copula estimation details. For semi-synthetic experiments, copula and marginal parameters were optimized using Adam [Kingma and Ba, 2017] for up to 3×10^4 full-batch epochs. Learning rates were 10^{-3} for the Weibull marginals and 10^{-2} for the copula parameter θ. Early stopping was triggered when the validation NLL failed to improve for 100 epochs and θ changed negligibly across successive epochs.

F ADDITIONAL RESULTS

F.1 COMPUTATIONAL ANALYSIS

Time complexity. Table F.1 reports the runtime of copula fitting, the dominant cost of dependent metric evaluation. For each candidate copula family, two Weibull marginals and a copula parameter are fitted jointly, followed by validation-based model selection. Runtime ranges from approximately 5 to 75 minutes. After fitting, the dependent metrics add less than 1ms relative to standard estimators across all datasets.

Space complexity. Table F.1 also reports the increase in peak PyTorch-allocated GPU tensor memory during fitting. This excludes CPU RAM, non-PyTorch allocations, and reserved but unallocated GPU memory. The increase ranges from approximately 0.3 to 11MiB. Fitted models are discarded afterward and therefore do not affect metric-evaluation memory.

F.2 ESTIMATION ERROR

Table F.2 reports the full bias and variability results corresponding to the main text estimation error analysis.

Table F.1: Copula fitting time and incremental peak PyTorch GPU tensor memory across datasets under the *Original* feature selection strategy. Fitting includes training Clayton and Frank copula models jointly with their Weibull marginals.

Dataset	Copula Time (min)	Copula Memory (MiB)
METABRIC	15.67	1.12
GBSG	5.22	0.43
NACD	14.76	1.91
SUPPORT	67.36	5.90
FLChain	55.98	5.40
WHAS	10.61	0.35
Employee	54.41	6.08
Churn	5.59	1.32
MIMIC-IV (all)	74.24	10.68
SEER (brain)	28.97	5.72
SEER (liver)	68.35	5.92
SEER (stomach)	38.01	5.92

Table F.2: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and proposed dependent metrics across semi-synthetic datasets and feature selection strategies, averaged over 5 survival learners and 10 experiments.

		WHAS	METABRIC	Churn	GBSG	NACD	FLCHAIN
Original	IPCW (KM)	0.073 ± 0.038	0.061 ± 0.012	0.117 ± 0.016	0.125 ± 0.039	0.092 ± 0.013	0.156 ± 0.008
	IPCW (CoxPH)	0.106 ± 0.070	0.051 ± 0.012	0.115 ± 0.019	0.129 ± 0.040	0.148 ± 0.106	0.162 ± 0.008
	Dep (KM)	0.055 ± 0.025	0.047 ± 0.011	0.120 ± 0.012	0.130 ± 0.038	0.091 ± 0.010	0.135 ± 0.008
	Dep (CG)	0.055 ± 0.025	0.047 ± 0.011	0.120 ± 0.013	0.126 ± 0.037	0.091 ± 0.010	0.135 ± 0.008
	Dep (CG)$^{\text{UW}}$	0.063 ± 0.025	0.049 ± 0.010	0.124 ± 0.014	0.134 ± 0.037	0.082 ± 0.011	0.090 ± 0.008
Top-5	IPCW (KM)	0.071 ± 0.037	0.062 ± 0.012	0.121 ± 0.018	0.124 ± 0.039	0.103 ± 0.014	0.133 ± 0.008
	IPCW (CoxPH)	0.065 ± 0.037	0.061 ± 0.013	0.114 ± 0.016	0.126 ± 0.038	0.098 ± 0.014	0.138 ± 0.008
	Dep (KM)	0.055 ± 0.024	0.048 ± 0.011	0.120 ± 0.013	0.130 ± 0.038	0.098 ± 0.010	0.108 ± 0.008
	Dep (CG)	0.055 ± 0.024	0.048 ± 0.011	0.119 ± 0.014	0.126 ± 0.037	0.097 ± 0.012	0.108 ± 0.008
	Dep (CG)$^{\text{UW}}$	0.065 ± 0.024	0.049 ± 0.010	0.125 ± 0.016	0.134 ± 0.037	0.089 ± 0.012	0.066 ± 0.008
Top-10	IPCW (KM)	0.071 ± 0.038	0.061 ± 0.011	0.119 ± 0.016	0.125 ± 0.039	0.098 ± 0.013	0.143 ± 0.013
	IPCW (CoxPH)	0.082 ± 0.051	0.051 ± 0.012	0.111 ± 0.015	0.130 ± 0.041	0.093 ± 0.017	0.150 ± 0.014
	Dep (KM)	0.054 ± 0.025	0.047 ± 0.011	0.121 ± 0.012	0.131 ± 0.038	0.095 ± 0.010	0.123 ± 0.017
	Dep (CG)	0.054 ± 0.025	0.047 ± 0.011	0.121 ± 0.012	0.126 ± 0.037	0.095 ± 0.010	0.123 ± 0.017
	Dep (CG)$^{\text{UW}}$	0.063 ± 0.025	0.049 ± 0.011	0.125 ± 0.014	0.134 ± 0.037	0.086 ± 0.011	0.078 ± 0.009
Rand. 25%	IPCW (KM)	0.090 ± 0.030	0.061 ± 0.018	0.129 ± 0.014	0.153 ± 0.032	0.109 ± 0.013	0.140 ± 0.020
	IPCW (CoxPH)	0.091 ± 0.031	0.056 ± 0.018	0.122 ± 0.013	0.157 ± 0.036	0.111 ± 0.017	0.147 ± 0.013
	Dep (KM)	0.081 ± 0.022	0.049 ± 0.017	0.126 ± 0.009	0.153 ± 0.026	0.101 ± 0.010	0.140 ± 0.012
	Dep (CG)	0.081 ± 0.022	0.049 ± 0.017	0.117 ± 0.009	0.147 ± 0.026	0.099 ± 0.010	0.140 ± 0.012
	Dep (CG)$^{\text{UW}}$	0.095 ± 0.026	0.050 ± 0.017	0.131 ± 0.009	0.161 ± 0.026	0.091 ± 0.010	0.032 ± 0.013
		SUPPORT	Employee	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stomach)
Original	IPCW (KM)	0.108 ± 0.052	0.152 ± 0.009	0.026 ± 0.016	0.077 ± 0.005	0.104 ± 0.013	0.119 ± 0.018
	IPCW (CoxPH)	0.109 ± 0.052	0.100 ± 0.006	0.039 ± 0.027	0.055 ± 0.005	0.097 ± 0.015	0.095 ± 0.012
	Dep (KM)	0.111 ± 0.051	0.148 ± 0.007	0.023 ± 0.012	0.089 ± 0.005	0.118 ± 0.004	0.125 ± 0.008
	Dep (CG)	0.109 ± 0.051	0.154 ± 0.007	0.024 ± 0.012	0.084 ± 0.006	0.106 ± 0.004	0.112 ± 0.008
	Dep (CG)$^{\text{UW}}$	0.093 ± 0.051	0.152 ± 0.008	0.024 ± 0.013	0.067 ± 0.005	0.092 ± 0.004	0.097 ± 0.008
Top-5	IPCW (KM)	0.109 ± 0.052	0.155 ± 0.010	0.027 ± 0.017	0.076 ± 0.005	0.109 ± 0.014	0.124 ± 0.018
	IPCW (CoxPH)	0.109 ± 0.052	0.103 ± 0.005	0.027 ± 0.018	0.083 ± 0.022	0.109 ± 0.017	0.128 ± 0.020
	Dep (KM)	0.112 ± 0.051	0.149 ± 0.007	0.026 ± 0.011	0.090 ± 0.005	0.120 ± 0.005	0.126 ± 0.008
	Dep (CG)	0.109 ± 0.051	0.156 ± 0.007	0.027 ± 0.013	0.085 ± 0.005	0.111 ± 0.005	0.116 ± 0.008
	Dep (CG)$^{\text{UW}}$	0.094 ± 0.051	0.154 ± 0.008	0.025 ± 0.013	0.068 ± 0.005	0.096 ± 0.005	0.101 ± 0.008
Top-10	IPCW (KM)	0.109 ± 0.052	0.152 ± 0.009	0.026 ± 0.017	0.077 ± 0.005	0.103 ± 0.013	0.119 ± 0.017
	IPCW (CoxPH)	0.109 ± 0.051	0.101 ± 0.005	0.031 ± 0.026	0.055 ± 0.005	0.097 ± 0.015	0.099 ± 0.015
	Dep (KM)	0.111 ± 0.051	0.147 ± 0.007	0.023 ± 0.011	0.090 ± 0.005	0.118 ± 0.004	0.124 ± 0.008
	Dep (CG)	0.109 ± 0.051	0.153 ± 0.007	0.024 ± 0.011	0.085 ± 0.006	0.107 ± 0.004	0.111 ± 0.008
	Dep (CG)$^{\text{UW}}$	0.093 ± 0.051	0.152 ± 0.008	0.025 ± 0.013	0.068 ± 0.005	0.093 ± 0.004	0.097 ± 0.009
Rand. 25%	IPCW (KM)	0.142 ± 0.058	0.161 ± 0.016	0.016 ± 0.009	0.111 ± 0.027	0.128 ± 0.015	0.132 ± 0.017
	IPCW (CoxPH)	0.142 ± 0.058	0.110 ± 0.010	0.017 ± 0.011	0.111 ± 0.028	0.197 ± 0.105	0.133 ± 0.011
	Dep (KM)	0.145 ± 0.057	0.137 ± 0.008	0.031 ± 0.013	0.113 ± 0.017	0.127 ± 0.004	0.135 ± 0.011
	Dep (CG)	0.141 ± 0.057	0.120 ± 0.008	0.032 ± 0.014	0.102 ± 0.016	0.109 ± 0.005	0.120 ± 0.013
	Dep (CG)$^{\text{UW}}$	0.125 ± 0.057	0.147 ± 0.009	0.021 ± 0.007	0.089 ± 0.017	0.098 ± 0.004	0.106 ± 0.013

Appendix I

Position: Stop Chasing the C-index when Evaluating Survival Analysis Models

Abstract: The current state of evaluation in survival analysis is plagued by the persistent use of evaluation metrics in ways that are misaligned with the stated modeling objective. In addition, many such evaluations are based on censoring assumptions that are left implicit or unjustified. This means that the reported performance can be misleading and may fail to answer the scientific or modeling question the evaluation was intended to address. In this position paper, we critically examine evaluation practices in survival analysis and highlight how censoring makes evaluation fundamentally different from standard regression or classification. We place particular focus on concordance-based measures, such as the C-index, which we show are heavily overused in the literature. To help identify appropriate metrics, we propose a set of key desiderata and introduce a double-helix ladder, in which valid evaluation requires alignment between metric and modeling assumptions. Through controlled experiments, we show that violations of this alignment can lead to misleading model comparisons. We conclude by providing practical guidance on how to evaluate a survival model.

Authors: Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen.

Status: Accepted to the ICML 2026 Position Paper Track (Spotlight).
Preprint: arXiv:2506.02075.

Position: Stop Chasing the C-index when Evaluating Survival Analysis Models

Christian Marius Lillelund [1 2] **Shi-ang Qi** [1 3 4] **Russell Greiner** [1 3] **Christian Fischer Pedersen** [2]

Abstract

The current state of evaluation in survival analysis is plagued by the persistent use of evaluation metrics in ways that are misaligned with the stated modeling objective. In addition, many such evaluations are based on censoring assumptions that are left implicit or unjustified. This means that the reported performance can be misleading and may fail to answer the scientific or modeling question the evaluation was intended to address. In this position paper, we critically examine evaluation practices in survival analysis and highlight how censoring makes evaluation fundamentally different from standard regression or classification. We place particular focus on concordance-based measures, such as the C-index, which we show are heavily overused in the literature. To help identify appropriate metrics, we propose a set of key desiderata and introduce a double-helix ladder, in which valid evaluation requires alignment between metric and modeling assumptions. Through controlled experiments, we show that violations of this alignment can lead to misleading model comparisons. We conclude by providing practical guidance on how to evaluate a survival model.

1. Introduction

Survival analysis – which studies ways to predict when an event will occur – is widely used in domains such as healthcare (Cohen et al., 2008; Lee et al., 2021), mechanical engineering (Zhang et al., 2019), and economics (Ahn, 2023). A defining characteristic of survival data is *censoring*, where the event time is only partially observed for some instances. The most common case is *right-censoring*, in which only a *lower bound* of the event time is known – *e.g.*, when a patient has not yet experienced disease progression by the end of a clinical study. Because censored observations appear in the test set (as well as the training set), performance metrics must match the prediction objective and account for censoring to be meaningful. In practice, however, these considerations are often overlooked:

First, metrics are often misaligned with the stated objective: many projects use measures that use naive discrimination by default, even when the real objective of the project is different – *e.g.*, accurate time-to-event or probabilistic prediction. **Second, these metric choices are often made without considering of censoring or correcting for it.** Many projects proceed as if censoring were "random" – that is, independent of the event process – an assumption we define precisely below, yet such assumptions are rarely stated, justified, or adjusted for empirically. Because censored observations appear in both training and test sets, evaluation metrics must explicitly account for censoring to avoid biased performance estimates. In this setting, censoring is not a technical nuisance but a structural feature that determines whether an evaluation metric is meaningful.

Zhou et al. (2023) found that over 80% of studies in survival analysis used the *C-index* as the primary metric. The C-index (CI) is a discrimination metric: it measures whether predicted risk scores correctly rank individuals by event time[1]. While the limitations of the C-index are well documented (Hartman et al., 2023), our findings suggest that its widespread use points to a broader issue: a review of 92 methodological and application-based survival papers, from 2023 to 2025 (Appendix A), reveals that **approximately 72%** rely on evaluation metrics that do not align with their stated objectives, and often do so without stating their censoring assumptions (*e.g.*, random) or attempting to correct

[1]Department of Computing Science, University of Alberta, Edmonton, Canada [2]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark [3]Alberta Machine Intelligence Institute, Edmonton, Canada [4]Vector Institute, Toronto, Canada. Correspondence to: Christian Marius Lillelund <clillelund@ualberta.ca>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]At a high level, each instance is represented by a feature vector $\boldsymbol{x}_i \in \mathcal{X}$, where typically $\mathcal{X} = \mathbb{R}^d$. A model $M(\cdot)$ predicts a scalar risk score $M(\boldsymbol{x}_i) \in \mathbb{R}$. For two instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we interpret $M(\boldsymbol{x}_i) > M(\boldsymbol{x}_j)$ as predicting that instance i will experience the event (*e.g.*, die) earlier than instance j. The C-index then measures how often this ranking is correct among comparable pairs, *i.e.*, pairs where one observed event precedes the other time. See Appendix B for a formal definition.

for censoring to make those same metrics valid. Figures 1 and 2 summarize these findings.

Position. We argue that a substantial portion of recent survival analysis research is evaluated incorrectly, leading to misleading model comparisons and overstated performance claims. This arises from a systematic mismatch between modeling objectives, evaluation metrics, and the censoring assumptions those metrics rely on. In practice, discrimination-based metrics (in particular the C-index) are often misused to support claims about time-to-event prediction or probabilistic calibration, while censoring is assumed to be random (Definition 2.1) with little justification. To make this mismatch explicit, we introduce a ladder hypothesis between models, metrics, and censoring assumptions. Through controlled experiments, we show that popular metrics become increasingly misleading as censoring deviates from the common assumption of randomness. While prior work has discussed why and when the C-index can be problematic, our goal is not only to critique existing practice but to offer practical guidance for meaningful evaluation. In particular, we argue that evaluation should start from the research objective, make explicit the censoring assumptions one is willing to defend, and ideally select metrics that are valid under those assumptions.

Section	Content
Sec. 2	Survival analysis and censoring assumptions.
Sec. 3	Examples of model-metric mismatch and five desiderata for evaluating survival models.
Sec. 4	Aspects of evaluation and metrics.
Sec. 5	The Ladder Hypothesis and practical experiments illustrating model-metric mismatch.
Sec. 6	Recommendations for aligning evaluation choices with study objectives and censoring assumptions.
Sec. 7	Alternative views on model evaluation.

Code to run the experiments is available at: `https://github.com/thecml/position-cindex`.

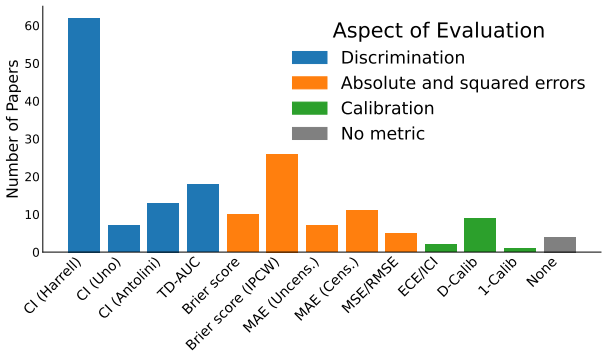


Figure 1. Overview of metric usage in survival analysis publications (2023-2025, see Appendix A). Papers reporting multiple metrics contribute one count to each metric. Note that most papers rely on metrics that only assess a model's discrimination (*i.e.*, ranking) ability, not how well it can make accurate time-to-event estimates nor provide calibrated survival probabilities.

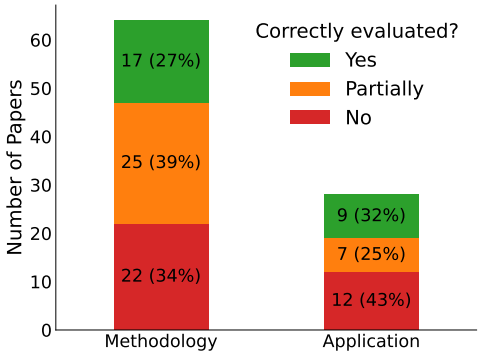


Figure 2. Overview of metric correctness in survival analysis publications (2023-2025; see Appendix A). Each paper is labeled **Yes**, **Partially**, or **No** depending on whether the reported evaluation metrics align with the stated modeling objective and whether censoring assumptions are stated, justified or corrected for to make those metrics valid. We find that 73% of surveyed methodology papers and 68% of application-focused papers do not fully satisfy these criteria.

2. Survival Analysis

2.1. Survival Data

We consider a survival dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes covariates, $t_i \in \mathbb{R}_+$ the observed time, and $\delta_i \in \{0, 1\}$ the event indicator. At the population level, let $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ denote the covariate vector, and let E and C denote nonnegative event-time and censoring-time random variables, with realizations e_i, c_i. We consider right-censored data, where

$$T := \min\{E, C\}, \qquad \Delta := \mathbb{1}[E \leq C].$$

Thus, each observed triplet $(\boldsymbol{x}_i, t_i, \delta_i)$ realizes $(\boldsymbol{X}, T, \Delta)$. We assume these triplets are drawn i.i.d. from $\Pr(\boldsymbol{X}, T, \Delta)$, equivalently induced by $\Pr(\boldsymbol{X}, E, C)$. Here, we focus on survival models that estimate an individual survival distribution (ISD), an instance-specific survival function describing the probability that the event has not occurred beyond each time point (Haider et al., 2020). Given $\boldsymbol{x}_i$, an ISD model $M(\cdot)$ produces $M(\boldsymbol{x}_i) = \widehat{S}(\cdot \mid \boldsymbol{x}_i)$, where

$$\widehat{S}(t \mid \boldsymbol{x}_i) \approx \Pr(E > t \mid \boldsymbol{X} = \boldsymbol{x}_i), \qquad t \geq 0.$$

ISDs support time-to-event estimates, pointwise event probabilities, and derived risk scores. We refer to these outputs as prediction targets.

2.2. Censoring Assumptions

We typically distinguish three assumptions about the relationship between E and C (see Figure 3), which is generally not identifiable from the observed data (Tsiatis, 1975).

These assumptions affect both the validity of a survival learner[2] and its evaluation metric.

Definition 2.1 (**Random Censoring**). Instances censored at time t are representative of all instances still at risk at t, in terms of their survival experience (Kleinbaum & Klein, 2012, Ch. 1). For example, in a medical study, patients are censored because the study ends, or because they move away for reasons unrelated to health. These are assumed to have the same subsequent death rate as uncensored patients who remain in the risk set[3]. Formally, $E \perp\!\!\!\perp C$[4].

Definition 2.2 (**Independent Censoring**). Within any subgroup of interest (*e.g.*, smokers or non-smokers), instances censored at time t are representative of those who remain at risk at t with respect to their survival experience (Kleinbaum & Klein, 2012). This is also called conditional independent censoring, since observed covariates are assumed to explain the event–censoring relationship. In a cancer study, tumor grade may affect both survival and loss to follow-up, so censoring may be associated with event time marginally. However, if loss to follow-up is unrelated to survival within each tumor-grade group, censoring is independent conditional on tumor grade. Formally, $E \perp\!\!\!\perp C \mid \boldsymbol{X}$[5].

Definition 2.3 (**Dependent Censoring**). Within any subgroup of interest, instances censored at time t are *not* representative of those who remain at risk at time t with respect to their survival experience. Equivalently, censoring is dependent when it remains related to the event process even after conditioning on the observed covariates. In the aforementioned cancer study, tumor grade may be recorded, but unobserved frailty, socioeconomic status, or treatment tolerance may affect both survival and dropout from follow-up. Then, even within each tumor-grade group, censored patients may have different survival prospects from those who remain under observation. Formally, $E \not\perp\!\!\!\perp C \mid \boldsymbol{X}$.

3. Why Evaluation Matters

Censoring is not a single mechanism, but may behave in various ways between datasets. After a learner has learned from a dataset that includes censored instances, the resulting model will then be evaluated against held-out data, which also includes censored instances[6]. The key distinction is whether censored individuals remain representative of those still under observation marginally, only after conditioning on observed covariates, or not even after such conditioning.

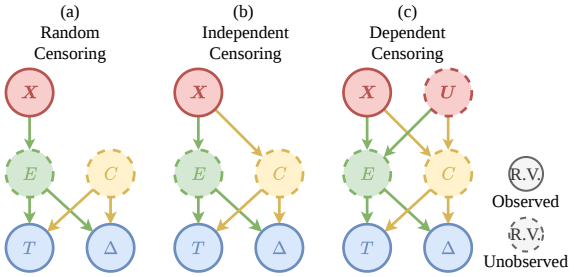


Figure 3. Comparison of three censoring assumptions, illustrated using directed acyclic graphs (DAGs) – where directed arcs show statistical dependence (Pearl, 2009). (a) Random censoring involves a random censoring time unrelated to the event time. (b) Independent censoring assumes that the censoring time is statistically independent of the event time given the observed covariates. (c) Dependent censoring violates this assumption as an unobserved confounding variable induces dependence between E and C.

3.1. Objective-Metric Mismatch

Our analysis shows that the community does not always carefully consider how their models are evaluated or which criteria are used to assess them. Here are three examples.

Steinberg et al. (2024) presented a foundation model called *MOTOR*, designed to predict time-to-event outcomes from 55M patient records. The authors focused on accurately predicting the time to a specific event, such as when a particular diagnosis code will be assigned to a patient, but did not directly evaluate time-to-event accuracy. Instead, they evaluated MOTOR using Brier score (Brier, 1950), 1-Calibration (D'Agostino & Nam, 2003), and two C-index variants (Harrell Jr. et al., 1996; Heagerty & Zheng, 2005). The latter only assess risk ranking[7].

Zisser & Aran (2024) presented a Transformer-based model to predict patient-specific survival probabilities for acute lymphoblastic leukemia. Although their goal was to predict the actual time to event, they again only evaluated their

[2]Here, a learner refers to a statistical or machine learning procedure that uses censored time-to-event data to fit a model, such as a Cox proportional hazards (CoxPH) model (Cox, 1972) or a Random Survival Forest (Ishwaran et al., 2008). Given an instance $\boldsymbol{x}_i$, the fitted model can then predict quantities such as hazard functions, survival functions, or risk scores.

[3]The *risk set* at time t consists of all individuals who are still under observation and have experienced neither the event nor censoring just prior to time t. These individuals are considered at risk of experiencing the event at time t, and thus contribute to the estimation of the hazard and related quantities.

[4]Meaning $P(E > t \mid C \geq t) = P(E > t)$ for all t.

[5]Meaning $P(E > t \mid C \geq t, \boldsymbol{X}) = P(E > t \mid \boldsymbol{X})$ for all t.

[6]Real datasets can have many censored instances. For example, in a 23-year prostate cancer study, approximately 68-80% of men did not die from prostate cancer and were therefore censored (Bill-Axelson et al., 2018).

[7]A model may achieve a high C-index by correctly ranking individuals while systematically misestimating event times, *e.g.*, predicting each event time to be twice as large as its true value; rankings are preserved, giving a high C-index, but absolute errors such as MAE are large. Conversely, a model may predict event times close to the population mean for all individuals, resulting in a good (low value) MAE but poor discrimination, since little individual-level variation is captured and rankings are largely random.

model using Harrell's C-index. This was a mismatch between the objective and evaluation: they wanted correct time-to-event predictions, but the C-index only told them how good their model was at discriminating patients. Because of this, we do not know if the model's predictions can be trusted for actual time-to-event prediction.

Liu et al. (2025) presented *HACSurv*, a copula-based survival model designed to capture dependence between competing risks and censoring. Although their goal was to exploit dependencies, they evaluated their method on datasets using standard metrics, such as Antolini's C-index (Antolini et al., 2005) and the integrated Brier score (IBS), where the latter uses IPCW[8] based on a Kaplan-Meier (KM) censoring estimate (Kaplan & Meier, 1958; Graf et al., 1999; Gerds & Schumacher, 2006). Since the KM estimator assumes independent censoring, this creates a mismatch: HACSurv is designed to model dependence involving censoring, while IBS and related IPCW-based measures are only valid under the independent censoring assumption. The authors acknowledge that IBS may be unreliable under dependent censoring; nevertheless, this example illustrates the broader issue that empirical claims based on standard survival metrics require explicit justification of the censoring assumptions under which those metrics are valid.

3.2. Metric Requirements

Below we propose five specific desiderata for metrics that evaluate survival models. These desiderata are desirable properties and formalize what it means for an evaluation metric to be appropriate for a given scientific goal. In the following sections, we use these desiderata to analyze commonly used metrics.

D1 Be a proper scoring rule. A proper scoring rule is defined as a scoring function (or loss) that is minimized (or maximized, depending on convention) exactly when the predicted probabilities match the true underlying probability distribution (see Appendix C for a formal definition). A metric should be a proper scoring rule because it rewards truthful predictions – *i.e.*, models that better reflect the underlying data-generating process should receive better scores.

D2 Be interpretable. A metric should be expressed in domain-relevant, human-understandable units (*e.g.*, days, weeks, months; or probabilities) and described using common, non-technical terminology. Its value should have a direct and intuitive connection to model behavior, without requiring additional transformations or assumptions. This allows practitioners and non-technical stakeholders to readily understand what is being measured and how differences in the metric reflect differences in model performance.

D3 Be model-agnostic. A metric should be independent of any model internals (*i.e.*, parameters) or information from the covariate distribution $P(\boldsymbol{X})$. This increases its comparability and robustness: we cannot fairly compare two models if the metric we are using depends on parameter choices or on quirks of $P(\boldsymbol{X})$. For example, a non-model-agnostic metric may assign different performance scores to two models that give identical predictions, simply because one has more parameters.

D4 Be sensitive to miscalibration. When a metric is used to evaluate probabilistic survival predictions, it should be sensitive to inaccurate survival probabilities. This is useful for detecting whether a model systematically overestimates or underestimates survival across the study period. For example, a model might predict that each of 100 patients has a 70% chance of surviving next year, while only 50 patients actually do.

D5 Be robust to censoring. A metric should support censoring in a manner consistent with its occurrence in the data. If our analysis suggests that censoring is indeed random (Definition 2.1), and censored individuals are equally likely to experience the event, the metric should reflect this observation. Conversely, if it suggests that censoring is dependent (Definition 2.3), the metric should account for this dependence.

4. Aspects of Evaluation and Common Metrics

We now introduce three aspects of evaluation in survival analysis and common metrics in the field; formal definitions are provided in Appendix B.

Discrimination. Discrimination metrics assess how well a model ranks individuals by predicted risk or event time. They measure relative ordering rather than the numerical accuracy of predicted risks, survival probabilities, or event-time estimates. Under censoring, this ordering must also be estimated under assumptions about which pairs are comparable and how censoring is handled. Figure 1 shows that the most common kind is Harrell's C-index, which estimates the proportion of correctly ordered pairs among comparable instances[9]. It can be biased under right-censoring because

[8] IPCW corrects for censoring by weighting observed outcomes by the inverse probability of remaining uncensored up to the relevant time. Marginal IPCW estimates the censoring survival function by $\widehat{G}(t) = \widehat{\Pr}(C > t)$, typically using the KM estimator. Conditional IPCW instead uses a covariate-dependent estimate $\widehat{G}(t \mid \boldsymbol{x}) = \widehat{\Pr}(C > t \mid \boldsymbol{X} = \boldsymbol{x})$. Using marginal IPCW implicitly assumes censoring does not depend on covariates; if censoring depends on $\boldsymbol{X}$, conditional censoring weights are required, otherwise the estimator can be biased.

[9] A pair is comparable if one instance experiences the event before the other; concordance means the instance with higher predicted risk has the event first.

Table 1. Overview of commonly used evaluation metrics and which desideratum each satisfies. A checkmark (✓) indicates that the metric fully satisfies the desideratum, a triangle (▲) indicates partial satisfaction, and a cross (✗) indicates that the metric does not satisfy the desideratum. For the interpretability desideratum, we rank the metrics on a scale from 1 to 7, where 1 indicates the lowest interpretability and 7 indicates the highest. The ranking follows desideratum **D2**, assigning higher ranks to metrics whose units and values are more directly understandable. The log-likelihood (LL) is included merely for comparison, as it is typically not used as a final evaluation metric.

Desideratum	CI (Harrell)	CI (Uno)	CI (Antolini)	IBS	MAE	D-Cal	LL
Proper scoring rule (D1)	✗	✗	✗	✓	▲	✗	✓
Interpretable (D2)	#2	#3	#5	#6	#1	#4	#7
Model-agnostic (D3)	▲	▲	▲	✓	✓	✓	✗
Sensitive to miscalibration (D4)	✗	✗	✗	✓	✗	✓	✓
Robust to censoring (D5)	✗	▲	✗	▲	✗	✗	▲

it estimates concordance over observed comparable pairs, which need not represent all pairs defined by the true event times. Uno's C-index (Uno et al., 2011) corrects this bias using IPCW, yielding an unbiased metric under random (using KM) or independent (using a covariate-based model) censoring. However, all commonly used C-index variants remain biased under dependent censoring, and none of the C-indices are proper scoring rules when the risk score is the predicted t-year event probability, $\widehat{r}_i(t) = 1 - \widehat{S}(t \mid \boldsymbol{x}_i)$ (Blanche et al., 2019). They are partially model-agnostic, as they do not rely on any model internals, but do require that the model can predict individual risk scores, not just ISDs[10].

Absolute and squared errors. Error-based metrics evaluate numerical discrepancies between model predictions and corresponding observed outcome summaries. The Brier score (Brier, 1950) measures the accuracy of predicted survival probabilities at a given time point using squared error: the prediction is a survival probability $\widehat{S}(t \mid \boldsymbol{x}_i)$, while the observed quantity is the survival status at time t, $\mathbb{1}[e_i > t]$, which equals 1 if individual i is event-free at t and 0 otherwise. To account for censoring, it is typically estimated using IPCW and averaged over time to yield the IBS, which summarizes prediction error across the study horizon. Another commonly used error metric is the MAE, which measures the mean absolute difference between a predicted event time $\widehat{e}_i$ and the underlying event time e_i. Because e_i is observed only when $\delta_i = 1$, MAE is either computed on uncensored observations (Schemper, 1990) or estimated using de-censoring methods such as MAE-Margin (Haider et al., 2020) or MAE-PO (Qi et al., 2023a). Although MAE and C-index may seem likely to favor the same models, this need not hold: MAE assesses accuracy of actual time predictions (*e.g.*, days, months), whereas C-index only checks rankings (Qi et al., 2023a). IBS is a proper scoring rule for survival distributions under random censoring (Rindt et al., 2022), while MAE is proper only for evaluating median survival time predictions, not full survival distributions (Qi et al., 2023a). IBS is sensitive to miscalibration, as it penalizes inaccurate survival probabilities over time, whereas MAE is not, since it evaluates only point predictions. When censoring is handled using KM, IBS and MAE are biased under dependent censoring (Emura & Chen, 2018).

Calibration. Calibration metrics evaluate whether predicted survival probabilities align with the event frequencies observed in the data. A common fixed-time approach is 1-Calibration (D'Agostino & Nam, 2003), which compares predicted event or survival probabilities at a prespecified horizon t^*, such as $\hat{S}(t^* \mid \boldsymbol{x}_i)$, with observed event frequencies by that time. It therefore assesses calibration at one clinically relevant horizon, not over the full survival curve. D-Calibration (Haider et al., 2020) takes a different approach: rather than fixing a time point, it evaluates whether predicted survival probabilities at the event times are distributionally calibrated[11]. Thus, D-Calibration assesses calibration of the predicted ISDs, rather than calibration at a single time horizon. However, neither 1-Calibration nor D-Calibration is a proper scoring rule, since each returns a goodness-of-fit p-value rather than a loss to be minimized. This p-value is also less directly interpretable than metrics expressed in probabilities or time units. Nevertheless, D-Calibration is model-agnostic because it only requires predicted ISDs. It is sensitive to distributional miscalibration, but when censoring is handled using KM, it is not robust to dependent censoring.

5. Model-Metric Consistency

We introduce a guiding principle called the *Ladder Hypothesis of Model-Metric Consistency* [12] (see Figure 4). In short,

[10] For ISD models, risk scores can be derived in several ways: Haider et al. (2020) use the negative predicted median survival time, whereas Blanche et al. (2019) use the predicted t-year event probability, $\widehat{r}_i(t) = 1 - \widehat{S}(t \mid \boldsymbol{x}_i)$.

[11] Under perfect distributional calibration, the values $\widehat{S}(e_i \mid \boldsymbol{x}_i)$ for uncensored instances are uniformly distributed on $[0, 1]$. D-Calibration tests this using a goodness-of-fit test; censored observations are distributed over compatible probability intervals.

[12] As an analogy, the distortion of the model-metric ladder can be compared to the emergence of a helical structure in physics: the ladder becomes "twisted" because introducing additional constraints (*e.g.*, dependent censoring or covariate-dependent mechanisms) renders a straight alignment between model and metric assumptions statistically unstable.

a model and its evaluation metric must stand on the same rung of the assumption ladder: if they rely on different censoring assumptions, neither the reported performance nor the metric estimate can be considered reliable.

In this section, we show that this principle is often overlooked. For example, models developed under an independent censoring assumption are often evaluated using metrics designed for random censoring. Even when these limitations are known, we speculate that authors may default to off-the-shelf metrics to maintain comparability with prior work, satisfy reviewer expectations, or because no widely accepted alternative exists (see also Alternative View 2 in Section 7).

Ladder Hypothesis of Model-Metric Consistency

A model can be considered trustworthy only when it is evaluated with metrics whose own validity rests on the same or weaker set of statistical assumptions that also govern the model. Conversely, an evaluation metric is itself trustworthy only when any auxiliary models or estimators on which it relies satisfy those same assumptions.

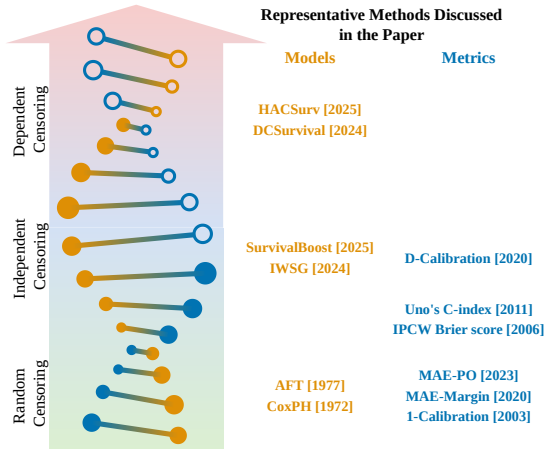


Figure 4. The double-helix ladder shows the current development of survival models (orange) and evaluation metrics (blue) under increasingly weakening censoring assumptions (going up). Filled dots indicate settings where valid methods exist, while hollow dots indicate unresolved gaps. The development of models has started to solve dependent censoring, while the development of metrics has not completely solved independent censoring yet. The resulting separation between the two strands highlights systematic model-metric mismatch across the assumption ladder.

5.1. A Simple Illustration

To illustrate the Ladder Hypothesis, we conduct a controlled experiment on synthetic data in which the predictive model (CoxPH) is held fixed and only the censoring mechanism is varied. We generate event times from the same covariate-dependent Weibull distribution across all settings, while varying censoring across three scenarios with progressively weaker assumptions: (1) random censoring, (2) independent censoring, and (3) dependent censoring induced via a Clayton copula[13] with Kendall's $\tau \in \{0.25, 0.5, 0.75\}$, where higher Kendall's τ corresponds to stronger dependence between event and censoring times. Note that standard CI and IBS implementations (Kvamme et al., 2019; Qi et al., 2024a) rely on global censoring corrections (*e.g.*, IPCW using the KM) that do not condition on $\boldsymbol{X}$, and therefore operate under random censoring. Full details of the data-generating process and experiment design are provided in Appendix D.

We evaluate the *same* predicted survival curves using both discrimination and error-based metrics. For discrimination, we consider the oracle[14] C-index, Harrell's C-index and Uno's C-index. For error-based metrics, we consider the oracle IBS, the naive IBS (no IPCW) and the IPCW-weighted IBS. Table 2 reports the oracle performance, which reflects the model's true predictive performance under each censoring mechanism. Since the predictive model is trained under an independent censoring assumption, stronger dependent censoring can also degrade the fitted model itself; the oracle metrics therefore capture genuine changes in predictive performance rather than evaluation bias. Figure 5 reports the resulting censored-data metric errors[15]. These errors isolate the additional bias introduced by evaluating censored test data without access to the true event times. As censoring deviates from the assumptions used by the metric estimators, CI and IBS estimates become increasingly biased, and may even suggest improvement when oracle performance is degrading.

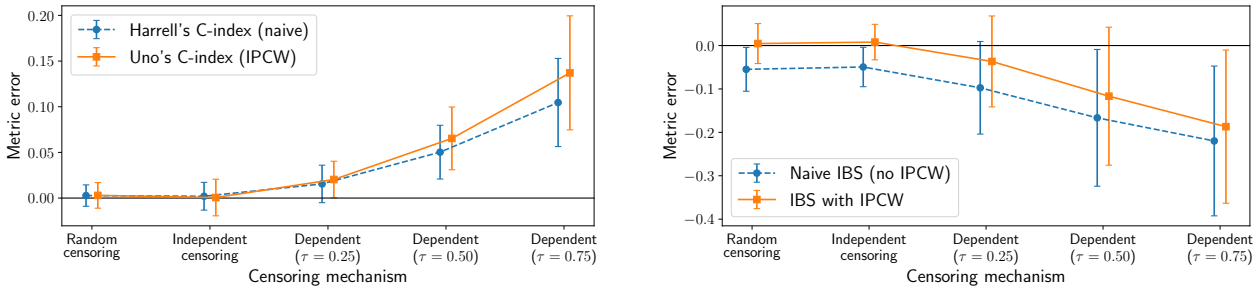
5.2. Random Censoring

Random censoring (*i.e.*, $E \perp\!\!\!\perp C$) is the first and most restrictive censoring assumption. Despite this limitation, almost all classical conditional survival algorithms (*i.e.*, methods that predict an individual outcome for instance i given covariates $\boldsymbol{x}_i$) were derived under random censoring, including CoxPH, Accelerated Failure Time (AFT) models (Farewell & Prentice, 1977), and Random Survival Forest (Ishwaran et al., 2008), etc. Evaluation practices evolved in parallel, and thus

[13] A copula is a link function that encodes cross-dependency between E and C, which applies to any marginal distribution.

[14] Oracle metrics are computed using *true* event times. This is possible in our simulation because the data-generating process records both the event time e_i and censoring time c_i for each instance i. The oracle metric therefore uses e_i directly and does not rely on censoring-related assumptions.

[15] The metric error is defined as the difference between the value reported by a given censored-data metric and its corresponding oracle value computed using the true event times.



(a) CI error. *(b)* IBS error.

Figure 5. Mean (± SD) metric errors averaged over 100 random seeds. Errors are relative to the model's oracle performance.

Table 2. Oracle model performance (± SD) under different censoring mechanisms, averaged over 100 random seeds.

Censoring	#Events (Cens.%)	$\text{CI}_{\text{oracle}}$ ↑	$\text{IBS}_{\text{oracle}}$ ↓
Random	2641 (73.6%)	0.634 ±0.018	0.090 ±0.040
Independent	3157 (68.4%)	0.634 ±0.018	0.084 ±0.037
Dependent ($\tau = 0.25$)	2969 (70.3%)	0.628 ±0.021	0.132 ±0.096
Dependent ($\tau = 0.50$)	2758 (72.4%)	0.618 ±0.025	0.199 ±0.144
Dependent ($\tau = 0.75$)	2536 (74.6%)	0.609 ±0.030	0.245 ±0.157

many popular metrics – such as the IPCW-adjusted time-specific Brier score (Graf et al., 1999; Gerds & Schumacher, 2006), time-specific calibration (D'Agostino & Nam, 2003), and Cox-Snell residuals (Cox & Snell, 1968) – were also introduced under the random-censoring framework. When the first models were developed, these metrics had not yet been proposed; researchers instead relied on calibration-style checks, for example, comparing average predictions with the KM estimator. Since the 1980s, these advances in evaluation have led researchers to validate models built under random censoring with metrics that make the same assumption (Brier score, Cox-Snell residuals) (Ishwaran et al., 2008; Van Belle et al., 2011; Yu et al., 2011). Consequently, these models and their evaluation procedures remain *model-metric consistent*.

5.3. Independent Censoring

Random censoring is rarely justifiable in practice; instead, modern survival analysis typically assumes that the censoring time C may depend on covariates $\boldsymbol{X}$ but is conditionally independent of the event time E given $\boldsymbol{X}$ (*i.e.*, $E \perp\!\!\!\perp C \mid \boldsymbol{X}$).

When independence holds only *given* $\boldsymbol{X}$, naively treating data as randomly censored can inflate or attenuate risk estimates. As an example, patients who leave follow-up early may differ systematically (*e.g.*, they are younger, healthier, wealthier) from those who remain, skewing both model training and evaluation. Ignoring this bias can yield biased estimates and misleading clinical conclusions. To mitigate censoring-induced bias, recent methods explicitly account for the censoring mechanism, either by modeling it directly or by reweighting censored observations (Robins & Rotnitzky, 1992; Cole & Hernán, 2004; Chapfuwa et al., 2021; Han et al., 2021; 2022; Alberge et al., 2025). Examples include the inverse-weighted survival game (IWSG) of Han et al. (2021) and SurvivalBoost (Alberge et al., 2025), both of which exploit covariate information in $\boldsymbol{X}$ to adjust for censoring.

Meanwhile, many popular evaluation metrics can differ from their oracle counterparts even under random censoring, especially in their commonly used unweighted forms, such as Harrell's C-index and the uncorrected Brier score (see Figure 5). This motivates IPCW-variants (see Section 4) that are unbiased if censoring weights are correctly estimated. Unfortunately, some papers still report *unweighted* C-index or Brier scores (see Figure 1); the resulting model-metric mismatch reintroduces the very same bias those models aim to eliminate. To our knowledge, consistently estimating IPCW weights under independent censoring is nontrivial in practice without additional modeling assumptions. As a result, evaluation procedures based on IPCW may fail to be consistent in practice. Therefore, these models and their evaluation procedures are not *model-metric consistent*.

5.4. Dependent Censoring

Dependent censoring (*i.e.*, $E \not\perp\!\!\!\perp C \mid \boldsymbol{X}$) is the most complicated form of censoring because the joint law of (E, C) is no longer identifiable from the observed-data distribution $\Pr(\boldsymbol{X}, T, \Delta)$ without further structural assumptions (Tsiatis, 1975). Existing approaches include sensitivity analyses for untestable dependence assumptions (Siannis et al., 2005; Huang & Zhang, 2008), partial-identification bounds under weaker assumptions, and model-based corrections that im-

pose structure on the joint distribution, for example through copulas (Huang & Zhang, 2008; Emura & Chen, 2018; Foomani et al., 2023; Zhang et al., 2024a; Liu et al., 2025) or proximal causal inference (Tchetgen Tchetgen et al., 2024; Ying, 2024). Representative copula-based approaches include DCSurvival (Zhang et al., 2024a) and HACSurv (Liu et al., 2025), which explicitly model dependence using a copula function. These developments mainly concern model estimation, however, not model evaluation.

Unfortunately, virtually all commonly used performance measures (Harrell's C-index, Uno's C-index, Brier score, MAE, D-Calibration) assume at least independent censoring. Under dependent censoring, they become biased, breaking model-metric consistency. Because there are no broadly accepted assumption-aligned metrics, most papers tackling dependent censoring report results on synthetic data where the ground truth is known (Qi et al., 2023a).Worryingly, some studies (see the last example in Section 3.1) nonetheless declare state-of-the-art performance on real datasets with alleged dependent censoring using independent-censoring metrics such as Antolini's C-index, IBS, or D-Calibration. Such scores are biased, rendering any claimed superiority an artifact of model-metric mismatch. One recent line of work by Lillelund et al. (2025a) addresses this gap by replacing the KM estimator used inside standard survival metrics with the Copula-Graphic estimator (Zheng & Klein, 1995). This estimator jointly models event and censoring times under an Archimedean copula (Emura & Chen, 2018), and yields copula-based analogues of common metrics that can empirically recover model error better under dependent censoring. However, such approaches require specifying a copula family, which introduces a strong and typically unverifiable assumption. Moreover, existing evidence is still mainly empirical, with few theoretical guarantees such as consistency or asymptotic normality for the resulting metrics.

In summary, survival analysis lacks broadly accepted evaluation metrics that remain unbiased under dependent censoring. Until such tools are developed and validated, performance claims for state-of-the-art models in this setting are, at best, speculative in real datasets. **Our central message is therefore a plea for recalibration:** researchers must scrutinize the assumptions behind the metrics they report, and the community should prioritize creating assumption-aligned evaluation procedures before proclaiming new modeling breakthroughs. This includes developing evaluation procedures that are robust to dependent censoring. In our opinion, rigorous evaluation should be the next frontier for progress under dependent censoring.

6. Practical Recommendations

Below we provide some practical recommendations for evaluation and consistency in survival analysis.

Recommendation 1 (Stick to your goal): In application-oriented studies, the primary evaluation metric should follow directly from the study objective, prediction target, and evaluation aspect: discrimination, absolute or squared error, calibration, or another aspect (Section 4). Risk ranking objectives (*e.g.*, prioritizing patients for a critical medical resource) call for discrimination metrics such as C-index or AUC variants (Hung & Chiang, 2010); time-to-event estimation favors error-based metrics such as MAE or MSE; and decision-making based on predicted survival probabilities is best assessed using calibration-oriented measures. Finally, among metrics that address the same aspect, the desiderata in Section 3.2 help clarify what the trade-offs are.

Recommendation 2 (Evaluate from multiple angles): In methodology-oriented studies (*e.g.*, developing a new general-purpose algorithm), we recommend reporting at least one representative metric from each of the aforementioned aspects (Section 4). Many new deep-learning methods increasingly predict the full survival curve (see Section 2.1); however, because the true survival curve is unobserved and its tail is often unreliable due to censoring, relying only on metrics that assess the full survival function can be misleading. We also recommend evaluating well-chosen scalar summaries[16], which can provide more stable and practically relevant feedback than comparing entire survival curves. For example, clinicians managing kidney disease usually act on a 2-year or 5-year predicted risk score (Stevens et al., 2024).

Recommendation 3 (Strive for consistency): Reported performance should be interpreted relative to the censoring assumptions supported by the chosen metric. Survival metrics encode specific assumptions about censoring, observability, and what constitutes a good prediction (see Section 4). We therefore recommend explicitly stating one's assumptions about censoring, choose metrics that align with these assumptions (*e.g.*, estimate censoring weights in Uno's C-index using a CoxPH model) and openly declare when model-metric consistency cannot be guaranteed. When such alignment is not possible, robustness checks, sensitivity analyses, or complementary metrics should be used to show that conclusions are not artifacts of the evaluation procedure itself. As alluded to, under conditionally independent censoring, a practical path forward is to replace marginal censoring corrections with covariate-dependent ones[17].

[16] Examples include landmark event probabilities $1-\widehat{S}(t^* \mid \boldsymbol{x}_i)$, the predicted median survival time $\inf\{t : \widehat{S}(t \mid \boldsymbol{x}_i) \leq 0.5\}$, and the restricted mean survival time (RMST) $\int_0^{t^*} \widehat{S}(t \mid \boldsymbol{x}_i)\,\mathrm{d}t$. The mean survival time is given by $\int_0^{\infty} \widehat{S}(t \mid \boldsymbol{x}_i)\,\mathrm{d}t$, but is often less stable because it depends on the tail of the predicted survival curve.

[17] For IPCW-based metrics, this means estimating the censoring survival function conditionally, $\widehat{G}(t \mid \boldsymbol{x}_i) = \widehat{\Pr}(C > t \mid \boldsymbol{X} = \boldsymbol{x}_i)$, rather than using a marginal KM estimate $\widehat{G}(t)$.

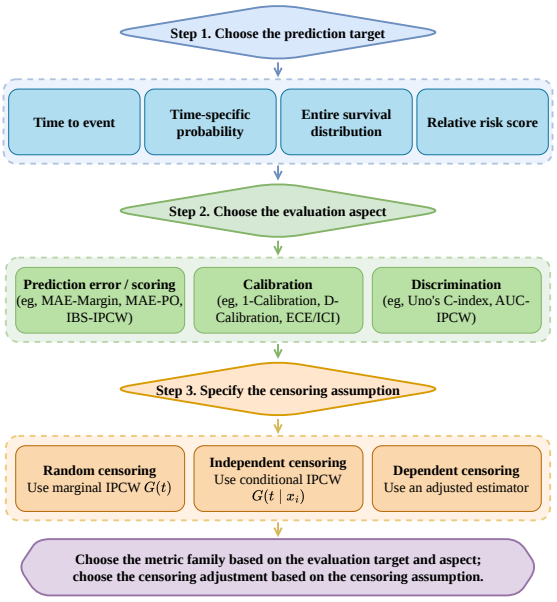


Figure 6. A workflow for selecting survival prediction metrics.

To make these recommendations actionable, we summarize them in the practical checklist (shown in Figure 6), which guides the selection of evaluation metrics based on the prediction target, evaluation aspect and censoring assumption.

7. Alternative Views

Alternative View 1 *In methodological papers, evaluation is often intended to be application-agnostic, and the C-index is therefore used as a general-purpose measure of predictive performance. Because it assesses a model's ability to correctly order individuals by risk without requiring well-calibrated absolute probabilities, it is robust to differences in baseline hazard specification and outcome prevalence. Empirically, improvements in C-index are frequently observed to coincide with improvements in other performance measures, making it a practical primary metric for comparing methods.*

Response Strong ranking performance does not imply calibrated survival probabilities or accurate time-to-event estimates (Hartman et al., 2023; Qi et al., 2023a). This is likely more pronounced for models that are trained only to optimize a ranking loss, *e.g.*, DeepSurv (Katzman et al., 2018) and the Hierarchical model (Tjandra et al., 2021), which have been shown to perform poorly on error and calibration-based metrics when reproduced (Lillelund et al., 2026a). We believe methodological contributions should be evaluated broadly rather than selectively, with metrics that cover the main aspects of performance, so that a method's strengths, weaknesses, and relevant baselines are clearly visible. Such transparency is essential for understanding how and when the proposed method offers real advantages.

Alternative View 2 *Even if standard evaluation metrics are biased under censoring, their use is still justified in practice. Authors may default to standard metrics to maintain comparability with prior work, to satisfy reviewer expectations or because no widely accepted alternatives exist under weaker censoring assumptions. Using familiar metrics also facilitates communication and interpretation within the community. Moreover, when comparing two models $M_A(\cdot)$ and $M_B(\cdot)$ on the same dataset, any censoring-induced bias from the dataset is shared across models, so relative performance comparisons remain meaningful.*

Response First, comparability with prior work does not mitigate the risk of propagating systematic error: consistent use of biased metrics propagates bias, not just noise. Biased metrics can make apparent progress an artifact of evaluation. Second, censoring-induced bias is not necessarily shared equally across models, especially when models rely on different censoring assumptions. For example, if $M_A(\cdot)$ implicitly assumes independent censoring while $M_B(\cdot)$ is designed to model dependent censoring, then evaluating both with a metric that assumes independence may fail to credit $M_B(\cdot)$ appropriately for its better predictive performance. Third, if no unbiased alternatives exist, that does not justify making strong empirical performance claims; instead, it means that some claims simply cannot yet be made reliably, especially when a method explicitly targets dependent censoring (see Section 3.1).

8. Conclusion

We argue that a substantial number of papers in survival analysis evaluate models using metrics that do not reflect the stated scientific goal or implicitly rely on censoring assumptions that are rarely stated, justified, or corrected for in practice. In a survey of 92 papers published between 2023 and 2025, we found approximately 72% fall under our criteria of misalignment. This suggests that this issue is widespread, and unfortunately can lead to overstated performance claims or prevent new models from being credited for methodological breakthroughs in modeling techniques. Therefore, we gave five desiderata to guide metric selection and showed why model-metric alignment is important with a concrete experiment. Going forward, we advocate for evaluation protocols in which metrics, assumptions, and data-generating mechanisms are made explicit and ideally aligned, and for further research on new metrics, especially under dependent censoring.

Acknowledgements

This research received support from the Natural Science and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), and the Alberta Machine Intelligence Institute (Amii).

References

Abdallah, W. B., Njah, H., and Jamoussi, S. New Dynamic Bayesian Network for Time-to-Event Prediction. In *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–7, 2024.

Abuhantash, F., Welsch, R., Finkelstein, S., and AlShehhi, A. Alzheimer's Disease Risk Prediction Using Machine Learning for Survival Analysis with a Comorbidity-Based Approach. *Scientific Reports*, 15:28723, 2025.

Ahn, H. J. Duration structure of unemployment hazards and the trend unemployment rate. *Journal of Economic Dynamics and Control*, 151:104664, 2023.

Alberge, J., Maladiere, V., Grisel, O., Abécassis, J., and Varoquaux, G. Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3619–3627, 2025.

Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24), 2005.

Apellániz, P. A., Parras, J., and Zazo, S. Leveraging the Variational Bayes Autoencoder for Survival Analysis. *Scientific Reports*, 14(1):24567, 2024.

Archetti, A. and Matteucci, M. Federated Survival Forests. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2023.

Baniecki, H., Sobieski, B., Szatkowski, P., Bombinski, P., and Biecek, P. Interpretable machine learning for time-to-event prediction in medicine and healthcare. *Artificial Intelligence in Medicine*, 159:103026, 2025.

Bian, Y., Yi, G., and He, W. Boosting Methods for Interval-censored Data with Regression and Classification. In *International Conference on Learning Representations*, volume 2025, pp. 52534–52565, 2025.

Bill-Axelson, A., Holmberg, L., Garmo, H., Taari, K., Busch, C., Nordling, S., Häggman, M., Andersson, S.-O., Andrén, O., Steineck, G., Adami, H.-O., and Johansson, J.-E. Radical Prostatectomy or Watchful Waiting in Prostate Cancer — 29-Year Follow-up. *New England Journal of Medicine*, 379(24):2319–2329, 2018.

Birolo, G., Rossi, I., Sartori, F., Rollo, C., Fariselli, P., and Sanavia, T. Beyond Cox Models: Assessing the Performance of Machine-Learning Methods in Non-Proportional Hazards and Non-Linear Survival Analysis. *Computers in Biology and Medicine*, 198:111176, 2025.

Blanche, P., Kattan, M. W., and Gerds, T. A. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

Bleistein, L., Nguyen, V.-T., Fermanian, A., and Guilloux, A. Dynamic survival analysis with controlled latent states. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4160–4204, 2024.

Blythe, R., Parsons, R., Barnett, A. G., Cook, D., McPhail, S. M., and White, N. M. Prioritising deteriorating patients using time-to-event analysis: prediction model development and internal–external validation. *Critical Care*, 28 (1):247, 2024.

Bone-Winkel, G. F. and Reichenbach, F. Improving credit risk assessment in P2P lending with explainable machine learning survival analysis. *Digital Finance*, 6(3):501–542, 2024.

Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Buyrukoğlu, G. Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction. *PeerJ Computer Science*, 10:e2147, 2024.

Chapfuwa, P., Assaad, S., Zeng, S., Pencina, M. J., Carin, L., and Henao, R. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145, 2021.

Chen, G. H. Survival Kernets: Scalable and Interpretable Deep Kernel Survival Analysis with an Accuracy Guarantee. *Journal of Machine Learning Research*, 25:1–78, 2024.

Chen, J., Liu, P., Chen, C., Su, Y., Zuo, E., Li, M., Wang, J., Yan, Z., Chen, X., Chen, C., and Lv, X. TDMFS: Tucker decomposition multimodal fusion model for pan-cancer survival prediction. *Artificial Intelligence in Medicine*, 162:103099, 2025.

Cohen, S. J., Punt, C. J. A., Iannotti, N., Saidman, B. H., Sabbath, K. D., Gabrail, N. Y., Picus, J., Morse, M., Mitchell, E., Miller, M. C., Doyle, G. V., Tissing, H., Terstappen, L. W. M. M., and Meropol, N. J. Relationship of Circulating Tumor Cells to Tumor Response, Progression-Free Survival, and Overall Survival in Patients With Metastatic Colorectal Cancer. *Journal of Clinical Oncology*, 26(19):3213–3221, 2008.

Cole, S. R. and Hernán, M. A. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Cox, D. R. and Snell, E. J. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265, 1968.

Cui, C., Tang, Y., and Zhang, W. Deep Survival Analysis With Latent Clustering and Contrastive Learning. *IEEE Journal of Biomedical and Health Informatics*, 28(5): 3090–3101, 2024a.

Cui, C., Tang, Y., and Zhang, W. Deep Contrastive Survival Analysis with Dual-View Clustering. *Electronics*, 13(24), 2024b.

Cui, Y., Dong, W., Li, Y., Janitz, A. E., Pokala, H. R., and Zhu, R. Explainable transformer-based deep survival analysis in childhood acute lymphoblastic leukemia. *Computers in Biology and Medicine*, 191:110118, 2025.

Curth, A. and van der Schaar, M. Understanding the Impact of Competing Events on Heterogeneous Treatment Effect Estimation from Time-to-Event Data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 7961–7980, 2023.

D'Agostino, R. B. and Nam, B.-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

De Bin, R. and Stikbakke, V. G. A boosting first-hitting-time model for survival analysis in high-dimensional settings. *Lifetime Data Analysis*, 29(2):420–440, 2023.

Do, H., Chang, Y., Cho, Y. S., Smyth, P., and Zhong, J. Fair Survival Time Prediction via Mutual Information Minimization. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, pp. 128–149, 2023.

Emura, T. and Chen, Y.-H. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer Singapore, 2018.

Fan, S., Xu, R., Dong, Q., He, Y., Chang, C., and Cui, P. Stable Cox regression for survival analysis under distribution shifts. *Nature Machine Intelligence*, 6(12):1525–1541, 2024.

Farewell, V. T. and Prentice, R. L. A study of distributional shape in life testing. *Technometrics*, 19(1):69–75, 1977.

Feng, J., Shi, C., Wu, Z., Yan, X., and Zhao, W. Beyond Average Value Function in Precision Medicine: Maximum Probability-Driven Reinforcement Learning for Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40322–40360, 2026.

Foomani, A. H. G., Cooper, M., Greiner, R., and Krishnan, R. G. Copula-based deep survival models for dependent censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 669–680, 2023.

Frauen, D., Schröder, M., Hess, K., and Feuerriegel, S. Orthogonal Survival Learners for Estimating Heterogeneous Treatment Effects from Time-to-Event Data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30247–30287, 2026.

Gerds, T. A. and Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6): 1029–1040, 2006.

Ghosh, D., Pal, S., Lutz, M., and Luo, S. Ensemble Survival Analysis for Preclinical Cognitive Decline Prediction in Alzheimer's Disease Using Longitudinal Biomarkers. *Journal of Alzheimer's Disease*, 107(3):1256–1266, 2025.

Gögl, M., Liu, Y., Yau, C., Watkinson, P., and Zhu, T. DoseSurv: Predicting Personalized Survival Outcomes under Continuous-Valued Treatments. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23836–23868, 2026.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Gupta, M., Kannan, G., Prasad, R., and Gupta, G. Deep Survival Analysis and Counterfactual Inference Using Balanced Representations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63, 2020.

Hajianfar, G., Avval, A. H., Hosseini, S. A., Nazari, M., Oveisi, M., Shiri, I., and Zaidi, H. Time-to-event Overall Survival Prediction in Glioblastoma Multiforme Patients Using Magnetic Resonance Imaging Radiomics. *La Radiologia Medica*, 128(12):1521–1534, 2023.

Han, X., Goldstein, M., Puli, A., Wies, T., Perotte, A., and Ranganath, R. Inverse-Weighted Survival Games. *Advances in Neural Information Processing Systems*, 34: 2160–2172, 2021.

Han, X., Goldstein, M., and Ranganath, R. Survival Mixture Density Networks. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182, pp. 224–248, 2022.

Haredasht, F. N., Dauda, K. A., and Vens, C. Exploiting Censored Information in Self-Training for Time-to-Event Prediction. *IEEE Access*, 11:96831–96840, 2023.

Harrell Jr., F. E., Lee, K. L., and Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Hartman, N., Kim, S., He, K., and Kalbfleisch, J. D. Pitfalls of the concordance index for survival outcomes. *Statistics in medicine*, 42(13):2179–2190, 2023.

Heagerty, P. J. and Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105, 2005.

Hernández–Pérez, C., Podlipnik, S., Ficapal, J., Puig, S., Malvehy, J., and Vilaplana, V. Comparative analysis and interpretability of survival models for melanoma prognosis. *Computers in Biology and Medicine*, 190:110027, 2025.

Hou, B., Li, H., Jiao, Z., Zhou, Z., Zheng, H., and Fan, Y. Deep Clustering Survival Machines with Interpretable Expert Distributions. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2023.

Hu, S. and Chen, G. H. Fairness in Survival Analysis with Distributionally Robust Optimization. *Journal of Machine Learning Research*, 25(246):1–85, 2024.

Huang, L., Xing, Y., Denœux, T., and Feng, M. An Evidential Time-to-Event Prediction Model Based on Gaussian Random Fuzzy Numbers. In *Belief Functions: Theory and Applications*, pp. 49–57, 2024.

Huang, X. and Zhang, N. Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4):1090–1099, 2008.

Huh, Y. K., Kim, M., Liu, K., and Zhou, S. An Integrated Uncertainty Quantification Model for Longitudinal and Time-to-Event Data. *IEEE Transactions on Automation Science and Engineering*, 22:5863–5876, 2025.

Hung, H. and Chiang, C.-T. Estimation methods for time-dependent AUC models with survival data. *The Canadian Journal of Statistics*, 38(1):8–26, 2010.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random Survival Forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Kaplan, E. L. and Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18 (1):1–12, 2018.

Kim, H. Survival Permanental Processes for Survival Analysis with Time-Varying Covariates. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72774–72793, 2023.

Kim, J.-M. Integrating Copula-Based Random Forest and Deep Learning Approaches for Analyzing Heterogeneous Treatment Effects in Survival Analysis. *Mathematics*, 13 (10):1659, 2025.

Kleinbaum, D. G. and Klein, M. *Survival Analysis: A Self-Learning Text*. Springer New York, NY, 2012.

Knottenbelt, W., McGough, W., Wray, R., Zhang, W. Z., Liu, J., Machado, I. P., Gao, Z., and Crispin-Ortuzar, M. CoxKAN: Kolmogorov-Arnold Networks for Interpretable, High-Performance Survival Analysis. *Bioinformatics*, 41(8), 2025.

Kvamme, H., Borgan, Ø., and Scheel, I. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20:1–30, 2019.

Lee, C., Light, A., Alaa, A., Thurtle, D., van der Schaar, M., and Gnanapragasam, V. J. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. *The Lancet Digital Health*, 3(3):e158–e165, 2021.

Lee, D., Park, H., and Lee, C. Toward a well-calibrated discrimination via survival outcome-aware contrastive learning. In *Advances in Neural Information Processing Systems*, pp. 30985–31014, 2024.

Li, B., Li, H., Chen, J., Xiao, F., Fang, X., Guo, R., Liang, M., Wu, Z., Mao, J., and Shen, J. A magnetic resonance imaging (MRI)-based deep learning radiomics model predicts recurrence-free survival in lung cancer patients after surgical resection of brain metastases. *Clinical Radiology*, 85, 2025.

Lillelund, C. M., Magris, M., and Pedersen, C. F. Uncertainty Estimation in Deep Bayesian Survival Models. In

2023 IEEE EMBS International Conference on Biomedical and Health Informatics, pp. 1–4, 2023.

Lillelund, C. M., ang Qi, S., and Greiner, R. Overcoming Dependent Censoring in the Evaluation of Survival Models. *preprint, arXiv:2502.19460*, 2025a.

Lillelund, C. M., Kalra, S., Greiner, R., and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis. *PLOS ONE*, 20(11):1–16, 2025b.

Lillelund, C. M., Magris, M., and Pedersen, C. F. Efficient Training of Probabilistic Neural Networks for Survival Analysis. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6157–6166, 2025c.

Lillelund, C. M., Gharari Foomani, A. H., Sun, W., Qi, S.-a., Greiner, R., and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). MENSA: A Multi-Event Network for Survival Analysis with Trajectory-Based Likelihood Estimation. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pp. 543–571, 2026a.

Lillelund, C. M., Harbo, M., and Pedersen, C. F. Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis. *Journal of Public Health*, 34: 905–917, 2026b.

Liu, J., Zhang, R., and Rudin, C. FastSurvival: Hidden Computational Blessings in Training Cox Proportional Hazards Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87712–87765, 2024.

Liu, X., Zhang, W., and Zhang, M.-L. HACSurv: A Hierarchical Copula-Based Approach for Survival Analysis with Dependent Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3079–3087, 2025.

Lu, Y., Reinholz, A. K., Till, S. E., Kalina, S. V., Saris, D. B., Camp, C. L., and Stuart, M. J. Predicting the Risk of Posttraumatic Osteoarthritis After Primary Anterior Cruciate Ligament Reconstruction: A Machine Learning Time-to-Event Analysis. *The American Journal of Sports Medicine*, 51(7):1673–1685, 2023.

Luo, J., Xiao, Z., Shi, W., Xie, L., Yang, H., Yin, X., and Zhang, Y. CVaDeS: A Conditional Variational Deep Survival Model for Survival Analysis. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 458–467, 2023.

Luo, Q., Zhang, Q., Liu, H., Chen, X., Yang, S., and Xu, Q. Time-dependent interpretable survival prediction model for second primary NSCLC patients. *International Journal of Medical Informatics*, 195:105771, 2025.

Massoua, A. B., Diallo, A. B., and Bouguessa, M. Towards Robust Time-to-Event Prediction: Integrating the Variational Information Bottleneck with Neural Survival Model. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024.

Monod, M., Micheli, A., and Bhatt, S. NeuralSurv: Deep Survival Analysis with Bayesian Uncertainty Quantification. In *Advances in Neural Information Processing Systems*, volume 38, pp. 51213–51265, 2026.

Nikolaou, N., Salazar, D., RaviPrakash, H., Gonçalves, M., Mulla, R., Burlutskiy, N., Markuzon, N., and Jacob, E. A Machine Learning Approach for Multimodal Data Fusion for Survival Prediction in Cancer Patients. *npj Precision Oncology*, 9:128, 2025.

Norcliffe, A., Cebere, B., Imrie, F., Lió, P., and van der Schaar, M. SurvivalGAN: Generating Time-to-Event Data for Survival Analysis. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 10279–10304, 2023.

Ogutu, S., Mohammed, M., and Mwambi, H. Deep Learning Models for the Analysis of High-Dimensional Survival Data with Time-Varying Covariates While Handling Missing Data. *Discover Artificial Intelligence*, 5:176, 2025.

Park, J., Kim, D., Kim, C., Park, H., Kang, S., and Kim, G. KSP: Kolmogorov-Smirnov metric-based Post-Hoc Calibration for Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24506–24568, 2026.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Pomsuwan, T. and Freitas, A. A. A Genetic Algorithm-based Auto-ML System for Survival Analysis. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pp. 370–377, 2024.

Pu, Y., Liu, J., and Chan, K. H. K. Comparative Study of Five-Year Cervical Cancer Cause-Specific Survival Prediction Models Based on SEER Data. *Scientific Reports*, 15:22465, 2025.

Qi, S.-a., Kumar, N., Farrokh, M., Sun, W., Kuan, L.-H., Ranganath, R., Henao, R., and Greiner, R. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28244–28276, 2023a.

Qi, S.-a., Kumar, N., Verma, R., Xu, J.-Y., Shen-Tu, G., and Greiner, R. Using Bayesian Neural Networks to Select Features and Compute Credible Intervals for Personalized Survival Prediction. *IEEE Transactions on Biomedical Engineering*, 70(12):3389–3400, 2023b.

Qi, S.-a., Sun, W., and Greiner, R. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. *Proceedings of the AAAI Symposium Series*, 2(1):453–457, 2024a.

Qi, S.-a., Yu, Y., and Greiner, R. Toward conditional distribution calibration in survival prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 86180–86225, 2024b.

Qi, S.-a., Yu, Y., and Greiner, R. Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 41303–41339, 2024c.

Rahman, S., Keats, A. I., Kabir, M. A., Newaz, A., and Islam, M. M. Survival Analysis of Breast Cancer Patients: A Population-Based Study from SEER. In *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, 2023.

Rechichi, I., Amprimo, G., Cicolin, A., and Olmo, G. Predicting Amyotrophic Lateral Sclerosis Progression: an EMG-based Survival Analysis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, 2024.

Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival regression with proper scoring rules and monotonic neural networks. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 1190–1205, 2022.

Robins, J. M. and Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology: methodological issues*, pp. 297–331. Springer, 1992.

Rollo, C., Pancotti, C., Birolo, G., Rossi, I., Sanavia, T., and Fariselli, P. SYNDSURV: A simple framework for survival analysis with data distributed across multiple institutions. *Computers in Biology and Medicine*, 172: 108288, 2024.

Schemper, M. The explained variation in proportional hazards regression. *Biometrika*, 77(1):216–218, 1990.

Seidi, N., Tripathy, A., and Das, S. K. Using Geographic Location-Based Public Health Features in Survival Analysis. In *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pp. 80–91, 2024.

Shen, X. and Chen, G. H. Deep Kernel Aalen-Johansen Estimator: An Interpretable and Flexible Neural Net Framework for Competing Risks. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pp. 1096–1125, 2026.

Siannis, F., Copas, J., and Lu, G. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1):77–91, 2005.

Sluiskes, M., Goeman, J., Beekman, M., Slagboom, E., van den Akker, E., Putter, H., and Rodríguez-Girondo, M. The AccelerAge framework: a new statistical approach to predict biological age based on time-to-event data. *European Journal of Epidemiology*, 39(6):623–641, 2024.

Song, A. H., Chen, R. J., Jaume, G., Vaidya, A. J., Baras, A., and Mahmood, F. Multimodal Prototyping for cancer survival prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 46050–46073, 2024.

Steinberg, E., Fries, J. A., Xu, Y., and Shah, N. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. In *The Twelfth International Conference on Learning Representations*, 2024.

Stevens, P. E., Ahmed, S. B., Carrero, J. J., Foster, B., Francis, A., Hall, R. K., Herrington, W. G., Hill, G., Inker, L. A., Kazancıoğlu, R., Lamb, E., Lin, P., Madero, M., McIntyre, N., Morrow, K., Roberts, G., Sabanayagam, D., Schaeffner, E., Shlipak, M., Shroff, R., Tangri, N., Thanachayanont, T., Ulasi, I., Wong, G., Yang, C.-W., Zhang, L., and Levin, A. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney International*, 105(4): S117–S314, 2024.

Supriya, K. and Anitha, A. Survival Analysis of Superficial Bladder Cancer patients using DeepSurv and Cox models. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pp. 1–7, 2024.

Tan, H., Chen, R., Qin, M., Tang, L., Wu, Z., Luo, Q., and Quan, Y. Tabular GAN-Based Oversampling of Imbalanced Time-to-Event Data for Survival Prediction. In *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 376–380, 2023.

Tang, J., Shao, W., and Zhang, D. Cancer Survival Analysis via Zero-shot Tumor Microenvironment Segmentation on Low-resolution Whole Slide Pathology Images. In *Advances in Neural Information Processing Systems*, volume 38, pp. 57043–57069, 2026.

Tang, Y., Zhang, J., Wang, H., and Su, J. Transformer-Based Dynamic Survival Probability Prediction for Longitudinal Observational Data. *Journal of Statistical Computation and Simulation*, 95(15):3257–3278, 2025.

Tang, Z., Liu, L., Shen, Y., Chen, Z., Ma, G., Dong, J., Sun, X., Zhang, X., Li, C., Zheng, Q., and Yang, L. Explainable Survival Analysis with Uncertainty Using Convolution-Involved Vision Transformer. *Computerized Medical Imaging and Graphics*, 110:102302, 2023.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. An Introduction to Proximal Causal Inference. *Statistical Science*, 39(3):375–390, 2024.

Tjandra, D., He, Y., and Wiens, J. A Hierarchical Approach to Multi-Event Survival Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):591–599, 2021.

Tsiatis, A. A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. K. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

Van Ness, M., Bosschieter, T., Din, N., Ambrosy, A., Sandhu, A., and Udell, M. Interpretable Survival Analysis for Heart Failure Risk Prediction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225, pp. 574–593, 2023.

Vauvelle, A., Wild, B., Eils, R., and Denaxas, S. Differentiable Sorting for Censored Time-to-Event Data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66539–66555, 2023.

Wan, G., Leung, B. W., DeSimone, M. S., Nguyen, N., Rajeh, A., Collier, M. R., Rashdan, H., Roster, K., Zhou, X., Moseley, C. B., Nirmal, A. J., Pelletier, R. J., Maliga, Z., Marko-Varga, G., Németh, I. B., Tsao, H., Asgari, M. M., Gusev, A., Stagner, A. M., Lian, C. G., Hurlbert, M. S., Liu, F., Yu, K.-H., Sorger, P. K., and Semenov, Y. R. Development and validation of time-to-event models to predict metastatic recurrence of localized cutaneous melanoma. *Journal of the American Academy of Dermatology*, 90(2):288–298, 2024.

Wang, L., Qin, X., Jiang, J., Li, Y., and Liaw, W. Interactive Explainable Deep Survival Analysis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, 2024.

Wang, X., Hu, R., and Li, M. Model averaging prediction for survival data with time-dependent effects. *Journal of Statistical Planning and Inference*, 238:106260, 2025a.

Wang, Z., Zhang, H., Wang, J., Jiang, C., He, H., and Ding, Y. Forecasting Time to Risk Based on Multi-Party Data: An Explainable Privacy-Preserving Decentralized Survival Analysis Method. *Information Processing & Management*, 62(1):103881, 2025b.

Woodward, A., Bandhey, H., Moore, J. H., and Urbanowicz, R. J. Survival-LCS: A Rule-Based Machine Learning Approach to Survival Analysis. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 431–439, 2024.

Wu, J., Ke, X., Jiang, X., Wu, H., Kong, Y., and Shao, L. Leveraging Tumor Heterogeneity: Heterogeneous Graph Representation Learning for Cancer Survival Prediction in Whole Slide Images. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64312–64337, 2024.

Wu, R., Qiao, J., Wu, M., Yu, W., Zheng, M., Liu, T., Zhang, T., and Wang, W. Neural Frailty Machine: Beyond proportional hazard assumption in neural survival regressions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5569–5597, 2023.

Xie, X., Wang, Y., Zhao, Z., Yang, Q., Huang, L., Zhou, F., and Wang, Y. Reliable Multimodal Cancer Survival Prediction With Confidence-Aware Risk Modeling. *IEEE Journal of Biomedical and Health Informatics*, 30(4): 3281–3292, 2026.

Xu, L. and Guo, C. CoxNAM: An interpretable deep survival analysis model. *Expert Systems with Applications*, 227:120218, 2023.

Xu, W., Wang, S., Shen, L., and Zhao, Y. Collaborative Survival Analysis on Predicting Alzheimer's Disease Progression. *Statistics in Biosciences*, 2024.

Yan, R., Lv, Z., Yang, Z., Lin, S., Zheng, C., and Zhang, F. Sparse and Hierarchical Transformer for Survival Analysis on Whole Slide Images. *IEEE Journal of Biomedical and Health Informatics*, 28(1):7–18, 2024.

Yang, X. and Qiu, H. Deep Gated Neural Network With Self-Attention Mechanism for Survival Analysis. *IEEE Journal of Biomedical and Health Informatics*, 29(4): 2945–2956, 2025.

Ying, A. Proximal survival analysis to handle dependent right censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1414–1434, 2024.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in neural information processing systems*, 24, 2011.

Zhang, J., Wang, S., Chen, L., Guo, G., Chen, R., and Vanasse, A. Time-Dependent Survival Neural Network for Remaining Useful Life Prediction. In *Advances in Knowledge Discovery and Data Mining*, volume 11439, pp. 441–452, 2019.

Zhang, W., Ling, C. K., and Zhang, X. Deep copula-based survival analysis for dependent censoring with identifiability guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20613–20621, 2024a.

Zhang, W., Xu, W., Chen, T., Sakal, C., and Li, X. Integrating Images and Genomics for Multi-Modal Cancer Survival Analysis via Mixture of Experts. *Information Fusion*, 126:103521, 2026.

Zhang, X., Mehta, D., Hu, Y., Zhu, C., Darby, D., Yu, Z., Merlo, D., Gresle, M., van der Walt, A., Butzkueven, H., and Ge, Z. Adaptive transformer modelling of density function for nonparametric survival analysis. *Machine Learning*, 114(2):31, 2025.

Zhang, Y., Wittrup, E., and Najarian, K. CoxFNN: Interpretable Machine Learning Method for Survival Analysis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5, 2024b.

Zhang, Y., Xu, Y., Chen, J., Xie, F., and Chen, H. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. In *The Twelfth International Conference on Learning Representations*, 2024c.

Zhao, Z., Li, W., Liu, P., Zhang, A., Sun, J., and Xu, L. X. Survival Analysis for Multimode Ablation Using Self-Adapted Deep Learning Network Based on Multisource Features. *IEEE Journal of Biomedical and Health Informatics*, 28(1):19–30, 2024.

Zheng, M. and Klein, J. P. Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula. *Biometrika*, 82(1):127–138, 1995.

Zhou, H., Wang, H., Wang, S., and Zou, Y. SurvMetrics: An R package for Predictive Evaluation Metrics in Survival Analysis. *R J.*, 14(4):252–263, 2023.

Zhou, H., Zhou, F., and Chen, H. Cohort-Individual Cooperative Learning for Multimodal Cancer Survival Analysis. *IEEE Transactions on Medical Imaging*, 44(2):656–667, 2025.

Zisser, M. and Aran, D. Transformer-based time-to-event prediction for chronic kidney disease deterioration. *Journal of the American Medical Informatics Association*, 31 (4):980–990, 2024.

A. Structured Survey

We conducted a structured survey to examine how evaluation metrics are used in recent survival analysis publications. Our review covered papers published between 2023.01.01 and 2025.12.31 across major scientific databases, including Google Scholar, ScienceDirect, Scopus, Springer Nature, and IEEE Xplore. We supplemented this database search with manual screening of major machine-learning venues (*e.g.*, NeurIPS, ICML, ICLR, AISTATS), selected biomedical conferences (*e.g.*, EMBC), and relevant journals. To identify relevant articles, we searched for combinations of survival-analysis and evaluation-related keywords (see Table 3). Papers were included if they addressed survival analysis using a learned predictive model and contained an empirical evaluation section; purely methodological statistics papers and epidemiological studies without predictive modeling were excluded. For each paper, we extracted the stated problem type, modeling objectives, the metrics used in the experimental section, and whether the chosen metrics aligned with the stated goals and with the censoring assumptions of the data. Our sampling strategy intentionally focuses on machine learning-oriented survival analysis and may therefore over-represent ML and biomedical venues relative to traditional biostatistics.

Table 3. Keywords used in the structured survey for identifying survival analysis papers with explicit evaluation protocols.

Aspect	**Keywords**
General	survival analysis, time-to-event, censoring, evaluation
Metrics (Discrimination)	concordance index, c-index, auc, harrell, uno, antolini
Metrics (Absolute and squared errors)	brier score, ibs, mae, mse
Metrics (Calibration)	calibration, d-calibration, ici, ece

Table 4 summarizes the frequency with which evaluation metrics appear across the articles included in our survey. These counts form the basis of the analysis presented in Figure 1.

Table 4. Metrics used in survival analysis papers (2023-2025). Total number of included articles: 92.

Evaluation Metric	Number of Papers (%)	Reference
CI (Harrell)	62 (67.4%)	(Pomsuwan & Freitas, 2024; Nikolaou et al., 2025; Lillelund et al., 2025b; Zhang et al., 2025; Abuhantash et al., 2025; Huang et al., 2024; Li et al., 2025; Birolo et al., 2025; Tang et al., 2026; Zhou et al., 2025; Xu et al., 2024; Hernández–Pérez et al., 2025; Qi et al., 2024c; Zhang et al., 2024b; Knottenbelt et al., 2025; Xu & Guo, 2023; Luo et al., 2023; Hou et al., 2023; Cui et al., 2024b; Gupta et al., 2023; Cui et al., 2024a; Vauvelle et al., 2023; Ghosh et al., 2025; Tang et al., 2023; Cui et al., 2025; Haredasht et al., 2023; Do et al., 2023; Liu et al., 2024; Bone-Winkel & Reichenbach, 2024; Kim, 2025; Zhang et al., 2026; Wang et al., 2024; Baniecki et al., 2025; Wu et al., 2024; Lillelund et al., 2026a; Wang et al., 2025a; Song et al., 2024; Monod et al., 2026; Abdallah et al., 2024; Rechichi et al., 2024; Lillelund et al., 2026b; Zhang et al., 2024c; Xie et al., 2026; Yan et al., 2024; Fan et al., 2024; Buyrukoğlu, 2024; Rahman et al., 2023; Supriya & Anitha, 2024; Woodward et al., 2024; Norcliffe et al., 2023; Rollo et al., 2024; Tan et al., 2023; Chen et al., 2025; Luo et al., 2025; Hajianfar et al., 2023; Lee et al., 2024; Qi et al., 2024b; Zisser & Aran, 2024; Lillelund et al., 2023; Qi et al., 2023b; Park et al., 2026; Seidi et al., 2024)
CI (Uno)	7 (7.6%)	(Lillelund et al., 2025b; Pu et al., 2025; Archetti & Matteucci, 2023; Bone-Winkel & Reichenbach, 2024; Lillelund et al., 2026b; Sluiskes et al., 2024; Lillelund et al., 2023)

Evaluation Metric	Number of Papers (%)	Reference
CI (Antolini)	13 (14.1%)	(Birolo et al., 2025; Yang & Qiu, 2025; Shen & Chen, 2026; Ogutu et al., 2025; Wan et al., 2024; Bleistein et al., 2024; Lillelund et al., 2025c; Hu & Chen, 2024; Apellániz et al., 2024; Zhao et al., 2024; Chen, 2024; Massoua et al., 2024; Tang et al., 2025)
TD-AUC	18 (19.6%)	(Zhang et al., 2025; Li et al., 2025; Birolo et al., 2025; Hernández–Pérez et al., 2025; Pu et al., 2025; Wan et al., 2024; Ghosh et al., 2025; Wang et al., 2025b; Baniecki et al., 2025; Van Ness et al., 2023; Lu et al., 2023; Blythe et al., 2024; Yan et al., 2024; Kim, 2023; Rollo et al., 2024; Tan et al., 2023; Luo et al., 2025; Massoua et al., 2024)
Brier score	10 (10.9%)	(Huang et al., 2024; Wan et al., 2024; Do et al., 2023; Baniecki et al., 2025; Lu et al., 2023; Buyrukoğlu, 2024; Woodward et al., 2024; Norcliffe et al., 2023; Tan et al., 2023; Luo et al., 2025)
Brier score (IPCW)	26 (28.3%)	(De Bin & Stikbakke, 2023; Lillelund et al., 2025b; Birolo et al., 2025; Pu et al., 2025; Qi et al., 2024c; Cui et al., 2024b; Shen & Chen, 2026; Ogutu et al., 2025; Cui et al., 2024a; Lillelund et al., 2025c; Hu & Chen, 2024; Liu et al., 2024; Archetti & Matteucci, 2023; Van Ness et al., 2023; Apellániz et al., 2024; Lillelund et al., 2026a; Wang et al., 2025a; Steinberg et al., 2024; Wu et al., 2023; Monod et al., 2026; Lillelund et al., 2026b; Zhao et al., 2024; Lee et al., 2024; Qi et al., 2024b; Lillelund et al., 2023; Park et al., 2026)
MAE (Uncens.)	7 (7.6%)	(Zhang et al., 2025; Qi et al., 2023a; Huh et al., 2025; Yang & Qiu, 2025; Kim, 2025; Wang et al., 2025a; Zisser & Aran, 2024)
MAE (Cens.)	9 (9.8%)	(Lillelund et al., 2025b; Zhang et al., 2025; Qi et al., 2023a; 2024c; Lillelund et al., 2025c; 2026a;b; Qi et al., 2024b; 2023b)
MSE/RMSE	5 (5.4%)	(Huh et al., 2025; Kim, 2025; Wang et al., 2025a; Sluiskes et al., 2024; Curth & van der Schaar, 2023)
ECE/ICI	2 (2.2%)	(Lillelund et al., 2025c; Do et al., 2023)
D-Calibration	9 (9.8%)	(Lillelund et al., 2025b; Qi et al., 2024c; Lillelund et al., 2025c; 2026a; Monod et al., 2026; Lillelund et al., 2026b; Lee et al., 2024; Qi et al., 2023b; Park et al., 2026)
1-Calibration	1 (1.1%)	(Steinberg et al., 2024)
None	4 (4.3%)	(Feng et al., 2026; Bian et al., 2025; Gögl et al., 2026; Frauen et al., 2026)

Table 5 provides a detailed qualitative evaluation of each paper included in our structured review. These counts form the basis of the analysis presented in Figure 2. For every work, we assess whether the evaluation metrics align with the stated modeling objectives and whether the censoring assumptions required by those metrics are respected, ignored, or implicitly violated.

To assign the labels **Yes**, **Partially**, and **No**, we apply the following criteria:

- **A paper is labeled Yes** if it satisfies *both*:
  - metrics used are appropriate for the stated objectives (*e.g.*, risk ranking, time-to-event prediction, calibration), and
  - censoring is stated, justified, or actually corrected for (*e.g.*, using IPCW with KM for random censoring).

- **A paper is labeled Partially** if it satisfies *exactly one* of the above:
  - objectives are evaluated using appropriate metrics but censoring bias is not addressed, *or*
  - censoring is handled correctly but key metrics for the stated objectives are missing.
- **A paper is labeled No** if it satisfies *neither* criterion. Typical examples include:
  - claiming to predict time-to-event outcomes while reporting only Harrell's C-index, or
  - using metrics that are invalid under censoring without correction, justification or acknowledging of the bias.

We consider the following classes of metrics appropriate for modern survival analysis:

- **Discrimination:** Harrell, Uno, and Antolini C-indices; time-dependent AUC (TD-AUC).
- **Absolute and squared errors:** Brier score, Brier score (IPCW), MAE/RMSE, and censored variants.
- **Calibration:** ECE/ICI, D-Calibration, 1-Calibration.

A paper's score is decremented when their stated goals (*e.g.*, predicting survival curves, estimating event times, or providing calibrated risk scores) are not evaluated using appropriate metrics, or when censoring renders the chosen metrics invalid without appropriate adjustment or acknowledgment. These criteria allow us to assess whether each paper's evaluation methodology genuinely supports its claims.

Table 5. Structured survey of survival analysis papers published between 2023-2025. Each entry includes the problem type, stated objectives, evaluation metrics used, and an assessment of whether the chosen metrics align with stated goals and censoring assumptions.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Abdallah et al., 2024)	Methodology	Predict time-to-event outcomes (disease progression, relapse, etc.), compare DBN performance against existing survival models, emphasize ranking / discrimination ability.	No	Wrong metrics for the stated TTE objective, censoring bias not addressed.
(Abuhantash et al., 2025)	Application	Predict time to conversion from CN to MCI, predict risk prediction / progression modeling, and clinical usefulness (risk stratification, early detection, etc.).	No	Missing TTE metrics, heavy censoring with unadjusted Harrell's CI and no discussion of assumptions.
(Apellániz et al., 2024)	Methodology	Estimate the full predictive distribution of time-to-event, handle right-censored data, compare survival models using standard survival metrics.	Yes	They evaluate discrimination and calibration, and address censoring by using IPCW.
(Archetti & Matteucci, 2023)	Methodology	Provide an efficient, federated survival model, achieve strong discrimination performance, achieve good calibration of survival probabilities, and handle censored, incomplete data across distributed clients.	Partially	Missing TTE metrics or actual calibration metrics (*e.g.*, D-calibration), but they correct for random censoring.
(Baniecki et al., 2025)	Methodology	Introduce time-dependent interpretability methods for survival analysis, demonstrate the usefulness of these interpretability tools on: Hospital Length of Stay (LoS) modeled as a time-to-event prediction problem, and Cancer survival prediction using TCGA multi-omics data.	No	Metrics do not match TTE or calibration objectives, censoring bias not addressed.
(Bian et al., 2025)	Methodology	Introduce nonparametric boosting algorithms for regression and classification when outcomes are interval-censored, using censoring-unbiased transformations within functional gradient descent to achieve accurate and scalable prediction under interval censoring.	No	No survival metrics used in real-world dataset.
(Birolo et al., 2025)	Methodology	Evaluate predictive performance of survival methods under PH / non-PH / linear / nonlinear conditions, compare survival metrics and show when Harrell CI is inappropriate, assess calibration, study sample-size effects and computational cost.	Partially	Metrics align with the goals, but censoring bias not addressed.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Bleistein et al., 2024)	Methodology	Model dynamic survival processes from longitudinal covariates by learning a controlled latent-state representation that drives the event intensity over time; goal is improved individualized risk ranking / time-varying risk prediction from irregularly sampled histories.	Yes	Metrics match their goals and censoring bias addressed.
(Blythe et al., 2024)	Application	Use a Cox time-to-event model to improve ranking of patients by risk, compare ranking performance of Cox vs discrete-time logistic regression, develop a model useful for triage prioritization.	Yes	Their only concern is ranking and Uno's AUC handles random censoring properly.
(Bone-Winkel & Reichenbach, 2024)	Application	Estimate hazard ratios and then assign borrowers into risk-rating groups, evaluate whether ratings separate default risk, IRR, and survival curves.	Yes	Metrics match risk prediction goal, and they correct for random censoring using Uno's CI.
(Buyrukoğlu, 2024)	Application	Predict time-to-relapse in breast cancer, compare the predictive ability of CoxPH vs. RSF vs. CForest, evaluate model discrimination and prediction error, provide variable importance explanations, illustrate survival probability curves for hypothetical patients.	No	Missing TTE metrics and censoring assumptions not addressed.
(Chen, 2024)	Methodology	Develop a deep kernel survival model (survival kernet) that scales to large datasets while remaining interpretable, and provide finite-sample accuracy guarantees for the predicted individual survival distributions.	Partially	Metrics match their ranking goals, but censoring bias not addressed.
(Chen et al., 2025)	Application	Develop a new multimodal survival model (TDMFS) using Tucker decomposition, improve survival prediction accuracy (*i.e.*, risk prediction, not explicit time-to-event estimation), perform risk stratification with Kaplan–Meier curves to show separation of high- vs. low-risk groups.	Partially	Proper metrics for risk ranking, but censoring handled poorly; no calibration; not suitable for true time-to-event evaluation.
(Cui et al., 2024b)	Methodology	Improve survival prediction performance for glioma patients, predict survival risk / rank patients, produce more accurate survival curves (they include predicted KM-like curves), compare their model to baseline models on multiple metrics.	No	Missing metrics for calibration and survival-curve accuracy. Harrell CI used without assumptions; classification metrics invalid under censoring.
(Cui et al., 2024a)	Methodology	Improve survival prediction accuracy, improve risk discrimination, improve calibration, handle censored data more effectively via contrastive learning, provide better patient-level survival curves (CIF).	Partially	Missing TTE error metrics, incomplete calibration, but censoring addressed under the random assumption.
(Cui et al., 2025)	Application	Predict patient-specific survival probabilities over time, model survival curves and hazard functions, assess model discrimination (ranking), provide interpretability (SHAP).	No	They want to evaluate survival probabilities, but use only Harrell CI. No use of censoring-robust metrics.
(Curth & van der Schaar, 2023)	Methodology	Analyze how competing events induce covariate shift for different causal estimands.	Yes	RMSE of estimated HTEs is the correct metric for their causal-identification goals.
(De Bin & Stikbakke, 2023)	Methodology	Develop non-PH model to predict first hitting time, evaluate predictive performance and calibration, no ranking.	Partially	IBS with IPCW is appropriate, but calibration only evaluated with plots.
(Do et al., 2023)	Methodology	Ensure fairness of survival-time predictions across sensitive groups, measure disparities in survival prediction.	Partially	Metrics are correct for their fairness objective, but Harrell's CI and BS are not corrected for censoring.
(Fan et al., 2024)	Methodology	Build a stable Cox model that improves generalization of risk ranking under distribution shifts, identify stable variables that maintain predictive value across cohorts, evaluate prognostic performance on real-world cancer datasets.	Partially	Metrics match risk prediction goal, but censoring bias not addressed.
(Feng et al., 2026)	Methodology	Develop a reinforcement-learning framework for alternating recurrent-event survival data that optimizes a probability-based objective (maximize the probability inter-event durations exceed a clinically meaningful threshold) rather than expected outcomes, enabling standard RL training via a lower-bound reformulation.	No	No survival metrics used in real-world dataset.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Frauen et al., 2026)	Methodology	Provide a toolbox of Neyman-orthogonal/meta-learning survival learners for estimating heterogeneous treatment effects from censored time-to-event data, improving robustness (e.g., under low overlap) and allowing plug-in of arbitrary base learners.	No	No survival metrics used in real-world dataset.
(Ghosh et al., 2025)	Application	Predict the time to clinical progression from CN to MCI/AD, early risk stratification using longitudinal biomarkers.	No	Missing TTE metrics, censoring bias not addressed.
(Gupta et al., 2023)	Methodology	Estimate ATE/ITE under censoring; reduce confounding; improve causal accuracy.	Partially	Causal metrics are correct, but the censoring bias is not addressed.
(Gögl et al., 2026)	Methodology	Estimate heterogeneous, dosage-dependent treatment effects for time-to-event outcomes by learning personalized survival predictions under continuous-valued interventions (dose-response), supporting individualized clinical decision-making from observational TTE data.	No	No survival metrics used in real-world dataset.
(Hajianfar et al., 2023)	Application	Predict time-to-event OS using radiomic + clinical features, identify strong preprocessing / feature selection / ML combinations, stratify patients into risk groups (Kaplan–Meier used for validation).	Partially	Harrell's CI measures risk ranking, but no IBS, MAE, RMSE, or other time-error metrics are reported.
(Haredasht et al., 2023)	Methodology	Predict time-to-event (survival time) and improve accuracy using a semi-supervised approach that uses censored observations more effectively, develop a self-training framework (STUART) that augments training data with confident predictions on censored instances, compare predictive performance against RSF, Cox-based models, and SP-AFT.	No	Missing TTE metrics and only use Harrell CI. No calibration tests and censoring bias left unaddressed.
(Hernández–Pérez et al., 2025)	Application	Evaluate and compare survival analysis models for melanoma prognosis, predict survival outcomes (OS, SS, DFS) for melanoma, assess both predictive performance and interpretability.	Partially	They use correct metrics for their ranking and calibration objectives, but censoring not properly handled (*e.g.*, no IPCW).
(Hou et al., 2023)	Methodology	Perform time-to-event prediction, perform survival clustering using interpretable expert distributions, compete with state-of-the-art deep methods (Deep Cox, DSM, SCA, VaDeSC), show that the mixture-of-experts formulation improves interpretability.	No	Missing TTE metrics and calibration, censoring bias is not addressed.
(Hu & Chen, 2024)	Methodology	Improve fairness across all sufficiently large subpopulations using a DRO-based method, be applicable to any survival model, including Cox, DeepHit, SODEN, maintain accuracy (discrimination & risk prediction quality), handle censoring within fairness metrics.	Yes	Metrics match stated goals and censoring handled appropriately.
(Huang et al., 2024)	Methodology	Predict continuous time-to-event values, quantify uncertainty using belief functions & random fuzzy numbers, handle right censoring correctly via modified evidential likelihood, achieve good predictive performance, achieve good calibration.	No	Missing TTE accuracy metrics and censoring bias not addressed.
(Huh et al., 2025)	Methodology	Jointly model longitudinal + TTE data; propagate uncertainty from both sources; improve risk prediction accuracy; quantify prediction uncertainty.	Partially	Metrics cover TTE accuracy but not calibration; censoring assumptions not addressed; no formal calibration test.
(Kim, 2025)	Methodology	Predict survival outcomes (vital status), estimate Heterogeneous Treatment Effects (HTE), evaluate differences in treatment effects across race, improve prediction with HT and IPTW weighting.	No	Metrics do not match their goals, and censoring bias ignored in evaluation.
(Kim, 2023)	Methodology	The paper proposed a novel Bayesian survival model to estimate a hazard function as a nonlinear regression function of covariates from survival data with time-varying covariates.	Yes	Metrics match goals and censoring bias addressed.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Knottenbelt et al., 2025)	Methodology	Predict log-partial hazard (risk ranking), recover symbolic formulae representing hazard contributions, automatic feature pruning using KAN structure.	Partially	They use correct metric for risk ranking, but censoring bias is not addressed.
(Lee et al., 2024)	Methodology	Improve discrimination (risk ranking) in deep survival models without sacrificing calibration by using survival-outcome-aware contrastive learning (paired with likelihood training) so that patients with similar event times are embedded closer while keeping well-calibrated survival probabilities.	Yes	They claim to address discrimination and calibration and evaluate both thoroughly; censoring assumptions stated and handled.
(Li et al., 2025)	Application	Predict RFS, a time-to-event outcome (RFS = time from surgery to recurrence/death), predict intracranial recurrence at fixed horizons (6, 12, 18 months), risk stratification using survival curves, construct a prognostic nomogram.	No	Metrics do not match TTE objective, censoring handling is incorrect for chosen metric.
(Lillelund et al., 2025b)	Application	Predict functional decline as an event in ALS patients, perform risk ranking, make counterfactual predictions.	Yes	Correct metrics for objectives and censoring bias addressed.
(Lillelund et al., 2025c)	Methodology	Compare probabilistic training approaches (VI vs. MCD vs. SNGP), evaluate prediction performance (ranking + time-to-event accuracy), evaluate calibration performance (distribution, coverage, probability calibration), provide computational efficiency comparisons.	Partially	Uses the appropriate metrics for ranking, time-to-event prediction, and calibration, but censoring bias not addressed.
(Lillelund et al., 2026a)	Methodology	Model multiple time-to-event distributions jointly, improve event ordering (local survival discrimination), improve global discrimination, provide population survival curves.	Yes	Correct metrics for objectives and censoring bias addressed.
(Lillelund et al., 2026b)	Application	Predict individual fall risk within 6 months, provide a ranking of individuals, predict time-to-event, provide calibrated survival curves for decision support.	Yes	Uses the appropriate metrics for ranking, time-to-event prediction, and calibration. Censoring bias addressed.
(Lillelund et al., 2023)	Methodology	Assess whether Bayesian NNs improve survival prediction under Cox PH, quantify uncertainty, compare deterministic vs Bayesian models on predictive performance.	Yes	Metrics align with objectives and appropriately handle censoring.
(Liu et al., 2024)	Methodology	Speed up and stabilize training of Cox proportional hazards models on modern high-dimensional data by introducing globally convergent surrogate-optimization methods, enabling practical training of sparse/high-quality Cox models and other Cox-related extensions.	Partially	Metrics match objective, but censoring assumptions never stated.
(Lu et al., 2023)	Methodology	Predict time-to-event outcomes, generate individualized survival curves, identify risk factors, compare RSF to KM.	Partially	Metrics do not fully address TTE prediction, and censoring bias is not addressed.
(Luo et al., 2023)	Methodology	Predict risk; evaluate feature selection; model uncertainty.	No	Only partial alignment with objectives and censoring bias is not addressed.
(Luo et al., 2025)	Application	Develop and validate machine-learning-based survival prediction models for overall survival (OS) in SP-NSCLC, predict survival probabilities over time, evaluate time-dependent predictor importance via permutation explanation, risk stratification into high- and low-risk groups via KM curves.	Partially	They use correct metrics for ranking goal, but censoring bias not addressed (no IPCW)
(Massoua et al., 2024)	Methodology	Predict the actual time-to-event (event-time) distribution, estimate survival density and survival function, improve accuracy of time-to-event estimation under noise.	No	Metrics do not match TTE objective, censoring bias not addressed.
(Monod et al., 2026)	Methodology	Provide deep survival prediction with principled Bayesian uncertainty quantification to produce well-calibrated survival functions and uncertainty estimates (especially in data-scarce regimes) while matching or improving discriminative performance.	Yes	They claim to address discrimination and calibration and evaluate both thoroughly, censoring assumptions stated and handled.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Nikolaou et al., 2025)	Application	Compare multimodal vs unimodal survival models for overall survival (OS) on TCGA; main focus on improved discrimination / predictive performance and multimodal fusion strategy (late fusion).	Partially	They use correct metric for risk ranking and acknowledges censoring-induced bias, but still uses a censoring-biased metric.
(Norcliffe et al., 2023)	Methodology	Formalize synthetic survival data generation, generate time-to-event data that incorporates censored and uncensored data, evaluate quality of generated data.	Partially	Missing TTE metrics, but censoring stated as assumed random.
(Ogutu et al., 2025)	Application	Apply survival models to predict time to HIV infection using high-dimensional data with time-varying cytokines.	Yes	Correct metrics for objectives and censoring bias addressed.
(Park et al., 2026)	Methodology	Develop a new calibration method for survival models (KSP), improve calibration across the entire survival distribution without harming discrimination, provide theoretical foundations for KS-cal and show consistency, empirically evaluate calibration performance across datasets and survival models.	Yes	Metrics align with claims and censoring assumptions addressed.
(Pomsuwan & Freitas, 2024)	Methodology	Predict time-to-event outcomes, cope with censored data, automatically select the best survival model + hyperparameters, improve predictive accuracy over baseline methods, evaluate performance across multiple datasets.	No	Missing TTE metrics and censoring bias not addressed.
(Pu et al., 2025)	Application	Predict 5-year cause-specific survival (CSS) for cervical cancer patients, compare five survival models, identify important clinical predictors of CSS using SHAP, provide risk stratification (KM curves + log-rank test).	Yes	Metrics match objectives and censoring addressed.
(Qi et al., 2023a)	Methodology	Propose a way to evaluate time-to-event accuracy under censoring.	Yes	They propose MAE variants for TTE prediction and explicitly assume independent censoring.
(Qi et al., 2024c)	Methodology	Improve calibration of individual survival distributions (D-cal, KM-cal), preserve discrimination, handle censoring correctly in a conformal-regression setting.	Yes	They claim to address discrimination and calibration and evaluate both, censoring assumptions stated and handled.
(Qi et al., 2024b)	Methodology	Improve survival probability prediction by post-hoc conformal calibration targeting conditional (covariate-dependent) distribution calibration—aiming to fix both marginal and conditional calibration while preserving discrimination, with theoretical guarantees.	Yes	They claim to address discrimination and calibration and evaluate both thoroughly, censoring assumptions stated and handled.
(Qi et al., 2023b)	Methodology	Accurate selection of pertinent features, computation of trustworthy credible intervals (uncertainty), precise estimation of event times or risk scores.	Partially	Metrics match the objective, but censoring bias is not addressed.
(Rahman et al., 2023)	Application	Predict risk of death, identify significant prognostic features, compare RSF to Cox and Gradient Boosting on predictive performance (predicting ranking, survivability).	No	Only ranking is evaluated, even though they imply broader prediction goals and censoring not handled.
(Rechichi et al., 2024)	Application	Predict disease progression in ALS using EMG-derived metrics, predict survival outcomes (i.e., event occurrence) in ALS, predict survival time (time-to-event), model evolution of ALS over time and generate survival curves to support patient follow-up.	No	Metrics do not match TTE objective, censoring bias not addressed.
(Rollo et al., 2024)	Methodology	Distributed synthetic-data survival modeling; claims TTE prediction.	No	Missing TTE metrics, no calibration, and incomplete censoring evaluation.
(Seidi et al., 2024)	Application	Determine whether adding geographic location-based public health features improves survival prediction quality, evaluate on SEER data using existing models (CoxPH, DSM), prediction goal is framed strictly in terms of agreement between predicted and observed survival times, quantified by CI (ranking), no claims about predicting time-to-event, calibration, or distribution accuracy.	Yes	They make no distributional or censoring-related claims.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Shen & Chen, 2026)	Methodology	Predict cumulative incidence functions (CIFs) for competing risks at the individual level, make interpretable predictions, provide a flexible neural estimator generalizing Aalen-Johansen.	Yes	Correct metrics for objectives and censoring bias addressed.
(Sluiskes et al., 2024)	Application	Predict residual life (survival time) via AFT models, convert residual life into predicted biological age, evaluate biological age clocks using discrimination and calibration metrics.	Yes	Metrics match risk prediction goal, and they correct for random censoring using Uno's CI.
(Song et al., 2024)	Application	Improve multimodal cancer survival prediction (prognostication/stratification) from whole-slide histology images and transcriptomics by compressing each modality into a small set of learned prototypes, enabling efficient fusion with lower compute/memory while improving interpretability and predictive performance.	No	They measure only discrimination aspects, and censoring bias not addressed.
(Steinberg et al., 2024)	Methodology	Predict the full time-to-event distribution. Improve discrimination, calibration.	Partially	Discrimination and calibration metrics match goals, but they do not evaluate TTE accuracy, despite claiming to predict TTE.
(Supriya & Anitha, 2024)	Application	Predict risk ranking for bladder cancer patients, rank patients by risk, provide treatment recommendations using hazard ratios/log risk differences, evaluate predictive performance of DeepMLPSurv vs Cox and other models, analyze tumor recurrence based on features.	Partially	Metrics align with goal, but censoring-robust metrics are missing.
(Tan et al., 2023)	Methodology	Evaluate whether GAN-based oversampling improves balanced survival prediction, assess impacts on survival-prediction performance.	No	Missing TTE metrics and calibration, censoring bias is not addressed.
(Tang et al., 2026)	Methodology	Enable efficient end-to-end cancer survival prediction from low-resolution whole-slide pathology images by performing zero-shot tumor-microenvironment (TME) segmentation to capture prognostic tissue information without expensive high-resolution patch processing, improving efficiency and interpretability.	No	They measure only discrimination aspects, censoring bias not addressed.
(Tang et al., 2023)	Methodology	Improve survival risk ranking using integrative deep model; stratify patients.	Partially	Ranking metrics match objective, but censoring bias not addressed.
(Tang et al., 2025)	Methodology	Predict survival probabilities dynamically over time, model longitudinal measurements, improve prediction accuracy over existing longitudinal survival models, interpretability via SHAP.	No	Missing TTE metrics and censoring bias not addressed.
(Van Ness et al., 2023)	Application	Predict incident heart failure risk using censored EHR time-to-event data, provide an interpretable survival model using EBMs, deliver accurate survival predictions compared with Cox and other survival models.	Yes	AUC fits the discrimination goal, IBS partially covers calibration, and IPCW applied.
(Vauvelle et al., 2023)	Methodology	Improve risk ranking performance over CoxPH, handle censoring properly, enable top-k risk prediction.	Partially	The CI fits the discrimination goal, but censoring bias left unaddressed and their datasets include heavy censoring.
(Wan et al., 2024)	Application	Develop time-to-event risk prediction models for melanoma metastatic recurrence, identify patients "at the highest risk of distant recurrence, who are therefore most likely to benefit from adjuvant therapy, focus on the methodological approach to time-to-event prediction.	Partially	Metrics match their TTE and risk goal, but censoring bias not addressed.
(Wang et al., 2025b)	Application	Forecast when a financial risk event will occur in a decentralized, privacy-preserving setting, provide explainability, via time-dependent coefficients across forecasting horizons.	No	Missing TTE metrics, censoring bias not addressed.
(Wang et al., 2024)	Methodology	Predict time-to-event (specifically time to death); estimate individualized survival curves; predict survival time "as close as possible" to true time; integrate explainability for understanding factors influencing survival time; support clinical time-dependent decision-making.	No	Metrics do not match TTE objective, censoring bias not addressed.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Wang et al., 2025a)	Methodology	Develop a model-averaging method for Cox models with time-varying covariate effects, improve prediction of survival outcomes when the true model structure is unknown or partially misspecified, provide theoretical justification (asymptotic optimality) and demonstrate predictive performance via simulations and real data.	Yes	Metrics align with their goals, and censoring handled stated via IPCW.
(Woodward et al., 2024)	Methodology	Predict survival times, predict full survival distributions, handle censoring with novel error and accuracy updates, detect complex genetic associations (epistasis, heterogeneity), provide interpretable rule-based models, perform well across varying censoring proportions, feature dimension, and interaction structure.	No	Missing TTE metrics, missing calibration, missing ranking, incomplete censoring evaluation.
(Wu et al., 2024)	Application	Improve cancer survival prediction from whole-slide images by explicitly modeling tumor/tissue heterogeneity as a heterogeneous graph, incorporating pathology domain knowledge about prognostic tissues and their spatial interactions.	No	They evaluate only discrimination, and censoring bias not addressed.
(Wu et al., 2023)	Methodology	Using frailty models to incorporate a multiplicative random effect to capture unobserved heterogeneity.	Yes	IBS match their goals, and they correct for censoring using IPCW.
(Xie et al., 2026)	Methodology	Improve risk prediction for cancer survival, provide reliable risk predictions via confidence-aware modeling, improve multimodal fusion through cross-modal alignment.	Partially	Metrics align with the goals, but censoring bias not addressed.
(Xu et al., 2024)	Application	Predict time-to-event outcomes, specifically time to Alzheimer's Disease onset, show improved time-to-event prediction when combining modalities.	No	Metrics do not match TTE objective, censoring bias not addressed.
(Xu & Guo, 2023)	Methodology	Build a nonlinear survival model compatible with Cox's partial likelihood, improve predictive ranking performance compared to CPH, RSF, DeepSurv, provide intrinsic interpretability through NAM shape functions.	Partially	Harrell CI is correct for ranking, but the censoring bias is not addressed.
(Yan et al., 2024)	Methodology	Show better survival prediction than baselines, and outperform SOTA based on CI.	Partially	Ranking metrics match objective, but censoring bias not addressed.
(Yang & Qiu, 2025)	Methodology	Model the time distributions of the first occurrences of events of interest, estimate the probability mass function (PMF) of the first hitting time.	No	TTE objective is only partially evaluated (MAE only on uncensored cases), censoring is not correctly handled as Antolini CI is biased under random censoring.
(Zhang et al., 2025)	Methodology	Predict the full time-to-event distribution via a PDF and survival function S(t), improve prediction of survival time via modeling mean lifetime, accurately predict censoring via novel loss terms, handle static + dynamic (longitudinal) covariates.	Yes	Metrics align with their goals, and they consider MAE-Hinge for censored cases.
(Zhang et al., 2024b)	Methodology	Predict time-to-event outcomes using a non-linear Cox extension; model intends to capture non-linear relationships in hazard while also providing interpretable rules; authors state goal of predicting time until an event occurs and emphasize learning hazards, survival function, and membership thresholds that reflect when patients are at high risk; aims to balance predictive performance with interpretability.	No	Metrics do not match TTE objective, censoring bias not addressed.
(Zhang et al., 2026)	Methodology	Develop a multimodal framework (MoE + cross-modal attention) to predict survival risk, improve integration of WSI features and pathway-based genomic features, achieve better risk stratification and prognostic discrimination across TCGA cancers.	Partially	They use correct metric for risk ranking, but censoring bias not addressed.
(Zhang et al., 2024c)	Application	Predict cancer survival, by integrating of pathological images and genomic data.	No	They claim to do survival prediction, but only use Harrell CI; censoring bias not addressed.
(Zhao et al., 2024)	Methodology	Benchmark predictive performance of survival models.	Partially	Metrics match objective, but censoring bias is not addressed.

Paper	Problem Type	Objectives	Correctly Evaluated?	Explanation
(Zhou et al., 2025)	Methodology	Risk ranking in cancer prognosis; Improve discrimination and generalization of multimodal survival models.	Partially	Metrics match ranking objective, but censoring not properly handled (*e.g.*, no IPCW).
(Zisser & Aran, 2024)	Methodology	Predict exact time-to-event (time to deterioration to CKD Stage 5), improve MAE vs. baseline TTE models (explicit claim), predict full survival curves, handle censored data correctly, enable ranking and risk stratification, enable clinical interpretability.	Partially	Metrics match some of their objectives (MAE for TTE accuracy, CI for ranking), but the censoring bias is unaddressed.

B. Formal Metric Definitions

Throughout, let $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, where $t_i = \min\{e_i, c_i\}$ and $\delta_i = \mathbb{1}[e_i \le c_i]$. Let $\widehat{S}(t \mid \boldsymbol{x}_i)$ denote the predicted survival function and let $G(t \mid \boldsymbol{x}) = \Pr(C > t \mid \boldsymbol{X} = \boldsymbol{x})$ denote the conditional censoring survival function. We write $\widehat{G}(t) = \widehat{\Pr}(C > t)$ for a marginal estimate of the censoring survival function, and $\widehat{G}(t \mid \boldsymbol{x})$ for a covariate-conditional estimate. When needed, let $\widehat{r}_i$ denote a scalar predicted risk score, where larger values indicate earlier predicted event.

Harrell's C-index. Harrell's C-index (Harrell Jr. et al., 1996) estimates the proportion of concordant pairs among observed comparable pairs:

$$\widehat{C}_{\text{Harrell}} = \frac{\sum_{i,j} \delta_i \, \mathbb{1}[t_i < t_j] \, \mathbb{1}[\widehat{r}_i > \widehat{r}_j]}{\sum_{i,j} \delta_i \, \mathbb{1}[t_i < t_j]}. \tag{1}$$

It evaluates discrimination only.

Uno's C-index. Uno's C-index (Uno et al., 2011) uses IPCW to reweight comparable pairs:

$$\widehat{C}_{\text{Uno}}(\tau) = \frac{\sum_{i,j} \delta_i \frac{1}{\widehat{G}(t_i)^2} \mathbb{1}[t_i < t_j, \; t_i < \tau] \, \mathbb{1}[\widehat{r}_i > \widehat{r}_j]}{\sum_{i,j} \delta_i \frac{1}{\widehat{G}(t_i)^2} \mathbb{1}[t_i < t_j, \; t_i < \tau]}. \tag{2}$$

With marginal $\widehat{G}(t) = \widehat{\Pr}(C > t)$, this targets random censoring; with conditional $\widehat{G}(t \mid \boldsymbol{x})$, it can target independent censoring if the censoring model is correctly specified.

Antolini's C-index. Antolini's C-index (Antolini et al., 2005) compares predicted survival probabilities at the earlier observed event time:

$$\widehat{C}_{\text{Antolini}} = \frac{\sum_{i,j} \delta_i \, \mathbb{1}[t_i < t_j] \, \mathbb{1}\left[\widehat{S}(t_i \mid \boldsymbol{x}_i) < \widehat{S}(t_i \mid \boldsymbol{x}_j)\right]}{\sum_{i,j} \delta_i \, \mathbb{1}[t_i < t_j]}. \tag{3}$$

It evaluates discrimination for predicted survival curves.

Brier score and IBS. At a fixed time t^*, the IPCW Brier score (Brier, 1950; Graf et al., 1999; Gerds & Schumacher, 2006) is

$$\widehat{\text{BS}}(t^*) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\widehat{S}(t^* \mid \boldsymbol{x}_i)^2 \mathbb{1}[t_i \le t^*, \delta_i = 1]}{\widehat{G}(t_i)} + \frac{(1 - \widehat{S}(t^* \mid \boldsymbol{x}_i))^2 \mathbb{1}[t_i > t^*]}{\widehat{G}(t^*)} \right]. \tag{4}$$

The integrated Brier score (Graf et al., 1999) averages this over a time interval:

$$\widehat{\text{IBS}} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \widehat{\text{BS}}(t) \, \mathrm{d}t. \tag{5}$$

Mean Absolute Error. Given an ISD, a point event-time prediction can be obtained from the predicted median survival time,

$$\widehat{e}_i = \inf\{t : \widehat{S}(t \mid \boldsymbol{x}_i) \le 0.5\}. \tag{6}$$

For uncensored observations, MAE is

$$\widehat{\mathrm{MAE}}_{\mathrm{uncens}} = \frac{1}{\sum_i \delta_i} \sum_{i=1}^{N} \delta_i \, |t_i - \widehat{e}_i|. \tag{7}$$

This ignores censored instances. Censored variants, such as MAE-Margin (Haider et al., 2020) and MAE-PO (Qi et al., 2023a), attempt to incorporate censored observations by accounting for the unobserved event times after censoring.

1-Calibration. 1-Calibration (D'Agostino & Nam, 2003) evaluates calibration at a fixed horizon t^*. Let

$$\widehat{F}(t^* \mid \boldsymbol{x}_i) = 1 - \widehat{S}(t^* \mid \boldsymbol{x}_i)$$

denote the predicted event probability by t^*. Instances are sorted by $\widehat{F}(t^* \mid \boldsymbol{x}_i)$ and partitioned into K groups $\mathcal{D}_1, \ldots, \mathcal{D}_K$. A Hosmer–Lemeshow-type statistic is then

$$\chi^2_{\mathrm{HL}} = \sum_{k=1}^{K} \frac{N_k \left(F_k(t^*) - \widehat{F}_k(t^*) \right)^2}{\widehat{F}_k(t^*) \left(1 - \widehat{F}_k(t^*) \right)}, \tag{8}$$

where $N_k = |\mathcal{D}_k|$, $F_k(t^*)$ is the observed event frequency by t^* in group k, and $\widehat{F}_k(t^*)$ is the average predicted event probability in that group.

D-Calibration. D-Calibration (Haider et al., 2020) evaluates whether the predicted survival probabilities at event times,

$$\widehat{S}(e_i \mid \boldsymbol{x}_i),$$

are uniformly distributed on $[0, 1]$ under a calibrated model. In practice, the interval $[0, 1]$ is partitioned into bins and the observed bin counts are compared to the expected uniform counts using a goodness-of-fit test. Censored observations are typically distributed over the range of survival probabilities compatible with their censoring time.

Log-likelihood. If the model provides both a predicted density $\widehat{f}(t \mid \boldsymbol{x}_i)$ and survival function $\widehat{S}(t \mid \boldsymbol{x}_i)$, the right-censored log-likelihood is

$$\ell = \sum_{i=1}^{N} \left[\delta_i \log \widehat{f}(t_i \mid \boldsymbol{x}_i) + (1 - \delta_i) \log \widehat{S}(t_i \mid \boldsymbol{x}_i) \right]. \tag{9}$$

It evaluates the full predictive distribution, but requires access to a density or hazard representation.

C. Proper Scoring Rules

This definition is adapted from (Rindt et al., 2022).

A scoring rule R takes a distribution P over some set $\mathcal{Y}$ and an observed value $y \in \mathcal{Y}$, and returns a score $R(P, y)$ assessing how well the predictive distribution P assigns probability to the observed value. For a positive scoring rule, a higher score indicates a better model fit.

A scoring rule is called *proper* if the true distribution achieves the optimal expected score, *i.e.*, if

$$\mathbb{E}_{y \sim P} R(P, y) \;\geq\; \mathbb{E}_{y \sim P} R(\widehat{P}, y)$$

for all predictive distributions $\widehat{P}$.

In the context of survival analysis, let $S(t \mid \boldsymbol{x})$ denote the true conditional survival function, let $\widehat{S}(t \mid \boldsymbol{x})$ denote a predicted survival function, and let $G(t \mid \boldsymbol{x}) = \Pr(C > t \mid \boldsymbol{X} = \boldsymbol{x})$ denote the conditional censoring survival function. We call the scoring rule R proper if, for every true conditional survival function $S(t \mid \boldsymbol{x})$, every conditional censoring survival function $G(t \mid \boldsymbol{x})$, and every covariate value $\boldsymbol{x} \in \mathcal{X}$, it holds that

$$\mathbb{E}_{E,C \mid \boldsymbol{X} = \boldsymbol{x}} R\big(S(\cdot \mid \boldsymbol{x}), (T, \Delta)\big) \geq \mathbb{E}_{E,C \mid \boldsymbol{X} = \boldsymbol{x}} R\big(\widehat{S}(\cdot \mid \boldsymbol{x}), (T, \Delta)\big)$$

for every predicted survival function $\widehat{S}(\cdot \mid \boldsymbol{x})$, where $E \mid \boldsymbol{X} = \boldsymbol{x}$ has survival function $S(t \mid \boldsymbol{x})$, $C \mid \boldsymbol{X} = \boldsymbol{x}$ has censoring survival function $G(t \mid \boldsymbol{x})$, and $T = \min\{E, C\}$, $\Delta = \mathbb{1}[E \leq C]$.

By taking the expectation with respect to $\boldsymbol{X}$, we obtain

$$\mathbb{E}_{E,C,\boldsymbol{X}} R\big(S(\cdot \mid \boldsymbol{X}), (T, \Delta)\big) \geq \mathbb{E}_{E,C,\boldsymbol{X}} R\big(\widehat{S}(\cdot \mid \boldsymbol{X}), (T, \Delta)\big)$$

for every family of predicted survival functions $\widehat{S}(\cdot \mid \boldsymbol{x})$, with $\boldsymbol{x} \in \mathcal{X}$.

D. Simulation Details

We evaluate model-metric consistency using a controlled synthetic data-generating process with known ground truth event times. Our simulation framework follows the linear Weibull construction of Foomani et al. (2023), extended to explicitly control the censoring mechanism under random, independent, and dependent censoring mechanisms.

Covariates. For each experiment, we generate N independent samples $\{\boldsymbol{x}_i\}_{i=1}^{N}$, where

$$\boldsymbol{x}_i \sim \mathrm{Unif}([0,1]^d),$$

with $d = 10$ covariates and $N = 10{,}000$ instances. Covariates are fixed across censoring mechanisms within each random seed (0-99) to isolate the effect of censoring.

Event and censoring time models. Both the event time E and censoring time C are generated from Weibull distributions with covariate-dependent hazards. Specifically, conditional on covariates $\boldsymbol{x}_i$, the hazard and survival functions for E (or censoring C respectively) are

$$h_E(t \mid \boldsymbol{x}_i) = \left(\frac{v}{\rho}\right)\left(\frac{t}{\rho}\right)^{v-1} \exp\left(g_\Psi(\boldsymbol{x}_i)\right), \tag{10}$$

$$S_E(t \mid \boldsymbol{x}_i) = \exp\left(-\left(\frac{t}{\rho}\right)^{v} \exp\left(g_\Psi(\boldsymbol{x}_i)\right)\right), \tag{11}$$

where v and ρ represent the shape and scale parameters of the Weibull distribution, respectively. In all experiments we use

$$(v_E, \rho_E) = (4, 17), \qquad (v_C, \rho_C) = (3, 12).$$

The risk functions $g_E(\boldsymbol{x})$ and $g_C(\boldsymbol{x})$ are linear:

$$g_E(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_E, \qquad g_C(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_C,$$

with coefficients $\boldsymbol{\beta}_E, \boldsymbol{\beta}_C \in \mathbb{R}^d$ sampled independently for each random seed from $\mathrm{Unif}([-1,1]^d)$.

Copula-based dependence. To control dependence between E and C, we adopt a copula-based construction. For each instance i, we first sample a pair of uniform random variables $(U_i, V_i) \in [0,1]^2$ from an Archimedean Clayton copula with Kendall's $\tau \in \{0.25, 0.5, 0.75\}$. The copula parameter θ is chosen to match the desired Kendall's τ via the standard Clayton relationship. Event and censoring times are then obtained via inverse transform sampling:

$$E = S_E^{-1}(V_i \mid \boldsymbol{x}_i), \qquad C = S_C^{-1}(U_i \mid \boldsymbol{x}_i).$$

When $\tau = 0$, U_i and V_i are sampled independently, yielding marginal independence between E and C.

Censoring mechanisms. We consider three censoring mechanisms:

- **Random censoring.** Censoring times are generated independently of both $\boldsymbol{X}$ and E. Formally, $E \perp\!\!\!\perp C$.
- **Independent censoring.** Censoring depends on covariates but remains conditionally independent of the event time: $E \perp\!\!\!\perp C \mid \boldsymbol{X}$.
  This is achieved by letting $g_C(\boldsymbol{x})$ depend on $\boldsymbol{x}$ while sampling (U, V) independently.
- **Dependent censoring.** Censoring and event times are dependent even after conditioning on covariates: $E \not\perp\!\!\!\perp C \mid \boldsymbol{X}$.
  This dependence is induced via the Clayton copula with $\tau > 0$.

Observed data. The observed time and event indicator are given by

$$t_i = \min(e_i, c_i), \qquad \delta_i = \mathbb{I}\{e_i \leq c_i\}.$$

For oracle evaluation, we retain access to the true event times e_i for all instances.

Model fitting and evaluation. For each dataset, we fit a CoxPH model using the observed $(t_i, \delta_i, \boldsymbol{x}_i)$ and obtain predicted survival curves $\widehat{S}(t \mid \boldsymbol{x}_i)$ via the Breslow estimator. Model performance is evaluated in three ways: (1) *oracle* metrics computed from true event times, (2) *naive* metrics computed directly from the uncensored instances without correction, and (3) *IPCW-corrected* metrics using estimated censoring weights from the KM estimator. All metrics are computed on held-out test sets and averaged over 100 random seeds. We additionally report the empirical censoring rate and the number of events to quantify censoring severity.

Bibliography

[1] Arthur Allen. Obama Unveils $215M 'Precision Medicine' Initiative to Study Genes, Disease. *Politico*, 2015. URL `http://www.politico.com/story/2015/01/obama-precision-medicine-gene-research-114760`. Accessed: 2025-03-11.

[2] Inke R. König, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V. Kopp. What is precision medicine? *European Respiratory Journal*, 50(4):1700391, 2017.

[3] Raju Vaishya, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. Artificial Intelligence (AI) Applications for COVID-19 Pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339, 2020.

[4] Joshua C. Denny and Francis S. Collins. Precision Medicine in 2030 - Seven Ways to Transform Healthcare. *Cell*, 184(6):1415–1419, 2021.

[5] Kevin B. Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E. Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L. Snowdon. Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science*, 14(1):86–93, 2021.

[6] George H. Chen. An Introduction to Deep Survival Analysis Models for Predicting Time-to-Event Outcomes. *Foundations and Trends in Machine Learning*, 17(6):921–1100, 2024.

[7] John Graunt. *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality*. John Martyn and James Allestry, 1662.

[8] Edmond Halley. An Estimate of the Degrees of the Mortality of Mankind; Drawn from Curious Tables of the Births and Funerals at the City of Breslaw; with an Attempt to Ascertain the Price of Annuities upon Lives. *Philosophical Transactions of the Royal Society of London*, 17:596–610, 1693.

[9] Nicolas Bacaër. *A Short History of Mathematical Population Dynamics*, volume 618. Springer, 2011.

[10] Krishnan Namboodiri and Chirayath M. Suchindran. *Life Table Techniques and Their Applications*. Academic Press, 2013.

[11] Elisa T. Lee and John Wang. *Statistical methods for survival data analysis*. John Wiley & Sons, 4 edition, 2003.

[12] John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[13] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Springer, 3 edition, 2012.

[14] Takeshi Emura and Yi-Hau Chen. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.

[15] Charles E. Ebeling. *An Introduction to Reliability and Maintainability Engineering*. Waveland Press, 2019.

[16] David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 4 edition, 2023.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[18] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):1–36, 2019.

[19] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

[20] Edwin L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[21] Wayne Nelson. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1:27–52, 1969.

[22] Odd Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701–726, 1978.

[23] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[24] David Faraggi and Richard Simon. A Neural Network Model for Survival Data. *Statistics in Medicine*, 14:73–82, 1995.

[25] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):1–9, 2019.

[26] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 2016.

[27] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, 21(1):1–63, 2020.

[28] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

[29] Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pages 45–57, 1975.

[30] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[31] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94 (446):496–509, 1999.

[32] Per K. Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115, 2002.

[33] Gregory T. Armstrong, Toana Kawashima, Wendy Leisenring, Kayla Stratton, Marilyn Stovall, Melissa M. Hudson, Charles A. Sklar, Leslie L. Robison, and Kevin C. Oeffinger. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. *Journal of Clinical Oncology*, 32(12):1218, 2014.

[34] Scott D. Solomon, Adel R. Rizkala, Jianjian Gong, et al. Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the PARAGON-HF trial. *JACC: Heart Failure*, 5(7):471–482, 2017.

[35] Evelyn Fix and Jerzy Neyman. A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients. *Human Biology*, 23(3):205–241, 1951.

[36] Frank E. Harrell Jr, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[37] Glenn W. Brier and Roger A. Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

[38] Shi-Ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28244–28276, 2023.

[39] Nicholas Hartman, Sehee Kim, Kevin He, and John D. Kalbfleisch. Pitfalls of the concordance index for survival outcomes. *Statistics in medicine*, 42(13): 2179–2190, 2023.

[40] Anne-Lene Kjældgaard, Katrine Pilely, Karsten S Olsen, Anders H Jessen, Anne Ø Lauritsen, et al. Prediction of survival in amyotrophic lateral sclerosis: a nationwide, Danish cohort study. *BMC Neurology*, 21(1):164, 2021.

[41] Danfang Zhang, Eva-Maria E. Hedlund, Sharon Lim, Fang Chen, Yin Zhang, Baocun Sun, and Yihai Cao. Antiangiogenic agents significantly improve survival in tumor-bearing mice by increasing tolerance to chemotherapy-induced toxicity. *Proceedings of the National Academy of Sciences*, 108(10):4117–4122, 2011.

[42] Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1):83–104, 1997.

[43] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[44] Christoph Molnar. *Interpretable Machine Learning*. Self-published, 3 edition, 2025. URL `https://christophm.github.io/interpretable-ml-book`.

[45] Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen. Uncertainty Estimation in Deep Bayesian Survival Models. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics*, pages 1–4, 2023.

[46] Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen. Efficient Training of Probabilistic Neural Networks for Survival Analysis. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6157–6166, 2025.

[47] Lena T. B. Christensen, Dikte Straadt, Stratos Vassis, Christian M. Lillelund, Peter B. Stoustrup, Ruben Pauwels, Thomas K. Pedersen, and Christian F. Pedersen. An Explainable and Conformal AI Model to Detect Temporomandibular Joint Involvement in Children Suffering from Juvenile Idiopathic Arthritis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1–4, 2024.

[48] Christian Marius Lillelund, Ali Hossein Gharari Foomani, Weijie Sun, Shi-ang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). MENSA: A multi-event network for survival analysis with trajectory-based likelihood estimation. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pages 543–571, 2026.

[49] Christian Marius Lillelund, Sanjay Kalra, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). A meaningful prediction of functional decline in amyotrophic lateral sclerosis based on multi-event survival analysis. *PLOS ONE*, 20(11):1–16, 2025.

[50] Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Position: Stop Chasing the C-index when Evaluating Survival Analysis Models. *arXiv preprint arXiv:2506.02075*, 2026. Accepted at ICML 2026.

[51] Christian Marius Lillelund, Shi-ang Qi, and Russell Greiner. Overcoming dependent censoring in the evaluation of survival models. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pages 3841–3866, 2026.

[52] Christian Marius Lillelund, Michael Harbo, and Christian Fischer Pedersen. Prognosis of fall risk in home care clients: A noninvasive approach using survival analysis. *Journal of Public Health*, 34:905–917, 2026.

[53] Christian Marius Lillelund, Fernando Pannullo, Morten Opprud Jakobsen, and Christian Fischer Pedersen. Predicting Survival Time of Ball Bearings in the Presence of Censoring. *arXiv preprint arXiv:2309.07188*, 2023. Presented at the AAAI Fall Symposium 2023: Survival Prediction Algorithms, Challenges & Applications (SPACA).

[54] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv preprint arXiv:1603.04467*, 2016.

[55] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[56] Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

[57] Roger B. Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, 2 edition, 2006.

[58] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random Survival Forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[59] Shi-ang Qi, Christian Marius Lillelund, Weijie Sun, and Russell Greiner. Current Trends and Techniques for Evaluating Time-to-Event Outcomes. Unpublished manuscript, 2026.

[60] David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[61] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[62] Maurice J Frank. On the simultaneous associativity off (x, y) and $x + y - f(x, y)$. *Aequationes Mathematicae*, 19(1):194–226, 1979.

[63] Anastasios Tsiatis. A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.

[64] Ali Hossein Gharari Foomani, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-Based Deep Survival Models for Dependent Censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 669–680, 2023.

[65] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477): 359–378, 2007.

[66] Susan S. Li, Rebecca Miller, Dmitry Tumin, William C. L. Stewart, Joseph D. Tobias, and Don Hayes Jr. Lung allocation score thresholds prioritize survival after lung transplantation. *Chest*, 156(1):64–70, 2019.

[67] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and Lee-Jen Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[68] James M. Robins. Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers. In *Proceedings of the Biopharmaceutical Section*, pages 24–33, 1993.

[69] James M. Robins and Dianne M. Finkelstein. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics*, 56(3):779–788, 2000.

[70] Mark J. Laan and James M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

[71] Anastasios A. Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[72] Michael Schemper. The explained variation in proportional hazards regression. *Biometrika*, 77(1):216–218, 1990.

[73] Pannagadatta K. Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining*, pages 655–660, 2007.

[74] Laura Manduchi, Ričards Marcinkevičs, Michela C. Massi, et al. A Deep Variational Approach to Clustering Survival Data. In *International Conference on Learning Representations*, 2022.

[75] Shi-ang Qi, Neeraj Kumar, Jian-Yi Xu, Jaykumar Patel, Sambasivarao Damaraju, Grace Shen-Tu, and Russell Greiner. Personalized breast cancer onset prediction from lifestyle and health history information. *Plos one*, 17(12):e0279174, 2022.

[76] Per K. Andersen and Maja P. Perme. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99, 2010.

[77] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

[78] Ralph B. D'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

[79] David W. Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

[80] John E. Angus. The probability integral transform and related results. *SIAM review*, 36(4):652–654, 1994.

[81] Shi-ang Qi, Yakun Yu, and Russell Greiner. Enhancing Survival Analysis: Improving Calibration without Compromising Discrimination. PowerPoint slides, AMII Artificial Intelligence Seminar, 2025. Presented March 14, 2025, Edmonton, Canada.

[82] Shi-ang Qi, Neeraj Kumar, Ruchika Verma, Jian-Yi Xu, Grace Shen-Tu, and Russell Greiner. Using Bayesian Neural Networks to Select Features and Compute Credible Intervals for Personalized Survival Prediction. *IEEE Transactions on Biomedical Engineering*, 70(12):3389–3400, 2023.

[83] Donna Tjandra, Yifei He, and Jenna Wiens. A Hierarchical Approach to Multi-Event Survival Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):591–599, 2021.

[84] Hans C. van Houwelingen Marta Fiocco, Hein Putter. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine*, 27:4340–4358, 2008.

[85] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Springer, 2 edition, 2021.

[86] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1145–1152, 2018.

[87] Sejin Kim, Michal Kazmierski, and Benjamin Haibe-Kains. Deep-CR MTLR: a Multi-Modal Approach for Cancer Survival Prediction with Competing Risks. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications*, volume 146, pages 223–231, 2021.

[88] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24, pages 1845—1853, 2011.

[89] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data with Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021.

[90] Erling Sverdrup. Estimates and Test Procedures in Connection with Stochastic Models for Deaths, Recoveries and Transfers Between Different States of Health. *Scandinavian Actuarial Journal*, 1965(3-4):184–211, 1965.

[91] Martin G. Larson. Covariate Analysis of Competing-Risks Data with Log-Linear Models. *Biometrics*, 40(2):459–469, 1984.

[92] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks. *arXiv preprint arXiv:1912.10481*, 2019.

[93] Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

[94] Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, volume 24, pages 2348—2356, 2011.

[95] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*, volume 29. Springer, 2005.

[96] Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[97] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. In *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512, 2020.

[98] Hrushikesh Loya, Pranav Poduval, Deepak Anand, Neeraj Kumar, and Amit Sethi. Uncertainty estimation in cancer survival prediction. *arXiv preprint arXiv:2003.08573*, 2020. Accepted at AI4AH Workshop at ICLR 2020.

[99] Martin Magris and Alexandros Iosifidis. Bayesian learning for neural networks: an algorithmic survey. *Artificial Intelligence Review*, 56(10):11773–11823, 2023.

[100] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pages 2951—2959, 2012.

[101] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

[102] Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A comprehensive open-source Python package for evaluating individual survival distributions. In *Proceedings of the 2023 AAAI Fall Symposia*, volume 2, pages 453–457, 2023.

[103] Peter C. Austin and Ewout W. Steyerberg. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38(21):4051–4065, 2019.

[104] Roberto Rongo, Ambrosina Michelotti, Thomas Klit Pedersen, Cory M. Resnick, and Peter Stoustrup. Management of temporomandibular joint arthritis in children and adolescents: An introduction for orthodontists. *Orthodontics & Craniofacial Research*, 26(S1):151–163, 2023.

[105] Christopher Schmidt, Taila Ertel, Martin Arbogast, Boris Hügle, Thekla von Kalle, and Andreas Neff. The Diagnosis and Treatment of Rheumatoid and Juvenile Idiopathic Arthritis of the Temporomandibular Joint. *Deutsches Ärzteblatt International*, 119(4):47–54, 2022.

[106] Kasper Dahl Kristensen, Peter Stoustrup, Annelise Küseler, Thomas Klit Pedersen, Marika Twilt, and Troels Herlin. Clinical predictors of temporomandibular joint arthritis in juvenile idiopathic arthritis: A systematic literature review. *Seminars in Arthritis and Rheumatism*, 45(6):717–732, 2016.

[107] Lukas Müller, Christian J. Kellenberger, Elvira Cannizzaro, Dominik Ettlin, Thomas Schraner, Isabel B. Bolt, Timo Peltomäki, and Rotraud K. Saurenmann. Early diagnosis of temporomandibular joint involvement in juvenile idiopathic arthritis: a pilot study comparing clinical examination and ultrasound to magnetic resonance imaging. *Rheumatology*, 48(6):680–685, 2009.

[108] Peter Stoustrup, Marinka Twilt, Lynn Spiegel, et al. Clinical Orofacial Examination in Juvenile Idiopathic Arthritis: International Consensus-based Recommendations for Monitoring Patients in Clinical Practice and Research Studies. *The Journal of Rheumatology*, 44(3):326–333, 2017.

[109] Isabel Cardoso, Peder Frederiksen, Ina Olmer Specht, Mina Nicole Händel, Fanney Thorsteinsdottir, Berit Lilienthal Heitmann, and Lars Erik Kristensen. Age and Sex Specific Trends in Incidence of Juvenile Idiopathic Arthritis in Danish Birth Cohorts from 1992 to 2002: A Nationwide Register Linkage Study. *International Journal of Environmental Research and Public Health*, 18(16), 2021.

[110] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[111] Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 549–581, 2023.

[112] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[113] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the Stratification of Multi-label Data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158, 2011.

[114] Jerry Morris. Amyotrophic lateral sclerosis (ALS) and related motor neuron diseases: an overview. *Neurodiagnostic Journal*, 55(3):180–194, 2015.

[115] Kevin Talbot. Motor neuron disease: the bare essentials. *Practical Neurology*, 9 (5):303–309, 2009.

[116] Eva L. Feldman, Stephen A. Goutman, Susanne Petri, Letizia Mazzini, Masha G. Savelieff, Pamela J. Shaw, and Gen Sobue. Amyotrophic lateral sclerosis. *The Lancet*, 400(10360):1363–1380, 2022.

[117] Jesse M. Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, and Arline Nakanishi. The ALSFRS-R: A Revised ALS Functional Rating Scale That Incorporates Assessments of Respiratory Function. *Journal of Neurological Sciences*, 169(1-2):13–21, 1999.

[118] Nazem Atassi, James Berry, Amy Shui, et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

[119] Mattia Prosperi, Yiqiang Guo, Matthew Sperrin, James S. Koopman, Jinsong Min, Xiaoqian He, Stephen Rich, Mingyang Wang, Iain E. Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

[120] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, 1966.

[121] Muzammil Arif Din Abdul Jabbar, Ling Guo, Sonakshi Nag, Yang Guo, Zachary Simmons, Erik P. Pioro, Savitha Ramasamy, and Crystal Jing Jing Yeo. Predicting Amyotrophic Lateral Sclerosis (ALS) Progression with Machine Learning. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 25(3-4):242–255, 2024.

[122] Nazem Atassi, Priyanka Yerramilli-Rao, Jackie Szymonifka, et al. Analysis of start-up, retention, and adherence in ALS clinical trials. *Neurology*, 81(15): 1350–1355, 2013.

[123] Anna Bill-Axelson, Lars Holmberg, Hans Garmo, et al. Radical Prostatectomy or Watchful Waiting in Prostate Cancer — 29-Year Follow-up. *New England Journal of Medicine*, 379(24):2319–2329, 2018.

[124] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NialiwI2V6`.

[125] Patrick J. Heagerty and Y Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105, 2005.

[126] Moshe Zisser and Dvir Aran. Transformer-based time-to-event prediction for chronic kidney disease deterioration. *Journal of the American Medical Informatics Association*, 31(4):980–990, 2024.

[127] Xin Liu, Weijia Zhang, and Min-Ling Zhang. HACSurv: A Hierarchical Copula-Based Approach for Survival Analysis with Dependent Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3079–3087, 2025.

[128] Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep Copula-Based Survival Analysis for Dependent Censoring with Identifiability Guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20613–20621, 2024.

[129] Ming Zheng and John P Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

[130] Louis-Paul Rivest and Martin T Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79:138–155, 2001.

[131] Carmen B Franse, Judith AC Rietjens, Alex Burdorf, Amy van Grieken, Ida J Korfage, Agnes van der Heide, Francesco Mattace Raso, Ed van Beeck, and Hein Raat. A prospective study on the variation in falling and fall risk among community-dwelling older citizens in 12 European countries. *BMJ Open*, 7(6), 2017.

[132] Marika Salminen, Tero Vahlberg, and Sirkka-Liisa Kivelä. The long-term effect of a multifactorial fall prevention programme on the incidence of falls requiring medical treatment. *Public Health*, 123(12):809–813, 2009.

[133] Sandra Klaperski-van Der Wal, Adam Bruton, Luke Felton, and Susan Turner. A mixed-method exploration of the effects and feasibility of an intergenerational fall-prevention gardening programme in older adults at risk of falling: a clinical trial. *Journal of Public Health*, 2023.

[134] David S. Lindberg, Mattia Prosperi, Ragnhildur I. Bjarnadottir, et al. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach. *International Journal of Medical Informatics*, 143:104272, 2020.

[135] Brian W. Patterson, Collin J. Engstrom, Varun Sah, Maureen A. Smith, Eneida A. Mendonça, Michael S. Pulia, Michael D. Repplinger, Azita G. Hamedani, David Page, and Manish N. Shah. Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits. *Medical Care*, 57(7):560–566, 2019.

[136] Hayao Nakatani, Masatoshi Nakao, Hidefumi Uchiyama, Hiroyoshi Toyoshiba, and Chikayuki Ochiai. Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study. *JMIR Medical Informatics*, 8(4):e16970, 2020.

[137] Amandine Dubois, Audrey Mouthon, Ranjith Steve Sivagnanaselvam, and Jean-Pierre Bresciani. Fast and automatic assessment of fall risk by coupling machine learning algorithms with a depth camera to monitor simple balance tasks. *Journal of NeuroEngineering and Rehabilitation*, 16(1):71, 2019.

[138] Yang Yang, John P. Hirdes, Joel A. Dubin, and Joon Lee. Fall Risk Classification in Community-Dwelling Older Adults Using a Smart Wrist-Worn Device and the Resident Assessment Instrument-Home Care: Prospective Observational Study. *JMIR Aging*, 2(1):e12153, 2019.

[139] Gerald Wilmink, Katherine Dupey, Schon Alkire, Jeffrey Grote, Gregory Zobel, Howard M. Fillit, and Satish Movva. Artificial Intelligence–Powered Digital Health Platform and Wearable Devices Improve Outcomes for Older Adults in Assisted Living Communities: Pilot Intervention Study. *JMIR Aging*, 3(2): e19554, 2020.

[140] Ali Chelli and Matthias Patzold. A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition. *IEEE Access*, 7:38670–38687, 2019.

[141] Noman Dormosh, Martijn C Schut, Martijn W. Heymans, Nathalie Van Der Velde, and Ameen Abu-Hanna. Development and Internal Validation of a Risk Prediction Model for Falls Among Older People Using Primary Care Electronic Health Records. *The Journals of Gerontology: Series A*, 77(7):1438–1445, 2022.

[142] Geoffrey J. Hoffman, Jinkyung Ha, Neil B. Alexander, Kenneth M. Langa, Mary Tinetti, and Lillian C. Min. Underreporting of Fall Injuries of Older Adults: Implications for Wellness Visit Fall Risk Screening: Accuracy of Self-Reported Fall Injuries. *Journal of the American Geriatrics Society*, 66(6):1195–1200, 2018.

[143] Chengyin Ye, Jinmei Li, Shiying Hao, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. *International Journal of Medical Informatics*, 137:104105, 2020.

[144] Sunyang Fu, Bjoerg Thorsteinsdottir, Xin Zhang, et al. A hybrid model to identify fall occurrence from electronic health records. *International Journal of Medical Informatics*, 162:104736, 2022.

[145] Judy A. Stevens, Michael F. Ballesteros, Karin A. Mack, Rose A. Rudd, Erin DeCaro, and Gerald Adler. Gender Differences in Seeking Care for Falls in the Aged Medicare Population. *American Journal of Preventive Medicine*, 43(1): 59–62, 2012.

[146] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777, 2017.

[147] Hülya Donat and Ayşe Özcan. Comparison of the effectiveness of two programmes on older adults at risk of falling: unsupervised home exercise and supervised group exercise. *Clinical Rehabilitation*, 21(3):273–283, 2007.

[148] Enas Fawzy Youssef and Alsayed Abd Elhameed Shanb. Supervised versus home exercise training programs on functional balance in older subjects. *Malaysian Journal of Medical Sciences*, 23(6):83–93, 2016.

[149] Manuel Montero-Odasso, Nathalie van der Velde, Finbarr C. Martin, et al. World guidelines for falls prevention and management for older adults: a global initiative. *Age and Ageing*, 51(9):afac205, 2022.

[150] Darcy J. McMaughan, Oluyomi Oloruntoba, and Matthew L. Smith. Socioeconomic Status and Access to Healthcare: Interrelated Drivers for Healthy Aging. *Frontiers in Public Health*, 8:231, 2020. doi: 10.3389/fpubh.2020.00231.

[151] Nam-Ho Kim, Dawn An, and Joo-Ho Choi. *Prognostics and Health Management of Engineering Systems*. Springer International Publishing, 2017.

[152] Liang Guo, Naipeng Li, Feng Jia, Yaguo Lei, and Jing Lin. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, 240:98–109, 2017.

[153] Caifeng Zheng, Weirong Liu, Bin Chen, Dianzhu Gao, Yijun Cheng, Yingze Yang, Xiaoyong Zhang, Shuo Li, Zhiwu Huang, and Jun Peng. A Data-driven Approach for Remaining Useful Life Prediction of Aircraft Engines. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 184–189, 2018.

[154] Zeina Al Masry, Patrick Schaible, Noureddine Zerhouni, and Christophe Varnier. Remaining useful life prediction for ball bearings based on health indicators. *MATEC Web of Conferences*, 261:02003, 2019.

[155] Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Transactions on Reliability*, 69(1):401–412, 2020.

[156] Yaping Wang, Jiajun Zhao, Chaonan Yang, Di Xu, and Jianghua Ge. Remaining useful life prediction of rolling bearings based on Pearson correlation-KPCA multi-feature fusion. *Measurement*, 201:111572, 2022.

[157] Weiyang Xu, Quansheng Jiang, Yehu Shen, Fengyu Xu, and Qixin Zhu. Rul Prediction for Rolling Bearings based on Convolutional Autoencoder and Status Degradation Model. *Applied Soft Computing*, 130:109686, 2022.

[158] Zifei Xu, Musa Bashir, Qinsong Liu, Zifan Miao, Xinyu Wang, Jin Wang, and Nduka Ekere. A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model. *Computers & Industrial Engineering*, 176:108999, 2023.

[159] Carlos Ferreira and Gil Gonçalves. Remaining Useful Life prediction and challenges: A literature review on the use of Machine Learning Methods. *Journal of Manufacturing Systems*, 63:550–562, 2022.

[160] Patrick Nectoux, Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, Brigitte Chebel-Morello, Noureddine Zerhouni, and Christophe Varnier. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *IEEE International Conference on Prognostics and Health Management*, pages 1–8, 2012.

[161] Daniel Zelterman, Chap T. Le, and Thomas A. Louis. Bootstrap techniques for proportional hazards models with censored observations. *Statistics and Computing*, 6(3):191–199, 1996.

[162] Qian Zhong, Jonas Mueller, and Jianshu Wang. Deep Learning for the Partially Linear Cox Model. *The Annals of Statistics*, 50(3):1348–1375, 2022.

[163] Paidamoyo Chapfuwa, Chunyuan Li, Nikhil Mehta, Lawrence Carin, and Ricardo Henao. Survival Cluster Analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL 20)*, pages 60–68, 2020.

[164] Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep Cox Mixtures for Survival Regression. In *Machine Learning for Healthcare Conference*, pages 674–708, 2021.

[165] Vincent Jeanselme, Brian Tom, and Jessica Barrett. Neural Survival Clustering: Non-Parametric Mixture of Neural Networks for Survival Clustering. In *Conference on Health, Inference, and Learning*, pages 92–102, 2022.

[166] Xiaobin Shen and George H. Chen. Deep Kernel Aalen-Johansen Estimator: An Interpretable and Flexible Neural Net Framework for Competing Risks. *preprint, arXiv:2512.08063*, 2025.

[167] Dimitris Bertsimas, Jack Dunn, Emma Gibson, and Agni Orfanoudaki. Optimal Survival Trees. *Machine Learning*, 111(8):2951–3023, 2022.

[168] Rui Zhang, Rui Xin, Margo Seltzer, and Cynthia Rudin. Optimal Sparse Survival Trees. In *International Conference on Artificial Intelligence and Statistics*, pages 352–360, 2024.

[169] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[170] Maxim S. Kovalev, Lev V. Utkin, and Ernest M. Kasimov. SurvLIME: A Method for Explaining Machine Learning Survival Models. *Knowledge-Based Systems*, 203:106164, 2020.

[171] Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. SurvSHAP(t): Time-dependent Explanations of Machine Learning Survival Models. *Knowledge-Based Systems*, 262:110234, 2023.

[172] Hanpu Zhou, Hong Wang, Sizheng Wang, and Yi Zou. SurvMetrics: An R package for Predictive Evaluation Metrics in Survival Analysis. *R J.*, 14(4): 252–263, 2023.

[173] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

[174] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[175] Boris Van Breugel and Mihaela Van Der Schaar. Position: Why Tabular Foundation Models Should Be a Research Priority. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 48976–48993, 2024.

[176] Chitwan Saharia, William Chan, Saurabh Saxena, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 36479–36494, 2022.

[177] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[178] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When Do Neural Nets Outperform Boosted Trees on Tabular Data? In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 76336–76369, 2023.

[179] Yegor Gorishniy, Ivan Rubachev, Vadim Khrulkov, and Alexey Babenko. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18932–18943, 2021.

[180] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-Tuned Simple Nets Excel on Tabular Datasets. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 23928–23941, 2021.

[181] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=KSugKcbNf9`.

[182] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=cp5PvcI6w8_`.

[183] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature*, 637:319–326, 2025.

[184] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[185] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

[186] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL `https://openreview.net/forum?id=iO4LZibEqW`.

[187] Marco Fornili, Federico Ambrogi, Patrizia Boracchi, and Elia Biganzoli. Piecewise Exponential Artificial Neural Networks (PEANN) for Modeling Hazard Function with Right Censored Data. In *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 125–136, 2014.

[188] Shi-ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. SurvivalPFN: Amortizing Survival Prediction via In-Context Bayesian Inference. *arXiv preprint arXiv:2605.15488*, 2026.

[189] Samuel Böhm, Lennart Purucker, Frank Hutter, and Pascal Schlosser. SurvPFN: Towards Foundation Models for Survival Predictions. *arXiv preprint arXiv:2606.04564*, 2026.