Probabilistic Time-to-Event Modeling Approaches for Risk Profiling

by

Paidamoyo Chapfuwa

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Lawrence Carin, Advisor

_____
Ricardo Henao

_____
Henry Pfister

_____
Michael J. Pencina

_____
Galen Reeves

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University
2021

Abstract

Probabilistic Time-to-Event Modeling Approaches for Risk Profiling

by

Paidamoyo Chapfuwa

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Lawrence Carin, Advisor

_____
Ricardo Henao

_____
Henry Pfister

_____
Michael J. Pencina

_____
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2021

Abstract

Modern health data science applications leverage abundant molecular and electronic health data, providing opportunities for machine learning to build statistical models to support clinical practice. Time-to-event analysis, also called survival analysis, stands as one of the most representative examples of such statistical models. Models for predicting the time of a future event are crucial for risk assessment, across a diverse range of applications, i.e., drug development, risk profiling, and clinical trials, and such data are also relevant in fields like manufacturing (e.g., for equipment monitoring). Existing time-to-event (survival) models have focused primarily on preserving the pairwise ordering of estimated event times (i.e., relative risk).

In this dissertation, we propose neural time-to-event models that account for calibration and uncertainty, while predicting accurate absolute event times. Specifically, we introduce an adversarial nonparametric model for estimating matched time-to-event distributions for probabilistically concentrated and accurate predictions. We consider replacing the discriminator of the adversarial nonparametric model with a survival-function matching estimator that accounts for model calibration. The proposed estimator can be used as a means of estimating and comparing conditional survival distributions while accounting for the predictive uncertainty of probabilistic models.

Moreover, we introduce a theoretically grounded unified counterfactual inference framework for survival analysis, which adjusts for bias from two sources, namely,

confounding (from covariates influencing both the treatment assignment and the outcome) and censoring (informative or non-informative). To account for censoring biases, a proposed flexible and nonparametric probabilistic model is leveraged for event times. Then, we formulate a model-free nonparametric hazard ratio metric for comparing treatment effects or leveraging prior randomized real-world experiments in longitudinal studies. Further, the proposed model-free hazard-ratio estimator can be used to identify or stratify heterogeneous treatment effects. For stratifying risk profiles, we formulate an interpretable time-to-event driven clustering method for observations (patients) via a Bayesian nonparametric stick-breaking representation of the Dirichlet Process.

Finally, through experiments on real-world datasets, consistent improvements in predictive performance and interpretability are demonstrated relative to existing state-of-the-art survival analysis models.

Acknowledgements

I would like to express my deepest appreciation to my dissertation committee: Drs. Lawrence Carin, Ricardo Henao, Henry Pfister, Michael Pencina, and Galen Reeves.

I am deeply indebted to my advisor Lawrence Carin who took a chance on me by enthusiastically accepting me into his exceptional research group. Dr. Carin introduced me to the rigors of conducting research and provided unwavering support, including the freedom to explore an under-explored area in machine learning. Dr. Carin also taught me life lessons in communication and gave me the courage to pursue a career in academia. I am extremely appreciative of the thoroughness Dr. Carin gave to my work.

The completion of my dissertation would not have been possible without the support and nurturing of my co-advisor, Ricardo Henao, who first introduced me to survival analysis. I am extremely appreciative to Dr. Henao for his patience in teaching me the fundamentals of probability and statistics, including providing the resources necessary to thrive during my time at Duke. Dr. Henao provided invaluable feedback on my writing, problem formulations, and presentations throughout my Ph.D. study.

I would also like to extend my deepest gratitude to Dr. Michael Pencina, whose constant guidance helped me gain a deeper understanding of individualized risk assessment in clinical decision making. I am extremely grateful to Henry Pfister and Galen Reeves. Dr. Pfister helped me realize the importance of model calibration and

connections to optimal transport. Through Dr. Reeves, I have learned to appreciate the role of information theory in machine learning. I am also grateful to Dr. Cynthia Rudin and Dr. Guillermo Sapiro for serving on my qualifying and preliminary exam committees, respectively.

Throughout my time at Duke, I have been fortunate to work with Dr. Chunyuan Li, Dr. Dinghan Shen, Dr. Chenyang Tao, Serge Assaad, Shuxi Zeng, Nikhil Mehta, Yamac Isik, Dr. Benjamin Goldstein, and Courtney Page. I would like to extend my sincere thanks to my Duke collaborators for their encouragement, intellectual insights, and diverse research skills. I am also grateful to Drs. Lisa Huettel, Matthew Reynolds, Tori Lodewick, Jeffrey Krolik, Jason Yu, and Dean Connie Simmons, for imparting foundational research skills and motivating me to pursue a doctoral degree.

I am extremely grateful to my Microsoft Research mentors: Drs. Ted Meeds, Julia Greissl, Jonathan Carlson, and Chunyuan Li (again). They allowed me to explore industry research including an introduction to immunomics and high-performance computing. I cannot begin to express my thanks to Drs. Meeds and Carlson for their continued investment in my academic career post-internship. I am also grateful to Black in AI, The Grace Hopper Celebration of Women in Computing, The Georgia Tech FOCUS Fellowship, and Duke African Graduate and Professional Students Association, for the instrumental role in providing opportunities necessary for me to thrive during my Ph.D. study.

I would like to extend my sincere thanks to my fellows who gave me great help during my Ph.D. study: Charlene Chabata, Dr. Rachel Draelos, Natalia Martinez, Martin Bertran, Gregory Spell, Dan Salo, Dr. Yitong Li, Dr. Liqun Chen, Shuyang Dai, Dr. Xinyuan Zhang, Ke Bai, Christy Yuan Li, Dr. Pengyu Cheng, and the many others from Dr. Carin's research group. I gratefully acknowledge the administrative support from Angela Chanh, Amy Kostrewa, and Delores Nolen during my Ph.D. milestones.

I am extremely appreciative of the support I have received from my family over the years in whatever I do. To Brian, thank you for cultivating my curiosity. To Simba, thank you for being my greatest advocate including in my moments of self-doubt.

To my parents, thank you for the gift of education.

Contents

List of Tables

List of Figures

1

Introduction

1.1 Dissertation Contributions

Below are the key contributions of this dissertation to risk profiling, counterfactual survival analysis, and comprehensive survival-specific metrics accounting for accuracy, calibration, and uncertainty.

1.1.1 Nonparametric probabilistic time-to-event models

This dissertation introduces a baseline parametric, accelerated failure time model, which is not limited by linear covariates effects on event times as the assumption [CTL+18]. However, parametric assumptions may be restrictive for large and high-dimensional datasets. Further, given the critical time-sensitive nature of time-to-event applications, it is highly desirable to design nonparametric models that are not only temporally accurate but that also produce interpretable, population-calibrated, and uncertainty-aware predictions. To address these challenges, [CTL+18] introduces an adversarial distribution matching approach for estimating population-matched time-to-event distributions for probabilistically concentrated and accurate predictions. Then, [CTL+20] proposes a novel covariate-conditional Kaplan-Meier estimator, accounting for the predictive uncertainty in survival model calibration. More-

over, [CTL+20] formulates an approach to directly match the conditional survival function of the model to that of the ground truth without the need for adversarial learning techniques. The survival function matching approach offers theoretical connections to the p-Wasserstein metric and enables an evaluation approach for comparing the estimated survival function against the ground truth. Importantly, the proposed framework can produce both accurate, uncertainty-aware, and calibrated risk scores, thus circumventing the need for an additional recalibration step often necessary in existing time-to-event approaches.

1.1.2 Counterfactual survival analysis with balanced representations

While balanced representation learning methods have been applied successfully to counterfactual inference from observational data, approaches that account for survival outcomes are relatively limited. This work introduces the first unified representation learning-based counterfactual framework for individualized treatment effect estimation for survival outcomes from observation data [CAZ+21]. Moreover, [CAZ+21] demonstrates the importance of censoring and that accounting for informative censoring is key to counterfactual survival analysis. The proposed counterfactual inference approach adjusts for bias from two sources, namely, confounding due to covariate-dependent selection bias and the censoring mechanism (informative or non-informative). Additionally, this work formulates a novel model-free nonparametric hazard ratio (HR) metric for comparing treatment effects or leveraging prior randomized real-world experiments in longitudinal studies. The proposed HR approach is an important improvement over the widely used Cox proportional hazard, which assumes a constant population HR over time. Importantly, the proposed approach significantly outperforms alternative approaches on the challenging FRAMINGHAM observational data, where without proper bias adjustments, naive approaches result in a counter-intuitive treatment effect from statins. This work will serve as an

important baseline for future work in real-world counterfactual survival analysis.

1.1.3 Bayesian nonparametric approach for risk profiling

In a clinical setting, the identification of high-, medium- and low-risk subpopulations along with accurate estimates of event times can potentially provide a more cost-effective way of targeting interventions, treatments, and care delivery plans. To this end, this work introduces an interpretable time-to-event driven clustering method of patients via a Bayesian nonparametric stick-breaking representation of the Dirichlet Process [CLM+20]. The proposed latent clustering approach improves the characterization of individual outcomes by leveraging regularities in subpopulations, thus accounting for population-level heterogeneity. Further, this work provides extensive experimental results on diverse healthcare datasets, including, (*i*) FRAMINGHAM: a longitudinal study of heart disease, and (*ii*) EHR: a large study from Duke University Health System centered around multiple inpatient visits for Type-2 diabetes comorbidities. Compared to existing baselines, experimental findings demonstrate that the clustering approach identifies interpretable and phenotypically heterogeneous subpopulations, which are critical for identifying subjects with diverse risk profiles. Moreover, the joint learning of individualized time-to-event predictions and clustering produces subpopulations consistent with survival time, while learning the number of clusters from the data.

1.2 Background

Assume a time-to-event dataset, $\mathcal{D} = \{(\boldsymbol{x}_n, t_n, y_n)\}_{n=1}^N$, consisting of N observations (or subjects). For the n-th observation, we have d covariates, $\boldsymbol{x}_n = [x_{1n}, \ldots, x_{dn}] \in \mathbb{R}^d$, a time point, t_n, and a censoring indicator, $y_n \in \{0, 1\}$. When $y_n = 1$, t_n represents the time-to-evento f interest, and when $y_n = 0$, t_n is the censoring time. Typically, events are right censored, meaning that when $y_n = 0$ the event of interest

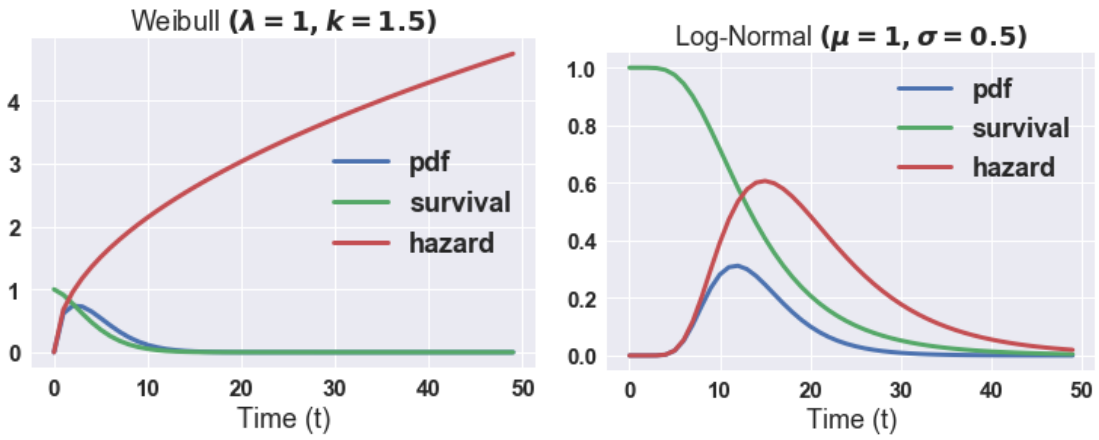


FIGURE 1.1: Parametric characterizations of time-to-event: Weibull (left) and log-normal (right) pdf $f(t|\boldsymbol{x})$, with corresponding survival function $S(t|\boldsymbol{x})$, and hazard function $\lambda(t|\boldsymbol{x})$, given the probability density function (pdf) parameters. The log-normal specification is used for the neural-based AFT model in Chapter 2.

has not been observed within time t_n. Though left and interval censored events are possible, these are far less common and are thus not usually considered in practice. Here we only consider right censoring, however, the proposed approaches are general and can be readily extended using ideas from [ADZ+20].

1.2.1 General Concepts

Time-to-event (or survival) models either characterize the conditional survival function $S(t|\boldsymbol{x})$, time density $f(t|\boldsymbol{x})$, or the hazards function $\lambda(t|\boldsymbol{x})$, where the conditioning is on covariates $\boldsymbol{x}$. The survival function $S(t|\boldsymbol{x}) = P(\tau > t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x}) = \exp\left(-\int_0^t \lambda(s|\boldsymbol{x})ds\right)$, is the fraction of the population that survives up to time t for $\tau > 0$, which can also be written as the complement of the conditional cumulative density function, $F(t|\boldsymbol{x})$; hence, $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$ is a monotonically decreasing function of time. The conditional hazards rate function $\lambda(t|\boldsymbol{x})$

$$\begin{aligned}\lambda(t|x) &= \lim_{dt\to 0} \frac{P(t < T < t + dt|X = x)}{P(T > t|X = x)dt} \\ &= -\frac{d\log S(t|x)}{dt} = \frac{f(t|x)}{S(t|x)}.\end{aligned} \tag{1.1}$$

is the instantaneous rate of occurrence of an event at time t given covariates $\boldsymbol{x}$, s.t. $\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(s|\boldsymbol{x})ds$ is the cumulative hazards.

Learning the time-to-event conditional distribution, $f(t|\boldsymbol{x})$, can in principle yield both $S(t|\boldsymbol{x})$ and $\lambda(t|\boldsymbol{x})$, provided

$$S(t|\boldsymbol{x}) = \exp\left(-\Lambda(t|\boldsymbol{x})\right) , \tag{1.2}$$

$$f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})S(t|\boldsymbol{x}) . \tag{1.3}$$

For some parametric choices of the conditional density, $f(t|\boldsymbol{x})$, the survival and hazards functions can be obtained in closed-form [KK10]. For instance, assuming the exponential density $f(t|\boldsymbol{x}) = \lambda_{\boldsymbol{x}} \exp(-\lambda_{\boldsymbol{x}} t)$, yields $\lambda(t|\boldsymbol{x}) = \lambda_{\boldsymbol{x}}$ and $S(t|\boldsymbol{x}) = \exp(-\lambda_{\boldsymbol{x}} t)$, where λ_x is a function of $\boldsymbol{x}$. Figure 1.1 provides examples of two common parametric characterizations.

In practice, we seek to approximate the time density $f(t|\boldsymbol{x})$ with $q(t|\boldsymbol{x})$, a function parametrically or nonparametrically specified and learned from data, $\mathcal{D}$. The dataset $\mathcal{D}$ represents the ground truth or, conceptually, the empirical joint distribution $p(t, y, \boldsymbol{x})$ with marginals $p(t)$, $p(y)$ and $p(\boldsymbol{x})$, from which $p(t)$ is of most interest in our case, as described below. Time-to-event models leverage the results in (1.2) and (1.3), to characterize the relationship between covariates $\boldsymbol{x}$ and time-to-event t, when estimating the conditional hazard function $\lambda(t|\boldsymbol{x})$. Two popular frameworks, Cox Proportional Hazards (CPH) [Cox92] and Accelerated Failure Time (AFT) [Wei92b] models, described below briefly for context , approach the estimation of $\lambda(t|\boldsymbol{x})$ using semi-parametric and parametric techniques, respectively.

1.2.2 Cox Proportional Hazard (CPH)

The CPH [Cox92] model is a semi-parametric, linear model where the conditional hazard function $\lambda(t|\boldsymbol{x})$ depends on time through the baseline hazard $\lambda_0(t)$ (independent of covariates $\boldsymbol{x}$) as

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\eta}) . \tag{1.4}$$

Provided with the N observation triplets in $\mathcal{D}$, CPH estimates the regression coefficients, $\boldsymbol{\eta} \in \mathbb{R}^p$, that maximize the partial likelihood [Cox92]:

$$\mathcal{L}(\boldsymbol{\eta}) = \prod_{i:y_i=1} \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\eta})}{\sum_{j:t_j \geqslant t_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\eta})}\,, \tag{1.5}$$

where $\mathcal{L}(\boldsymbol{\eta})$ is independent of the baseline hazard in (1.4). Note that (1.5) only depends on the ordering of t_i, for $i, \ldots, N$, and not their actual values. CPH is nonparametric in that it estimates the ordering of the events, not their times, thus avoiding the need to specify a distribution for $\lambda_0(t)$. Several techniques have been developed that assume a parametric distribution for $\lambda_0(t)$, in order to estimate the actual time-to-event. See [BAB05] for specifications of $\lambda_0(t)$ that result in exponential, Weibull or Gompertz survival density functions.

1.2.3 Accelerated Failure Time (AFT)

The AFT model [Wei92b] is a popular alternative to the widely used CPH model. In this model, similar to CPH, it is assumed that $\lambda(t|\boldsymbol{x}) = \psi(\boldsymbol{x})\lambda_0(\psi(\boldsymbol{x})t)$, where $\psi(\boldsymbol{x})$ is the total effect of covariates, $\boldsymbol{x}$, usually through a linear relationship $\psi(\boldsymbol{x}) = \exp(-\boldsymbol{x}^\top \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ represents the regression coefficients. If the conditional survival density function satisfies $f(t|\boldsymbol{x}) = \psi(\boldsymbol{x}) f_0(t)$, i.e., $S(t)$ independent of $\boldsymbol{x}$ like in CPH, then we can write

$$\log t = \log(t_0) - \log \psi(\boldsymbol{x}) = \xi - \boldsymbol{x}^\top \boldsymbol{\eta}\,, \tag{1.6}$$

where $t_0 \sim p_0(t)$ is the unmoderated time, thus ξ characterizes the baseline survival density distribution. Note the similarity between (1.4) and (1.6) despite the differences in their motivation. Different choices of the baseline distribution yield a variety of AFT distributions, including Weibull, log-normal, gamma and inverse Gaussian [KM05]. Intuitively, AFT assumes the effect of the covariates, $\psi(\boldsymbol{x})$, accelerates or delays the life course, which is often meaningful in a clinical or pharmaceutical setting, and sometimes easier to interpret compared with CPH [Wei92b]. Empirical

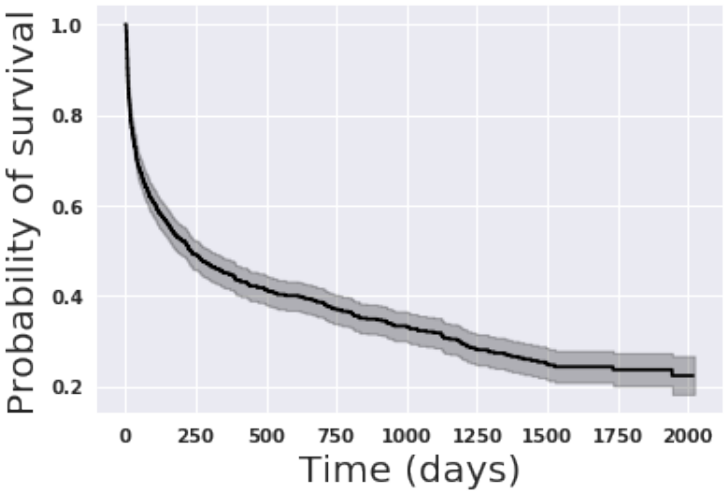


FIGURE 1.2: KM-estimate of $S(t)$ for the SUPPORT dataset (see Chapter 3 for details), evaluated at distinct and ordered observed event times (censored and non-censored). Error bars (shaded area) are calculated according to the Greenwood's formula [G$^+$26].

evidence has shown that AFT is more robust to missing values and misspecification of the survival function than CPH [KAK97]. Further, the AFT formulation lends naturally to the estimation of survival functions [KK02].

1.2.4 The Kaplan-Meier Estimator

The standard Kaplan-Meier (KM) estimator [KM58] is a widely-used frequentist approach to estimate the (marginal) survival function, $S(t)$, using samples from $p(t)$, i.e., the time-to-event empirical distribution. Let $\mathcal{T} = \{t_i | t_i > t_{i-1} > \ldots > t_0\}$ be the set of distinct and ordered observed event times (censored and non-censored). The KM estimate for time t_i can be evaluated recursively as

$$\hat{S}_{\text{KM}}(t_i) = \left(1 - \frac{d_i}{n_i}\right) \hat{S}_{\text{KM}}(t_{i-1}) \,, \tag{1.7}$$

where n_i is the number of subjects at risk at the beginning of the follow-up interval $[t_i, t_{i+1})$, d_i is the number of non-censored events that occur within the same interval, $[t_i, t_{i+1})$, and $\hat{S}_{\text{KM}}(t_0) = 1$, indicating that at t_0 there are no observed events so $d_n = 0$

and $n_0 = N$. It has been shown [HLM11] that the KM estimator can be interpreted as a random process, where the number of events, d_i, within each discrete interval $[t_i, t_{i+1})$ can be modeled as a draw from a Binomial distribution $d_i \sim \text{Binomial}(n_i, \pi)$, with mean event rate π. Moreover, it has been proven that KM is a consistent estimator [PJ77], i.e., $\sqrt{N}(\hat{S}_{\text{KM}}(t) - S(t))$ converges to a Gaussian process [BC74], with zero mean and covariance function approximated recursively by Greenwood's formula [G$^+$26]. For illustration, Figure 1.2 shows the estimated $S(t)$ according to the Kaplan-Meier estimator for the SUPPORT dataset (see Chapter 3 for details). We see that the estimated population-wide probability of survival $S_{\text{KM}}(t = 500) = 0.4$, can be interpreted as expecting 60% of individuals in the SUPPORT cohort dying within $t = 500$ days. Moreover, by the end of the study, only 20% of the population is expected to be alive, i.e., $S_{\text{KM}}(t = 2,000) = 0.2$.

2

Adversarial Time-to-Event Modeling

2.1 Introduction

Time-to-event modeling is one of the most widely used statistical analysis tools in biostatistics and, more broadly, health data science applications. For a given subject, these models estimate either a risk score or the time-to-event distribution, from the time at which a set of covariates (predictors) are observed. In practice, the model is parameterized as a weighted, often linear, combination of covariates. Time is estimated parametrically or nonparametrically, the former by assuming an underlying time distribution, and the latter as proportional to observed event times. These models have been used widely in risk profiling [ROC+18], drug development [FRG+87], and prevention of online fraudulent activities [ZYW19]. Time-to-event modeling, and in a larger context, marked temporal point processes [XFY+17, DDT+16, ME17, XXY+18, LXZ+18, UDR18, LHR+15], constitute the fundamental analytical tools in applications for which the future behavior of a system or individual is to be characterized statistically.

The principal time-to-event modeling tool is the Cox Proportional Hazards (Cox-

PH) model [Cox92]. Cox-PH is a semi-parametric model that assumes the effect of covariates is a fixed, time-independent, multiplicative factor on the hazard rate, which characterizes the instantaneous death rate of the surviving population. By optimizing a partial likelihood formulation, Cox-PH circumvents the difficulty of specifying the unknown, time-dependent, baseline hazard function. Consequently, Cox-PH results in point-estimates proportional to the event times. Further, estimation of Cox-PH models depends heavily on event ordering and not the time-to-event itself, which is known to compromise the scalability of the estimation procedure to large datasets. This poor scaling behavior is manifested because the formulation is not amenable to stochastic training with minibatches.

It is well accepted that the fixed-covariate-effects assumption made in Cox-PH is strong, and unlikely to hold in reality [Aal94]. For instance, individual heterogeneity and other sources of variation, often likely to be dependent on time, are rarely measured or totally unobservable. This unobservable variation has been gradually recognized as a major concern in survival analysis and cannot be safely ignored [Col15, AG$^+$01]. When these sources of variation are independent of time, they can be modeled via fixed or random effects [Aal94, Hou95]. However, in cases for which they render the hazard rate time-dependent, such variation is difficult to control, diagnose, or model parametrically. Cox-PH is known to be sensitive to such assumption violations [AG$^+$01, KK10]. Moreover, Cox-PH focuses on the estimation of the covariate effects rather than the survival time distribution, i.e., time-to-event prediction. The motivation behind Cox-PH and its shortcomings make it less appealing in applications where prediction is of highest importance.

An alternative to the Cox-PH model is the Accelerated Failure Time (AFT) model [Wei92b]. AFT makes the simplifying assumption that the effect of covariates either accelerates or delays the event progression, relative to a parametric baseline time-to-event distribution. However, by not making the baseline hazard a constant, as

in standard Cox-PH, AFT is often a more reasonable assumption in clinical settings when predictions are important [Wei92b]. AFT also encompasses a wide range of popular parametric proportional hazards models and proportional odds models, when the event baseline time distribution is specified properly [KM05]. Learning in AFT models falls into the category of maximum likelihood estimation, and therefore it scales well to large datasets, when trained via stochastic gradient descent. Further, AFT is also more robust to unobserved variation effects, relative to Cox-PH [KAK97].

From a machine learning perspective, recent advances in deep learning are starting to transform clinical practice. Equipped with modern learning techniques and abundant data, machine-learning-driven diagnostic applications have surpassed human-expert performance in a wide array of health care applications [CYA13, CNC+16, HDWF+17, DZAD17, GPC+16]. However, applications involving time-to-event modeling have been largely under-explored. From the existing approaches, most focus on extending Cox-PH with nonlinear neural-network-based covariate mappings [KSC+16, FS95, ZYH16], casting the time-to-event modeling as a discretized-time classification problem [YGLB11, Fot18], or introducing a nonlinear map between covariates and time via Gaussian processes [FRT16, AvdS17]. Interestingly, all of these approaches focus their applications toward relative risk, fixed-time risk (e.g., 1-year mortality) or competing events, rather than event-time estimation, which is key to individualized risk assessment.

Generative Adversarial Networks (GANs) [GPAM+14] have recently demonstrated unprecedented potential for generative modeling, in settings where the goal is to estimate complex data distributions via implicit sampling. This is done by specifying a flexible generator function, usually a deep neural network, whose samples are adversarially optimized to match in distribution to those from real data. Succesful examples of GAN include generation of images [RMC15, SGZ+16], text [YZWY17, ZGF+17] and data conditioned on covariates [IZZE17, RAY+16]. How-

ever, ideas from adversarial learning are yet to be exploited for the challenging task that is time-to-event modeling.

Previous work often represents time-to-event distributions using a limited family of parametric forms, i.e., log-normal, Weibull, Gamma, Exponential, etc. It is well understood that parametric assumptions are often violated in practice, largely because of the model is unable to capture unobserved (nuisance) variation. This fundamental shortcoming is one of the main reasons why non-parametric methods, e.g., Cox proportional hazards, are so popular. Adversarial learning leverages a representation that implicitly specifies a time-to-event distribution via sampling, rather than learning the parameters of a pre-specified distribution. Further, GAN-learning penalizes unrealistic samples, which is a known issue in likelihood-based models [KALL18].

The work presented here seeks to improve the quality of the predictions in non-parametric time-to-event models. We propose a deep-network-based nonparametric time-to-event model called a *Deep Adversarial Time-to-Event* (DATE) model. Unlike existing approaches, DATE focuses on the estimation of time-to-event distributions, rather than event ordering, thus emphasizing predictive ability. Further, this is done while accounting for missing values, high-dimensional data and censored events. The key contributions associated with the DATE model are: (*i*) The first application of GANs to nonlinear and nonparametric time-to-event modeling, conditioned on covariates. (*ii*) A principled censored-event-aware cost function that is distribution-free and independent of time ordering. (*iii*) Improved uncertainty estimation via deep neural networks with stochastic layers. (*iv*) An alternative, parametric, non-adversarial time-to-event AFT model to be used as baseline in our experiments. (*v*) Results on benchmark and real data demonstrate that DATE outperforms its parametric counterpart by a substantial margin.

2.2 Baseline Deep Regularized Accelerated Failure Time (DRAFT)

DRAFT is a semi-parametric survival-analysis model that uses deep networks and regularization to improve the basic Accelerated Failure Time (AFT) framework. We estimate the conditional survival density function $f(t|\boldsymbol{x})$, given observed covariates $\boldsymbol{x} \in \mathbb{R}^p$. The survival time is assumed to be distributed log-normal, and is estimated via MLE with a deep neural network specification that learns parameters $\mu_{\boldsymbol{\beta}}(\boldsymbol{x})$ and $\sigma^2_{\boldsymbol{\beta}}(\boldsymbol{x})$, i.e., the mean and standard deviation of the log-normal distribution. For convenience, we adopt the log-normal distribution for event time t, because we found that it is considerably more stable during optimization, compared to other popular survival distributions, e.g., Weibull or Gamma. The restriction on the support for the time-to-event target, e.g., nonnegative real numbers, compounded by the tail behavior of some distributions, such as the Weibull, posed significant numerical challenges when evaluating derivatives for the optimization procedure.

Starting from (1.6), we consider the following MLP-based log-normal AFT:

$$\log t = \mu_{\boldsymbol{\beta}}(\boldsymbol{x}) + \xi\,, \quad \xi \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\beta}}(\boldsymbol{x}))\,, \tag{2.1}$$

where $\mu_{\boldsymbol{\beta}}(\boldsymbol{x})$ and $\sigma^2_{\boldsymbol{\beta}}(\boldsymbol{x})$ are MLPs parameterized by $\boldsymbol{\beta}$, representing the mean and variance of the log-transformed time-to-event as a function of covariates $\boldsymbol{x}$.

The likelihood function of the log-normal AFT model in (2.1) for all events (censored and non-censored) is then

$$\begin{aligned}\prod_i^N p(t_i|\boldsymbol{x}_i) &= \prod_{i:l_i=1} f_{\boldsymbol{\beta}}(t_i|\boldsymbol{x}_i) \prod_{i:l_i=0} S_{\boldsymbol{\beta}}(t_i|\boldsymbol{x}_i) \\ &= \prod_{i:l_i=1} \phi(\nu(t_i, \boldsymbol{x}_i)) \prod_{i:l_i=0} (1 - \Phi(\nu(t_i, \boldsymbol{x}_i)))\,,\end{aligned} \tag{2.2}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the Gaussian density and cumulative density functions, respectively, $\nu(t_i, \boldsymbol{x}_i) = (\log t_i - \mu_{\boldsymbol{\beta}}(\boldsymbol{x}_i))/\sigma_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$. The likelihood in (2.2) is convenient,

because it allows estimation of time-to-event, while seamlessly accounting for censored events. The latter comes as a benefit of having a parametric model with closed-form cumulative density function.

The loss function $\mathcal{L}(\boldsymbol{\beta}; \mathcal{D})$ for the Deep Regularized AFT (DRAFT) model is

$$\mathcal{L}(\boldsymbol{\beta}; \mathcal{D}) = -\log p(t|\mathcal{D}) + \eta R(\boldsymbol{\beta}; \mathcal{D}) \,, \tag{2.3}$$

where the first term is the negative log-likelihood loss from (2.2), $\eta > 0$ is a tuning parameters, and $R(\boldsymbol{\beta}; \mathcal{D})$ is a regularization loss that encourages event times to be properly ordered. Specifically, we use the following $R(\boldsymbol{\beta}; \mathcal{D})$ adapted from [SKDo$^+$08]:

$$R(\boldsymbol{\beta}; \mathcal{D}) = \frac{1}{|\mathcal{E}|} \sum_{i:l_i=1} \sum_{j:t_j>t_i} 1 + \frac{\log \sigma(\mu_{\boldsymbol{\beta}}(\boldsymbol{x}_j) - \mu_{\boldsymbol{\beta}}(\boldsymbol{x}_i))}{\log 2}, \tag{2.4}$$

where $\mathcal{E}$ is the set of all pairs $\{i, j\}$ in $\{1, \ldots, N\}$ for which the second argument is observed, i.e., $l_i = 1$, and $\sigma(\cdot)$ is the sigmoid function. The loss function above is a lower bound on the Concordance Index (CI) [HJLC$^+$84], which constitutes a difficult-to-optimize discrete objective, that is widely used as a performance metric for survival analysis, precisely because it captures time-to-event order. Further, it is reminiscent of the partial-likelihood of CPH in (1.5), but is more amenable to stochastic training. The loss function in (2.3) is optimized using stochastic gradient descent on minibatches from $\mathcal{D}$.

2.3 Deep Adversarial Time-to-Event (DATE)

We develop a nonparametric model for $p(t|\boldsymbol{x})$, where t is the (non-censored) time-to-event from the time at which covariates $\boldsymbol{x}$ were observed. More precisely, we learn the ability to sample from $p(t|\boldsymbol{x})$ via approximation $q(t|\boldsymbol{x})$, the conditional time-to-event distribution. Further, we do so without specifying a distribution for the marginal

(baseline survival distribution), $p_0(t)$, which in AFT is usually assumed log-normal. Like in CPH and AFT, we assume that $p_0(t)$ is independent of covariates $\boldsymbol{x}$.

For censored events, $l_i = 0$, we wish the model to have a high likelihood for $p(t > t_i|\boldsymbol{x}_i)$, while for non-censored events, $l_i = 1$, we wish that the pairs $\{\boldsymbol{x}_i, t_i\}$ be consistent with data generated from $p(t|\boldsymbol{x})p_0(\boldsymbol{x})$, where $p_0(\boldsymbol{x})$ is the (empirical) marginal distribution for covariates, from which we can sample but whose explicit form is unknown.

We consider a conditional generative adversarial network (cGAN) [MO14], in which we draw approximate samples from $p(t|\boldsymbol{x})$, for $l_i = 1$, as

$$t = G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1), \quad \boldsymbol{\epsilon} \sim p_\epsilon(\boldsymbol{\epsilon}), \tag{2.5}$$

where $p_\epsilon(\boldsymbol{\epsilon})$ is a simple distribution, e.g., isotropic Gaussian or uniform (discussed below). The generator, $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1)$ is a deterministic function of $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$, specified as a deep neural network with model parameters $\boldsymbol{\theta}$, that implicitly defines $q_\theta(t|\boldsymbol{x}, l = 1)$ in a nonparametric manner. We explicitly note that $l = 1$, to emphasize that all t drawn from this model are event times (non-censored times). Ideally the pairs $\{\boldsymbol{x}, t\}$ manifested from the model in (2.5) are indistinguishable from the observed data $\{\boldsymbol{x}, t, l = 1\} \in \mathcal{D}$, i.e., the non-censored samples.

Let $\mathcal{D}_{nc} \subset \mathcal{D}$ and $\mathcal{D}_c \subset \mathcal{D}$ be the disjoint subsets of non-censored and censored data, respectively. Given a discriminator function $D_{\boldsymbol{\phi}}(\boldsymbol{x}, t)$ specified as a deep neural network with model parameters $\boldsymbol{\phi}$, the loss function based on the non-censored data has the form

$$\begin{aligned} \ell_1(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}_{nc}) = \mathbb{E}_{(t,\boldsymbol{x})\sim p_{nc}}[D_{\boldsymbol{\phi}}(\boldsymbol{x}, t)] \\ + \mathbb{E}_{\boldsymbol{x}\sim p_{nc}, \boldsymbol{\epsilon}\sim p_\epsilon}[1 - [D_{\boldsymbol{\phi}}(\boldsymbol{x}, G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1))]\,, \end{aligned} \tag{2.6}$$

where $p_{nc}(t, \boldsymbol{x})$ is the empirical joint distribution responsible for $\mathcal{D}_{nc}$, and the expectation terms are estimated through samples $\{t, \boldsymbol{x}\} \sim p_{nc}(t, \boldsymbol{x})$ and $\boldsymbol{\epsilon} \sim p_\epsilon(\boldsymbol{\epsilon})$ only.

We seek to maximize $\ell_1(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}_{nc})$ wrt discriminator parameters $\boldsymbol{\phi}$, while seeking to minimize it wrt generator parameters $\boldsymbol{\theta}$. For non-censored data, the loss in (2.6) is the standard cGAN.

We also leverage the censored data $\mathcal{D}_c$ to inform the parameters $\boldsymbol{\theta}$ of generative model $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 0)$. We therefore consider the additional loss function

$$\ell_2(\boldsymbol{\theta}; \mathcal{D}_c) = \mathbb{E}_{(t,\boldsymbol{x})\sim p_c, \boldsymbol{\epsilon}\sim p_\epsilon}[\max(0, t - G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 0)]\,, \tag{2.7}$$

where $\max(0, \cdot)$ encodes that $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 0)$ incurs no loss as long as the sampled time is larger than the censoring point. Further, $p_c(t, \boldsymbol{x})$ is the empirical joint distribution responsible for $\mathcal{D}_c$, from which samples $\{t, \boldsymbol{x}\}$ are drawn to approximate the expectation in (2.7). Note that $\max(0, \cdot)$ is one of many choices; smoothed or margin-based alternatives may be considered, but are not addressed here, for simplicity.

For cases in which the proportion of observed events is low, the loss in (2.6) and (2.7) under-represent the desire that time-to-events must be as close as possible to the ground truth, t. For this purpose, we also impose a distortion loss $d(\cdot, \cdot)$

$$\ell_3(\boldsymbol{\theta}; \mathcal{D}_{nc}) = \mathbb{E}_{(t,\boldsymbol{x})\sim p_{nc}}[d(t, G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1))]\,, \tag{2.8}$$

that penalizes $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1)$ for not being close to the event time t for non-censored events only. In the experiments in Section 2.5, we set $d(a, b) = \|a - b\|_1$, i.e., absolute error.

The complete loss function is

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}) = \ell_1(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}_{nc}) \\ + \lambda_2 \ell_2(\boldsymbol{\theta}; \mathcal{D}_c) + \lambda_3 \ell_3(\boldsymbol{\theta}; \mathcal{D}_{nc})\,, \end{aligned} \tag{2.9}$$

where $\{\lambda_2, \lambda_3\} > 0$ are tuning parameters controlling the trade-off between non-censored and censored loss functions relative to the discriminator objective in (2.6). In Section 2.5 we set $\lambda_2 = \lambda_3 = 1$, provided that (2.7) and (2.8) are written in terms

of expectations, thus already accounting for the proportion differences in $\mathcal{D}_c$ and $\mathcal{D}_{nc}$. However, this may not be sufficient in heavily imbalanced cases or when the time domains for $\mathcal{D}_c$ and $\mathcal{D}_{nc}$ are very different.

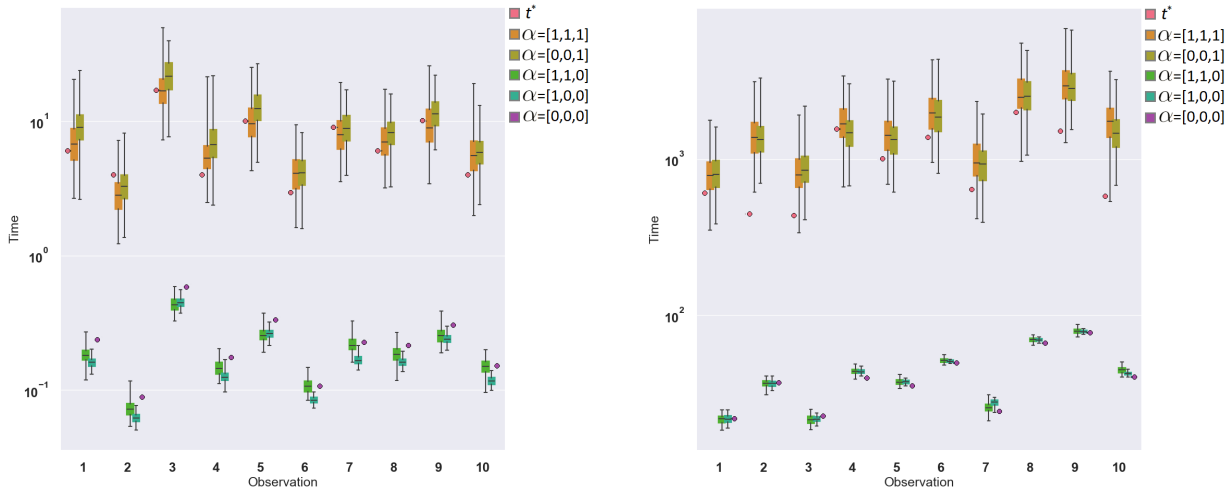
The loss function in (2.9) is optimized using stochastic gradient descent on mini-batches from $\mathcal{D}$. We maximize $\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D})$ wrt $\boldsymbol{\phi}$ and minimize it wrt $\boldsymbol{\theta}$. The terms $\ell_1(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}_{nc})$ and $\ell_3(\boldsymbol{\theta}; \mathcal{D}_{nc})$ reward $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1)$ if it synthesizes data that are consistent with $\mathcal{D}_{nc}$, and the term $\ell_2(\boldsymbol{\theta}; \mathcal{D}_c)$ encourages $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon}; l = 1)$ to generate event times that are consistent with the data $\mathcal{D}_c$, i.e., larger than censoring times.

2.3.1 Time-to-event uncertainty

The generator in (2.5) has a single source of stochasticity, $\boldsymbol{\epsilon}$, which in GAN-based models has been traditionally applied as input to the model, independent of covariates $\boldsymbol{x}$. In an MLP architecture, $\boldsymbol{h}_1 = g(\boldsymbol{W}_{10}\boldsymbol{x} + \boldsymbol{W}_{11}\boldsymbol{\epsilon})$, where $\boldsymbol{h}_1$ denotes the vector of layer-1 hidden units, $g(\cdot)$ is the activation function (RELU in the experiments), $\boldsymbol{W}_{10}$ and $\boldsymbol{W}_{11}$ are weight matrices for covariates and noise, respectively, and bias terms have been omitted for simplicity.

In a model with multiple layers, the noise term applied to the input tends to have a small effect on the distribution of sampled event times (see Section 2.5). More specifically, samples from $q_\theta(t|\boldsymbol{x})$ tend to have small variance. This results in a model with underestimated uncertainty, hence overconfident predictions. This is due to many factors, including compounding effects of activation nonlinearities, layer-wise regularizers (e.g., dropout), and cancelling terms when the support of the noise distribution is the real line (both positive and negative). Although the loss function in (2.9) rewards the generator for producing (non-censored) event times close to the ground-truth, thus in principle encouraging event time distributions to cover it, this rarely happens in practice. This issue is well-known in the GAN literature [SGZ+16].

Here we take a simple approach, consisting of adding sources of stochasticity to



(a) Non-Censored (b) Censored

FIGURE 2.1: Effects of stochastic layers on uncertainty estimation on 10 randomly selected test-set subjects from the SUPPORT dataset. Ground truth times are denoted as t^* and box plots represent time-to-event distributions from a 2-layer model, where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]$ indicates whether the corresponding noise source, $\{\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$, is active. For example $\boldsymbol{\alpha} = [1, 0, 0]$ indicates noise on the input layer only.

every layer of the generator as $\boldsymbol{h}_j = g(\boldsymbol{W}_{j0}\boldsymbol{h}_{j-1} + \boldsymbol{W}_{j1}\boldsymbol{\epsilon}_j)$, where $j = 1, \ldots, L$ and L is the number of layers. By doing this, we encourage increased coverage on the event times produced by the generator, without substantially changing the model or the learning procedure. In the experiments, we use a multivariate uniform distribution, $\boldsymbol{\epsilon}_j \sim \text{Uniform}(0, 1)$, for $j = 1, \ldots, L$, over Gaussian to reduce cancelling effects. As we show empirically, this approach produces substantially better coverage compared to having noise only on the input layer and without the convergence issues associated with the additional stochasticity.

In Figure 2.1 we illustrate the contribution of the noise on each layer to the distribution of event times. In this example, we show 10 test-set estimated time-to-event distributions using a 2-layer model with noise sources in all layers, including the input. We see that ground-truth times are nicely covered by the estimated

distributions. Also, that the combination of noise sources, rather than any individual source, jointly contribute to the desired distribution coverage.

2.4 Related Work

Deep learning models, specifically MLPs, have been successfully integrated with Cox-PH-based objectives to improve risk estimation in time-to-event models. [FS95] proposed an neural-network-based model optimized using the standard partial-likelihood cost function from Cox-PH. [KSC+16] is similar to [FS95], but leverages modern deep learning techniques such as weight decay, batch normalization and dropout. [LSC+17] replaced the partial-likelihood formulation in (1.5) with Efron's approximation [Efr77] and an isotonic regression cost function adapted from [MJV+12] to handle censored events. [ZYH16] proposed a time-to-event model for image covariates based on convolutional networks.

From the Gaussian process literature, [FRT16] proposed a time-to-event model inspired by a Poisson process, where the nonlinear map between covariates and time is modeled as a Gaussian process on the Poisson rate. More recently, [AvdS17] proposed a deep multi-task Gaussian process model for survival analysis with competing risks, and learned via variational inference. Following a different path, other approaches recast the time-to-event problem as a classification task. [YGLB11] proposed a linear model where (discretized) time is estimated using a sequence of dependent regressors. More recently, [Fot18] extended their approach to a nonlinear mapping of covariates using deep neural networks. Generative approaches have also been proposed to infer survival-time distributions with variational inference. Deep Survival Analysis (DSA) [RPEB16] specifies a latent model that leverages deep exponential family distribution, however their approach does not handle censored events.

All the above methods focus on relative risk quantified as the CI on the ordering of event times, or fixed-time risk, e.g., 1-year mortality. However, relative risk is most

useful when associated with covariate effects, which is difficult in nonlinear models based either on neural networks or Gaussian processes. Fixed-time risk, although very useful in practice, can be recast as a classification problem rather than a substantially more complex time-to-event model. Importantly, none of these approaches consider the task of time-to-event estimation, despite the fact that Gaussian process and generative approaches can be repurposed for such task.

2.5 Experiments

The loss functions in (2.9) and (2.3) for DATE and DRAFT, respectively, are minimized via stochastic gradient descent. At test time, we draw 200 samples from (2.5) and (2.1) for DATE and DRAFT, respectively, and use medians for quantitative results requiring point estimates, i.e., $\hat{t} = \text{median}(\{t_s\}_{s=1}^{200})$, where t_s is a sample from the trained model. Detailed network architectures, optimization parameters and initialization settings are in the Supplementary Material. TensorFlow code to replicate experiments can be found at `https://github.com/paidamoyo/adversarial_time_to_event`.

Comparison Methods For non-deep learning based models, we considered arguably the two most popular approaches to time-to-event modeling, namely, (regularized) Cox-Efron and RSF. For deep learning models, we considered DRAFT, which generalizes existing neural-network-based methods by using both a parametric log-normal AFT objective and a non-parametric ordering cost function. Extending DRAFT to a mixture of log-normal distributions with different variances but shared mean, did not result in improved performance. We did not consider (variational) models for non-censored events only because learning from censored events is one of the main defining characteristics of time-to-event modeling (otherwise, the model essentially becomes non-negative regression). Further, in most practical situations the propor-

Table 2.1: Summary of datasets used in experiments.

	EHR	FLCHAIN	SUPPORT	SEER
Events (%)	23.9	27.5	68.1	51.0
N	394,823	7,894	9,105	68,082
p (cat)	729 (106)	26 (21)	59 (31)	789 (771)
NaN (%)	1.9	2.1	12.6	23.4
$t_{\max}$	365 days	5,215 days	2,029 days	120 months

tion of non-censored events is low, e.g., 24% in the EHR.

Datasets Our model is evaluated on 4 diverse datasets: *i*) FLCHAIN: a public dataset introduced in a study to determine whether non-clonal serum immunoglobin free light chains are predictive of survival time [DKK+12]. *ii*) SUPPORT: a public dataset introduced in a survival time study of seriously-ill hospitalized adults [KHL+95]. *iii*) SEER: a public dataset provided by the Surveillance, Epidemiology, and End Results Program. See [RYJK+07] for details concerning the definition of the 10-year follow-up breast cancer subcohort used in our experiments. *iv*) EHR: a large study from Duke University Health System centered around inpatient visits due to comorbidities in patients with Type-2 diabetes.

The datasets are summarized in Table 2.1, where p denotes the number of covariates to be analyzed, after one-hot-encoding for categorical (cat) variables. Events indicates the proportion of the observed events, i.e., those for which $l_i = 1$. NAN indicates the proportion of missing entries in the $N \times p$ covariate matrix and $t_{\max}$ is the time range for both censored and non-censored events.

Details about the public datasets: FLCHAIN, SUPPORT and SEER, including pre-processing procedures, can be found in the references provided above. EHR is a study designed to track primary care encounters of $19,064$ Type-2 diabetes patients over a period of 10 years (2007-2017). The purpose of the analysis is to predict diabetes-related causes of hospitalization within 1 year of an EHR-recorded primary care encounter. Data is processed and analyzed at the patient encounter-level. The

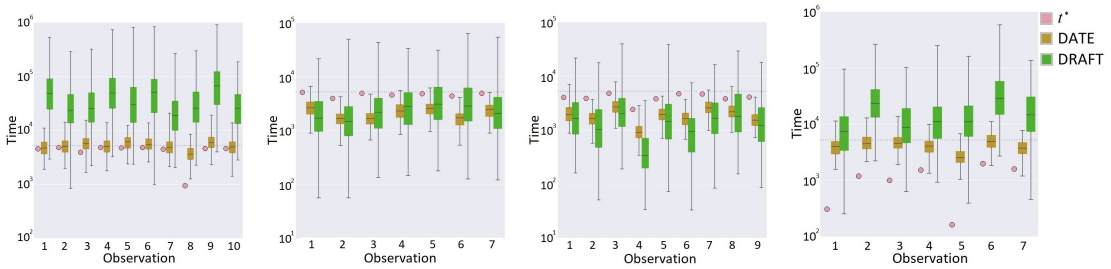


FIGURE 2.2: Example test-set predictions on FLCHAIN data. Top best (left) and worst (middle-left) predictions on censored events, and top best (middle-right) and worst (right) predictions on non-censored events. Circles denote ground-truth events or censoring points, while box-plots represent distributions over 200 samples for both DATE and DRAFT. The horizontal dashed line represents the range ($t_{\max} = 5,215$ days) of the events.

total number of encounters is $N = 394,823$. To avoid bias due to multiple encounters per patient, we split the training, validation and test sets so that a given patient can only be in one of the sets. The covariates, collected over a period of a year before the primary care encounter of interest, consist of a mixture of continuous and categorical summaries extracted from electronic health records: vitals and labs (minimum, maximum, count and mean values); comorbidities, medications and procedures ICD-9/10 codes (binary indicators and counts); and demographics (age, gender, race, language, smoking indicator, type of insurance coverage, as either continuous or categorical variables).

2.5.1 Qualitative results

First, we visually compare the test-set time-to-event distributions by DATE and DRAFT on FLCHAIN data. In Figure 2.2 we show the top best (left) and worst (middle-left) predictions on censored events, and the top best (middle-right) and worst (right) predictions on non-censored events. Circles denote ground-truth events or censoring points, while box-plots represent distributions over 200 samples for both DATE and DRAFT models. We see that: (*i*) in nearly every case, DATE is more accurate than DRAFT. (*ii*) DRAFT tends to make predictions outside the event range ($t_{\max} = 5,215$ days), denoted as a horizontal dashed line. (*iii*) DRAFT

tends to overestimate the variance of its predictions, approximately by one order of magnitude relative to DATE. This is not very surprising as DRAFT has an MLP dedicated to estimate, conditioned on the covariates, the variance of the time-to-event distribution. However, note that variances estimated well over the domain of the events ($t_{\max}$) are not necessarily meaningful or desirable. Figures with similar findings for the other 3 datasets can be found in the Supplementary Material.

To provide additional insight into the performance of DATE compared to DRAFT, we report the Normalized Relative Error (NRE) defined as $(\hat{t}-t)/t_{\max}$ and $\min(0, \hat{t}-t)/t_{\max}$ for non-censored and censored events, respectively, where t, $\hat{t}$ and $t_{\max}$ denote the ground-truth time-to-event, median time estimated (from samples) and event range, as indicated in Table 2.1. The NRE distribution provides a visual representation of the extent of test-set errors, while revealing whether the models are biased toward either overestimating ($\hat{t} > t^*$) or underestimating ($t^* > \hat{t}$) the event times. Although models with unbiased NREs are naturally preferred, in most clinical applications where being conservative is important, overestimated time-to-events must be avoided as much as possible. Figure 2.3 shows NRE distributions for test-set non-censored events on SUPPORT and EHR data. We see that DRAFT results in a considerable amount of errors beyond the event range ($|\text{NRE}| > 1$), $t_{\max} = 120$ months or $t_{\max} = 365$ days for SUPPORT and EHR, respectively. Further, we see that the NRE distribution for DRAFT is heavily skewed toward $\text{NRE} > 1$, thus tending to overestimate event times. On the other hand, DATE produces errors substantially more concentrated around 0 and within $|\text{NRE}| < 1$, relative to DRAFT. This demonstrates the advantage of the adversarial method DATE over the likelihood-based method DRAFT in generating realistic samples. Similar results were observed on the other datasets for both censored and non-censored events. See Supplementary Material for additional figures.

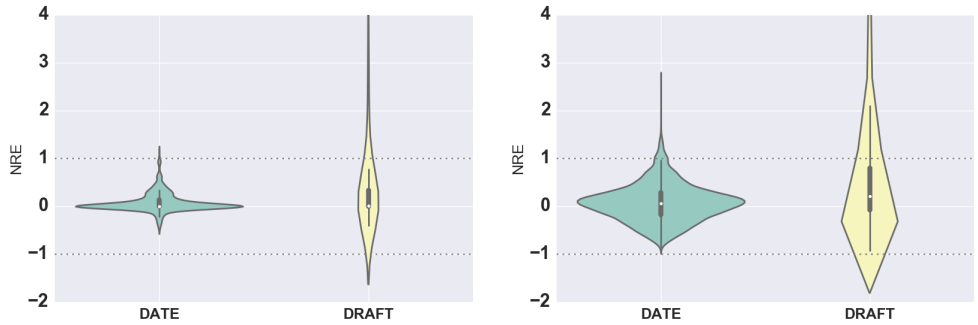


FIGURE 2.3: Normalized Relative Error (NRE) distribution for SUPPORT (top) and EHR (bottom), test-set non-censored events. The horizontal dashed lines represent the range of the events, $t_{\max} = 120$ months and $t_{\max} = 365$ days, respectively.

2.5.2 Quantitative results

Relative absolute error The perfromance of DATE is evaluated in terms of absolute error relative to the event range, i.e., $|\hat{t} - t|/t_{\max}$. For censored events, the relative error is defined as $\max(0, t - \hat{t})/t_{\max}$, to account for the fact that no error is made as long as $t \leqslant \hat{t}$. Table A.4 shows median and 50% empirical intervals for relative absolute errors on non-censored events, on all test-data. Results on censored data are small and comparable across approaches, and are thus presented in the Supplementary Material. Specifically, we see that DATE outperforms DRAFT in 3 our of 4 cases by a substantial margin, and is comparable on the SUPPORT data. For instance, on EHR data 75% of all DATE test-set predictions have a relative absolute error less than 43% (approx. 156 days) which is substantially better than the 81% (approx. 295 days) by DRAFT.

Missing data Since missing data are common in clinical data, e.g., SEER data contains 23.4% missing values, we also consider a modified version of DATE, where the generator in (2.5) takes the form $t = G_\theta(\boldsymbol{z}, \boldsymbol{\epsilon}, l = 1)$, where $\boldsymbol{z}$ is modeled as an adversarial autoencoder [DBP+16, LLC+17, CDP+18] with an encoder/decoder pair specified similar to DATE. See Supplementary Material for additional details. This model, denoted in Table 2.2 as DATE-AE, does not require missing covariates to be

Table 2.2: Median relative absolute errors (as percentages of $t_{\max}$), on non-censored data. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	DATE	DATE-AE	DRAFT
EHR	$\mathbf{23.6}_{(11.1,43.0)}$	$24.5_{(12.4,44.0)}$	$36.7_{(16.1,81.3)}$
FLCHAIN	$19.5_{(9.5,31.1)}$	$\mathbf{19.3}_{(8.9,32.4)}$	$26.2_{(9.0,53.5)}$
SUPPORT	$2.7_{(0.4,16.1)}$	$\mathbf{1.5}_{(0.4,19.2)}$	$2.0_{(0.2,35.3)}$
SEER	$\mathbf{18.6}_{(8.3,34.1)}$	$20.2_{(10.3,35.8)}$	$23.7_{(9.9,51.2)}$

imputed before hand, which is the case of DATE and DRAFT as specified originally. The results show no substantial performance improvement by DATE-AE, relative to DATE; results indicate that all of these approaches (DRAFT, DATE and DATE-AE) are robust to missing data. As a benchmark, we took FLCHAIN and SUPPORT, to then artificially introduced missing values ranging in proportion from 20% to 50%. Results in Supplementary Material support the idea of robustness, since all three approaches resulted in median relative absolute errors within 1% of those in Table 2.2.

We also tried to quantify statistically the match between time-to-event samples generated from DATE and those from the empirical distribution of the data, using the distribution-free two-sample test based on Maximum Mean Discrepancy (MMD) proposed by [STS+17]. Due to sample size limitations (number of non-censored events in the test-set) and high-variances on the p-value estimates, we could not reliably reject the hypothesis that real and DATE samples are drawn from the same distribution. We did confirm it for DRAFT, which is not surprising considering both qualitative and quantitative results discussed above.

Concordance Index The concordance Index (CI) [HJLC+84], which quantifies the degree to which the order of the predicted times is consistent with the ground truth, is the most well-known performance metric in survival analysis. Although not the

Table 2.3: Median of 95% intervals for all test-set time-to-event distributions on SUPPORT data. Ranges in parentheses are 50% empirical quantiles.

	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
Non-censored			
All	$60.0_{(3.9,176.5)}$	$149.9_{(8.5,926.8)}$	$37.9_{(3.5,237.4)}$
Input	$28.9_{(1.8,114.8)}$	$22.4_{(1.5,91.2)}$	$33.7_{(1.6,127.6)}$
Output	-	$168.8_{(16.6,844.3)}$	-
Censored			
All	$231.3_{(177.2,332.1)}$	$1397.3_{(990.9,2000.1)}$	$350.5_{(254.4,539.3)}$
Input	$137.3_{(99.4,205.0)}$	$86.9_{(64.4,135.0)}$	$155.8_{(106.7,229.3)}$
Output	-	$1158.6_{(873.8,1670.4)}$	-

focus of our approach, we compared DATE to DATE-AE, DRAFT, Random Survival Forests [IKBL08], and Cox-PH (with Efron's approximation [Efr77]). We found all of these models to be largely comparable. The results, presented in the Supplementary Material, show DATE(-AE) and DRAFT being the best-performing models on EHR and SUPPORT, respectively. On SEER, DATE(-AE) and DRAFT outperform Cox-PH and RSF. Finally, on FLCHAIN, the smallest dataset, all methods perform about the same. Note that unlike Cox-PH and DRAFT in (1.5) and (2.3), DATE(-AE) does not explicitly encourage proper time-ordering on the objective function; it is consequently deemed a strength of the proposed GAN-based DATE model that it properly learns ordering, without needing to explicitly impose this condition when training. DATE does not have a clear advantage in terms of learning the correct order, however, we verified empirically that adding an ordering cost function, $R(\boldsymbol{\beta}; \mathcal{D})$ in (2.3), to DATE does not improve the results.

Distribution coverage We now demonstrate that the DATE model, with noise sources on all layers, has time-to-event distributions with larger variances than versions of DATE with noise only on the input of the neural network. Table 2.3 shows the median of the 95% intervals for all test-set time-to-event distributions on SUPPORT

data. DATE with Uniform(0,1) has larger variance and coverage compared to the other alternatives, while keeping relative absolute errors and CIs largely unchanged (see Supplementary Material for details). We did not run models with Uniform(-1,1) and Gaussian(0,1) only on the output layer, because from the other results presented above it is clear that these two options are not nearly as good as having Uniform(0,1) noise on all layers. Note also that we did not include DRAFT in these comparisons. DRAFT has naturally good coverage due to the variance of the time-to-event distributions being modeled independent for each observation as a function of the covariates (see for instance Figure 2.2). However, DRAFT has difficulties keeping good coverage while maintaining good performance, i.e., small absolute relative error.

2.6 Conclusions

We have presented an adversarially-learned time-to-event model that leverages a distribution-free cost function for censored events. The proposed approach extends GAN models to time-to-event modeling with censored data, and it is based on deep neural networks with stochastic layers. The model yields improved uncertainty estimation relative to alternative approaches. As a baseline model for our experiments, we also proposed a parametric AFT-based with a parametric log-normal distribution on the time of event. To the best of our knowledge, this work is the first to leverage adversarial learning to improve estimation of time-to-event distributions, conditioned on covariates. Experimental results on challenging time-to-event datasets showed that DATE, our adversarial solution, consistently outperforms DRAFT, its parametric (log-normal) counterpart. As future work, we will extend DATE to models with competing risks and longitudinally measured covariates.

3

Calibration and Uncertainty in Neural Time-to-Event Modeling

3.1 Introduction

Estimating temporally accurate event times typically involves the use of parametric Maximum Likelihood Estimation (MLE) approaches [KK10] or recently-developed nonparametric sampling based methods, e.g., via adversarial learning [CTL+18] or normalizing flows [MPER18]. Conventional nonparametric time-to-event (also called survival) models primarily involve methods that target the Concordance Index (C-Index) [HJLC+84], a metric related to the receiver operating characteristic, that quantifies the degree to which estimated event times result in pairwise orderings that are consistent with observed event times, i.e., the ground truth. Consequently, any model that is able to estimate properly ordered but proportional event times can score high in terms of C-Index. However, in many applications one is not only interested in the relative risk, but also in the absolute time of an event.

Classical survival models include the Cox Proportional Hazards (CPH) semiparametric model [Cox92] that learns relative risk (proportional to time-to-event) as a

function of covariates, and the Accelerated Failure Time (AFT) model [Wei92b], a parametric specification for temporally accurate event times, that assumes covariates either accelerate or decelerate the progression of event time. AFT often assumes log-normal distributed event times, however, other likelihood functions have been considered, e.g., exponential, Gamma, Weibull, etc. [BAB05, KK10]. These classical approaches assume a linear relationship between event times and covariates, which may be limiting for modern, large and highly heterogeneous datasets.

Time-to-event methods based on deep-learning are often direct extensions of classical models, aiming to learn more flexible, non-linear mappings between event times and covariates. CPH-based deep learning methods [KSC+16] have in some settings demonstrated improvements in C-Index relative to classical approaches. Parametric extensions include the Deep Regularized Accelerated Failure Time (DRAFT) model [CTL+18], Deep Survival Analysis (DSA) [RPEB16], and the Survival Continuous Ranked Probability Score (S-CRPS) model [ADZ+20]. Nonparametric extensions include Deep Adversarial Time-to-Event (DATE) [CTL+18], nonparametric DSA [MPER18] and Gaussian-process-based models [FRT16, AvdS17, LZAvdS19]. As an alternative to a strict time-to-event formulation, some approaches discretize event times and specify models that predict the probability of survival at discrete intervals, including Multi-Task Logistic Regression (MTLR) [YGLB11], neural-MTRL [Fot18], and DeepHit which accounts for competing risks [LZYvdS18].

Methods that produce uncertainty-aware predictions aim to estimate time-to-event distributions, rather than point estimates. Most approaches, parametric or not, result in either a parametric time-to-event distribution, e.g., log-normal in AFT, DRAFT and S-CRPS and Weibull in DSA, or samples from an implicitly defined distribution, e.g., DATE and nonparametric DSA. The latter uses normalizing flows. Importantly, uncertainty-aware predictions are only useful if the time-to-event distributions are concentrated, i.e., their probability masses have coverage much smaller

than the observed time range. This is key, because only in that case can uncertainty be leveraged effectively for ranking or prioritizing events or subjects. However, only a few approaches have considered the uncertainty of the predictions when assessing performance: [CTL+18] via distribution coverage and [ADZ+20] via coefficient-of-variation metrics.

Calibration, a descriptor of a predictive model that characterizes the statistical consistency of the predictions relative to the distribution of the observations on a population level, has been studied in forecasting [DF83], Bayesian analysis [Daw82], and in machine learning, for classification [GPSW17] and regression [KFE18] problems. Unfortunately, it is relatively under-explored in time-to-event models. Exceptions include [VLR17, ZZ18, LZAvdS19] that use (time horizon) thresholded time-to-event Brier scores to asses calibration [Bri50], [ADZ+20] that uses calibration slope as a way to compare model performance, and D-calibration [HHDG20] which accounts for calibration in only non-censored observations. Note that although Brier scores are often used to assess calibration, most commonly in classification models, summaries of calibration curves such as the calibration slope are usually considered more informative [SVC+10].

We present an approach that implicitly defines time-to-event distributions conditioned on covariates via a neural network specification, from which we can synthesize temporally accurate, concentrated and calibrated time-to-event distributions. To this end, *i*) we present a parametric time-to-event AFT model that serves as a baseline in our experiments. *ii*) We present an adversarial formulation for nonparametric time-to-event modeling. Note that early work on *i*) and *ii*) has been previously described in [CTL+18]. *iii*) We introduce a reinterpretation of the Kaplan-Meier estimator for survival functions, which we extend to estimate survival functions conditional on covariates. *iv*) We introduce an approach to directly match the conditional survival function of the model to that of the ground truth, without the need of adversar-

ial techniques [GPAM+14], and demonstrate that the proposed survival function matching approach is related to minimizing the earth mover's distance. Finally, v) we present extensive quantitative and qualitative results, showing that our survival function matching outperforms existing time-to-event models in terms of calibration, while being competitive in terms of C-Index and concentration (sharpness) of the predicted time-to-event distributions.

3.2 Survival Function Matching (SFM)

We extend DATE by replacing the adversarial loss with a calibration regularizer derived from the KM estimator. Below we construct the calibration objective by generalizing the KM estimator to (estimated) distributions of event times.

3.2.1 Distribution-Based Kaplan-Meier Estimator

The standard KM estimator is a population statistic that approximates the marginal survival distribution $S(t)$. Consequently, KM does not explicitly accommodate use of individualized (subject-level) conditional survival functions. Considering that time-to-event methods are primarily tasked with individualized predictions of conditional time densities, $f(t|\boldsymbol{x})$, which can be used to obtain conditional survival functions $S(t|\boldsymbol{x})$, below we present a modified KM estimator that accounts for individualized time-to-event predictions.

We first consider a KM estimator for point estimates of $S(t|\boldsymbol{x})$, directly formulated from the standard KM in (1.7). It is then extended to probabilistic, distribution estimates of $S(t|\boldsymbol{x})$. The point-estimate-based KM, denoted PKM, estimates the population survival function accounting for covariates using predictions $\hat{T}_n \sim g(\boldsymbol{x}_n)$, where $\hat{T}_n$ represents an observed (ground truth) time-to-event from $p(t)$ and $g(\boldsymbol{x}_n)$ is some predictive function, or a summary from a probabilistic estimate of the conditional density $f(t|\boldsymbol{x}_n)$, e.g., $\hat{T}_n \sim g(q(t|\boldsymbol{x}_n))$, where $g(\cdot) = \text{mean}(\cdot)$ and $q(t|\boldsymbol{x}_n)$ is the

approximated conditional time-to-event distributions learned from dataset $\mathcal{D}$. We then write

$$\hat{S}_{\text{PKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} \mathbb{I}(t_{i-1} \leqslant \hat{T}_n < t_i)}{N - \sum_{n=1}^{N} \mathbb{I}(\hat{T}_n < t_{i-1})}\right) \times \hat{S}_{\text{PKM}}(t_{i-1}) \,, \tag{3.1}$$

where $\hat{S}_{\text{PKM}}(t_0) = 1$, $\mathbb{I}(a)$ is an indicator function such that $\mathbb{I}(a) = 1$ if a holds or $\mathbb{I}(a) = 0$ otherwise. It follows from (3.1) that $\hat{S}_{\text{PKM}}(t_i) = \hat{S}_{\text{KM}}(t_i)$.

To account for predictive uncertainly, i.e., for probabilistic estimates $q(t|\boldsymbol{x}_n)$, we extend (3.1) to distribution-based Kaplan-Meier (DKM) estimator. Specifically, we write

$$\hat{S}_{\text{DKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} F_n(t_i|\boldsymbol{x}_n) - F_n(t_{i-1}|\boldsymbol{x}_n)}{N - \sum_{j=1} F_n(t_{i-1}|\boldsymbol{x}_n)}\right) \times \hat{S}_{\text{DKM}}(t_{i-1}) \,, \tag{3.2}$$

where $F_n(t_i|\boldsymbol{x}_n)$ is the estimated cumulative density function for subject n conditioned on covariates $\boldsymbol{x}_n$ and evaluated at t_i. Note that $\hat{S}_{\text{DKM}}(t_i) = \mathbb{E}_{q(t|\boldsymbol{x}_1)\ldots q(t|\boldsymbol{x}_N)}[\hat{S}_{\text{PKM}}(t_i)]$, which follows from statistical independence, so (3.2) averages over (samples of) $q(t|\boldsymbol{x}_n)$ rather than being evaluated on summaries (e.g., averages) of $q(t|\boldsymbol{x}_n)$ as in (3.1). For probabilistic estimates $q(t|\boldsymbol{x}_n)$ of $f(t|\boldsymbol{x}_n)$, the estimator in (3.2) is attractive because it accounts for the predictive uncertainty of the model, thus on a population level, it comprehensively captures the uncertainty of the estimated conditional survival distribution.

3.2.2 Calibration in Time-to-Event Models

For evaluating calibration in time-to-event models, we propose the KM estimator as it is a consistent estimator [PJ77] of $S(t)$, which is also asymptotically D-calibrated [HHDG20]. These desirable KM estimator properties cannot be guaranteed with

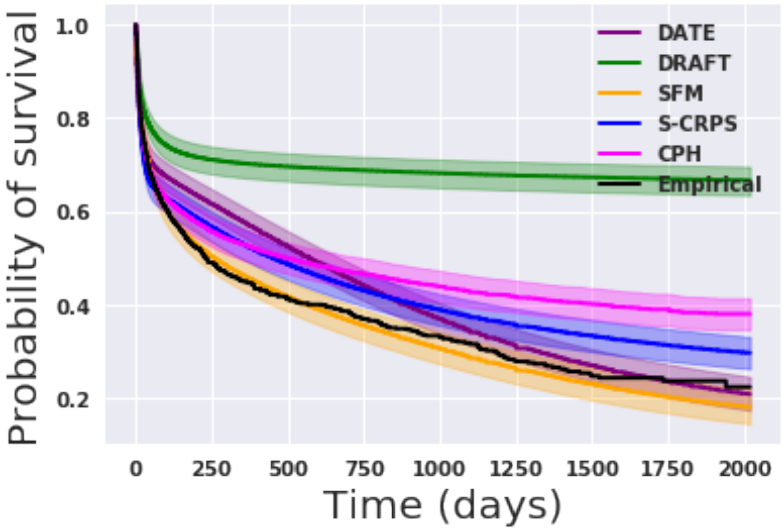


FIGURE 3.1: Survival function estimates for SUPPORT data. Survival functions estimated for ground truth (Empirical) and four models (DATE, DRAFT, SFM and S-CRPS), estimated via $\hat{S}_{\text{KM}}(t)$(1.7) and $\hat{S}_{\text{DKM}}(t)$ (3.2), respectively. Error bars (shaded regions) are calculated according to the Greenwood's formula [G$^+$26].

other discretizing (binning) based calibration methods. Figure 3.1 shows estimated survival distributions on the test set of the SUPPORT dataset (see Section 3.5 for details) for five different models (DATE, DRAFT, SFM, CPH and S-CRPS) using DKM in (3.2), as well as the ground truth (Empirical) using KM in (1.7). Error bars (shaded regions) are calculated using the exponential Greenwood's formula [HLM11].

From Figure 3.1, we see that DKM in (3.2) can be used to visually assess the calibration of estimated event times from different models relative to the ground truth. Specifically, we see that one of the models, SFM (described below) matches the ground truth (Empirical) substantially better than the alternatives (see Section 3.5 for details). Strikingly, the other three models underestimate survival almost everywhere. In the experiments, we use KM and DKM to more directly visualize calibration, and summarize it in terms of calibration slope. Further, below we leverage DKM to encourage calibration during model training, i.e., that DKM for a given model that approximates $q(t, \boldsymbol{x}_n)$ matches as well as possible the true survival distribution estimated via KM.

We propose a nonparametric model for survival-function matching. Specifically, we approximate draws from the density $f(t|\boldsymbol{x})$ via deterministic function $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$, which we specify as a neural network parameterized by $\boldsymbol{\theta}$ and where $\boldsymbol{\epsilon}$ is a source of stochasticity, distributed according to some simple distribution, e.g., uniform or Gaussian. In this manner, we do not impose or assume an explicit form on $q(t|\boldsymbol{x})$, we only seek to efficiently synthesize the draw of samples from it. This specification is similar to DATE, from Chapter 2, but below we will show that we can circumvent the need for an adversarial objective while accounting for calibration.

3.2.3 Calibration objective

Assume as above that $\mathcal{T}$ is the set of distinct and ordered observed event times (censored or non-censored). To estimate the parameters of the model $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$ that generates time-to-event samples on a population level, we match synthesized samples to the empirical survival function, $S(t)$, thus producing calibrated predictions. We propose optimizing the following objective

$$\ell_{\text{cal}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \left\| \hat{S}^{p(t)}_{\text{PKM}}(t_i) - \hat{S}^{G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})}_{\text{PKM}}(t_i) \right\|_1 , \tag{3.3}$$

where $|\mathcal{T}|$ is the cardinality of $\mathcal{T}$, and $\hat{S}^{p(t)}_{\text{PKM}}(t_i)$ and $\hat{S}^{G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})}_{\text{PKM}}(t_i)$ are obtained from $p(t)$ and samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$, respectively. This is connected to KM, because $\hat{S}^{p(t)}_{\text{PKM}}(t_i)$ are obtained from $p(t)$.

The objective in (3.3) seeks to obtain model parameters, $\boldsymbol{\theta}$, for which model and empirical survival functions match. Note that the objective accounts for both censoring and non-censored events. Provided that the conditional survival distribution $S(t|\boldsymbol{x}) = P(\tau > t|\boldsymbol{x})$ for $\tau \geqslant 0$ is the complement of conditional cumulative density function $F(t|\boldsymbol{x})$, matching the conditional survival function also matches the conditional $f(t|\boldsymbol{x})$, i.e., the time-to-event distribution.

Learning with (3.3) is challenging because $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathcal{D})$ is a discrete function, and

thus backpropagation is difficult. Several techniques have been developed to efficiently obtain unbiased and low-variance gradients for backpropagation with discrete objectives or sampling distributions, thus alleviating some of its challenges. Such techniques include REINFORCE [Wil92], reparameterization tricks [RMW14, KW14], and more recently RELAX [GCW+18], a technique that combines REINFORCE and reparameterization tricks via a variance-reduction neural network.

To circumvent the challenges of optimizing over the discrete function in (3.1) and to favor simplicity, we instead optimize over its expectation in (3.2), which is continuous. However, replacing (3.1) with (3.2) is not only inefficient, as it requires generating multiple samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$, but also challenging because $F(t|\boldsymbol{x})$, the conditional cumulative function for $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$ is not available in closed-form. Conveniently, we can replace the indicator functions $\mathbb{I}(a)$ in (3.1) with Heaviside step functions, $H(b) = \frac{1}{2}(\operatorname{sign}(b) + 1)$, therefore obtaining a differentiable formulation:

$$\hat{S}_{\mathrm{PKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} H(\hat{T}_n - t_{i-1}) - H(\hat{T}_n - t_i)}{N - \sum_{n=1}^{N} H(t_{i-1} - \hat{T}_n)}\right) \times \hat{S}_{\mathrm{PKM}}(t_{i-1})\,, \tag{3.4}$$

where, the derivative of signum function $\frac{d\,\operatorname{sign}(b)}{db} = 2\delta(b)$ and $\delta(\cdot)$ is the Dirac delta function. When evaluating the objective, $\ell_{\mathrm{cal}}(\boldsymbol{\theta}; \mathcal{D})$ in (3.3), $\hat{T}_n$ is either a sample from the model, $\hat{T}_n = G_{\boldsymbol{\theta}}(\boldsymbol{x}_n, \boldsymbol{\epsilon})$, or an observed time $\hat{T}_n \sim p(t)$, for $\hat{S}_{\mathrm{PKM}}^{G_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\epsilon})}(t_i)$ or $\hat{S}_{\mathrm{PKM}}^{p(t)}(t_i)$, respectively.

Accuracy objective

The objective $\ell_{\mathrm{cal}}(\boldsymbol{\theta}; \mathcal{D})$ in (3.3) optimizes over a population estimate that encourages calibration. However, calibration alone does not result in time-to-event samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\epsilon})$ that are accurate or concentrated wrt the ground truth at the individual level. This happens because, for a given problem, there exist many so-

lutions that yield well-calibrated predictions that are not necessarily accurate, thus not practically useful. For instance, take the extreme case for which a model learns to estimate $p(t)$ independent of (ignoring) the covariates, $\boldsymbol{x}$, thus effectively recovering the KM estimator in (1.7). So motivated, we also specify accuracy-enforcing objective functions for censored and non-censored observations by borrowing from the DATE formulation. Specifically, we split dataset $\mathcal{D}$ into two disjoint sets $\mathcal{D}_c$ and $\mathcal{D}_{nc}$, for censored and non-censored observations, respectively, and let $(t, \boldsymbol{x}) \sim p_c$ and $(t, \boldsymbol{x}) \sim p_{nc}$ represent, respectively, empirical distributions for these sets. We write objective functions for $\mathcal{D}_c$ and $\mathcal{D}_{nc}$ as

$$\ell_{\text{acc}}(\boldsymbol{\theta}; \mathcal{D}_c, \mathcal{D}_{nc}) = \ell_2(\boldsymbol{\theta}; \mathcal{D}_c) + \ell_3(\boldsymbol{\theta}; \mathcal{D}_{nc}) \,, \tag{3.5}$$

where the first term, taken from (2.7), encourages that time-to-event samples from the model, evaluated on censored observations, are larger than the censoring time. The second term, or absolute error taken from (2.8), encourages time-to-event samples to be accurate, i.e., as close as possible to the ground truth, for non-censored (observed) observations.

3.2.4 Consolidated objective

The complete objective function for the proposed Survival Function Matching (SFM) model is $\ell(\boldsymbol{\theta}; \mathcal{D}) = \ell_{\text{cal}}(\boldsymbol{\theta}; \mathcal{D}) + \lambda \ell_{\text{acc}}(\boldsymbol{\theta}; \mathcal{D}_c, \mathcal{D}_{nc})$, where $\lambda > 0$ is a free parameter controlling the trade-off between the accuracy objective and the survival function matching objective in (3.3). In the experiments we let $\lambda = 1$, however, λ can be optimized by grid search if desired. Refer to the ablation study in Appendix B.1.2 for the effects of calibration (3.3), accuracy (3.5), and consolidated (observed events only) objectives. The complete objective is optimized using stochastic gradient descent on minibatches from $\mathcal{D}$. Note that $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathcal{D})$ is a population-level objective that may be affected by the minibatch size. However, empirically we did not observe substantial differences in the performance metrics when varying the minibatch

size (see Appendix B.1.1). We attribute the model being insensitive to minibatch size to the insight that learning with minibatches can be understood as encouraging the model to be calibrated for every minibatch, thus consequently also encouraging global calibration.

3.2.5 Theoretical Motivation

Optimal mass transport approaches for distribution matching in machine learning tasks have received considerable attention recently [CZZ+19, SZRM18]. For one-dimensional problems, it has been shown that the characterization of the p-Wasserstein metric has a simple form [KPT+17] $W_p(P,Q) = (\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p dz)^{1/p}$ where, $F(z)^{-1}$ and $G(z)^{-1}$, for $z \in (0,1)$, are the quantile functions of $p(t)$ and $q(t)$, respectively, and $F(t)$ and $G(t)$, their corresponding cumulative density functions. Interestingly for $p = 1$, $W_p(P,Q)$ is also known as the *Monge-Rubenstein metric* [Vil08] or the *earth mover's distance* [RTG00], and it is essentially the absolute difference between the quantile functions for $p(t)$ and $q(t)$. By contrast, the SFM objective in (3.3) is the absolute difference between the cumulative density functions for $p(t)$ and $q(t)$, provided that $F(t) = 1 - S(t)$. As a result, minimizing (3.3) and $W_p(P,Q)$ are closely related approaches to matching $p(t)$ and $q(t)$. However, the latter explicitly imposes survival-distribution matching, which we consider more appropriate considering the goal is to obtain calibrated predictions in the context of time-to-event modeling.

3.3 Related Work

Calibration, also known as reliability [Daw82] in Bayesian analysis, is a statistical concept that refers to the consistency (in distribution) of estimates relative to (population-wide) observed measurements, e.g., event times. A model is considered well-calibrated if predicted population probabilities are statistically con-

sistent with ground truth. Ideally, a well-calibrated binary classification model $U(\cdot) : \mathcal{X} \in \mathbb{R}^d \rightarrow \mathcal{Y} \in \{0, 1\}$, given covariates $\boldsymbol{x}_n \in \mathcal{X}$ and binary labels $y_n \in \mathcal{Y}$ is one for which $P(Y = 1|U(X) = p) = p$, $\forall p \in [0, 1]$. Further, a regression model $U(\cdot) : \mathcal{X} \in \mathbb{R}^d \rightarrow \mathcal{Y} \in \mathbb{R}$ is considered well-calibrated if $N^{-1} \sum_{n=1}^{N} \mathbb{I}\{y_n \leqslant F_n^- 1(p)\} \rightarrow p$, $\forall p \in [0, 1]$ as $N \rightarrow \infty$, s.t. $F_n^{-1}(p) : [0, 1] \in \mathcal{Y}$, where $F_n(y)$ is the cumulative density function and $F_n^{-1}(p)$ is the quantile function [KFE18].

In the context of time-to-event modeling, calibration refers to the concept of obtaining a predictor of event times (that may or may not be probabilistic) whose predictions match, on a population level, the survival distribution $S(t)$. A well calibrated model will not only accurately estimate relative times between subjects, but also absolute time of events. In Section 3.2, we present a framework for evaluating calibration in time-to-event models. The proposed framework leverages the desirable properties of the KM estimator described in Section 1.2.4.

Existing calibration literature in predictive models has primarily focused on recalibration techniques for predictions from classification [GPSW17] or regression models [KFE18]. For classification tasks, the Brier score [Bri50] is a commonly used proper score metric, quantifying the accuracy of probabilistic predictions, and thus it is often used to assess calibration. The Brier score has also been used to asses calibration in time-to-event models [VLR17, ZZ18, LZAvdS19], however, this score has to be evaluated at pre-specified (thresholded) time horizons. Alternatively, S-CRPS [ADZ$^+$20] considers the integral of the Brier score evaluated at all possible thresholds [GR07], which is a more principled and comprehensive approach than calibration at pre-specifying time horizon thresholds.

The approach presented here is inspired by [ADZ$^+$20]. They considered calibration slope as a metric for evaluating performance in time-to-event models. However, our formulation is very different from that of [ADZ$^+$20], in the sense that they encourage calibration by optimizing a proper score rule, the Continuous Ranked

Probability Score (CRPS), whereas we tackle it directly as a survival-distribution-matching problem. In the experiments in Section 3.5, we show empirically that our more direct approach to calibration consistently outperforms CRPS. Interestingly, excluding approaches that address thresholded calibration with Brier-scores [VLR17, ZZ18, LZAvdS19], only S-CRPS [ADZ+20] considers global calibration as a performance metric. All the others focus on accuracy-centric performance estimates, e.g., C-Index and relative absolute error.

3.4 Extensions

3.4.1 Interpreting Time-to-Event Models using Attention

Inspired by attention based models used in image [XBK+15] and natural language processing [RCW15] applications, we introduce an attention module to estimate the relative importance of covariates at the observation level. Unlike traditional risk-based models, e.g., ℓ_1-regularized generalized regression models (including CPH) [FHT10, BFJ11], that select a fixed subset of covariates from high-dimensional data for the entire population, our attention module captures individual covariate predictive ability, allowing for a precision-medicine-centric [CV15] approach to risk prediction in time-to-event models.

The covariates attention (importance) vector, $\boldsymbol{a}_i$, for observation $\boldsymbol{x}_i$ is obtained from the covariates themselves as

$$\boldsymbol{a}_i = \text{softmax}(\sigma(\mathbf{W}_0 \boldsymbol{x}_i + \boldsymbol{b}_0)) \,, \tag{3.6}$$

where $\mathbf{W}_0 \in \mathbb{R}^{p \times p}$ is the weight matrix, $\boldsymbol{b_0} \in R^p$ is the bias term, and $\sigma(\cdot)$ is the sigmoid function. This approach, where sigmoid and softmax links are used jointly, has been proposed before [XBK+15] and encourages sparsity of the attention vector.

The attention-transformed data $\boldsymbol{x}'_i$ is obtained as $\boldsymbol{x}'_i = \boldsymbol{x}_i \odot \boldsymbol{a}_i$, where $\odot$ is the Hadamard product. Note, as desired, covariates with near-zero attention values do

not contribute to the latent representation $\boldsymbol{z}$ or the prediction of the time-to-event t. Specifically, we add the attention mechanism to the input of DRAFT for interpretable ranking of global (population) and local (subpopulation) covariate importances; the latter is based on the clustering structure of trained attention vectors. See Section 3.5 for a qualitative analysis on a real-world dataset.

3.4.2 Competing Risks

Conventional survival analysis applications focus on single time-to-event outcomes, e.g., death. However, in practice, events can be caused by different reasons, commonly called competing risks, some of which may be of interest, or otherwise usually coded as censored events. In clinical applications, for example, events may be caused by one or more outcomes, often captured as diagnoses. As a result, there is a need in clinical decision making for predicting cause-specific survival times, while jointly accounting for the risk of multiple likely-to-occur outcomes. Further, this joint approach is advantageous from a modeling perspective, as it allows for cross-outcome information sharing and it informs the model about event type heterogeneity. This approach may likely result in a better characterization of the individual outcomes, compared to building independent models for each outcome or event type.

Below we extend the approach in Section 3.2, SFM, to cause-specific time-to-event modeling. Assume K mutually exclusive outcomes or event types, thus $y_i \in \{0, 1, 2, ..., K\}$, where $y_i = 0$ still indicates a censored event. We can specify an objective to match K survival functions and optimize the accuracy of each outcome individually, using a joint model. This is done by first specifying $G_{\boldsymbol{\theta}_k}(\boldsymbol{z}, \boldsymbol{\epsilon}_k)$, an outcome-specific generator conditioned on a shared latent representation $\boldsymbol{z}$, also specified as a neural network $q(\boldsymbol{z}|\boldsymbol{x}) = G_{\boldsymbol{\theta}_z}(\boldsymbol{x}, \boldsymbol{\epsilon})$ parameterized by $\boldsymbol{\theta}_z$. The complete model is then parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ and optimized using $\ell_{\text{CR}}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{k=1}^{K} \ell(\boldsymbol{\theta}_k; \mathcal{D})$,that implicitly depends on $\boldsymbol{\theta}_z$ via $G_{\theta_z}(\boldsymbol{z}, \boldsymbol{\epsilon})$.

3.5 Experiments

We qualitatively and quantitatively present experimental comparisons considering SFM, DATE, DRAFT, CPH [Cox92], S-CRPS [ADZ+20], MTLR [YGLB11], RSF [IKBL08] and two existing neural extensions of CPH: *i*) MLP-Rank: a minor modification of the DeepSurv model [KSC+16]; a partial-likelihood (CPH) model specified with MLPs, similar to DRAFT. Since the partial-likelihood approach is prohibitive for large datasets, we modify it to use the rank-based objective in (2.4). In this way, DeepSurv can be trained via stochastic gradient descent. *ii*) MLP-Efron: that extends the MLP-Rank model with Efron's approximation [Efr77] for handling tied event times. This model is similar and comparable to [LSC+17], where the MLP is specified as an isotonic regression model.

All neural-network-based models are specified in terms of two-layer MLPs of 50 hidden units with Rectified Linear Unit (ReLU) activation functions, batch normalization and apply dropout of $p = 0.2$ on all layers. We set the minibatch size to $M = 350$ and use the Adam optimizer with the following hyperparameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and $\epsilon = 1 \times 10^{-8}$.We initialize all network weights according to *Xavier*[GB10]. SFM and DATE inject noise in all layers; see Section 2.3.1 for more details. Datasets are split into training, validation and test sets as 80%, 10% and 10% partitions, respectively, stratified by non-censored event proportion. The validation set is used for early stopping and learning model hyperparameters. All models are trained using one NVIDIA P100 GPU with 16GB memory. Source code for the proposed models is available at `https://github.com/paidamoyo/calibration_uncertainty_t2e`.

Table 3.1: Summary statistics of the datasets for the experiments. Time range, $t_{\max}$, is noted in days except for SEER for which time is measured in months.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP
Events (%)	23.9	27.5	68.1	51.0	23.8
N	394,823	7,894	9,105	68,082	5026
d (cat)	729 (106)	26 (21)	59 (31)	789 (771)	206
Missing (%)	1.9	2.1	12.6	23.4	18.2
$t_{\max}$	365	5,215	2,029	120	5,794

3.5.1 Datasets

We consider five diverse datasets: *i*) FLCHAIN: a public dataset investigating non-clonal serum immunoglobin free light chains effects on survival time [DKK$^+$12]. *ii*) SUPPORT: a public dataset for a survival-time study of seriously-ill hospitalized adults [KHL$^+$95]. *iii*) SEER: a public dataset provided by the Surveillance, Epidemiology, and End Results (SEER) Program. We restrict the dataset to a 10-year follow-up breast cancer subcohort with three competing risks (breast cancer, cardiovascular and others). See [RYJK$^+$07] for preprocessing details. *iv*) EHR: a large study from the Duke University Health System centered around multiple inpatient visits due to comorbidities in patients with Type-2 diabetes [CTL$^+$18]. *v*) SLEEP: a subset of the Sleep Heart Health Study (SHHS) [QHI$^+$97], a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing.

Table 3.1 presents summary statistics of the datasets, where d denotes the size of the individual covariate vector $\boldsymbol{x}$ after one-hot encoding for categorical (cat) variables. Events indicates the proportion of the non-censored events, i.e., the events of interest for which $y_i = 1$. Missing indicates the proportion of missing entries in the $N \times d$ covariate matrix, and $t_{\max}$ is the time range for both censored and non-censored events. For all datasets except SEER, that uses months, events are measured in days. We do not convert time to a common scale and model it as is.

Details of the public datasets: FLCHAIN, SUPPORT and SEER, including preprocessing procedures, are provided in the above references. The other two datasets, EHR and SLEEP are not public but can be obtained upon request; see [CTL+18] and [ZCM+18], respectively. For SLEEP we focus on the baseline clinical visit and aggregated demographics, medications and questionnaire data as covariates.

As shown in Table 3.1, survival datasets often contain substantial missingness, e.g. up to 23% in SEER data. Interestingly, [MPER18] showed via the information-theoretic data processing inequality that there is no additional information to be gained by actively imputing missing values during training with an autoencoding arm, when compared to a simpler pre-imputation approach in which missing values are imputed with median and mode for continuous and categorical covariates, respectively. In view of this, here we adopt a pre-imputation strategy.

3.5.2 Comparison with Basic Neural Extensions of CPH and AFT

We first compare DRAFT against basic neural network based methods, CPH and RSF, in terms of C-Index [HJLC+84] and Relative Absolute Error (RAE) [YGLB11]. Results in Table 3.2 show that DRAFT outperforms the other approaches across all datasets, in both metrics. Note that for datasets of the size considered here, CI differences in the order of 10^{-2} are statistically significant. RSF complexity scales with N^2, where N is the size of the dataset, as a result, we were unable to obtain results for EHR, even after running the model for more than 5 days. DMGP [AvdS17] is an AFT-based model similar to ours, but using Gaussian processes instead of MLPs. We were unable to run their model; however, we replicated their preprocessing to be able to compare results directly. DRAFT outperforms DMGP by approximately 0.03 CI units; see [AvdS17] results for comparison.

Table 3.2: C-Index and RAE results on test data. C-Index Differences in the order of 10^{-2} are statistically significant.

	EHR	FLCHAIN	SUPPORT	SEER
C-Index				
DRAFT	**0.78**	**0.83**	**0.86**	**0.83**
CPH	0.75	0.83	0.83	0.82
RSF	—	0.82	0.80	0.82
MLP-Rank	0.71	0.77	0.67	0.77
MLP-Efrons	0.67	0.78	0.58	0.73
RAE				
DRAFT	**0.20**	**0.24**	**0.66**	**0.37**
CPH	0.99	1.0	0.92	0.96
RSF	-	0.95	0.80	0.76
MLP-Rank	0.62	1.0	0.78	0.57
MLP-Efron's	1.0	1.0	0.98	0.99

3.5.3 Comparison with Advanced Nonparametric Models

We now qualitatively and quantitatively compare SFM, DATE, DRAFT, CPH [Cox92], S-CRPS [ADZ+20], and MTLR [YGLB11].

Quantitative evaluation

For a comprehensive quantitative evaluation of time-to-event models, we consider three metrics that highlight different aspects of model performance: *i*) Concordance Index (C-Index) [HJLC+84] to quantify preservation of pairwise orderings wrt ground truth events, *ii*) Coefficient of Variation (CoV) to assess uncertainty concentration by quantifying the dispersion of estimated time-to-event distributions, and *iii*) Calibration to asses the statistical consistency of the conditional survival distribution learned by a model relative to that of the ground truth. As discussed previously, a high-performing model is one that not only preserves pairwise ordering of event times, but also results in concentrated and well-calibrated time-to-event distributions. Ablation study results in Appendix B.1.2 demonstrate that: *i*) accounting for censored data is important in improving both accuracy and calibration; *ii*) consol-

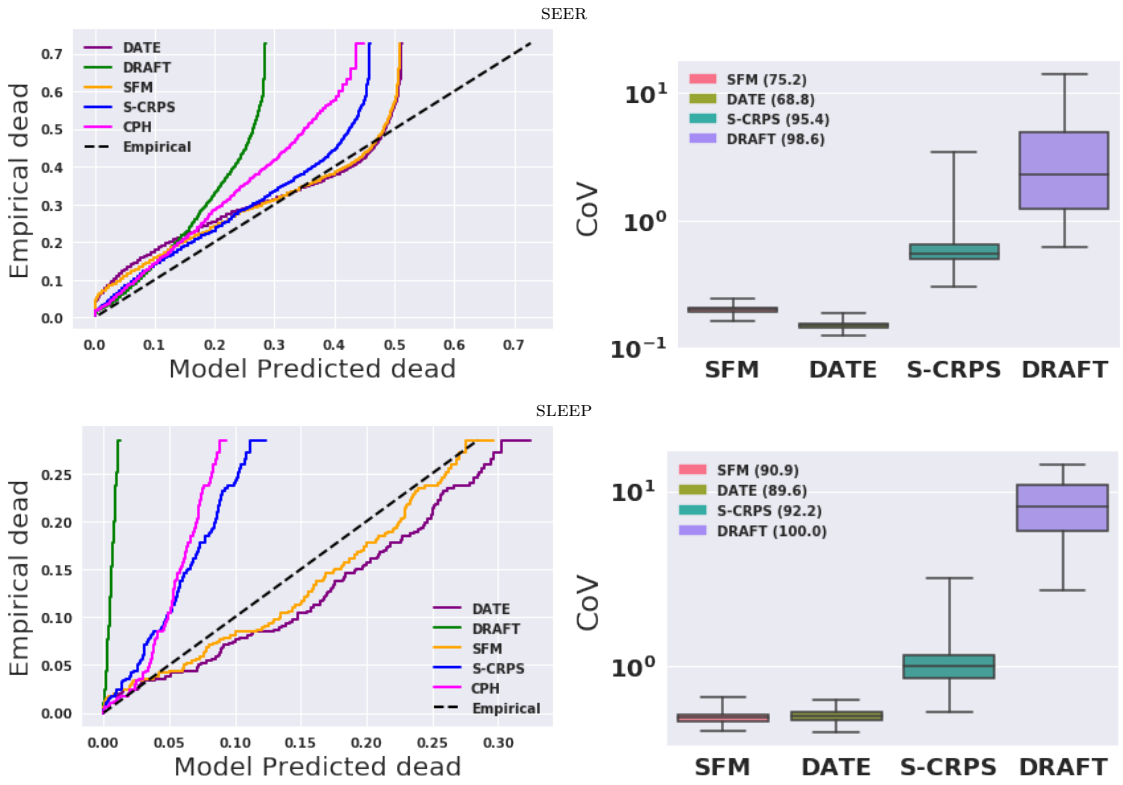


FIGURE 3.2: Test set calibration and variation visualized for two datasets: SEER and SLEEP (rows). Left: proportion of events of interest vs. predicted events (calibration curve). A perfectly calibrated model will follow the (dashed) diagonal line. Right: coefficient of variation (CoV) distributions. The legend shows the percentage of test set events covered by 95% intervals from predicted time-to-event distributions.

idated (accuracy plus calibration) yields better calibrated and accurate predictions than with either accuracy (3.5) or calibration (3.3) only objectives. As discussed below, SFM outperforms other approaches in terms of calibration while being competitive in terms of C-Index (time ordering) and CoV (concentration).

Calibration: We evaluate calibration both visually and quantitatively. For the visual assessment, we plot the conditional survival distributions estimated from the model predictions using DKM in (3.2) and compare it with the empirical survival distribution (ground truth) using KM in (1.7), as shown in Figure 3.1. Alternatively, we plot the estimated conditional cumulative density function for each model using $1 - \hat{S}_{DKM}(t_i)$ against the marginal cumulative density function for the ground truth

Table 3.3: Performance metrics. SFM is the proposed model.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP
Calibration slope					
DATE	0.7537	0.9668	0.9068	0.9161	0.9454
DRAFT	3.2138	5.4183	2.9640	2.0763	25.2855
S-CRPS	1.6246	1.9662	1.1795	1.1613	2.5746
CPH	2.5543	1.9116	1.3909	1.4358	3.8278
MTLR	2.1957	1.9449	1.2017	1.2476	2.4792
SFM	**0.7734**	**0.9807**	**0.9405**	**0.9540**	**1.0235**
Mean CoV					
DATE	**0.2477**	**0.3585**	**0.2987**	**0.1485**	0.5168
DRAFT	5.0305	6.2952	3.8689	3.4501	8.4918
S-CRPS	0.8585	0.9412	0.7351	0.6036	1.0240
CPH	-	-	-	-	-
MTLR	-	-	-	-	-
SFM	0.2953	0.4484	0.3930	0.1993	**0.5045**
C-Index					
DATE	0.7756	0.8264	0.8421	**0.8320**	0.7416
DRAFT	**0.7796**	0.8341	0.8560	0.8310	**0.7617**
S-CRPS	0.7704	0.8286	**0.8685**	0.8298	0.7529
CPH	0.7542	**0.8344**	0.8389	0.8223	0.6435
MTLR	-	-	-	-	-
SFM	0.7786	0.8318	0.8319	0.8314	0.7491

using $1-\hat{S}_{KM}(t_i)$. In both cases, $t_i \in \mathcal{T}$. If the estimated cumulative density matches the ground truth, the plotted calibration curve will describe a diagonal line with unit slope. See Figure 3.2 below for examples on SEER and SLEEP datasets. Curves above and below the diagonal underestimate and overestimate risk, respectively. Thus, for the quantitative assessment we calculate the calibration slope, which is obtained from the curve described by $1 - \hat{S}_{DKM}(t_i)$ vs. $1 - \hat{S}_{KM}(t_i)$. Since the cumulative density $F(t)$ is unknown for sampling-based approaches, e.g., DATE and SFM, we use a Gaussian Kernel Density Estimator (KDE) [Sil18] on samples from the model, $\{t_{ns}\}_{s=1}^{200}$.

Results in Table 3.3 show that in terms of calibration slope, fully nonparametric models, specifically SFM and DATE, are better calibrated than S-CRPS and DRAFT, both parametrized as log-normally distributed models. SFM is the best

performing model across all datasets, followed by DATE, S-CRPS, MTLR, CPH then DRAFT. We attribute these results to the fact that we directly match the survival function as part of model training. However, it is surprising that DATE and S-CRPS do not perform nearly as well, considering that DATE adversarially matches the time-to-event distribution, thus indirectly matching the cumulative distribution, and S-CRPS optimizes a proper scoring rule (the integral of Brier score at all possible thresholds [GR07]) that in principle should produce calibrated predictions.

For the EHR data it is not surprising that none of the models are well calibrated, because observations in this dataset are not i.i.d., due to patients having multiple encounters. Since the models and KM-based estimators considered implicitly assume datasets are composed of i.i.d. observations, calibration does not necessarily hold. This necessitates further investigation, which we leave as interesting future work. However, to test the hypothesis that the model should be better calibrated in the i.i.d. case, we restricted the EHR dataset to the first encounter per patient (N=19,064), which results in a better calibrated SFM model; see Figure 3.3 for more details.

Concordance Index: C-Index is arguably the most commonly used performance metric in survival analysis. This metric is useful to assess relative risk because it quantifies ordering rather than temporal accuracy. Models with high C-Index are good for the purpose of ranking observations into different risk categories, especially in medical settings. Since the C-Index is evaluated on point estimates, we summarize time-to-event distributions as medians, i.e., $\hat{t} = \text{median}(\{t_{ns}\}_{s=1}^{200})$, where t_{ns} is a sample from the trained model, $t_{ns} \sim G_{\boldsymbol{\theta}}(\boldsymbol{x}_n, \boldsymbol{\epsilon}_s)$, on the test set.

Results in Table 3.3 show that none of the models has a clear advantage over the others, as the C-Index is largely comparable for the remaining four datasets. Apart from the small and high event rate SUPPORT dataset where S-CRPS and DRAFT (both parametric log-normally distributed models) achieve (statistically) significantly higher C-Index compared to SFM (and CPH). While it is possible to compute non-

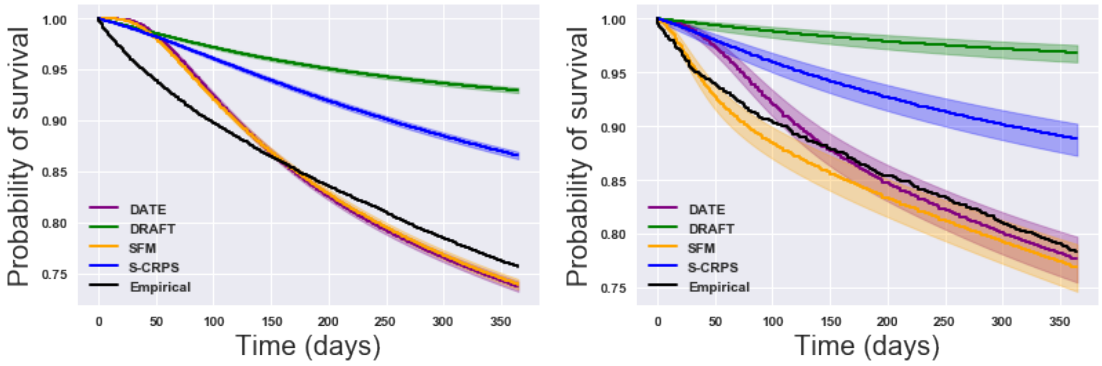


FIGURE 3.3: Estimated survival functions for EHR using all non iid data (left) and a subset of iid observations (right).

global C-Index at pre-specified time horizons, we are unable to compute CoV, as we cannot recover $f(t|\boldsymbol{x}) = S(t|x)h(t|\boldsymbol{x})$ from MTLR which does not specify the conditional hazards, $h(t|\boldsymbol{x})$.

Coefficient of Variation: The Coefficient of Variation (CoV) quantifies the dispersion of a probability distribution. It is defined formally as $\sigma\mu^{-1}$, where σ and μ are respectively the standard deviation and mean of the distribution being tested. To summarize the variation of the time-to-event distributions estimated by different models on the test set, we use Mean CoV, which is defined across all time-to-event predictions, i.e., $N_{\text{te}}^{-1}\sum_{n=1}^{N_{\text{te}}}\sigma_n\mu_n^{-1}$, where N_{te} is the size of the test set and σ_i and μ_i are sample standard deviations and means over $\{t_{ns}\}_{s=1}^{200}$. A model with concentrated time-to-event distributions is one for which mean CoV is as small as possible.

Figure 3.2 shows test set CoV distributions. We see that *i*) DRAFT and S-CRPS have considerably wider variation in CoV, thus better 95% posterior coverage (see legend) compared to SFM and DATE; and *ii*) SFM and DATE are comparable, though DATE is slightly better. Note that we cannot evaluate CoV or coverage for CPH and MTLR since in their standard form they only produce point estimates.

Table 3.3 shows that across all datasets DATE, SFM and S-CRPS are on average low-variance models while DRAFT is a considerably higher-variance model. DATE

Table 3.4: DATE model with different choices of noise sources with varying layer. Median of 95% intervals for all test-set time-to-event distributions on SUPPORT data. Ranges in parentheses are 50% empirical quantiles.

	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
Non-censored			
All	$60.0_{(3.9,176.5)}$	$149.9_{(8.5,926.8)}$	$37.9_{(3.5,237.4)}$
Input	$28.9_{(1.8,114.8)}$	$22.4_{(1.5,91.2)}$	$33.7_{(1.6,127.6)}$
Output	-	$168.8_{(16.6,844.3)}$	-
Censored			
All	$231.3_{(177.2,332.1)}$	$1397.3_{(990.9,2000.1)}$	$350.5_{(254.4,539.3)}$
Input	$137.3_{(99.4,205.0)}$	$86.9_{(64.4,135.0)}$	$155.8_{(106.7,229.3)}$
Output	-	$1158.6_{(873.8,1670.4)}$	-

and SFM are the best-performing in terms predicting concentrated event times given that mean CoV < 0.5. High-variance time-to-event distributions are not desirable because when prediction uncertainty is large relative to the time range, they cannot be used to inform decision making.

Distribution coverage: We now demonstrate that the DATE model, with noise sources on all layers, has time-to-event distributions with larger variances than versions of DATE with noise only on the input of the neural network. Table 3.4 shows the median of the 95% intervals for all test-set time-to-event distributions on SUPPORT data. DATE with Uniform(0,1) has larger variance and coverage compared to the other alternatives, while keeping relative absolute errors and CIs largely unchanged. We did not run models with Uniform(-1,1) and Gaussian(0,1) only on the output layer, because from the other results presented above it is clear that these two options are not nearly as good as having Uniform(0,1) noise on all layers. Note also that we did not include DRAFT or SFM in these comparisons. DRAFT has naturally good coverage due to the variance of the time-to-event distributions being modeled independent for each observation as a function of the covariates (see for instance Figure 3.4). However, DRAFT has difficulties keeping good coverage while maintaining good performance, i.e., small absolute relative error. SFM leverages a

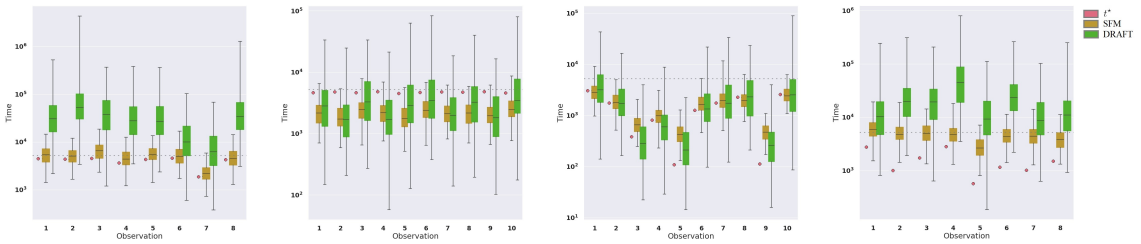


FIGURE 3.4: Example test-set predictions on FLCHAIN data. Top best (left) and worst (middle-left) predictions on censored events, and top best (middle-right) and worst (right) predictions on non-censored events. Circles denote ground-truth events or censoring points, while box-plots represent distributions over 200 samples for both SFM and DRAFT. The horizontal dashed line represents the range ($t_{\max} = 5,215$ days) of the events.

similar DATE architecture of injecting Uniform(0,1) noise on all layers thus results are comparable.

Qualitative analysis of predicted event times

We visually compare the test-set time-to-event distributions by SFM and DRAFT on FLCHAIN data. In Figure 3.4 we show the top best (left) and worst (middle-left) predictions on censored events, and the top best (middle-right) and worst (right) predictions on non-censored events. Circles denote ground-truth events or censoring points, while box-plots represent distributions over 200 samples for both SFM and DRAFT models. We see that: (*i*) in nearly every case, SFM is more accurate than DRAFT. (*ii*) DRAFT tends to make predictions outside the event range ($t_{\max} = 5,215$ days), denoted as a horizontal dashed line. (*iii*) DRAFT tends to overestimate the variance of its predictions, approximately by one order of magnitude relative to SFM. This is not very surprising as DRAFT has an MLP dedicated to estimate, conditioned on the covariates, the variance of the time-to-event distribution. However, note that variances estimated well over the domain of the events ($t_{\max}$) are not necessarily meaningful or desirable.

To provide additional insight into the performance of DATE compared to DRAFT, we report the Normalized Relative Error (NRE) defined as $(\hat{t} - t)/t_{\max}$ and $\min(0, \hat{t} -$

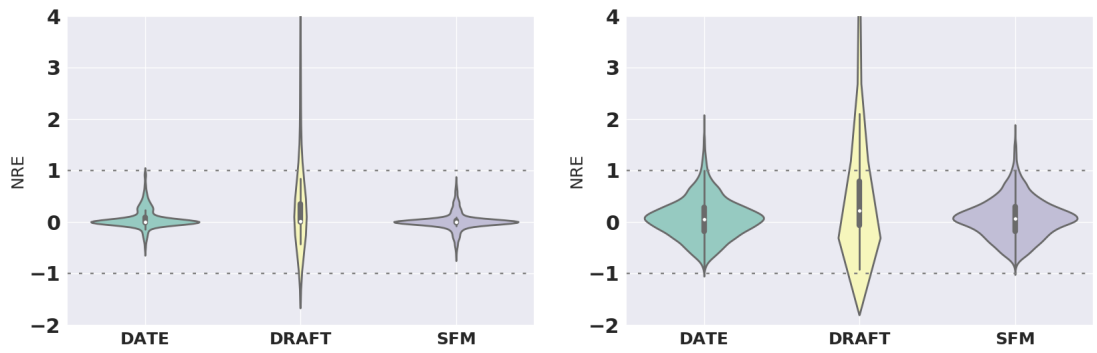
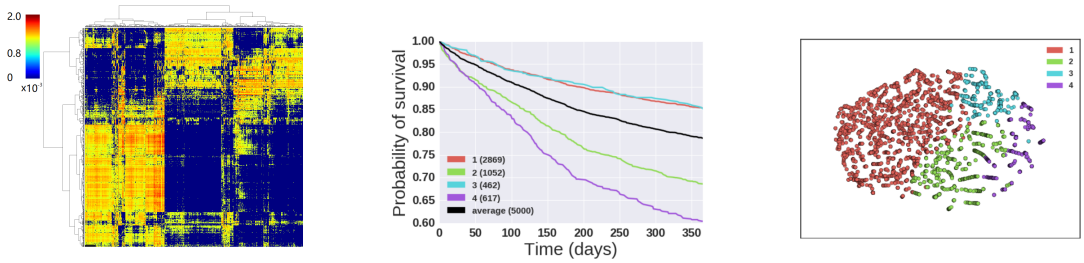


FIGURE 3.5: Normalized Relative Error (NRE) distribution for SUPPORT (left) and EHR (right), test-set non-censored events. The horizontal dashed lines represent the range of the events, $t_{\max} = 120$ months and $t_{\max} = 365$ days, respectively.

$t)/t_{\max}$ for non-censored and censored events, respectively, where t, $\hat{t}$ and $t_{\max}$ denote respectively the ground-truth time-to-event, median time estimated (from samples) and event range, as indicated in Table 3.1. The NRE distribution provides a visual representation of the extent of test-set errors, while revealing whether the models are biased toward either overestimating ($\hat{t} > t^*$) or underestimating ($t^* > \hat{t}$) the event times. Although models with unbiased NREs are naturally preferred, in most clinical applications, where being conservative is important, overestimated time-to-events must be avoided as much as possible. Figure 3.5 shows NRE distributions for test-set non-censored events on SUPPORT and EHR data. We see that DRAFT results in considerable errors beyond the event range ($|\text{NRE}| > 1$), $t_{\max} = 120$ months or $t_{\max} = 365$ days for SEER and EHR, respectively. Further, we see that the NRE distribution for DRAFT is heavily skewed toward NRE > 1, thus tending to overestimate event times. On the other hand, SFM and DATE produce errors substantially more concentrated around 0 and within $|\text{NRE}| < 1$, relative to DRAFT. This demonstrates the advantage of the distribution matching methods, SFM and DATE, over the likelihood-based method DRAFT in generating realistic samples. Similar results were observed on the other datasets for both censored and non-censored events.



(a) Attention cluster map. (b) Kaplan-Meier curves. (c) Latent features *t*-SNE.

FIGURE 3.6: Attention results on 5,000 randomly selected patients from the EHR dataset. Hierarchical clustering on the attention matrix (a) reveals 4 tightly grouped clusters (c) with different survival profiles (b).

Qualitative evaluation of calibration

There are several metrics for measuring the quality of calibration, e.g., calibration slope and Brier score [Mur73]. However, none of these summaries of calibration are as richly informative as visually comparing survival functions or cumulative density functions as described above. In Figure 3.2 we show calibration curves for two different datasets, SEER and SLEEP, the largest and smallest dataset, respectively. From these results (consistent across all datasets) we see that *i*) SFM performs better than the other approaches considered; *ii*) DRAFT is the worst performer; and *iii*) all approaches are poorly calibrated on SEER data once half of the population has had events.

Under further examination of the SEER data, we found there is a large subset of the population that gets administratively censored at $t = 80$ months, which explains the generalized sudden divergence of calibration in Figure 3.2. This type of informative censoring is not random and needs to be modeled appropriately. However, this extension is beyond the current scope and thus left as future work. Nonetheless, to test this idea, we truncated the data beyond $t = 88$ months and verified that the model is considerably better calibrated (results not shown).

Table 3.5: Top 5 population level (All) and cluster-specific (Figure 3.6) covariates for the EHR dataset.

Cluster	Top 5 covariates
1	Hypertension, cardiac arrhythmia, stroke, chronic pulmonary disease (count); obesity.
2	Number of systolic blood pressure measurement, hypertension; chronic kidney disease; diabetes mellitus with complications, chemistry and hematology.
3	Number of hbA1c measurements, acute cerebrovascular disease, pulmonary heart disease, secondary malignancies, other respiratory therapy.
4	Past year ACE Inhibitor, max hbA1c, depression, pneumonia, mean HDL.
All	Cardiac arrhythmia, hypertension, coronary artery disease, renal disease, chronic kidney disease.

3.5.4 Interpretable time-to-event using attention

To demonstrate the interpretation capabilities of the proposed attention mechanism, below we describe how to leverage the estimated importance vectors in (3.6) to get a better understanding of *i*) the population characteristics of the data being analyzed, and *ii*) global and local subsets of predictive covariates. For the former, using a randomly selected subset of 5,000 patients from the EHR dataset, we performed two-way hierarchical clustering on the 729-dimensional importance vectors estimated using (3.6); after the model has been trained. The resulting cluster map is shown in Figure 3.6(a), where warmer colors indicate higher covariate importance, thus associated with predictive ability. For better visualization, we set to zero (dark blue) importance values lower than the probability of selecting a covariance uniformly at random, i.e., $1/p \approx 0.0014$. The block structure of the clustered importance vectors in Figure 3.6(a) reveals well-defined subsets of patients that share similar important covariates. In fact, these subsets have unique risk profiles, as evidenced by their Kaplan-Meier curves [KM58], shown in Figure 3.6(b). Further, a t-SNE embedding

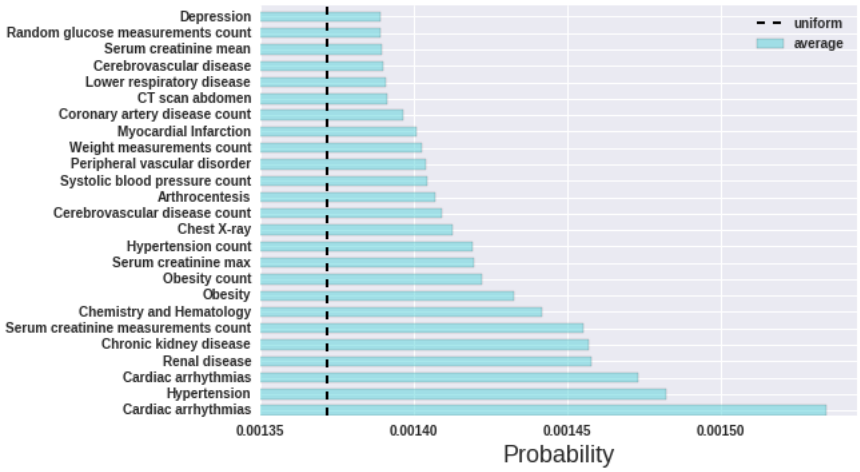


FIGURE 3.7: Top 25 covariates for the EHR dataset ordered by population-level importance.

of the latent features, $\boldsymbol{z}$, in Figure 3.6(c), is consistent with the attention-derived clusters in 3.6(b).

On a global (population) scale, Figure 3.7 shows importance values averaged over the entire test set, for the top 25 most predictive covariates. Not surprisingly, we see that existing cardiovascular complications (cardiac arrhythmia, hypertension and coronary artery disease, CAD) and renal complications (renal disease and chronic kidney disease, CKD) make it to the top of the list. On a local (sub-population) scale, Table 3.5 shows the top 5 most predictive covariates for the patients in the 4 clusters identified by hierarchical clustering and shown in Figure 3.6.

3.5.5 Competing Risks

Quantitative and qualitative results of SEER competing risks are shown in Table 3.6. These demonstrate the importance of estimating cause-specific time-to-event distributions, i.e., time-to-event predictions generated jointly by the SFM-CR model for CANCER and CVD events outperform the SFM model in Table 3.3 in terms of C-Index and calibration slope.

Table 3.6: SEER competing risks quantitative results.

	Events (%)	Calibration Slope	Mean CoV	C-Index
CANCER	26.7	1.0182	0.2129	0.8806
CVD	7.66	1.0019	0.2272	0.8378
OTHER	16.6	0.9373	0.2193	0.8107

3.6 Conclusions

We have presented an adversarially-learned time-to-event model, that leverages a distribution-form-free loss function for censored events. The proposed approach extends GAN models to time-to-event modeling with censored data, and it is based on deep neural networks with stochastic layers. The model yields improved uncertainty estimation relative to the AFT-based parametric alternative. Additionally, we introduced a distribution-based Kaplan-Meier (DKM) estimator for evaluating calibration in time-to-event predictions. Leveraging this estimator, we have replaced the adversarial objective in DATE with SFM, a survival-function-matched neural-network-based regularizer for synthesizing calibrated time-to-event predictions. It was demonstrated that our survival-distribution-matching approach is related to earth mover's minimization. Extensive experiments demonstrated that SFM outperforms other methods in estimating concentrated and calibrated time-to-event distributions, while remaining competitive in terms of concordance index. As future work, we plan to extend SFM to calibration in the non-i.i.d. setting, and to account for informative missingness.

4

Enabling Counterfactual Survival Analysis with Balanced Representations

4.1 Introduction

Survival analysis or time-to-event studies focus on modeling the time of a future event, such as death or failure, and investigate its relationship with covariates or predictors of interest. Specifically, we may be interested in the *causal effect* of a given intervention or treatment on survival time. A typical question may be: will a given therapy increase the chances of survival of an individual or population? Such causal inquiries on survival outcomes are common in the fields of epidemiology and medicine [Rob86, HKH$^+$96, YBD$^+$16]. As an important current example, the COVID-19 pandemic is creating a demand for methodological development to address such questions, specifically, when evaluating the effectiveness of a potential vaccine or therapeutic outside randomized controlled trial settings.

Traditional causal survival analysis is typically carried out in the context of a randomized controlled trial (RCT), where the treatment assignment is controlled by researchers. Though they are the gold standard for causal inference, RCTs are

usually long-term engagements, expensive and limited in sample size. Alternatively, the availability of observational data with comprehensive information about patients, such as electronic health records (EHRs), constitutes a more accessible but also more challenging source for estimating causal effects [HSN08, JDC+09]. Such observational data may be used to augment and verify an RCT, after a particular treatment is approved and in use [GCC+19, FLS11, LHS14]. Moreover, the wealth of information from observational data also allows for the estimation of the individualized treatment effect (ITE), namely, the causal effect of an intervention at the individual level. In this work, we develop a novel framework for counterfactual time-to-event prediction to estimate the ITE for survival or time-to-event outcomes from observational data.

Estimating the causal effect for survival outcomes in observational data manifests two principal challenges. First, the treatment assignment mechanism is not known a priori. Therefore, there may be variables, known as confounders, affecting both the treatment and survival time, which lead to selection bias [BP12], i.e., that the distributions across treatment groups are not the same. In this work, we focus on selection biases due to confounding, but other sources may also be considered. For instance, patients who are severely ill are likely to receive more aggressive therapy, however, their health status may also inevitably influence survival. Traditional survival analysis neglects such bias, leading to incorrect causal estimation. Second, the exact time-to-event is not always observed, i.e., sometimes we only know that an event has not occurred up to a certain point in time. This is known as the censoring problem. Moreover, censoring might be informative depending on the characteristics of the individuals and their treatment assignments, thus proper adjustment is required for accurate causal estimation [CH04, Día19].

Traditional causal survival-analysis approaches typically model the effect of the treatment or covariates (not time or survival) in a parametric manner. Two commonly used models are the Cox proportional hazards (CoxPH) model [Cox72] and

the accelerated failure time (AFT) model [Wei92b], which presume a linear relationship between the covariates and survival probability. Further, proper weighting for each individual has been employed to account for confounding bias from these models [Aus07, Aus14, HCM$^+$05]. For instance, probability weighting schemes that account for both selection bias and covariate dependent censoring have been considered for adjusted survival curves [CH04, Día19]. Moreover, such probability weighting schemes have been applied to causal survival-analysis under time-varying treatment and confounding [Rob86, HBR00]. See [vdLR03, Tsi07, VdLR11, HR20] for an overview. Such linear specification makes these models interpretable but compromises their flexibility, and makes it difficult to adapt them for high-dimensional data or to capture complex interactions among covariates. Importantly, these methods lack a counterfactual prediction mechanism, which is key for ITE estimation (see Section 2).

Fortunately, recent advances in machine learning, such as representation learning or generative modeling, have enabled causal inference methods to handle high-dimensional data and to characterize complex interactions effectively. For instance, there has been recent interest in tree-based [CGM$^+$10, WA18] and neural-network-based [SJS17, ZBvdS20, AZT$^+$21] approaches. For pre-specified time-horizons, the nonparametric Random Survival Forest (RSF) [IKBL08] and Bayesian Additive regression trees (BART) [CGM$^+$10] have been extended to causal survival analysis. RSF has been applied to causal survival forests with weighted bootstrap inference [SWD$^+$18, CKWZ20] while a BART is extended to account for survival outcomes in Surv-BART [SLML16], and AFT-BART [HLRV20]. See [HJL20] for an extensive investigation of the causal survival tree-based methods.

Alternatively, when estimating the ITE, neural-network-based methods propose to regularize the transformed covariates or representations for an individual to have balanced distributions across treatment groups, thus accounting for the confounding

bias and improving ITE prediction. However, most approaches employing representation learning techniques for counterfactual inference deal with continuous or binary outcomes, instead of time-to-event outcomes with censoring (informative or non-informative). Moreover, while recent neural-network-based survival analysis methods [NLD21, RPEB16, CTL+18, ADZ+20, MPER18, LZAvdS19, LZYvdS18, XTH20] have improved survival predictions when censoring is non-informative, they lack mechanisms for accounting for informative censoring or confounding biases. Hence, a principled generalization to the context of *counterfactual survival analysis* is needed.

In this work we leverage balanced (latent) representation learning to estimate ITEs via counterfactual prediction of survival outcomes in observational studies. We develop a framework to predict event times from a low-dimensional transformation of the original covariate space. To address the specific challenges associated with counterfactual survival analysis, we make the following contributions:

- We develop an optimization objective incorporating adjustments for informative censoring, as well as a balanced regularization term bounding the generalization error for ITE prediction. For the latter, we repurpose a recently proposed bound [SJS17] for our time-to-event scenario.

- We propose a generative model for event times to relax restrictive survival linear and parametric assumptions, thus allowing for more flexible modeling. Our approach can also provide nonparametric uncertainty quantification for ITE predictions.

- We provide survival-specific evaluation metrics, including a new *nonparametric hazard ratio* estimator, and discuss how to perform model selection for survival outcomes. The proposed model demonstrates superior performance relative to the commonly used baselines in real-world and semi-synthetic datasets.

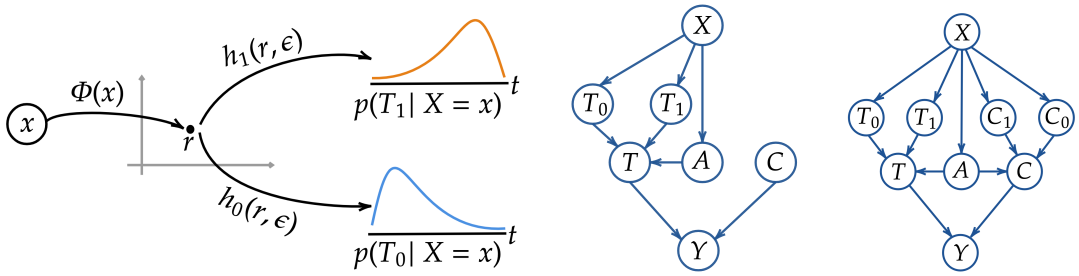
- We introduce a survival-specific semi-synthetic dataset and demonstrate an approach for leveraging prior randomized experiments in longitudinal studies for model validation.

4.2 Problem Formulation

We first introduce the basic setup for performing causal survival analysis in observational studies. Suppose we have N units, with N_1 units being treated and N_0 in the control group ($N = N_1 + N_0$). For each unit (individual), we have covariates X, which can be heterogeneous, e.g., a mixture of categorical and continuous covariates which, in the context of medicine, may include labs, vitals, procedure codes, etc. We also have a treatment indicator A, where $A = 0$ for the controls and $A = 1$ for the treated, as well as the outcome (event) of interest T. Under the potential-outcomes framework [Rub05], let T_0 and T_1 be the potential event times for a given subject under control and treatment, respectively. In practice we only observe one realization of the potential outcomes, i.e., the factual outcome $T = T_A$, while the counterfactual outcome T_{1-A} is unobserved.

In survival analysis, the problem becomes more difficult because we do not always observe the exact event time for each individual, but rather the time up to which we are certain that the event has not occurred; specifically, we have a (right) censoring problem, most likely due to the loss of follow-up. We denote the censoring time as C and censoring indicator as $\delta \in \{0, 1\}$. The actual observed time is $Y = \min(T_A, C)$, i.e., the outcome is observed (non-censored) if $T_A < C$ and $\delta = 1$.

In this work, we are interested in the expected difference between the T_1 and T_0 conditioned on X for a given unit (individual), which is commonly known as the *individualized treatment effect* (ITE). Specifically, we wish to perform inference on the conditional distributions of T_1 and T_0, i.e., $p(T_1|X)$ and $p(T_0|X)$, respectively, as shown in Figure 4.1. In practice, we observe N realizations of (Y, δ, X, A) for



(a) Counterfactual inference (b)Non-informative (c) Informative

FIGURE 4.1: (a) Illustration of the proposed counterfactual survival analysis (CSA). Covariates $X = x$ are mapped into latent representation r via deterministic mapping $r = \Phi(x)$. The potential outcomes are sampled from $t_a \sim p(T_A|X = x)$ for $A = a$ via stochastic mapping $h_A(r, \tilde{\epsilon})$, where stochasticity is induced with a planar-flow-based transformation, $\tilde{\epsilon}$, of a simple distribution $p(\epsilon)$, i.e., uniform or Gaussian. (b) and (c) show the proposed causal graphs for non-informative and informative censoring, respectively.

observed time, censoring indicator, covariates and treatment indicator, respectively; hence, from an observational study the dataset takes the form $\mathcal{D} = \{(y_i, \delta_i, x_i, a_i)\}_{i=1}^N$. Below, we discuss several common choices of estimands in survival analysis.

4.2.1 Estimands of Interest

We begin by considering survival analysis in the absence of an intervening treatment choice, A. Let $F(t|x) \triangleq P(T \leqslant t|X = x)$ be the cumulative distribution function of the event (failure) time, t, given a realization of the covariates, x. Survival analysis is primarily concerned with characterization of the survival function conditioned on covariates $S(t|x) \triangleq 1 - F(t|x)$, and the hazard function or risk score, $\lambda(t|x)$, defined below. $S(t|x)$ is a monotonically decreasing function indicating the probability of survival up to time t. The hazard function measures the instantaneous probability of the event occurring between $\{t, t + \Delta t\}$ given $T > t$ and $\Delta t \to 0$. From standard definitions [KK10], the relationship between cumulative and hazard function is

formulated as

$$\begin{aligned}\lambda(t|x) &= \lim_{dt\to 0} \frac{P(t < T < t + dt|X = x)}{P(T > t|X = x)dt} \\ &= -\frac{d \log S(t|x)}{dt} = \frac{f(t|x)}{S(t|x)}.\end{aligned} \tag{4.1}$$

From (4.1) we see that $f(t|x) \triangleq P(T = t|X = x) = \lambda(t|x)S(t|x)$, is the conditional event time density function [KK10].

Given the binary treatment A, we are interested in its impact on the survival time. For ITE estimation, we are also interested in the difference between the two potential outcomes T_1, T_0. Let $S_A(t|x)$ and $\lambda_A(t|x)$ denote the survival and hazard functions for the potential outcomes T_A, i.e., T_1 and T_0. Several common estimands of interest include [ZTU+12, TJCP16]:

- *Difference in expected lifetime*:
  $\text{ITE}(x) = \int_0^{t_{\max}} \{S_1(t|x) - S_0(t|x)\} \mathrm{d}t = \mathbb{E}\{T_1 - T_0|X = x\}$.
- *Difference in survival function*: $\text{ITE}(t, x) = S_1(t|x) - S_0(t|x)$.
- *Hazard ratio*: $\text{ITE}(t, x) = \lambda_1(t|x)/\lambda_0(t|x)$.

The inference difficulties associated with the above estimands from observational data are two-fold. First, there are confounders affecting both the treatment assignment and outcomes, which stem from selection bias, i.e., the treatment and control covariate distributions are not necessarily the same. Also, we do not have direct knowledge of the conditional treatment assignment mechanism, i.e., $P(A = a|X = x)$, also known as the propensity score. Let $\perp\!\!\!\perp$ denote statistical independence. For estimands to be identifiable from observational data, we make two assumptions: (i) $\{T_1, T_0\} \perp\!\!\!\perp A|X$, i.e. no unobserved confounders or ignorability, and (ii) overlap in the covariate support $0 < P(A = 1|X = x) < 1$ almost surely

if $p(X = x) > 0$. Second, the censoring mechanism is also unknown and may lead to bias without proper adjustment. We consider two censoring mechanisms in our work, (i) conditionally independent or informative censoring: $T \perp\!\!\!\perp C|X, A$, and (ii) random or non-informative censoring: $T \perp\!\!\!\perp C$. Note that for informative censoring, we also have to consider potential censoring times C_1 and C_0 and their conditionals $p(C_1|X)$ and $p(C_0|X)$, respectively. Figure 4.1 shows causal graphs illustrating these modeling assumptions.

4.3 Modeling

To overcome the above challenges and adjust for observational biases, we propose a unified framework for *counterfactual survival analysis* (CSA). Specifically, we repurpose the counterfactual bound in [SJS17] for our time-to-event scenario and introduce a nonparametric approach for stochastic survival outcome predictions. Below we formulate a theoretically grounded and unified approach for estimating (i) the encoder function $r = \Phi(x)$, which deterministically maps covariates x to their corresponding latent representation $r \in \mathbb{R}^d$, and (ii) two stochastic time-to-event generative functions, $h_A(\cdot)$, to implicitly draw samples from both potential outcome conditionals $t_a \sim p_{h,\Phi}(T_A|X = x)$, for $A = \{1, 0\}$, and where t_a indicates the sample from $p_{h,\Phi}(T_A|X = x)$ is for $A = a$. Further, we formulate a general extension that accounts for informative censoring by introducing two stochastic censoring generative functions, $\nu_A(\cdot)$, to draw samples for potential censoring times $c_a \sim p_{\nu,\Phi}(C_A|X = x)$. The model-specifying functions, $\{h_A(\cdot), \nu_A(\cdot), \Phi(\cdot)\}$, are parameterized via neural networks. See the Supplementary Material (SM) for details. Figure 4.1 summarizes our modeling approach.

4.3.1 Accounting for selection bias

We wish to estimate the potential outcomes, i.e., event times, which are sampled by distributions parameterized by functions $\{h_A(\cdot), \Phi(\cdot)\}$, i.e.,

$$t \sim p_{h,\Phi}(T|X = x, A = a) \tag{4.2}$$

$$t_a \sim p_{h,\Phi}(T_a|X = x) \tag{4.3}$$

We obtain (4.3) from (4.2) via the strong ignorability assumption, i.e., $\{T_0, T_1\} \perp\!\!\!\perp A|X$ (consistent with the causal graphs in Figure 4.1(b) and 4.1(c)) and $0 < P(A = a|X = x) < 1$, and the consistency assumption, i.e., $T = T_A|A = a$. A similar argument can be made for informative censoring based on Figure **??**, so we can also write $c_a \sim p_{\nu,\Phi}(C_A|X = x)$. Given (4.3), model functions $\{h_A(\cdot), \Phi(\cdot)\}$ and $\nu_A(\cdot)$ for informative censoring can be learned by leveraging standard statistical optimization approaches, that minimize a loss hypothesis $\mathcal{L}$ given samples from the empirical distribution $(y, \delta, x, a) \sim p(Y, \delta, X, A)$, i.e., from dataset $\mathcal{D}$. Specifically, we write $\mathcal{L}$ as

$$\mathcal{L} = \mathbb{E}_{(y,\delta,x,a)\sim p(Y,\delta,X,A)} \left[\ell_{h,\Phi}(t_a, y, \delta)\right] , \tag{4.4}$$

where $\ell_{h,\Phi}(t_a, y, \delta)$ is a loss function that measures the agreement of $t_a \sim p_{h,\Phi}(T_A|X = x)$ (and $c_a \sim p_{\nu,\Phi}(C_A|X = x)$ for informative censoring) with ground truth $\{y, \delta\}$, the observed time and censoring indicator, respectively.

For some parametric formulations of event time distribution $p_{h,\Phi}(T_A|X = x)$, e.g., exponential, Weibull, log-Normal, etc., and provided the censoring mechanism is non-informative, $-\ell_{h,\Phi}(t_a, y, \delta)$ is the closed form log likelihood. Specifically, $-\ell_{h,\Phi}(t_a, y, \delta) \triangleq \log p_{h,\Phi}(T_a|X = x) = \delta \cdot \log f_{h,\Phi}(t_a|x) + (1 - \delta) \cdot \log S_{h,\Phi}(t_a|x)$, which implies that the conditional event time density and survival functions can be calculated in closed form from transformations $\{h_A(\cdot), \Phi(\cdot)\}$ of x. See the SM for parametric examples of (4.4) accounting for informative censoring.

We further define the expected loss for a given realization of covariates x and treatment assignment a over observed times y (censored and non-censored), and the censoring indicator δ as $\zeta_{h,\Phi}(x,a) \triangleq \mathbb{E}_{(y,\delta,x)\sim p(Y,\delta|X)}\ell_{h,\Phi}(t_a,y,\delta)$ as in [SJS17]. For a given subject with covariates x and treatment assignment a, we wish to minimize both the factual and counterfactual losses, $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{CF}}$, respectively, by decomposing $\mathcal{L} = \mathcal{L}_{\mathrm{F}} + \mathcal{L}_{\mathrm{CF}}$ as follows

$$\begin{aligned} \mathcal{L}_{\mathrm{F}} &= \mathbb{E}_{(x,a)\sim p(A,X)}\zeta_{h,\Phi}(x,a)\,, \\ \mathcal{L}_{\mathrm{CF}} &= \mathbb{E}_{(x,a)\sim p(1-A,X)}\zeta_{h,\Phi}(x,a)\,. \end{aligned} \tag{4.5}$$

Let $u \triangleq P(A=1)$ denote the marginal probability of treatment assignment. We can readily decompose the losses in (4.5) according to treatment assignments. The decomposed factual $\mathcal{L}_{\mathrm{F}} = u\cdot\mathcal{L}_{\mathrm{F}}^{A=1} + (1-u)\cdot\mathcal{L}_{\mathrm{F}}^{A=0}$, and similarly, the decomposed counterfactual $\mathcal{L}_{\mathrm{CF}} = (1-u)\cdot\mathcal{L}_{\mathrm{CF}}^{A=1} + u\cdot\mathcal{L}_{\mathrm{CF}}^{A=0}$. In practice, only factual outcomes are observed, hence, for a non-randomized non-controlled experiment, we cannot obtain an unbiased estimate of $\mathcal{L}_{\mathrm{CF}}$ from data due to selection bias (or confounding). Therefore, we bound $\mathcal{L}_{\mathrm{CF}}$ and $\mathcal{L}$ below following [SJS17].

Corollary 1. *Assume $\Phi(\cdot)$ is an invertible map, and $\alpha^{-1}\zeta_{h,\Phi}(x,a) \in G$, where G is a family of functions, $p_\Phi^{A=a} \triangleq p_\Phi(R|A=a)$ is the latent distribution for group $A=a$, and $\alpha>0$ is a constant. Then, we have:*

$$\begin{aligned} \mathcal{L}_{\mathrm{CF}} &\leqslant (1-u)\cdot\mathcal{L}_{\mathrm{F}}^{A=1} + u\cdot\mathcal{L}_{\mathrm{F}}^{A=0} + \alpha\cdot\mathrm{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0}) \\ \mathcal{L} &\leqslant \mathcal{L}_{\mathrm{F}}^{A=1} + \mathcal{L}_{\mathrm{F}}^{A=0} + \alpha\cdot\mathrm{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0})\,. \end{aligned} \tag{4.6}$$

The integral probability metric (IPM) [Mül97, SFG+12] measures the distance between two probability distributions p and q defined over M, i.e., the latent space of R. Formally,
$\mathrm{IPM}_G(p,q) \triangleq \sup_{g\in G}\left|\int_M g(m)\,(p(m)-q(m))\,dm\right|$, where $g : m \to \mathbb{R}$, represents a class of real-valued bounded measurable functions on M [SJS17]. Therefore,

model functions $\{h_a(\cdot), \Phi(\cdot)\}$ can be learned by minimizing the upper bound in (4.6) consisting of (i) only factual losses under both treatment assignments and (ii) an IPM regularizer enforcing latent distributional equivalence between the treatment groups. Note that if the data originates from a RCT it follows (by construction) that $\text{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0}) = 0$.

4.3.2 Accounting for censoring bias

Below we formulate an approach for estimating functions $h_A(\cdot)$ and $\nu_A(\cdot)$ for synthesizing (sampling) non-censored $t_a \sim p_{h,\Phi}(T_A|X = x)$ and censored $c_a \sim p_{\nu,\Phi}(C_A|X = x)$ times, respectively. While some parametric assumptions for $p_{h,\Phi}(T_A|X = x)$ yield easy-to-evaluate closed forms for $S_{h,\Phi}(t_a|x)$ that can be used as likelihood for censored observations, they are restrictive, and have been shown to generate unrealistic high variance samples [CTL$^+$18]. So motivated, we seek a nonparametric likelihood-based approach that can model a flexible family of distributions, with an easy-to-sample approach for event times $t_a \sim p_{h,\Phi}(T_a|X = x)$. We model the event time generation process with a source of randomness, $p(\epsilon)$, e.g. Gaussian or uniform, which is obtained from a neural-network-based nonlinear transformation. In the experiments we use a *planar flow* formulation parameterized by $\{U_h, W_h, b_h\}$ [RM15], however, other specifications can also be used. Note that [MPER18] has previously leveraged normalizing flows for survival analysis, however, our approach is very different in that it focuses on i) formulating a counterfactual survival analysis framework that accounts for informative or non-informative censoring mechanisms and confounding, and ii) modeling event times as a continuous variable instead of discretizing them. Specifically, we transform the source of randomness, ϵ, using a single layer specification as follows

$$\begin{aligned} \tilde{\epsilon}_h &= \epsilon + U_h \tanh(W_h \epsilon + b_h)\,, \quad \epsilon \sim \text{Uniform}(0, 1)\,, \\ t_a &= h_A(r, \tilde{\epsilon}_h)\,, \quad r = \Phi(x) \end{aligned} \tag{4.7}$$

where $\{U_h, W_h\} \in \mathbb{R}^{d\times d}$, $\{b_h, \epsilon\} \in \mathbb{R}^d$, d is the dimensionality of the planar flow; each component of ϵ is drawn independently from Uniform$(0, 1)$, and $\tilde{\epsilon}_h$ may be viewed as a skip connection with stochasticity in ϵ. Further, $h_A(r, \tilde{\epsilon}_h)$ and $\Phi(x)$ are time-to-event generative and encoding functions, respectively, parameterized as neural networks. For simplicity, the dimensions of r and ϵ are set to d, however, they can be set independently if desired. In practice, we are interested in generating realistic event-time samples; therefore, we account for both censored and non-censored observations by adopting the objective from [CTL+18], formulated as

$$\begin{aligned}\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}} \triangleq \mathbb{E}_{(y,\delta,x,a)\sim p(Y,\delta,X,A),\epsilon\sim p(\epsilon)} \big[\delta \cdot (|y - t_a|) \\ + (1-\delta)\cdot(\max(0, y - t_a))\big] ,\end{aligned} \tag{4.8}$$

where the first term encourages sampled event times t_a to be close to y, the ground truth for observed events, i.e., $\delta = 1$, while penalizing t_a for being smaller than the censoring time when $\delta = 0$. Further, the expectation is taken over samples (a minibatch) from empirical distribution $p(Y, \delta, X, A)$.

Informative censoring We model informative censoring similar to (4.8) but mirroring the censoring indicators to encourage accurate censoring time samples c_a for $\delta = 0$, while penalizing c_a for being smaller than y for $\delta = 1$ (observed events). Specifically, we set an independent source of randomness like in (4.7) but parameterized by $\{U_\nu, W_\nu, b_\nu\}$ and censoring generative functions $\nu_A(r, \tilde{\epsilon}_\nu)$, parameterized as neural networks, where $c_a \sim p_{\nu,\Phi}(C_A|X = x)$ formulated as

$$\begin{aligned}\ell_c(\nu, \Phi) = \mathbb{E}_{(y,\delta,x,a)\sim p(y,\delta,X,A),\boldsymbol{\epsilon}\sim p(\epsilon)} \big[(1-\delta)\cdot(|y - c_a|) \\ + \delta \cdot (\max(0, y - c_a))\big] .\end{aligned} \tag{4.9}$$

Further, we introduce an additional time-order-consistency loss that enforces the correct order of the observed time relative to the censoring indicator, i.e., $c_a < t_a$ if

$\delta = 0$ and $t_a < c_a$ if $\delta = 1$, thus

$$\begin{aligned}\ell_{\mathrm{TC}}(h, \nu, \Phi) = \mathbb{E}_{(\delta,x,a)\sim p(\delta,X,A),\epsilon\sim p(\epsilon)} \left[\delta \cdot (\max(0, t_a - c_a)) \right. \\ \left. + (1-\delta) \cdot (\max(0, c_a - t_a))\right] .\end{aligned} \tag{4.10}$$

Note that $\ell_{\mathrm{TC}}(h, \nu, \Phi)$ does not depend on the observed event times but only on the censoring indicators. Finally, we write the consolidated CSA loss for informative censoring (CSA-INFO) by aggregating (4.8), (4.9) and (4.10) as

$$\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA-INFO}} \triangleq \mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}} + \ell_{\mathrm{c}} + \ell_{\mathrm{TC}} \,. \tag{4.11}$$

4.3.3 Learning

Model functions $\{h_A(\cdot), \Phi(\cdot), \nu_A(\cdot)\}$ are learned by minimizing the bound (4.6), via stochastic gradient descent on minibatches from $\mathcal{D}$, with $\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}}$ for non-informative censoring and $\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA-INFO}}$ for informative censoring. Further, for the IPM regularization loss in (4.6), we optimize the dual formulation of the *Wasserstein distance*, via the regularized *optimal transport* [Vil08, Cut13]. Consequently, we only require $\alpha^{-1}\zeta_{h,\Phi}(x, a)$ to be 1-Lipschitz [SJS17] and α is selected by grid search on the validation set using only factual data (details below).

4.4 Metrics

We propose a comprehensive evaluation approach that accounts for both factual and causal metrics. Factual survival outcome predictions are evaluated according to standard survival metrics that measure diverse performance characteristics, such as concordance index (C-Index) [HJLC$^+$84], mean coefficient of variation (COV) and calibration slope (C-slope) [CLM$^+$20]. See the SM for more details on these metrics. For causal metrics, defined below, we introduce a nonparametric hazard ratio (HR) between treatment outcomes, and adopt the conventional precision in estimation of heterogeneous effect (PEHE) and average treatment effect (ATE) performance

metrics [Hil11]. Note that PEHE and ATE require ground truth counterfactual event times, which is only possible for (semi-)synthetic data. For HR, we compare our findings with those independently reported in the literature from gold-standard RCT data.

Nonparametric Hazard Ratio In medical settings, the population hazard ratio $\mathrm{HR}(t)$ between treatment groups is considered informative thus has been widely used in drug development and RCTs [YBD$^+$16, MEB$^+$12]. For example, $\mathrm{HR}(t) < 1$, > 1, or ≈ 1 indicate population positive, negative and neutral treatment effects at time t, respectively. Moreover, $\mathrm{HR}(t)$ naturally accounts for both censored and non-censored outcomes. Standard approaches for computing $\mathrm{HR}(t)$ rely on the restrictive proportional hazard assumption from CoxPH [Cox72], which is constituted as a semi-parametric linear model $\lambda(t|a) = \lambda_{\mathrm{b}}(t)\exp(a\beta)$. However, the constant covariate (time independent) effect is often violated in practice (see Figure 4.2). For CoxPH, the marginal HR between treatment and control can be obtained from regression coefficient β learned via maximum likelihood without the need for specifying the baseline hazard $\lambda_{\mathrm{b}}(t)$:

$$\mathrm{HR}_{\mathrm{CoxPH}}(t) = \frac{\lambda(t|a=1)}{\lambda(t|a=0)} = \exp(\beta)\,. \tag{4.12}$$

So motivated, we propose a nonparametric, model-free approach for computing $\mathrm{HR}(t)$, in which we do not assume a parametric form for the event time distribution or the proportional hazard assumption from CoxPH. This approach only relies on samples from the conditional event time density functions, $f(t_1|x)$ and $f(t_0|x)$, via $t_a = h_A(\cdot)$ from (4.7).

Definition 1. *We define the nonparametric marginal Hazard Ratio and its approximation,* $\hat{\text{HR}}(t)$*, as*

$$\begin{aligned} \text{HR}(t) &= \frac{\lambda_1(t)}{\lambda_0(t)} = \frac{S_0(t)}{S_1(t)} \cdot \frac{S_1'(t)}{S_0'(t)}\,, \\ \hat{\text{HR}}(t) &= \frac{\hat{S}_0^{\text{PKM}}(t)}{\hat{S}_1^{\text{PKM}}(t)} \cdot \frac{m_1(t)}{m_0(t)}\,, \end{aligned} \tag{4.13}$$

where for $\text{HR}(t)$ we leveraged (4.1) to obtain (4.13) and $S'(t) \triangleq dS(t)/dt$. The nonparametric assumption for $S(t)$ makes the computation of $S'(t)$ challenging. Provided that $S(t)$ is a monotonically decreasing function, for simplicity, we fit a linear function $S(t) = m{\cdot}t{+}c$, and set $S'(t) \approx m$. Note that the linear model is only used for estimating $S'(t)$ from the nonparametric estimation of $S(t)$. Bias from $S'(t)$ can be reduced by considering more complex function approximations for $S(t)$, e.g., polynomial or spline. For the nonparametric estimation of $S(t)$ we leverage the model-free population point-estimate-based nonparametric Kaplan-Meier [KM58] estimator of the survival function $\hat{S}^{\text{PKM}}(t)$ in [CLM+20], also detailed in Chapter 3, to marginalize both factual and counterfactual predictions given covariates x. The approximated hazard ratio, $\hat{\text{HR}}(t)$, is thus obtained by combining the approximations $\hat{S}_a^{\text{PKM}}(t)$ and m_a. A similar formulation for the conditional, $\hat{\text{HR}}(t|x)$, can also be derived. See the SM for full details on the evaluation of $\hat{\text{HR}}(t)$ and $\hat{\text{HR}}(t|x)$. Note that for some AFT- or CoxPH-based parametric formulations, $\text{HR}(t|x)$, can be readily evaluated because $f(t_a|x)$ and $S(t_a|x)$ are available in closed form.

In the experiments, we will use $\text{HR}(t)$ to compare different approaches against results reported in RCTs (see Tables 4.1 and 4.3). Further, we will use $\text{HR}(t|x)$ to illustrate stratified treatment effects (see Figure 4.2). Note that though a neural-network-based survival recommender system [KSC+18] has been previously used to estimate $\text{HR}(t|x)$, their approach does not account for confounding or informative censoring thus it is susceptible to bias.

Table 4.1: Performance comparisons on ACTG-SYNTHETIC data, with 95% HR(t) confidence interval. The ground truth, test set, hazard ratio is HR(t) = $0.52_{(0.39,0.71)}$.

Method	**Causal metrics**			**Factual metrics**		
	ϵ_{PEHE}	ϵ_{ATE}	HR(t)	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	NA	NA	$0.97_{(0.86,1.09)}$	NA	NA	NA
CoxPH-IPW	NA	NA	$0.48_{(0.03,7.21)}$	NA	NA	NA
CoxPH-OW	NA	NA	$0.60_{(0.53,0.68)}$	NA	NA	NA
Surv-BART	352.07	77.89	$0.0(0.0, 0.0)$	(0.706, 0.686)	0.001	(0.398, ∞)
AFT-Weibull	367.92	133.93	$0.47_{(0.47,0.47)}$	(0.21, 0.267)	6.209	(0.707, 0.729)
AFT-log-Normal	377.76	157.64	$0.47_{(0.47,0.47)}$	(0.675, 0.556)	6.971	(0.707, 0.729)
SR	369.47	88.55	$0.38_{(0.33,0.65)}$	(0.791, 0.744)	0	(0.985, 1.027)
CSA (proposed)	358.72	**0.8**	$0.45_{(0.39,0.65)}$	(0.787, 0.767)	0.131	(0.985, 1.026)
CSA-INFO (proposed)	**344.3**	31.19	$\mathbf{0.53}_{(0.41,0.67)}$	(0.78, 0.764)	0.13	(0.999, 1.029)

Precision in Estimation of Heterogeneous Effect (PEHE) A general individualized estimation error is formulated as

$$\epsilon_{\text{PEHE}} = \sqrt{\mathbb{E}_X[(\text{ITE}(x) - \hat{\text{ITE}}(x))^2]},$$

where ITE(x) is the ground truth, $\hat{\text{ITE}}(x) = \mathbb{E}_T\left[\gamma\left(T_1\right) - \gamma\left(T_0\right)|X = x\right]$ and $\gamma(\cdot)$ is a deterministic transformation. In our experiments, $\gamma(\cdot)$ is the average over samples from $t_a \sim p_{h,\Phi}(T_A|X = x)$. Alternative estimands, e.g., thresholding survival times $\gamma(T_A) = I\{T_A > \tau\}$, can also be considered as described in Section 4.2.1.

Average Treatment Effect (ATE) The population treatment effect estimation error is defined as

$$\epsilon_{\text{ATE}} = |\text{ATE} - \hat{\text{ATE}}|,$$

where $\text{ATE} = \mathbb{E}_X[\text{ITE}(x)]$ (ground truth) and $\hat{\text{ATE}} = \mathbb{E}_X[\hat{\text{ITE}}(x)]$.

Note that both PEHE and ATE require ground truth (population and individual) treatment effects to be available, which is only possible in synthetic and semi-synthetic data (benchmarking) scenarios.

4.5 Experiments

We describe the baselines and datasets that will be used to evaluate the proposed counterfactual survival analysis methods (CSA and CSA-INFO). Detailed architec-

Table 4.2: Summary statistics of the datasets.

	FRAMINGHAM	ACTG	ACTG-SYNTHETIC
Events (%)	26.0	26.9	48.9
Treatment (%)	10.4	49.5	55.9
N	3,435	1,054	2,139
p	32	23	23
Missing (%)	0.23	1.41	1.38
$t_{\max}$ (days)	7,279	1,231	1,313

ture information of the proposed methods (CSA and CSA-INFO) and baselines (AFT-log-Normal, AFT-Weibull, Semi-supervised Regression(SR)) are provided in the SM. Pytorch code to replicate experiments can be found at `https://github.com/paidamoyo/counterfactual_survival_analysis`. Throughout the experiments, we use the standard HR(t) for CoxPH based methods in (4.12) and (4.13) for all others. The bound in (4.6) is sensitive to α, thus we propose approximating proxy counterfactual outcomes $\{Y_{\text{CF}}, \delta_{\text{CF}}\}$ for the validation set, according to the covariate Euclidean nearest-neighbour (NN) from the training set. We select the α that minimizes the validation loss $\mathcal{L} = \mathcal{L}_{\text{F}} + \mathcal{L}_{\text{CF}}$ from the set $(0, 0.1, 1, 10, 100)$.

Baselines We consider the following competitive baseline approaches: (i) propensity weighted CoxPH [SWH09, BHC$^+$14, RR83]; (ii) IPM (4.6) regularized AFT (log-Normal and Weibull) models; (iii) an IPM (4.6) regularized deterministic semi-supervised regression (SR) model with accuracy objective from [CTL$^+$18], as a contrast for the proposed stochastic predictors (CSA and CSA-INFO); and (iv) survival Bayesian additive regression trees (Surv-BART) [SLML16]. For CoxPH, we consider three normalized weighting schemes: (i) inverse probability weighting (IPW) [HT52, CTD09], where $\text{IPW}_i = \frac{a_i}{\hat{e}_i} + \frac{1-a_i}{1-\hat{e}_i}$; ii) overlapping weights (OW) [CHIM06, LMZ18], where $\text{OW}_i = a_i \cdot (1 - \hat{e}_i) + (1 - a_i) \cdot \hat{e}_i$; and iii) the standard RCT uniform assumption. A simple linear logistic model $\hat{e}_i = \sigma(x_i; w)$, is used as an approximation, $\hat{e}_i$, to the unknown propensity score $P(A = 1|X = x)$. See the SM

Table 4.3: Performance comparisons on FRAMINGHAM data, with 95% HR(t) confidence interval. Test set NN assignment of y_{CF} and δ_{CF} yields biased HR(t) = $1.23_{(1.17,1.25)}$, while previous large scale longitudinal RCT studies estimated HR(t) = $0.75_{(0.64,0.88)}$ [YBD$^+$16].

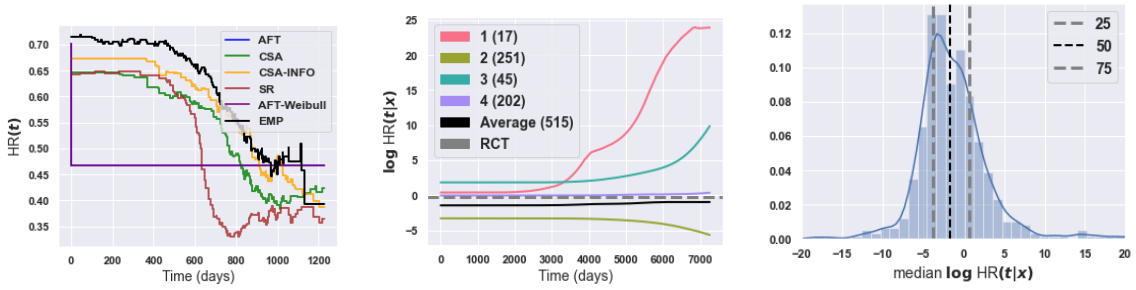
Method	**Causal metric**	**Factual metrics**		
	HR(t)	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	$1.69_{(1.38,2.07)}$	NA	NA	NA
CoxPH-IPW	$1.09_{(0.76,1.57)}$	NA	NA	NA
CoxPH-OW	$0.88_{(0.73,1.08)}$	NA	NA	NA
Surv-BART	$14.99_{(14.9,14.9e8)}$	(0.629, 0.630)	0.003	(0.232, 0.084)
AFT-Weibull	$1.09_{(1.09,1.09)}$	(0.734, 0.395)	8.609	(0.857, 0.89)
AFT-log-Normal	$1.55_{(1.46,1.55)}$	(0.68, 0.56)	10.415	(0.979, 0.732)
SR	$0.58_{(0.53,0.71)}$	(0.601, 0.57)	0	(0.491, 0.63)
CSA (proposed)	$1.04_{(1.00,1.09)}$	(0.763, 0.728)	0.161	(0.891, 0.81)
CSA-INFO (proposed)	$\mathbf{0.81}_{(0.77,0.83)}$	(0.752, 0.651)	0.156	(0.907, 0.881)

for more details of the baselines.

Datasets We consider the following datasets summarized in Table 4.2: (i) FRAMINGHAM, is an EHR-based longitudinal cardiovascular cohort study that we use to evaluate the effect of statins on future coronary heart disease outcomes [BLV$^+$94]; (ii) ACTG, is a longitudinal RCT study comparing monotherapy with Zidovudine or Didanosine with combination therapy in HIV patients [HKH$^+$96]; and (iii) ACTG-SYNTHETIC, is a semi-synthetic dataset based on ACTG covariates. We simulate potential outcomes according to a Gompertz-Cox distribution [BAB05] with selection bias from a simple logistic model for $P(A = 1|X = x)$ and AFT-based censoring mechanism. The generative process is detailed in the SM. Table 4.2 summarizes the datasets according to (i) covariates of size p; (ii) proportion of non-censored events, treated units, and missing entries in the $N \times p$ covariate matrix; and (iii) time range $t_{\max}$ for both censored and non-censored events. Missing entries are imputed with the median or mode if continuous or categorical, respectively.

Quantitative Results Experimental results for two data-sets in Tables 4.1 and 4.3, illustrate that AFT-based methods have high variance, inferior in calibration and C-

Index than accuracy-based methods (SR, CSA, CSA-INFO). Surv-BART is the least calibrated but low variance method. CSA-INFO and CSA outperform all methods across all factual metrics, whereas CSA-INFO is better calibrated, has low variance but slightly lower C-Index than CSA. Note that we fit CoxPH using the entire dataset; since it does not support counterfactual inference, we do not present factual metrics. By properly adjusting for both informative censoring and selection bias, CSA-INFO significantly outperforms all methods in treatment effect estimation according to HR(t) and $\boldsymbol{\epsilon}_{\text{PEHE}}$, across non-RCT datasets, while remaining comparable to AFT-Weibull on the RCT dataset (see the SM). Further, RCT-based results on ACTG data in the SM illustrate comparable HR(t) across all models except for AFT-log-Normal and Surv-BART, which overestimate, and SR, which underestimates risk. For non-RCT datasets (ACTG-SYNTHENTIC and FRAMINGHAM), CoxPH-OW has a clear advantage over all CoxPH based methods, mostly credited to the well-behaved bounded propensity weights $\in [0, 1]$. Interestingly, the FRAMINGHAM observational data exhibits a common paradox, where without proper adjustment of selection and censoring bias, naive approaches would result in a counter-intuitive treatment effect from statins. However, there is severe confounding from covariates such as age, BMI, diabetes, CAD, PAD, MI, stroke, etc., that influence both treatment likelihood and survival time. Table 4.3, demonstrates that CSA-INFO is clearly the best performing approach. Specifically, its HR(t), reverses the biased observational treatment effect, to demonstrate positive treatment from statins, which is consistent with prior large RCT longitudinal findings [YBD$^+$16]. Consequently, our experiments are comprehensive and we are confident that the CSA-INFO performance benefits are attributed to (i) accounting for informative censoring bias; (ii) accounting for selection bias (optimal IPM regularizer with $\alpha > 0$); and (iii) flexible and non-parametric generative modeling of event times from the stochastic planar flow.



(a) ACTG-SYNTHENTIC HR(t) (b) FRAMINGHAM $\log \mathrm{HR}(t|x)$ (c) FRAMINGHAM $\log \mathrm{HR}(t|x)$ pdf

FIGURE 4.2: (a) Inferred population HR(t) compared against ground truth (EMP) on ACTG-SYNTHETIC data. CSA-INFO-based (b) cluster-specific average $\log \mathrm{HR}(t|x)$ curves and (c) estimated density of median $\log \mathrm{HR}(t|x)$ values on the test set of the FRAMINGHAM dataset. Clusters assignment were obtained via hierarchical clustering of individualized $\log \mathrm{HR}(t|x)$ traces.

Qualitative Results Figure 4.2(a) demonstrates that CSA-INFO matches the ground truth population hazard, HR(t), better than alternative methods on ACTG-SYNTHETIC data. See the SM for ACTG and FRAMINGHAM. Figure 4.2(b) shows sub-population log hazard ratios for four patient clusters obtained via hierarchical clustering on the individual log hazard ratios, $\log \mathrm{HR}(t|x)$, of the test set of FRAMINGHAM data. Interestingly, these clusters stratify treatment effects into: positive (2), negative (1 and 3), and neutral (4) sub-populations. Moreover, the estimated density of median $\log \mathrm{HR}(t|x)$ values in Figure 4.2(c) illustrates that nearly 70% of the testing set individuals have $\log \mathrm{HR}(t|x) < 0$, thus may benefit from taking statins. Further, we isolated the extreme top and bottom quantiles, $\mathrm{HR}(t|x) < 0.024$ and $\mathrm{HR}(t|x) > 1.916$, respectively, of the median $\log \mathrm{HR}(t|x)$ values for the test set of FRAMINGHAM, as shown in Figure 4.2(c). After comparing their covariates, we found that individuals with the following characteristics may benefit from taking statins: young, male, diabetic, without prior history (CAD, PAD, stroke or MI), high BMI, cholesterol, triglycerides, fasting glucose, and low high-density lipoprotein. Note that individuals with contrasting covariates experience may not benefit from taking statins. There seem to be consensus that diabetics and high-cholesterol patients benefit from statins

[CLLK04, WBM+04]. See SM for additional results.

4.6 Conclusions

We have proposed a unified counterfactual inference framework for survival analysis. Our approach adjusts for bias from two sources, namely, confounding (covariates influence both the treatment assignment and the outcome) and censoring (informative or non-informative). Relative to competitive alternatives, we demonstrate superior performance for both survival-outcome prediction and treatment-effect estimation, across three diverse datasets, including a semi-synthetic dataset which we introduce. Moreover, we formulate a model-free nonparametric hazard ratio metric for comparing treatment effects or leveraging prior randomized real-world experiments in longitudinal studies. We demonstrate that the proposed model-free hazard-ratio estimator can be used to identify or stratify heterogeneous treatment effects. Finally, this work will serve as an important baseline for future work in real-world counterfactual survival analysis. In future work, we plan to understand the sensitivity of our estimates to unobserved confounding [CHH+59] and the effect of both censoring bias and selection bias on causal identifiability.

5

Survival Cluster Analysis

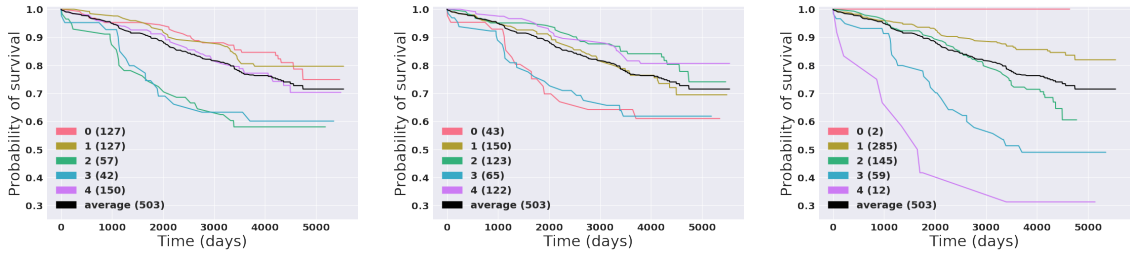
5.1 Introduction

Time-to-event models have primarily focused on either estimating a (point estimate) risk score or individualized time-to-event distributions. Parametric models estimate the time-to-event distribution conditional on covariates by assuming a parametric form of the event distribution, i.e., exponential, Weibull, log-normal, etc. Parametric models fall under the Accelerated Failure Time (AFT) [Wei92a] framework, provided they assume covariates either accelerate or decelerate the time-to-event. Assuming a parametric distribution is inflexible as the hazard function depends on the selected baseline distribution, for example assuming an exponential distribution, yields a constant hazard rate function. Alternatively, Cox Proportional Hazards (CoxPH) [Cox92], a semi-parametric, linear model for estimating relative risks is widely used in practice, as it does not require one to specify the baseline distribution. For pre-specified time-horizons, the non-parametric Random Survival Forest (RSF) [IKBL08] was proposed to estimate the cumulative hazard function based on an ensemble of binary decision trees, albeit often limited by scaling problems for large and high-

dimensional datasets.

With recent advances in machine learning, deep learning methods have improved classical survival analysis methods by leveraging non-linear relationship between covariates, for improved time-to-event or risk score predictions. Deep learning methods inspired by CoxPH or AFT have been proposed, e.g., DeepSurv [KSC+18], Deep Survival Analysis (DSA) [RPEB16], Deep Regularized Accelerated Failure Time (DRAFT) [CTL+18], Gaussian-process-based models [FRT16, AvdS17], and the Survival Continuous Ranked Probability Score (S-CRPS) [ADZ+20]. Sampling-based nonparametric methods have been proposed as well, e.g., normalizing-flow-based DSA [MPER18], adversarial-learning-based Deep Adversarial Time to Event (DATE) [CTL+18] and Survival Function Matching (SFM) [CTL+20]. Another class of nonparametric methods discretize time-to-event to predict survival probability within pre-specified discrete-interval event times with logistic-regression-based methods [YGLB11, Fot18, LZYvdS18]. Further, deep learning methods have also successfully addressed calibration [CTL+20, ADZ+20, LZAvdS19] and competing risks [ZZ18, AvdS17].

Clustering based on risk-profiles in survival analysis is relatively under-explored in machine learning, but is critical in applications such as (clinical) decision making. Identifying phenotypically heterogeneous subpopulations in the context of risk prediction is an important step toward machine-learning-based models for precision medicine [CV15, DZAD17]. Existing clustering methods for stratifying risks in survival analysis include feature based K-means (see Figure 5.1(a)) or hierarchical clustering [ESBB98, SKS+15, ASK+18]. Principal component cluster analysis has also been considered [APS+14]. However, it is well understood that feature-based clustering in covariate space may produce clusters that are not consistent with survival outcomes [BT04, GB13], particularly for high-dimensional datasets, such as gene expression data.



(a) K-means (b) SSC-Bair (c) SCA

FIGURE 5.1: Cluster-specific Kaplan-Meier survival profiles for three clustering methods on the SLEEP dataset (See Section 5.4 for details). (a) Standard K-means. (b) CoxPH-based covariate selection followed by K-means. (c) Proposed approach for joint learning of individualized time-to-event predictions and clustering. By jointly learning clustering with respect to both the covariates $\boldsymbol{x}$ and predicted time-to-event t, our model (SCA) can identify high-, medium- and low-risk individuals. Demonstrating the need to account for time information via a non-linear transformation of covariates when clustering survival datasets.

Methods that account for survival outcomes in clustering include CoxPH-inspired techniques [BT04, GB13], implemented as a two-step process: first, high CoxPH scoring covariates are selected, then a classical clustering approach like K-means is applied (see Figure 5.1(b)). However, CoxPH-based approaches are limited by the proportional hazards assumption. Alternatively, [XDM+19] proposed an outcome driven attention-based multi-task deep learning model for classification and then applied K-means on the latent representations to cluster subjects with acute coronary syndrome. More recently, [MTRN19] introduced DeepCLife, a method that learns clusters by maximizing the pairwise differences between the survival functions of all cluster pairs. This is done by indirectly maximizing the logrank score [Man66]. Unlike DeepCLife, which aims to optimize clusters but not predictions, our goal is to jointly characterize time-to-event predictive distributions from a clustered latent space conditioned on covariates (see Figure 5.1(c)).

We propose a model for time-to-event predictions equipped with a structured latent representation that allows for clustering via a prior for infinite mixture of

distributions. We circumvent the challenges associated with infinite mixtures in stochastic learning by leveraging a truncated Dirichlet process (DP) with a stick breaking representation. The proposed model, termed Survival Clustering Analysis (SCA), is specified as: *i*) a deterministic encoder that maps covariates into a latent representation; *ii*) a stochastic time-to-event predictor that feeds from the latent representation; and *iii*) a distribution matching objective that encourages latent representations to behave as a mixture of distributions following a DP structure. This approach allows identification and analysis of phenotypically heterogeneous subpopulations. Our experiments demonstrate that SCA yields consistent improvements in predictive performance and cluster quality relative to existing methods.

5.2 Background

In a conventional time-to-event (survival analysis) setup, we are given N observations. Individual observation are described by triplets $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, l_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i = \{x_i, ..., x_p\}$ is a p-dimensional vector of covariates, t_i is the time-to-event and $l_i \in \{0, 1\}$ is the censoring indicator. When $l_i = 0$ (censored) the subject has not experienced an event up to time t_i, while $l_i = 1$ indicates observed (ground truth) event times.

Time-to-event models are conditional on covariates: the event time density function $f(t|\boldsymbol{x})$, the hazard rate (risk score) function $\lambda(t|\boldsymbol{x})$ or the survival function $S(t|\boldsymbol{x}) = \ P(T > t) = 1 - F(t|\boldsymbol{x})$, also known as the probability of failure occurring after time t, where $F(t|\boldsymbol{x})$ is the cumulative density function. From standard survival function definitions [KK10], the relationship between these three characterizations is formulated as $f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x})S(t|\boldsymbol{x})$.

In practice, modern (often large) datasets are not homogeneous but composed of phenotipically heterogeneous subpopulations, i.e., subsets of observations that cluster according to both covariates and time-to-event similarities. In a clinical setting for

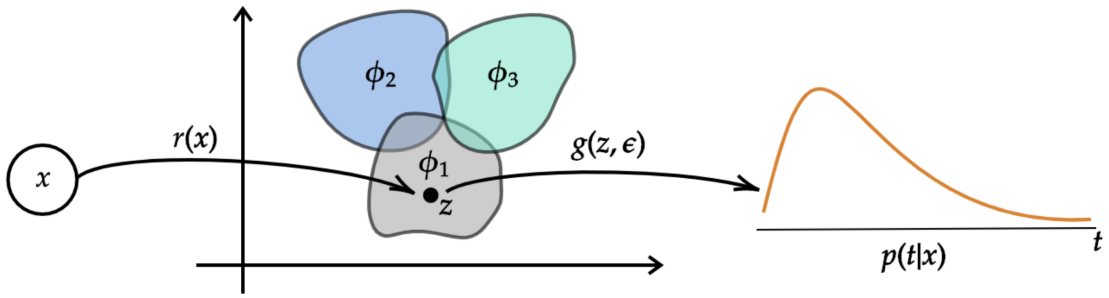


FIGURE 5.2: Illustration of Survival Clustering Analysis (SCA). The latent space has a mixture-of-distributions structure, illustrated as three mixture components $\{\phi_k\}_{k=1}^3$. Observation $\boldsymbol{x}$ is mapped into its latent representation via a deterministic encoding $\boldsymbol{z} = r_{\boldsymbol{\psi}}(\boldsymbol{x})$ belonging to ϕ_1, which is then used to stochastically predict (via sampling) the time-to-event via $t = g_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{\epsilon})$.

instance, identification of, e.g., high-, medium- and low-risk subpopulations that are equipped with accurate estimates of time-to-event has the potential to result in a more cost effective way of targeting interventions, treatments or care delivery. We formulate an approach to jointly learn individualized time-to-event distributions and clusters informed by time-to-event profiles.

5.3 Survival Cluster Analysis

The Bayesian nonparametrics approach formulated below encourages latent representations to behave as a mixture of distributions, following a Dirichlet Process (DP) structure via a distribution matching approach. Further, we learn to cluster the latent space in a stochastic manner for which the number of clusters is unknown. To demonstrate the efficacy of our clustering algorithm, we also present a time-to-event prediction formulation, leveraging current state-of-the-art time-to-event prediction models. See the Supplementary Material for the list of variable definitions used in our formulation.

5.3.1 Clustering with Dirichlet Process

A DP is formally defined as $G \sim \mathrm{DP}(\gamma_o, G_o)$ and parametrized by the base probability measure G_o and concentration parameter $\gamma_o > 0$ [Fer73]. With probability one [Set94]:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \qquad \pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l)\,, \tag{5.1}$$

where $\phi_k \sim G_o$, δ_{ϕ_k} represents a probability measure concentrated at ϕ_k and $V_k \sim \mathrm{Beta}(1, \gamma_o)$ are stick-breaking weights with statistics that depend on parameter γ_0. The sequence $\boldsymbol{\pi} = \{\pi_k\}_{i=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$, with probability one, such that $\boldsymbol{\pi} \sim \mathrm{GEM}(\gamma_o)$ [Pit02]. Further, note that π_k represents the likelihood that $G = \phi_k$.

Discrete distribution G is suitable as a prior for mixture components in infinite mixture models [Ras00]. Further, the stick-breaking process [Set94] that generates $\boldsymbol{\pi}$ results in a mechanism that allows one to learn the number of mixture components (clusters) from data. In fact, the number of distinct atoms, $\{\phi_k\}_{k=1}^{\infty}$, has been shown to grow with the size of the data as $O(\log N)$ [ASR88]. So motivated, we assume that data embedded in a latent space are distributed according to a mixture of distributions with parameters specified by the base probability measure G_o, as described below.

Assuming exchangeable latent representations $\{\boldsymbol{z}\}_{i=1}^{N}$, we propose generating event times following the generative process below

$$p(\boldsymbol{c}) = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{c}_k} \tag{5.2}$$

$$\boldsymbol{z}_n \sim st(\boldsymbol{c}_{u_n}, \nu) \tag{5.3}$$

$$t_n \sim g_{\boldsymbol{\theta}}(\boldsymbol{z}_n, \boldsymbol{\epsilon}_n)\,, \tag{5.4}$$

where $g_{\boldsymbol{\theta}}(\boldsymbol{z}, \epsilon)$ is a function that implicitly represents the conditional time-to-event density, $f(t|\boldsymbol{x})$, specified as a neural network with parameters $\boldsymbol{\theta}$. The source of

stochasticity, $\boldsymbol{\epsilon}$, for $g_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{\epsilon})$, is set to a simple distribution $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}$, e.g., uniform or Gaussian. The latent representation for the n-th observation, $\boldsymbol{z}_n$ is distributed according to $\phi_{u_n} = st(\boldsymbol{c}_{u_n}, \nu)$, where u_n is the mixture component membership indicator for $\boldsymbol{z}_n$. Lastly, together with (5.3), $p(\boldsymbol{c})$ in (5.2) represents an infinite mixture of Student's t-distributions with ν degrees of freedom and means $\{\boldsymbol{c}_k\}_{k=1}^{\infty}$, each of which is drawn independently from the base probability measure G_o as $\boldsymbol{c}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The Student's t distribution in (5.3) is a general yet parametrically simple distribution, robust to outliers and amenable to efficient computations and gradient estimates. It has been widely used in machine learning for mixture modeling [Ras00], clustering [XGF16] and visualization [MH08]. Further, we formulate the t distribution according to the normal-inverse-gamma likelihood, where marginalizing out the variance yields a Student-t distribution, see [Bis06] for details. Interestingly, as special cases, when $\nu = 1$, $\boldsymbol{z}_n$ is Cauchy distributed while for $\nu > 3$, $\boldsymbol{z}_n$ approaches a Gaussian distribution.

The generative process above further requires learning a mapping function from covariates to latent space, $\boldsymbol{z}_n = r_{\boldsymbol{\psi}}(\boldsymbol{x}_n)$ with parameters $\boldsymbol{\psi}$, that is globally consistent with the mixture model prior in (5.2) and (5.3), parameterized by $\{\pi_k, \boldsymbol{c}_k\}_{k=1}^{\infty}$. In addition, we also need to learn the parameters $\boldsymbol{\theta}$ of the time-to-event generating function $g_{\boldsymbol{\theta}}(\boldsymbol{z}_n, \epsilon_n)$ in (5.4). This specification, illustrated in Figure 5.2, constitutes the proposed Survival Clustering Analysis (SCA).

Note that unlike existing unsupervised and supervised autoencoding approaches [VLL$^+$10, KW14, JZT$^+$17], we do not seek to model the covariates, $\boldsymbol{x}$. Rather, we make time-to-event predictions based on a latent representation specified as a function of observed covariates, required to be consistent with a mixture of distributions prior. Consequently, we need not specify a decoding arm to reconstruct the covariates, $\boldsymbol{x}$.

In practice, learning the mixture component assignments u_n and a potentially

infinite number of mixture components with minibatches (stochastically) is challenging, because the former constitutes a discrete random variable and the latter requires keeping track of the number of non-empty mixture components during learning. To circumvent this, we learn the mixture component assignments probabilistically as $q(u_n = k|\boldsymbol{x}_n)$, and use a truncated representation of the DP formulation [IJ01, BJ$^+$06], which for large enough truncation number, denoted as K, is virtually indistinguishable from a standard DP [IJ01].

5.3.2 Latent-Space Representation

Following the conventional maximum likelihood formulation for mixture models [Bis06], we can approximate the distributions for the mixture assignments and mixture proportions as follows

$$\begin{aligned} q(u_n = k|\boldsymbol{x}_n) &= \frac{\alpha_{nk}}{\sum_{k=1}^{K} \alpha_{nk}} \\ \alpha_{nk} &\propto \pi_k p(r_{\boldsymbol{\psi}}(\boldsymbol{x}_n)|\boldsymbol{c}_k, \nu) \\ q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M}) &= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\xi}) \\ \xi_k &= \frac{1}{K} + \sum_{n=1}^{M} q(u_n = k|\boldsymbol{x}_n)\,, \end{aligned} \tag{5.5}$$

where $\boldsymbol{\xi} = \{\xi_k\}_{k=1}^{K}$ is a variational parameter for expectation $\mathbb{E}[\boldsymbol{\pi}]$, M is the minibatch size and we have replaced $\boldsymbol{z}_n$ in (5.3) with the encoding of covariates into latent space, i.e., $\boldsymbol{z}_n = r_{\boldsymbol{\psi}}(\boldsymbol{x}_n)$. However, (5.5) is not necessarily consistent with the DP in

(5.2) and its stick-breaking prior, $\boldsymbol{\pi} \sim \text{GEM}(\gamma_0)$, which from (5.1) should result in

$$
\begin{aligned}
p(u_n = k|\boldsymbol{x}_n) &= \frac{\beta_{nk}}{\sum_{k=1}^{K} \beta_{nk}} \\
\beta_{nk} &\propto V_k \prod_{l=1}^{k-1} (1 - V_l) p(r_{\boldsymbol{\psi}}(\boldsymbol{x}_n)|\boldsymbol{c}_k, \nu) \\
p(\boldsymbol{\pi}|\boldsymbol{\gamma}, \{\boldsymbol{x}_n\}_{n=1}^{M}) &= \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\gamma}) \\
\gamma_k &= \gamma_0 + \sum_{n=1}^{M} p(u_n = k|\boldsymbol{x}_n)\,,
\end{aligned}
\tag{5.6}
$$

where $V_k \sim \text{Beta}(1, \gamma_0)$, which in practice is complicated by the need to sample from the mixture proportion weights $\{V_k\}_{k=1}^{K}$. In our implementation, instead of sampling from V_k, we use its expectation, i.e., $\mathbb{E}[V_k] = (1 + \gamma_0)^{-1}$. Alternatively, we could also use a reparameterizable distribution such as the Kumaraswamy distribution, which is closely related to the Beta distribution as in [NS17]. However, we found that using expectations, which is common in variational formulations [BJ+06, JGJS99], works well in practice.

In order to make $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M})$ in (5.5) and $p(\boldsymbol{\pi}|\boldsymbol{\gamma}, \{\boldsymbol{x}_n\}_{n=1}^{M})$ in (5.6) consistent, we want their distributions to match, i.e., we seek to learn $\boldsymbol{\psi}$ of $\boldsymbol{z}_n = r_{\psi}(\boldsymbol{x}_n)$, $\{\pi_k\}_{k=1}^{K}$ and $\{\boldsymbol{c}_k\}_{k=1}^{K}$, so the approximation $q(\boldsymbol{\pi}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ matches the desired stick breaking behavior of (5.6). For this purpose, we minimize

$$
\begin{aligned}
&\ell_{\text{dp}}(\boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^{K}; \mathcal{D}) = \\
&\qquad \text{KL}\left(q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_n\}_{n=1}^{M}) || p(\boldsymbol{\pi}|\boldsymbol{\gamma}, \{\boldsymbol{x}_n\}_{n=1}^{M})\right)\,.
\end{aligned}
\tag{5.7}
$$

The KL Divergence between the two Dirichlet distributions q and p with respect to their corresponding parameters ξ and γ, has a desirable closed form formulation

defined as

$$\mathrm{KL}\left(q||p\right)) = \ln\Gamma(\xi_0) - \ln\Gamma(\gamma_0) + \sum_{k=1}^{K}\left(\ln\Gamma(\gamma_k) - \ln\Gamma(\xi_k)\right)$$

$$+ \sum_{k=1}^{K}(\xi_k - \gamma_k)\left(\Phi(\xi_k) - \Phi(\xi_0)\right) , \tag{5.8}$$

where $\xi_0 = \sum_{k=1}^{K}\xi_k$, $\gamma_0 = \sum_{k=1}^{K}\gamma_k$, $\Phi(\cdot)$ is the *digamma function* and $\Gamma(\cdot)$ is the *Gamma function*.

This loss function is used during learning to update $\boldsymbol{\psi}$ and $\{\boldsymbol{c}_k\}_{k=1}^{K}$. For $\{\pi_k\}_{k=1}^{K}$, the mixture proportions, we use a simple updating procedure akin to online expectation-maximization (EM) [CM09]. In particular, we update iteratively as

$$\boldsymbol{\pi}^{t+1} = \eta\boldsymbol{\pi}^{t} + (1-\eta)\mathbb{E}[q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_n\}_{n=1}^{M})], \tag{5.9}$$

where $0 < \eta < 1$ is the step size and we initialize $\boldsymbol{\pi}_k^0 = 1/K$. In practice, we set $\eta = 0.9$; however, η can also be selected using grid search. The online approach in (5.9) is widely used to update global parameters in stochastic learning procedures. For instance, it has been used to learn the population mean and variance in batch normalization [IS15].

5.3.3 Time-to-Event Distributions

In addition to the clustered, mixture representation of the latent space, we also seek a high-performing time-to-event model that yields concentrated, accurate and calibrated time-to-event predictions, while accounting for censored event times ($l_n = 0$). We borrow the accuracy objective from DATE [CTL$^+$18] (detailed in Chapter 2) and the calibration objective from SFM [CTL$^+$20] (detailed in Chapter 3). Below we describe these objectives in the context of our formulation.

Accuracy Objective The dataset $\mathcal{D}$ is split into two disjoint sets $(t, \boldsymbol{x}) \sim p_c$ and $(t, \boldsymbol{x}) \sim p_{nc}$, where p_c and p_{nc} represent censored and non-censored empirical distri-

butions for these sets, respectively. We leverage the accuracy objective from DATE [CTL+18] formulated as

$$\ell_{\text{acc}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) = \mathbb{E}_{(t,\boldsymbol{x})\sim p_c, \boldsymbol{\epsilon}\sim p_\epsilon}[\max(0, t - g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon}))] + \mathbb{E}_{(t,\boldsymbol{x})\sim p_{nc}, \boldsymbol{\epsilon}\sim p_\epsilon}[||t - g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon})||_1], \tag{5.10}$$

where $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}$ has a simple distribution (uniform or Gaussian). $\ell_{\text{acc}}(\boldsymbol{\theta}; \mathcal{D})$ encourages that time-to-event samples from the model, evaluated on censored observations, $l_n = 0$, are larger than the censoring time, while close to the ground truth for non-censored (observed) events, $l_n = 1$.

Calibration Objective We desire that samples generated from the model $g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon})$ match the empirical marginal distribution $p(t)$. We borrow the calibration objective from SFM [CTL+20] defined over the set of distinct and ordered observed event times (censored and non-censored), $\mathcal{T} = \{t_i | t_i > t_{i-1} > \ldots > t_0\}$,

$$\ell_{\text{cal}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) = \frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \left\| \hat{S}^{p(t)}_{\text{PKM}}(t_i) - \hat{S}^{g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon})}_{\text{PKM}}(t_i) \right\|_1, \tag{5.11}$$

where $\hat{S}_{\text{PKM}}$ is formulated as:

$$\hat{S}_{\text{PKM}}(t_i) = \left(1 - \frac{\sum_{n:l_n=1} H(\hat{T}_n - t_{i-1}) - H(\hat{T}_n - t_i)}{M - \sum_{n=1}^{M} H(t_{i-1} - \hat{T}_n)}\right) \times \hat{S}_{\text{PKM}}(t_{i-1}), \tag{5.12}$$

and $H(b) = \frac{1}{2}(\text{sign}(b) + 1)$ is the Heaviside step function. When evaluating the objective, $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathcal{D})$ in (5.11), $\hat{T}_n$ is either a sample from the model, $\hat{T}_n = g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon})$, or an observed time $\hat{T}_n \sim p(t)$, for $\hat{S}^{g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}), \boldsymbol{\epsilon})}_{\text{PKM}}(t_i)$ or $\hat{S}^{p(t)}_{\text{PKM}}(t_i)$, respectively. Expression $\hat{S}_{\text{PKM}}$ represents the point-estimate-based formulation of the Kaplan Meier estimator, see [CTL+20] for details.

5.3.4 Learning

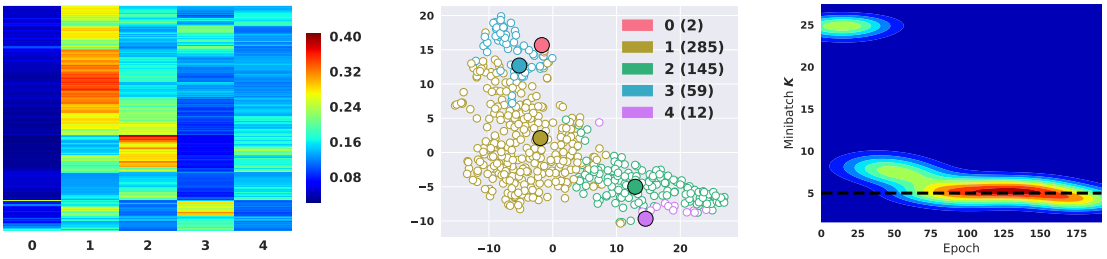
For joint learning of all model parameters, $\{\boldsymbol{c}_k\}_{k=1}^K$, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, we optimize both the latent representation and time-to-event (accuracy and calibration) objectives. The complete objective function for the proposed Survival Cluster Analysis (SCA) model is

$$\begin{aligned}\ell(\boldsymbol{\theta}, \boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^K; \mathcal{D}) = \ell_{\text{dp}}(\boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^K; \mathcal{D}) \\ + \lambda_2 \ell_{\text{acc}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) + \lambda_3 \ell_{\text{cal}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) \,,\end{aligned} \tag{5.13}$$

where $\lambda_2, \lambda_3 > 0$ are hyper-parameters controlling the trade-off between accuracy and calibration objectives, relative to the clustering objective in (5.7). For simplicity and comparability with SFM, we set $\lambda_2 = \lambda_3 = 1$. The objective in (5.13) is optimized using stochastic gradient descent on minibatches from $\mathcal{D}$.

In practice, $\{\boldsymbol{c}_k\}$ is updated according to stochastic gradient descent by optimizing the KL objective (5.7), and is initialized with K-means after pretraining (5.13) without the clustering objective. During inference, we assign a new observation, $\boldsymbol{x}_\star$, to a cluster by first evaluating $q(u_\star = k|\boldsymbol{x}_\star)$ for $k = 1, \ldots, K$, then obtaining a hard assignment according to $u_\star = \text{argmax}_k \; q(u_\star = k|\boldsymbol{x}_\star)$.

The maximum number of mixture components K is fixed during learning. However, provided that the KL divergence (5.7) encourages mixture proportions to follow a stick-breaking process, the effective number of mixture components, i.e., those with non-empty observation assignments, will be smaller than K, thus effectively resulting in the model learning the number of mixture components. This is illustrated in Figure 5.3 and described below in the experiments. The number of degrees of freedom, ν is a hyperparameter, set to 1 in our experiments.



(a) Cluster assignments (b) t-SNE plot (c) Inferred minibatch K

FIGURE 5.3: Inferred clusters on the testing set of SLEEP dataset, with $K = 25$ and $\gamma_o = 3$ where: (a) corresponding individual probability distribution $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M})$, are approximated according to (5.5), (b) joint t-SNE plot of centroids $\boldsymbol{c}_k$ with latent representation $\boldsymbol{z}$ and (c) density plot of inferred number of clusters K during training.

Table 5.1: Summary statistics of the datasets used in the experiments. The time range, $t_{\max}$, is noted in days except for SEER for which time is measured in months.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
Events (%)	23.9	27.5	68.1	51.0	23.8	11.14
N	394,823	7,894	9,105	68,082	5026	40,078
d (categorical)	729 (106)	26 (21)	59 (31)	789 (771)	206	12 (8)
Missing (%)	1.9	2.1	12.6	23.4	18.2	0.33
$t_{\max}$	365	5,215	2,029	120	5,794	6,000

5.4 Experiments

The comparisons presented below are made across a diverse range of six datasets, as summarized in Table 5.1. Refer to the Supplementary Material for all details concerning the experimental setup. Throughout the experiments, we set $K = 25$ and select $\gamma_o = \{2, 3, 4, 8\}$ via grid search cross-validation from the training sets. TensorFlow code to replicate experiments can be found at `https://github.com/paidamoyo/survival_cluster_analysis`.

Datasets Table 5.1 shows the summary statistics of the six datasets considered. The datasets are diverse in number of observations N, varying amounts of categorical (cat) and continuous covariates d, proportions of non-censored events, missingness rates in the $N \times d$ covariate matrix, and time horizon $t_{\max}$ (measured in days, ex-

cept for SEER which is measured in months). Following information-theoretic data processing inequality conclusions from [MPER18], demonstrating insignificant performance change relative to pre-imputation, we impute continuous and categorical covariates with the median and mode, respectively. In our experiments we do not convert time to a common scale but model it as is.

Publicly accessible datasets include: *i*) FLCHAIN: a study of non-clonal serum immunoglobin free light chains effects on survival time [DKK+12]. *ii*) SUPPORT: investigates the survival time of seriously-ill hospitalized adults [KHL+95]. *iii*) SEER: accessible from the Surveillance, Epidemiology, and End Results (SEER) Program. The dataset is preprocessed according to the details described in [RYJK+07]. We restrict the dataset to a 10-year follow-up breast cancer subcohort.

The following datasets are available upon request: *iv*) EHR: a large study from the Duke University Health System centered around multiple inpatient visits due to comorbidities in patients with Type-2 diabetes [CTL+18]. *v*) SLEEP: a subset of the Sleep Heart Health Study (SHHS) [QHI+97], a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing. We focus on the baseline clinical visit and aggregated demographics, medications and questionnaire data as covariates. *vi*) FRAMINGHAM: a subset (Framingham Offspring) of the longitudinal study of heart disease [BLV+94] dataset, initially for predicting 10-year risk for future coronary heart disease (CHD).

Clustering Baselines We consider the standard K-means and CoxPH based SSC-Bair [BT04] as strong clustering baselines for SCA. We provide quantitative evaluations in terms of the logrank score [Man66], and qualitative visualization of the clustering-based Kaplan-Meier sub-population survival curves.

Table 5.2: Inferred cluster specific covariate information on the testing set for the FRAMINGHAM dataset. The inferred cluster assignments are according to the corresponding individual probability distribution $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M})$, approximated according to (5.5). Ranges in parentheses are 50% empirical ranges over (median) test-set predictions for the continuous and proportions for categorical covariates.

Covariates	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6
Continous							
Age	$56_{(48,62)}$	$50_{(43,58)}$	$59_{(52,63)}$	$55_{(48,61)}$	$47_{(35,54)}$	$55_{(49,62)}$	$58_{(50,65)}$
HDL (mg/dL)	$43_{(37,53)}$	$52_{(44,63)}$	$67_{(59,85)}$	$54_{(45,66)}$	$62_{(55,70)}$	$41_{(35,48)}$	$42_{(36,52)}$
Total Cholesterol	$198_{(193,207)}$	$176_{(168,183)}$	$266_{(250,285)}$	$220_{(207,236)}$	$148_{(138,157)}$	$251_{(235,275)}$	$173_{(158,188)}$
Systolic Blood Pressure	$126_{(117,137)}$	$110_{(102,119)}$	$141_{(130,153)}$	$115_{(106,125)}$	$110_{(102,117)}$	$126_{(115,139)}$	$132_{(120,147)}$
Categorical							
Hypertension medication (Yes)	25.5%	4.97%	40.1%	11.3%	1.1%	41.6%	41.0%
Diabetic (Yes)	6.9%	2.63%	3.0%	3.3%	0.0%	20.8%	16.7%
Gender (Female)	36.9%	82.5%	63.6%	69.6%	74.5%	33.4%	36.4%
Current smoker (Yes)	23.9%	14.6%	28.0%	16.6%	22.3%	45.6%	25.1%
Race (Black)	16.4%	3.5%	29.5%	1.5%	8.5%	21.7%	27.7%
Race (Chinese)	4.2%	2.6%	0.0%	1.5%	2.1%	1.1%	3.0%
Race (Hispanic)	5.0%	2.3%	2.3%	2.3%	1.0%	4.0%	5.0%
Race (White)	74.4%	91.5%	68.2%	85.7%	88.3%	73.2%	64.2%

Time-to-Event Baselines We compare SCA to the following time-to-event baselines: SFM [CTL$^+$20], DATE [CTL$^+$18], S-CRPS [ADZ$^+$20], CoxPH [Cox92], MTLR [YGLB11] and DRAFT [CTL$^+$18]. From these, SFM and DATE are key to our comparisons because we leverage components from their formulation into SCA; namely, the accuracy loss from DATE and the distribution matching loss from SFM. In that sense, we expect SCA to perform as good as SFM and DATE, but with the added benefit of producing clusters with distinct risk profiles. We present quantitative evaluations in terms of C-index, Calibration slope, Relative Absolute Error (RAE), and mean Coefficient of Variation (CoV). Details of these metrics are provided in the Supplementary Material.

5.4.1 Qualitative Results

Figure 5.3 shows for the SLEEP dataset *a*) estimated individualized cluster assignment probability distributions (rows) evaluated according to (5.5); *b*) t-SNE plots of the model inferred centroids, $\boldsymbol{c}_k$, as well as the individual latent representation $\boldsymbol{z} = r_{\boldsymbol{\psi}}(\boldsymbol{x})$; and *c*) density plot of the inferred number of (non-empty) clusters K

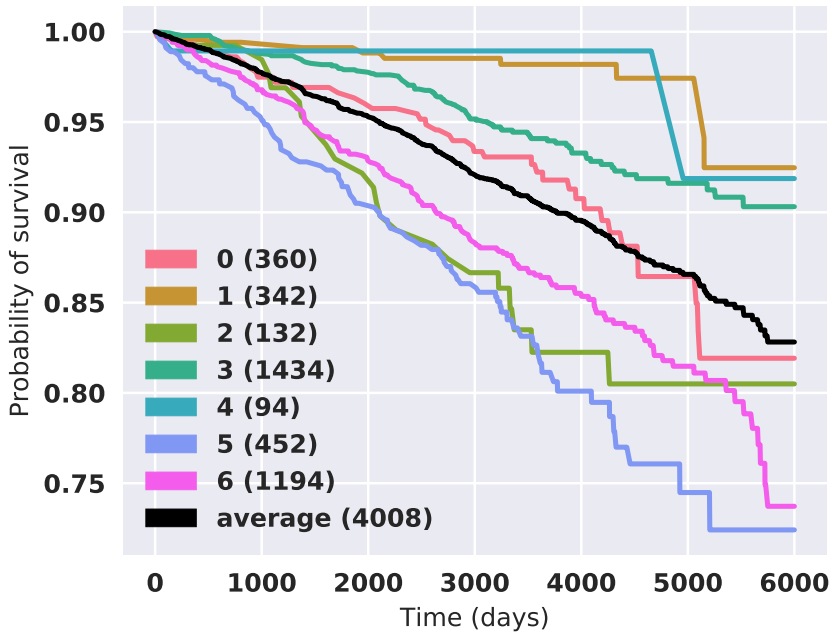


FIGURE 5.4: Inferred Cluster specific Kaplan-Meir Curves on the testing set of FRAMINGHAM dataset, with $K = 25$ and $\gamma_o = 8$. The inferred clusters assignment is according to the corresponding individual probability distribution $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^M)$, approximated according to (5.5).

during training. See the Supplementary Material for similar figures for all the other datasets, where we also include corresponding Kaplan-Meier curves, as in Figure 5.1.

Interestingly, the cluster-specific covariate statistics for the FRAMINGHAM dataset, which has the least number of covariates, are shown in Table 5.2 and are consistent with findings from the Framingham Heart Study [BLV+94], which identified high blood cholesterol and high blood pressure as major risk factors for cardiovascular disease.

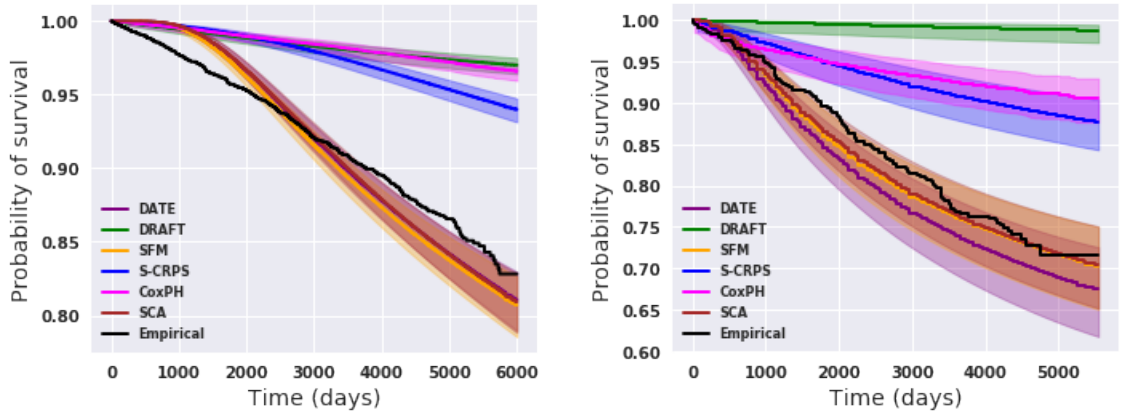
We obtain the cluster specific Kaplan-Meir curves illustrated in Figure 5.4 with corresponding cluster specific covariate information shown in Table 5.2. The inferred individual cluster assignment is obtained according to the individual probability distribution $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^M)$, approximated according to (5.5). We consider curves above the population average low-risk while the curves below to be high-risk.

Therefore, our model identifies three high-risk clusters, indexed by 2, 5, 6: i) clus-

ter 6 and cluster 5 have similar statistics, as they both consists disproportionately of diabetic individuals on hypertension medication with elevated total cholesterol (normal is below 200), high systolic blood pressure (normal is below 120), and noticeably low HDL (normal is greater than 60); *ii*) cluster 2, is also driven by age, which is expected, where about 40% of the population is on hypertension medication, and with the worst systolic blood pressure and cholesterol compared to other clusters; *iii*) lower-risk clusters 1, 3, 4 are mostly comprised of females with normal levels of HDL, total cholesterol and systolic blood pressures; *iv*) cluster 0 represents the average statistics of the Framingham dataset, thus the survival curves directly follows the empirical population survival. Finally, note that the three high-risk clusters (2, 5 and 6) have a substantial over-representation of African Americans, known to have an increased risk for cardiovascular disease [BLV+94]. See the Supplementary Material for additional inferred cluster specific Kaplan-Meir curves on all datasets.

We demonstrate that by jointly learning clustering with respect to both the covariates $\boldsymbol{x}$ and predicted time-to-event t, our model SCA can identify high-, medium- and low-risk individuals , which is essential for clinical decision making. During inference, both the risk profile and individualized time-to-event can provide a comprehensive prediction mechanism for identifying cluster-based risk factors, cluster-based risk profiles and individualized time predictions. Further, the advantage of matching the empirical mixture distribution with a (truncated) DP yields sparse predictions of cluster assignment probabilities, $q(\boldsymbol{\pi}|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M})$, manifested as high confidence cluster assignments illustrated as a heatmap Figure 5.3(a).

Calibration Curves We visually compare calibration curves from DATE, DRAFT, SCA, SFM, S-CRPS and CoxPH. Figure 5.5 shows the estimated populations-based model survival functions according to [CTL+20] and empirical Kaplan-Meier for the FRAMINGHAM and SLEEP datasets. Error bars (shaded area) are calculated accord-



(a) FRAMINGHAM (b) SLEEP

FIGURE 5.5: Survival function estimates for (a) FRAMINGHAM and (b) SLEEP data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CoxPH). Error bars (shaded area) are calculated according to the Greenwood's formula [G$^+$26].

ing to the Greenwood's formula [G$^+$26]. For all datasets, SCA- and SFM-estimated population survival functions closely match the empirical ground truth survival function, which is consistent with the high calibration slopes results in Table 5.3. See the Supplementary Material for additional calibration and survival function results on all datasets.

5.4.2 Quantitative Results

Below we describe performance metrics across all datasets and models. Specifically, calibration slope, mean CoV (coefficient of variation), C-index [HJLC$^+$84] and Relative Absolute Error [RAE, YGLB11] provide a comprehensive evaluation, as they offer insights into consistency of time-to-event predictions, concentration of predicted distributions, pairwise ranking consistency, and accuracy of event time predictions, respectively. The results demonstrate that by jointly modeling the time-to-event and cluster assignments we obtain a better calibrated model, that is competitive in C-index, concentrated and accuracy of predictions. Table 5.3 shows the calibration slopes and RAE across all datasets and models. See the Supplementary Material for

Table 5.3: Calibration slope and RAE metrics on test data.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
Calibration slope						
DATE	0.7537	0.9668	0.9068	0.9161	0.9454	0.7737
DRAFT	3.2138	5.4183	2.9640	2.0763	25.2855	5.7345
S-CRPS	1.6246	1.9662	1.1795	1.1613	2.5746	2.6114
CoxPH	2.5543	1.9116	1.3909	1.4358	3.8278	4.9945
MTLR	2.1957	1.9449	1.2017	1.2476	2.4792	5.4878
SFM	0.7734	0.9807	0.9405	**0.9540**	1.0235	0.7626
SCA (proposed)	**0.8006**	**0.9900**	**1.0086**	0.9290	**1.0223**	**0.8044**
RAE (non-censored)						
DATE	**0.6107**	0.5222	0.6691	0.5289	0.5224	0.5122
DRAFT	0.7099	0.6399	0.7109	0.6097	0.7465	0.6697
S-CRPS	0.7240	0.6378	**0.4851**	0.5323	0.7330	0.8369
CoxPH	-	-	-	-	-	-
MTLR	-	-	-	-	-	-
SFM	0.6146	**0.5111**	0.6398	0.5294	**0.5162**	**0.5074**
SCA (proposed)	0.6186	0.5134	0.6295	**0.5193**	0.5424	**0.5074**

Table 5.4: Logrank score and standard errors in parentheses. The best performing K-means and SSC-Bair models were selected from the set $K = \{2, 3, 4, 5, 6\}$.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
SCA	-	**278.49 (0.0)**	496.50 (0.0)	**4803.59 (0.0)**	**63.74 (0.0)**	123.31 (0.0)
SSC-Bair	409.93 (0.0)	4.27 (0.37)	**1204**.14 (0.0)	4084.78 (0.0)	20.37 (0.0)	**125.87** (0.0)
K-means	417.00 (0.0)	5.33 (0.38)	99.06 (0.0)	3985.67 (0.0)	21.83 (0.0)	88.87 (0.0)

detailed RAE, mean CoV and C-index results.

Table 5.4 presents the clustering performances of the best performing K-means, SSC-Bair and SCA algorithms, measured in terms of the logrank score [Man66], for SSC-Bair and K-means we selected the best performing model from the set $K = \{2, 3, 4, 5, 6\}$.

Calibration Slope For calibration we use the framework developed in SFM to evaluate the models [CTL+20]. An ideal calibration slope is 1, while a slope < 1 and slope > 1 indicates whether the model tends to underestimate or overestimate risk, respectively. The clustering objective in SCA augments the calibration objective we borrow from SFM, thus improving the calibration even for non-iid observations, such

as FRAMINGHAM and EHR, which are considered poorly calibrated, as illustrated in SFM. Given that SCA leverages the calibration objective of SFM, it is not surprising that both SCA and SFM are competitive, followed by DATE, S-CRPS, MTLR, CoxPH and lastly DRAFT. See Supplementary Material for qualitative calibration plots.

Relative Absolute Error (RAE) We compute RAE for both censored and non-censored events. In Table 5.3 we present the RAE for non-censored event times ($l_n = 1$) for models that predict absolute event times, thus excluding scoring based models (CoxPH and MTLR). The results demonstrate that DATE, SFM and SCA (nonparametric) methods outperform DRAFT and S-CRPS (parametric) methods, which is expected since they all use a similar accuracy-aware objective function. For censored events ($l_n = 0$), RAE provides the lower bound error given the censored time provides tail information of $p(t|\boldsymbol{x})$; parametric methods (DRAFT and S-CRPS) have small advantage over nonparametric methods (SFM, DATE and SCA). See the Supplementary Material for additional results on censored event times.

Concordance Index (C-index) C-index is a ranking metric that does not account for uncertainty in time-to-event predictions. Therefore to evaluate the time-to-event models (except CoxPH) in terms of C-index, we use point summaries of the individualized time-to-event distributions, specifically, $\hat{t} = \text{median}(\{t_{ns}\}_{s=1}^{200})$, where t_{ns} is a sample from the trained model, $t_{ns} = g_{\boldsymbol{\theta}}(r_{\boldsymbol{\psi}}(\boldsymbol{x}_n), \boldsymbol{\epsilon}_s)$ on the test set. Apart from the small covariates, the very low event rate FRAMINGHAM dataset and the small high event rate SUPPORT dataset, none of the models have a clear advantage on the C-index metric. This is not surprising because C-index with very low event rate is heavily influenced by the censored observations. Note, for MTLR, although we can compute the C-index at prespecified thresholds, we are unable to compute a global

C-index.

Coefficient of Variation (CoV) Models that characterize the event time density function $f(t|\boldsymbol{x})$ result in uncertainty-aware time-to-event predictions. In practice, it is highly desirable for a model to generate concentrated time-to-event predictions. The CoV (coefficient of variation) measures the dispersion in a distribution; a Cov > 1 indicates high variance, while CoV < 1 indicates low variance distributions. Cov results provided in the Supplementary Material demonstrate that DATE, SCA and SFM are consistently low-variance distributions, followed by S-CRPS and lastly DRAFT. We cannot compute CoV for both MTLR and CoxPH. CoxPH estimates risk score, and therefore cannot be evaluated on CoV. MTLR does not specify the conditional hazards, $\lambda(t|\boldsymbol{x})$, and thus we cannot recover $f(t|\boldsymbol{x}) = S(t|\boldsymbol{x})\lambda(t|\boldsymbol{x})$.

Logrank Score The logrank score is a nonparametric statistic that evaluates the similarity between a pair of survival functions, yielding high values for curves that are highly unlikely to be similar [Man66]. Further, the logrank statistic is especially powerful for measuring differences between survival functions that follow the Cox proportional hazard assumption, i.e., the survival functions do not cross. For K clusters, we compute $\binom{K}{2}$ pairwise comparisons. Table 5.4 demonstrates that our proposed SCA is the best performing method, followed by SS-Bair and lastly K-means. Interestingly, SCA is unable to recover any clustering structure from the EHR dataset, as it is a homogeneous population of Type-2 diabetes subjects, whereas K-means and SSC-Bair are always able to produce clusters (which may be misleading for homogeneous datasets). This supports the need to account for survival information when clustering survival datasets, as both SCA and SSC-Bair incorporate time information in their clustering approaches.

5.5 Conclusions

We have developed the first time-to-event model for inferring individualized risk-based cluster assignments, while jointly predicting the time-to-event. Leveraging a Bayesian nonparametric stick-breaking representation of the Dirichlet Process, we have presented a method for learning a clustering structure in a latent representation, for which the number of clusters is unknown. We have demonstrated the need to account for time information when clustering survival datasets. Our model identifies interpretable and phenotyopically heterogeneous subpopulations, which are critical in a clinical setting for identifying subjects with diverse risk profiles. Extensive experiments demonstrate that the joint modeling approach yields substantial performance gains in calibration and logrank scores, while remaining competitive in preserving pairwise ordering, predicting concentrated and accurate distributions. In the future, we plan to extend this work to account for locally-consistent, calibrated and accurate predictions within identified subpopulations.

6

Conclusions

Machine learning research tailored for clinical decision-making has the potential to improve clinical care. This dissertation introduces nonparametric probabilistic time-to-event models that account for population calibration (also known as reliability [Daw82]) and uncertainty while predicting accurate absolute event times. Extensive experiments show that the proposed distribution matching methods [CTL+18, CTL+20] outperform existing approaches in terms of calibration and concentration of time-to-event distributions.

However, societal pitfalls associated with automated decision-making, i.e., exacerbating health disparities [VEJ20], need to be addressed before these models can be used safely in practice. Therefore, developing the methodology to address such critical safety issues, i.e., accounting for algorithmic bias and calibration is crucial. While this dissertation focuses on population calibration and uncertainty-aware predictions [CTL+20], extensions applicable to subpopulation calibration and non-i.i.d calibration remain under-explored.

For interpretability, we propose identifying subpopulations with distinct risk profiles [CLM+20], while jointly accounting for accurate individualized time-to-event

predictions. Therefore, we present a Bayesian nonparametrics approach that leverages regularities in subpopulations, thus accounting for population-level heterogeneity. The proposed approach: (i) represents observations (subjects) in a clustered latent space, for which the number of clusters is unknown; and (ii) encourages accurate time-to-event predictions and clusters (subpopulations) with distinct risk profiles. Further, we demonstrate the need to account for time information when clustering survival datasets. Experiments on real-world datasets show consistent improvements in predictive performance and interpretability relative to existing state-of-the-art survival analysis models. Extending this work to account for locally consistent, calibrated, and accurate predictions within identified subpopulations is left to future work.

In practice, we may be interested in the causal effect of a given intervention or treatment on survival time. However, counterfactual inference approaches that account for survival outcomes are relatively limited. When the outcome of interest is a time-to-event, special precautions for handling censored events need to be taken, as ignoring censored outcomes may lead to biased estimates. In this dissertation, we propose a theoretically grounded unified framework for counterfactual inference applicable to survival outcomes [CAZ+21]. The proposed approach adjusts for bias from two sources, namely, confounding (covariates influence both the treatment assignment and the outcome) and censoring (informative or noninformative).

Further, we formulate a nonparametric hazard ratio metric for evaluating average and individualized treatment effects [CAZ+21]. Experimental results on real-world and semi-synthetic datasets, the latter of which we introduce, demonstrate that the proposed approach significantly outperforms competitive alternatives in survival-outcome prediction and treatment-effect estimation. In future work, we plan to understand the sensitivity of our estimates to unobserved confounding [CHH+59] and the effect of both censoring bias and selection bias on causal identifiability.

Finally, this dissertation formulates probabilistic machine learning approaches for risk profiling from single event type observational data. Many real-world longitudinal datasets are recorded over time as asynchronous irregular event sequences, e.g., electronic health records. Fortunately, the machine learning approaches from this dissertation offer the strong foundation needed for advancing research focused on continuous-time modeling, specifically: (*i*) modeling complex structured observational event data, e.g., recurrent [SSC+18], related among individuals, multi-state [GSG20], and multivariate [ME17]; and (*ii*) exploring connections to statistical causality, e.g., time-varying confounding or treatment [Rob86, HBR00].

Appendix A

Supplemental Material for "Adversarial Time-to-Event Modeling"

A.1 Missing data and DATE-AE

DATE-AE extends DATE by jointly learning the mapping $\boldsymbol{x} \rightarrow \boldsymbol{z} \rightarrow t$, where $\boldsymbol{z}$ is modeled as an adversarial autoencoder. For imputation, the covariates (entries of $\boldsymbol{x}$) in the encoder are set to zero if the entry is missing. When evaluating the reconstruction loss γ_3 in (A.1), we only do so for observed covariates; in this way the autoencoder can learn the correlation structure of the observed data despite missingness and without the need for imputation, while letting the decoder, $\boldsymbol{x} = \text{decoder}(\boldsymbol{z})$, handle the imputation if needed. Note that for time-to-event prediction, at test time, we do not have to impute missing values as we can directly evaluate $\boldsymbol{x} \rightarrow \boldsymbol{z} \rightarrow t$. DATE-AE, extends DATE formulation with additional autoencoder

discriminator and generator losses shown below:

$$\begin{aligned}
\gamma_1(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}, \boldsymbol{\psi}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{x}, \tilde{\boldsymbol{z}})}[D_{\boldsymbol{\psi}}(\boldsymbol{x}, \tilde{\boldsymbol{z}})] \\
&\quad + \mathbb{E}_{(\tilde{\boldsymbol{x}}, \boldsymbol{z})}[1 - D_{\boldsymbol{\psi}}(\tilde{\boldsymbol{x}}, \boldsymbol{z})], \\
\gamma_2(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{z} \sim p(\boldsymbol{z}), \hat{\boldsymbol{z}})}[d(\boldsymbol{z}, \hat{\boldsymbol{z}})], \\
\gamma_3(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}; \mathcal{D}) &= \mathbb{E}_{(\boldsymbol{x} \sim p(\boldsymbol{x}), \hat{\boldsymbol{x}})}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})], \\
\min_{\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}} \max_{\boldsymbol{\psi}} \gamma(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}, \boldsymbol{\psi}; \mathcal{D}) &= \gamma_1(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}, \boldsymbol{\psi}; \mathcal{D}) \\
&\quad + \zeta_2 \gamma_2(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}; \mathcal{D}) \\
&\quad + \zeta_3 \gamma_3(\boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_{\boldsymbol{z}}; \mathcal{D}) \,, \qquad \text{(A.1)}
\end{aligned}$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$, $\tilde{\boldsymbol{z}} = G_{\boldsymbol{\theta}_{\boldsymbol{x}}}(\boldsymbol{x}, \boldsymbol{\epsilon}_{\boldsymbol{x}})$, $\boldsymbol{z} \sim p(\boldsymbol{z})$, $\tilde{\boldsymbol{x}} = G_{\boldsymbol{\theta}_{\boldsymbol{z}}}(\boldsymbol{z}, \boldsymbol{\epsilon}_{\boldsymbol{z}})$, $\boldsymbol{\epsilon}$ is the noise source, d is the distortion measure and $\{\zeta_2, \zeta_3\}$ are reconstruction tuning parameters.

Tables A.1 and A.2 compares the effects of randomly introducing missing values on the Flchain relative absolute error and concordance-index respectively.

A.2 Concordance index and relative absolute error

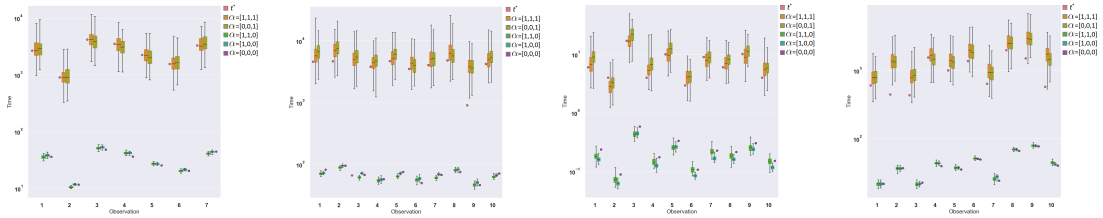
Tables A.3 and A.4 show comparisons on concordance-index and relative absolute error across all datasets.

A.3 Normalized Relative Error (NRE)

Figures A.2, A.3, A.4 and A.5, show comparison on NRE distributions for both censored and non-censored events.

A.4 Test set time-to-event distributions

We randomly draw best and worst observation samples based on the NRE metric. Figures A.6 , A.7, A.8 and A.9, show the corresponding distributions comparisons relative to the ground truth or censored time $t^{\star}$.



(a) Non-Censored (b) Censored (c) Non-Censored (d) Censored

FIGURE A.1: Effects of stochastic layers on uncertainty estimation on 10 randomly selected test-set subjects from the FLCHAIN ((a) and (b)) and SUPPORT ((c) and (d)) datasets. Ground truth times are denoted as t^* and box plots represent time-to-event distributions from a 2-layer model, where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]$ indicates whether the corresponding noise source, $\{\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$, is active. For example $\boldsymbol{\alpha} = [1, 0, 0]$ indicates noise on the input layer only.

A.5 Effects of noise source and stochastic layers

Figure A.1 shows the contribution effects of stochastic layers for noise Uniform(0,1) on both censored and non-censored time-to-event distributions. Tables A.5 and A.6 compares noise sources on relative absolute error and CI.

Table A.1: Introduced proportion of missing values comparison on Flchain relative absolute error. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	0.10	0.20	0.30	0.50
Non-Censored				
DATE	$19.9_{(9.6,32.7)}$	$\mathbf{19.8}_{(9.1,33.7)}$	$\mathbf{19.7}_{(10.8,33.2)}$	$19.7_{(10.3,33.5)}$
DATE-AE	$\mathbf{19.2}_{(9.6,34.9)}$	$21.9_{((9.5,33.4)}$	$20.6_{(9.7,32.8)}$	$\mathbf{18.3}_{(9.5,32.9)}$
DRAFT	$32.9_{(10.0,92.3)}$	$34.1_{(11.5,119.8)}$	$\mathbf{19.7}_{(10.3,33.5)}$	$19.7_{(10.3,33.5)}$
Censored				
DATE	$\mathbf{0}_{(0,20.4)}$	$1.9_{(0,19.4)}$	$2.7_{(0,20.1)}$	$7.3_{(0,21.8)}$
DATE-AE	$\mathbf{0}_{(0,12.9)}$	$3_{(0,19)}$	$\mathbf{2.1}_{(0,16.5)}$	$\mathbf{6}_{(0,21.3)}$
DRAFT	$\mathbf{0}_{(0,0)}$	$\mathbf{0}_{(0,0)}$	$7.3_{(0,21.8)}$	$7.3_{(0,21.8)}$

A.6 Architecture of the neural network

In all experiments, DATE and DRAFT are specified in terms of two-layer MLPs of 50 hidden units with Rectified Linear Unit (ReLU) activation functions and batch normalization [IS15]. The discriminator for DATE is a similarly defined MLP. As an

Table A.2: Introduced proportion of missing values comparison on FLCHAIN Concordance-Index.

	0.10	0.20	0.30	0.50
DATE	0.815	0.803	**0.803**	0.784
DATE-AE	0.814	0.804	0.799	**0.785**
DRAFT	**0.822**	**0.807**	0.801	0.783

Table A.3: Concordance-Index results on test data.

	DATE	DATE-AE	DRAFT	Cox-Efron	RSF
EHR	**0.78**	**0.78**	0.76	0.75	–
FLCHAIN	**0.83**	**0.83**	**0.83**	**0.83**	0.82
SUPPORT	0.84	0.83	**0.86**	0.84	0.80
SEER	**0.83**	**0.83**	**0.83**	0.82	0.82

Table A.4: Median relative absolute errors (as percentages of $t_{\max}$), on non-censored and censored data. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	DATE	DATE-AE	DRAFT
Non-censored			
EHR	$\mathbf{23.6}_{(11.1,43.0)}$	$24.5_{(12.4,44.0)}$	$36.7_{(16.1,81.3)}$
FLCHAIN	$19.5_{(9.5,31.1)}$	$\mathbf{19.3}_{(8.9,32.4)}$	$26.2_{(9.0,53.5)}$
SUPPORT	$2.7_{(0.4,16.1)}$	$\mathbf{1.5}_{(0.4,19.2)}$	$2.0_{(0.2,35.3)}$
SEER	$\mathbf{18.6}_{(8.3,34.1)}$	$20.2_{(10.3,35.8)}$	$23.7_{(9.9,51.2)}$
Censored			
EHR	$12.4_{(0,38.7)}$	$1.6_{(0,34.)}$	$\mathbf{0}_{(0,0)}$
FLCHAIN	$0_{(0,18.8)}$	$0_{(0,15.6)}$	$\mathbf{0}_{(0,0)}$
SUPPORT	$0_{(0,13.0)}$	$0_{(0,8.8)}$	$\mathbf{0}_{(0,0)}$
SEER	$\mathbf{0}_{(0,0)}$	$\mathbf{0}_{(0,0)}$	$\mathbf{0}_{(0,0)}$

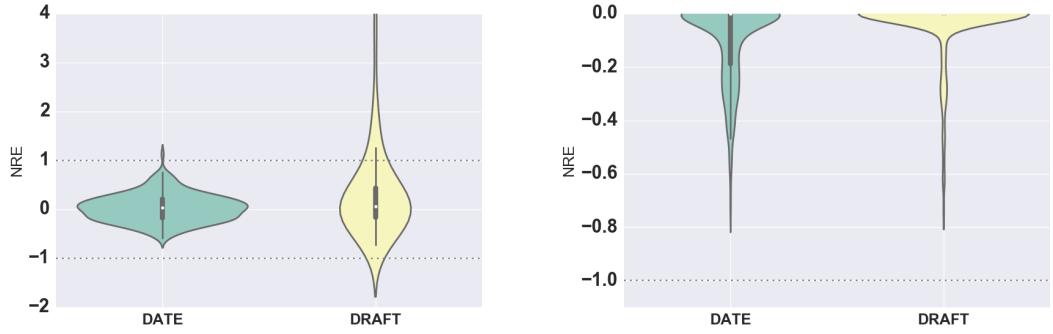
optimizer, we use Adam [KA15] with the following hyperparameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and epsilon 1×10^{-8}. Further, we set the minibatch size to $M = 350$ and use dropout with $p = 0.8$ on all layers. All the network weights are initialized using *Xavier* [GB10]. Datasets are split into training, validation and test sets as 80%, 10% and 10% partitions, respectively, stratified by non-censored event proportion. We use the validation set for early stopping and

Table A.5: Effects of noise source and stochastic layers on SUPPORT Median relative absolute error. Ranges in parentheses are 50% empirical ranges over (median) test-set predictions.

	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
Non-censored			
All	$2.4_{(0.4,19.9)}$	$2.2_{(0.5,19.2)}$	$1.9_{(0.4,17.)}$
Input	$2.2_{(0.4,18.)}$	$\mathbf{1.8}_{(0.4,16.1)}$	$1.9_{(0.4,14.9)}$
Output		$2.6_{(0.4,21.1)}$	
Censored			
All	$\mathbf{0}_{(0,14.6)}$	$\mathbf{0}_{(0,13.7)}$	$\mathbf{0}_{(0,16.4)}$
Input	$\mathbf{0}_{(0,15.3)}$	$1.2_{(0,22.4)}$	$0.8_{(0,21.2)}$
Output		$\mathbf{0}_{(0,8.2)}$	

Table A.6: Effects of noise source and stochastic layers on SUPPORT concordance-index.

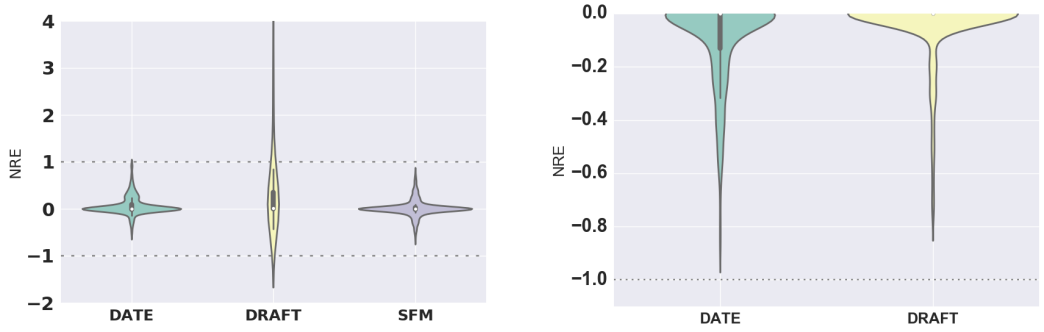
	Uniform(-1,1)	Uniform(0,1)	Gaussian(0,1)
All	0.825	**0.835**	0.826
Input	0.841	0.829	0.825
Output		**0.836**	



(a) Non-Censored (b) Censored

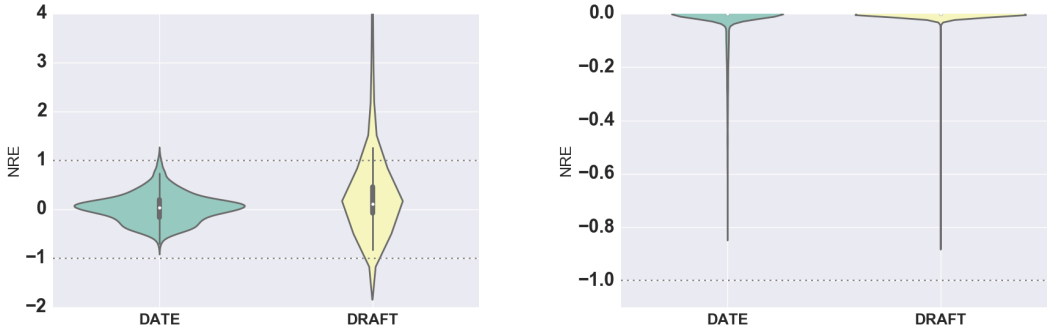
FIGURE A.2: Normalized relative error on FLCHAIN test data.

learning model hyperparameters. DATE is executed using one NVIDIA P100 GPU with 16GB memory.



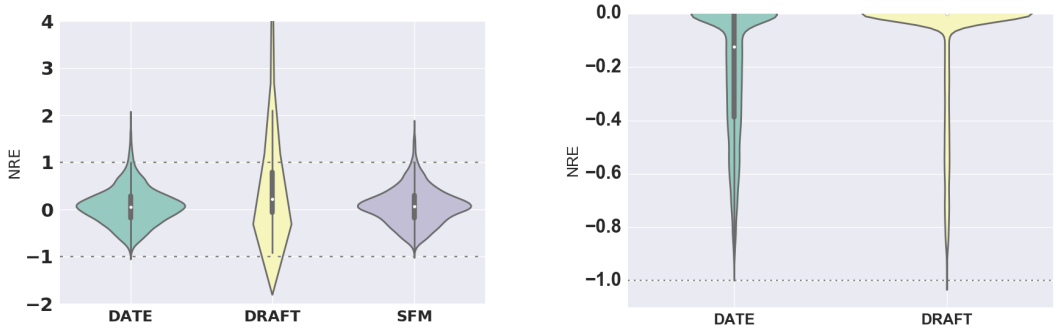
(a) Non-Censored (b) Censored

FIGURE A.3: Normalized relative error on SUPPORT test data.



(a) Non-Censored (b) Censored

FIGURE A.4: Normalized relative error on SEER test data.



(a) Non-Censored (b) Censored

FIGURE A.5: Normalized relative error on EHR test data.

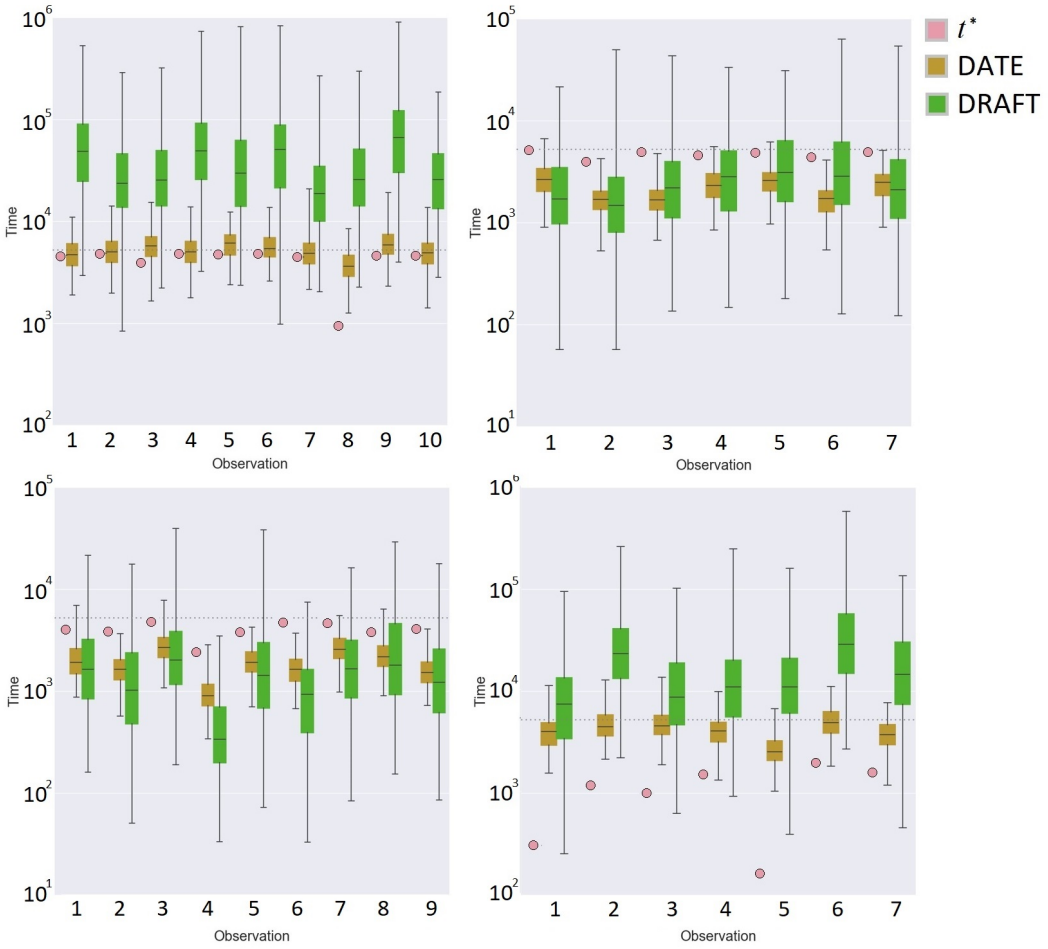


FIGURE A.6: Comparison on FLCHAIN Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

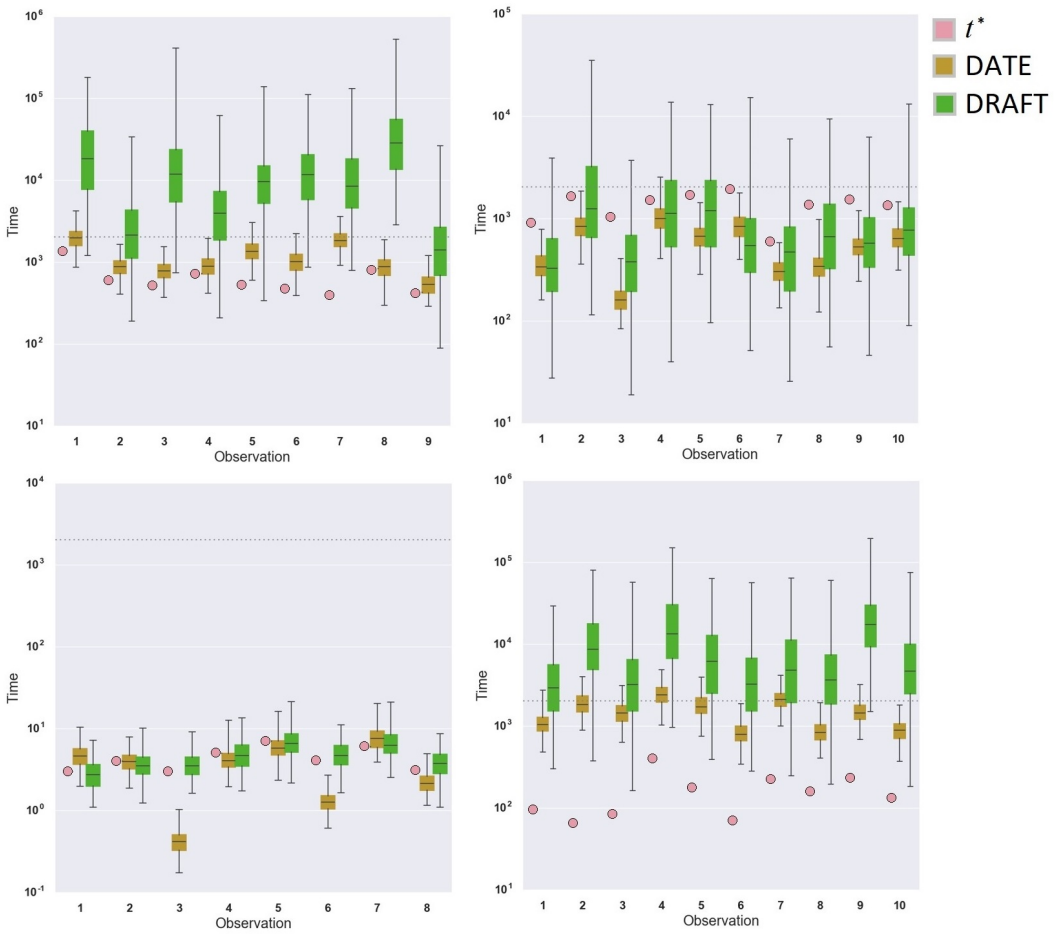


FIGURE A.7: Comparison on SUPPORT Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

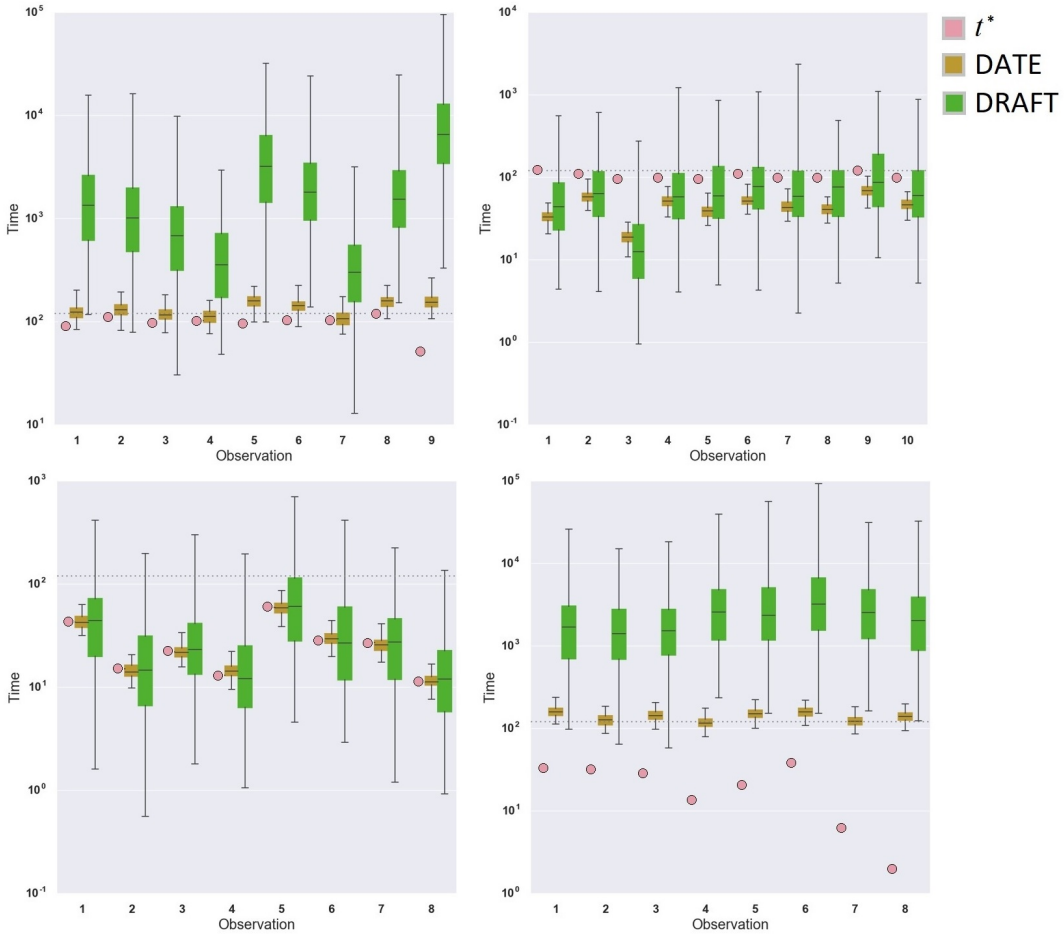


FIGURE A.8: Comparison on SEER Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

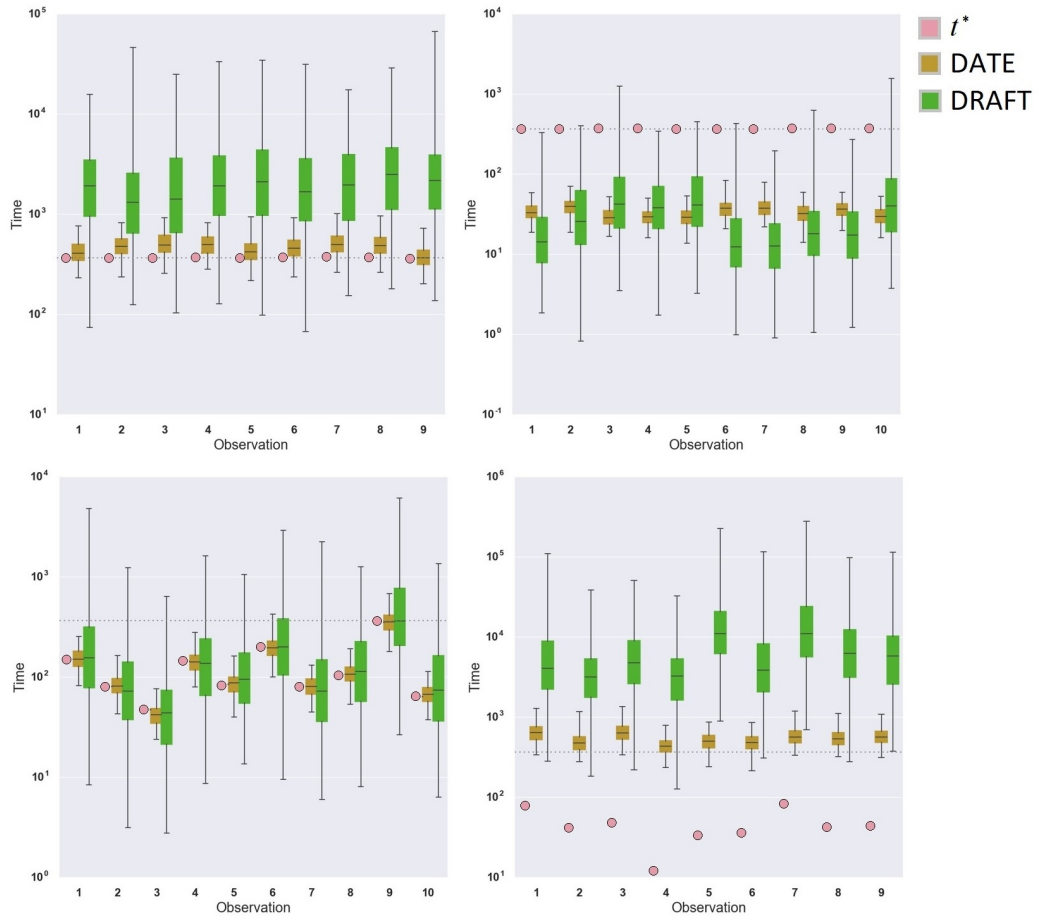


FIGURE A.9: Comparison on EHR Censored best (top-left), worst (top-right) and Non-Censored best (bottom-left), worst (bottom-right).

Appendix B

Supplemental Material for "Calibration and Uncertainty in Neural Time-to-Event Modeling"

B.1 Experiments

B.1.1 Effects of Batch Size on Performance

Table B.1 shows SFM performance metrics across a range of batch sizes for the FLCHAIN dataset, which confirm that the choice of batch-size does not substantially influence any of the performance metrics being considered.

Table B.1: SFM batch size sensitivity on FLCHAIN dataset.

	100	250	500	750	1000
Calibration slope					
	1.0110	0.9766	0.9807	0.9864	**0.9916**
Mean CoV					
	0.4740	**0.4026**	0.4484	0.4332	0.4672
C-Index					
	0.8302	0.8294	**0.8318**	0.8296	0.8287

B.1.2 Ablation Study

Table B.2 shows the ablation study performance results. To evaluate the effect of calibration (CAL) objective, accuracy (ACC) objective (for both censored and observed events), and accuracy only for observed data (ignoring censored data), we compare SFM against ACC, CAL and CAL + ACC (observed only) events objectives.

Table B.2: Ablation study performance results.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP
Calibration slope					
CAL	0.3281	0.4088	0.7561	0.4535	0.6002
ACC	**0.7819**	0.9670	**1.0277**	0.9490	1.0330
CAL + ACC (observed only)	0.2836	0.3867	0.5867	0.5212	0.5055
SFM	0.7734	**0.9807**	0.9405	**0.9540**	**1.0235**
Mean CoV					
CAL	0.4023	0.4501	0.3960	0.3860	0.4067
ACC	0.3226	0.4251	0.2906	0.2042	0.4759
CAL + ACC (observed only)	0.2670	0.3538	0.2363	0.2182	0.4243
SFM	0.2953	0.4484	0.3930	0.1993	0.5045
C-Index					
CAL	0.4778	0.4623	0.4667	0.5450	0.5024
ACC	0.7762	0.8263	**0.8446**	0.8309	**0.7660**
CAL + ACC (observed only)	0.6832	0.5742	0.8289	0.6771	0.5459
SFM	**0.7786**	**0.8318**	0.8319	**0.8314**	0.7491

Appendix C

Supplemental Material for "Enabling Counterfactual Survival Analysis with Balanced Representations"

C.1 General log-likelihood

The general likelihood-based loss hypothesis that accounts for informative censoring is formulated as:

$$-\ell_{h,\Phi,\nu}(t_a, c_a, y, \delta) = \log p_{h,\Phi,\nu}(T_A, C_A | X = x) \tag{C.1}$$

$$= \log p_{h,\Phi}(T_A | X = x) + \log p_{\nu,\Phi}(C_A | X = x)\,, \tag{C.2}$$

where (C.2) follows from the conditional independence (informative censoring) assumption $T \perp\!\!\!\perp C | X, A$. For some parametric formulations of event $p_{h,\Phi}(T_A | X = x)$ and censoring $p_{\nu,\Phi}(C_A | X = x)$ time distributions, e.g., exponential, Weibull, log-Normal, etc., then $-\ell_{h,\Phi,\nu}(t_a, c_a, y, \delta)$ is the closed-form log-likelihood, where:

$$\log p_{h,\Phi}(T_A | X = x) \triangleq \delta \cdot \log f_{h,\Phi}(t_a | x) + (1 - \delta) \cdot \log S_{h,\Phi}(t_a | x), \tag{C.3}$$

$$\log p_{\nu,\Phi}(C_A | X = x) \triangleq (1 - \delta) \cdot \log e_{\nu,\Phi}(c_a | x) + \delta \cdot \log G_{\nu,\Phi}(c_a | x), \tag{C.4}$$

where $\{S_{h,\Phi}(\cdot), G_{\nu,\Phi}(\cdot)\}$ and $\{f_{h,\Phi}(\cdot), e_{\nu,\Phi}(\cdot)\}$ are survival and density functions respectively.

C.2 Metrics

C.2.1 Estimands of Interest

Several common estimands of interest include [ZTU+12, TJCP16]:

- *Difference in expected lifetime*:
  $\mathrm{ITE}(x) = \int_0^{t_{\max}} \{S_1(t|x) - S_0(t|x)\} \mathrm{d}t = \mathbb{E}\{T_1 - T_0 | X = x\}$.
- *Difference in survival function*: $\mathrm{ITE}(t, x) = S_1(t|x) - S_0(t|x)$.
- *Hazard ratio*: $\mathrm{ITE}(t, x) = \lambda_1(t|x)/\lambda_0(t|x)$.

In our experiments, we consider both the hazard ratio and difference in expected lifetime. The difference of expected lifetime is expressed in terms of both survival functions and expectations:

$$\begin{aligned}
\mathbb{E}[T|X = x] &= \int_{-\infty}^{\infty} t f(t|x) dt \\
&= \int_0^{\infty} (1 - F(t|x))\, dt - \int_{-\infty}^{0} F(t|x) dt \qquad \text{(C.5)} \\
&= \int_0^{t_{\max}} S(t|x) dt\,, \qquad \text{(C.6)}
\end{aligned}$$

where (C.5) follows from standard properties of expectations and (C.6) from $1 - F(t|x) = S(t|x)$ and $\int_{-\infty}^{0} F(t|x)dt = 0$. Below we formulate an approach for estimating the individualized and population hazard ratio.

C.2.2 Nonparametric Hazard Ratio

To estimate the proposed *nonparametric hazard ratio* $\mathrm{HR}(t)$ in (4.13) we leveraged (4.1) and $S'(t) \triangleq dS(t)/dt$. For the estimator $\hat{\mathrm{HR}}(t)$, provided that $S(t)$ is a monotonically decreasing function, for simplicity, we fit a linear function $S(t) = m \cdot t + c$ and set $S'(t) \approx m$. Further, we leverage $\hat{S}^{\mathrm{PKM}}(t)$ in [CLM+20] (also detailed in

Table C.1: Performance comparisons on ACTG data, with 95% HR(t) confidence interval. Test set NN assignment of y_{CF} and δ_{CF} yields unbiased ground truth estimator $\text{HR}(t) = 0.54_{(0.51,0.61)}$, since study is a RCT.

Method	**Causal metric**	**Factual metrics**		
	HR(t)	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	$0.49_{(0.38,0.64)}$	NA	NA	NA
CoxPH-IPW	$0.49_{(0.36,0.68)}$	NA	NA	NA
CoxPH-OW	$0.49_{(0.36,0.68)}$	NA	NA	NA
Surv-BART	$3.93_{(3.93,4.90)}$	(0.665, 0.845)	0.001	(0.394, 0.517)
AFT-Weibull	$\mathbf{0.53}_{(0.53,0.53)}$	(0.53,0.351)	3.088	(0.847,0.813)
AFT-log-Normal	$3.75_{(3.75,3.75)}$	(0.717, 0.619)	7.995	(0.847, 0.321)
SR	$0.21_{(0.21,0.28)}$	(0.628, 0.499)	0	(1.388, 0.442)
CSA (proposed)	$0.63_{(0.59,0.68)}$	(0.831, 0.814)	0.132	(1.042, 1.129)
CSA-INFO (proposed)	$0.6_{(0.54,0.66)}$	(0.786, 0.822)	0.13	(0.875, 0.938)

Chapter 3) , defined as the model-free population point-estimate-based nonparametric Kaplan-Meier [KM58] estimator. We denote J distinct and ordered observed event times (censored and non-censored) by the set $\mathcal{T} = \{t_j | t_j > t_{j-1} > \ldots > t_0\}$ from N realizations of Y. Formally, the population survival $\hat{S}_A^{\text{PKM}}(t)$ is recursively formulated as

$$\hat{S}_A^{\text{PKM}}(t_j) = \left(1 - \frac{\sum_{n:\delta_n=1} \mathbb{I}\left(t_{j-1} \leqslant \gamma(T_A^{(n)}) < t_j\right)}{N - \sum_{n=1}^{N} \mathbb{I}\left(\gamma(T_A^{(n)}) < t_{j-1}\right)}\right) \hat{S}_A^{\text{PKM}}(t_{j-1})\,, \tag{C.7}$$

where $\hat{S}_A^{\text{PKM}}(t_0) = 1$, and $\mathbb{I}(b)$ represent an indicator function such that $\mathbb{I}(b) = 1$ if b holds or $\mathbb{I}(b) = 0$ otherwise. Further, $\gamma(\cdot)$ is a deterministic transformation for summarizing T_A, in our experiments, $\gamma(\cdot) = \text{median}(\cdot)$, computed over samples from $t_a \sim p_{h,\Phi}(T_A|X = x)$. Note from (C.7), we marginalize both factual and counterfactual predictions given covariates x.

A similar formulation for the conditional, individualized HR($t|x$), can also be derived, where the cumulative density $F_A(t|x) = 1 - S_A(t|x)$, is estimated with a Gaussian Kernel Density Estimator (KDE) [Sil86] on samples from the model, $t_a \sim p_{h,\Phi}(T_A|X = x)$. Then we have:

Definition 2. *Nonparametric conditional Hazard Ratio and its approximation,* $\hat{\text{HR}}(t|x)$*,*

as

$$\begin{aligned} \mathrm{HR}(t|x) &= \frac{\lambda_1(t|x)}{\lambda_0(t|x)} = \frac{S_0(t|x)}{S_1(t|x)} \cdot \frac{S_1'(t|x)}{S_0'(t|x)} \\ \hat{\mathrm{HR}}(t|x) &= \frac{\hat{S}_0^{\mathrm{KDE}}(t|x)}{\hat{S}_1^{\mathrm{KDE}}(t|x)} \cdot \frac{m_1(t|x)}{m_0(t|x)}\,, \end{aligned} \tag{C.8}$$

where, $S'(t|x) \triangleq dS(|t|x)/dt$ is also approximated with fitting a linear function $S(t|x) = m \cdot t + c$, and setting $S'(t|x) \approx m$. Note that for some parametric formulations, $\mathrm{HR}(t|x)$, can be readily evaluated because $f(t_a|x)$ and $S(t_a|x)$ are available in closed form.

C.2.3 Factual Metrics

Concordance Index C-Index (also related to receiver operating characteristic) is a widely used survival ranking metric which naturally handles censoring. It quantifies the consistency between the order of the predicted times or risk scores relative to ground truth. C-Index is evaluated on point estimates, we summarize individualized predicted samples from CSA and CSA-INFO, i.e., $\hat{t}_a = \mathrm{median}\left(\{t_s\}_{s=1}^{200}\right)$, where t_s is a sample from the trained model.

Calibration Slope Calibration quantifies distributional statistical consistency between model predictions relative to ground truth. We measure population calibration by comparing population survival curves from model predictions against ground truth according to [CLM+20]. We desire a high calibrated model, with calibration slope of 1, while a slope < 1 and slope > 1 indicates underestimation or overestimation risk, respectively.

Coefficient of Variation The coefficient of variation (COV) $\sigma\mu^{-1}$, the ratio between standard deviation and mean, quantifies distribution dispersion. A COV > 1 and < 1 indicates a high or low variance distribution, in practice, we desire low variance

distribution. We use Mean COV $N^{-1}\sum_{i=1}^{N}\sigma_i\mu_i^{-1}$, where for subject i we compute $\{\mu_i, \sigma_i\}$ from samples $\{t_s\}_{s=1}^{200}$.

C.3 Baselines

Cox proportional hazard (CoxPH) CoxPH assumes a semi-parametric linear model $\lambda(t|a) = \lambda_{\mathrm{b}}(t)\exp(a\beta)$, thus the hazard ratio between treatment and control can be obtained without specifying the baseline hazard $\lambda_{\mathrm{b}}(t)$ as in (4.12). A simple logistic model $\hat{e}_i = \sigma(x_i;\eta)$, is used to approximate the unknown propensity score $P(A=1|X=x)$. Methods that adjust for selection bias (or confounding) learn β by maximizing a propensity weighted partial likelihood [SWH09, BHC+14, RR83]

$$\mathcal{L}(\beta) = \prod_{i:\delta_i=1}\left(\frac{\exp(a_i\beta)}{\sum_{j:t_j\geqslant t_i}\hat{w}_j\cdot\exp(a_j\beta)}\right)^{\hat{w}_i}. \tag{C.9}$$

We consider three normalized weighting schemes for w , namely, (i) inverse probability weighting (IPW) [HT52, CTD09], where $\mathrm{IPW}_i = \frac{a_i}{\hat{e}_i} + \frac{1-a_i}{1-\hat{e}_i}$, (ii) overlapping weights (OW) [CHIM06, LMZ18], where $\mathrm{OW}_i = a_i\cdot(1-\hat{e}_i) + (1-a_i)\cdot\hat{e}_i$, and (iii) the standard RCT Uniform assumption. Note that this modeling approach requires fitting over the entire dataset, thus has no inference capability.

Accelerated Failure Time (AFT) We implement IPM regularized neural-based log-Normal and Weibull AFT baselines. Both approaches have a desirable closed form $S_{h,\Phi}(t_a|x)$, thus enabling maximum likelihood based estimation, where

$$\begin{aligned}-\mathcal{L}_{\mathrm{F}}^{\mathrm{AFT}} \triangleq \mathbb{E}_{(y,\delta,x,a)\sim p(y,\delta,X,A)}\big[&\delta\cdot\log f_{h,\Phi}(t_a|x)\\ &+(1-\delta)\cdot\log S_{h,\Phi}(t_a|x)\big]\,.\end{aligned} \tag{C.10}$$

The log-Normal mean and variance parameters are learned such that, $\log t_a = \mu_{h,\Phi}(h(r,a)) + \epsilon$, where $\epsilon\sim\mathcal{N}(0,\sigma^2_{h,\Phi}(h(r,a)))$ and $r=\Phi(x)$. Further, we learn the Weibull scale and shape parameters, where $t_a = \lambda_{h,\Phi}(h(r,a))\cdot(-\log U)^{\left(k_{h,\Phi}(h(r,a))\right)^{-1}}$

and $U \sim \text{Uniform}(0, 1)$. We regularize (C.10) with the IPM loss, for maximum likelihood optimization.

Semi-supervised regression (SR) To demonstrate the effectiveness of our flow-based uncertainty estimation approach we contrast CSA with a deterministic accuracy objective from [CTL+18], where $t_a = h(r, a)$ and:

$$\mathcal{L}_{\text{F}}^{\text{SR}} \triangleq \mathbb{E}_{(y,\delta,x,a)\sim p(y,\delta,X,A)} \left[\delta \cdot (|y - t_a|) + (1 - \delta) \cdot (\max(0, y - t_a))\right], \tag{C.11}$$

where (C.11) is regularized according to the IPM loss.

	age6	ascvd_hx6	bmi6	bpmeds6	chol5	dbp6	diab6	female
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	55.558140	0.069767	29.326328	0.317829	202.147287	75.922481	0.116279	0.465116
std	9.412348	0.255748	4.800124	0.467448	40.368053	7.184632	0.321809	0.500726
min	35.000000	0.000000	20.777429	0.000000	118.000000	55.000000	0.000000	0.000000
25%	50.000000	0.000000	25.995640	0.000000	176.000000	71.000000	0.000000	0.000000
50%	54.000000	0.000000	28.835150	0.000000	196.000000	76.000000	0.000000	0.000000
75%	61.000000	0.000000	31.847777	1.000000	225.000000	80.000000	0.000000	1.000000
max	84.000000	1.000000	45.135681	1.000000	312.000000	100.000000	1.000000	1.000000

(a) FRAMINGHAM $\text{HR}(t|x) > 1.916$

	age6	ascvd_hx6	bmi6	bpmeds6	chol5	dbp6	diab6	female
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	60.666667	0.139535	26.375823	0.286822	197.581395	71.627907	0.046512	0.604651
std	10.185263	0.347855	5.325558	0.454041	28.029197	11.657260	0.211411	0.490832
min	37.000000	0.000000	17.676632	0.000000	118.000000	49.000000	0.000000	0.000000
25%	53.000000	0.000000	22.687889	0.000000	180.000000	62.000000	0.000000	0.000000
50%	60.000000	0.000000	25.285077	0.000000	198.000000	70.000000	0.000000	1.000000
75%	69.000000	0.000000	29.230393	1.000000	217.000000	81.000000	0.000000	1.000000
max	78.000000	1.000000	45.992112	1.000000	290.000000	105.000000	1.000000	1.000000

(b) FRAMINGHAM $\text{HR}(t|x) > 1.916$

FIGURE C.1: Covariate statistics for top (a) and bottom (b) quantiles, of the median $\log \text{HR}(t|x)$ values for the test set of FRAMINGHAM.

	gluc5	hdl5	pad_hx6	sbp6	smoke6	stk_hx6	mi_hx6	trigly5
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	102.201550	43.953488	0.007752	123.782946	0.170543	0.015504	0.038760	164.139535
std	34.450912	11.543979	0.088045	13.923879	0.377575	0.124027	0.193774	78.358625
min	75.000000	26.000000	0.000000	99.000000	0.000000	0.000000	0.000000	46.000000
25%	90.000000	35.000000	0.000000	114.000000	0.000000	0.000000	0.000000	119.000000
50%	95.000000	43.000000	0.000000	122.000000	0.000000	0.000000	0.000000	143.000000
75%	103.000000	50.000000	0.000000	131.000000	0.000000	0.000000	0.000000	200.000000
max	289.000000	95.000000	1.000000	170.000000	1.000000	1.000000	1.000000	468.000000

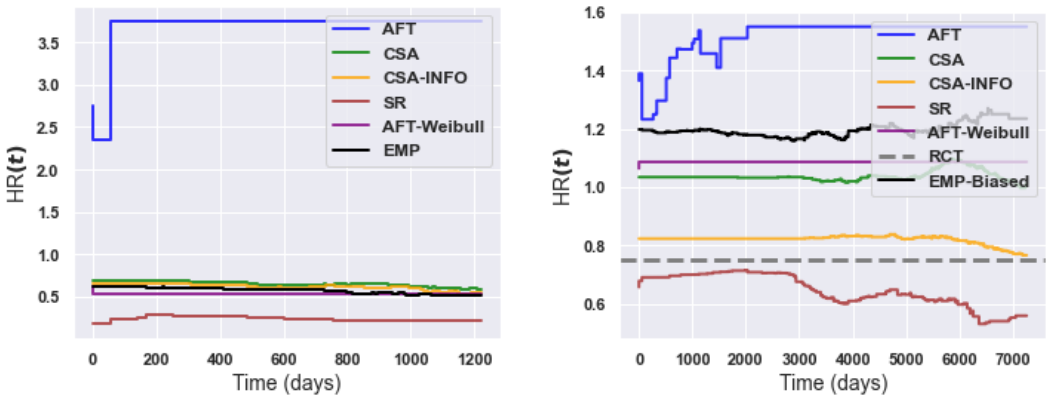
(a) FRAMINGHAM $\text{HR}(t|x) > 1.916$

	gluc5	hdl5	pad_hx6	sbp6	smoke6	stk_hx6	mi_hx6	trigly5
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	96.023256	58.550388	0.054264	128.007752	0.217054	0.046512	0.054264	119.937984
std	16.391912	16.147253	0.227420	22.184417	0.413847	0.211411	0.227420	133.107261
min	48.000000	22.000000	0.000000	88.000000	0.000000	0.000000	0.000000	33.000000
25%	88.000000	49.000000	0.000000	111.000000	0.000000	0.000000	0.000000	63.000000
50%	95.000000	59.000000	0.000000	126.000000	0.000000	0.000000	0.000000	87.000000
75%	101.000000	69.000000	0.000000	140.000000	0.000000	0.000000	0.000000	109.000000
max	228.000000	101.000000	1.000000	214.000000	1.000000	1.000000	1.000000	1149.000000

(b) FRAMINGHAM $\text{HR}(t|x) > 1.916$

FIGURE C.2: Covariate statistics for top (a) and bottom (b) quantiles, of the median $\log \text{HR}(t|x)$ values for the test set of FRAMINGHAM.

Survival Bayesian additive regression trees (Surv-BART) Surv-BART [SLML16] is a nonparametric tree-based approach for estimating individualized survivals $\hat{S}(t_a^{(j)}|X = x)$ (defined at pre-specified J time-horizons) from an ensemble of regression trees.



(a)ACTG HR(t) (b) FRAMINGHAM HR(t)

FIGURE C.3: Inferred population HR(t) comparisons on (a) ACTG and (b) FRAMINGHAM datasets.

Note, Surv-BART does not adjust for both selection bias and informative censoring. While, we fit two separate models based on factual treatment and control data, causal metrics are estimated with both factual and counterfactual predictions.

C.4 Experiments

C.4.1 Generating ATCG-Synthetic Dataset

The ACTG-SYNTHETIC, is a semi-synthetic dataset based on ACTG covariates [HKH+96]. We simulate potential outcomes according to a Gompertz-Cox distribution [BAB05] with selection bias from a simple logistic model for $P(A = 1|X = x)$ and AFT-based

censoring mechanism. Below is our generative scheme:

$$
\begin{aligned}
X &= \text{ ACTG covariates} \\
P(A=1|X=x) &= \frac{1}{b} \times \left(a + \sigma\left(\eta(\text{AGE} - \mu_{\text{AGE}} + \text{CD40} - \mu_{\text{CD40}})\right)\right) \\
T_A &= \frac{1}{\alpha_A} \log\left[1 - \frac{\alpha_A \log U}{\lambda_A \exp\left(x^T \beta_A\right)}\right] \\
U &\sim \text{Uniform}(0,1) \\
\log C &\sim \text{Normal}(\mu_c, \sigma_c^2) \\
Y &= \min(T_A, C)\,, \quad \delta = 1 \text{ if } T_A < C, \text{ else } \delta = 0\,,
\end{aligned}
$$

where $\{\beta_A, \alpha_A, \lambda_A, b, a, \eta, \mu_c, \sigma_c\}$ are hyper-parameters and $\{\mu_{\text{AGE}}, \mu_{\text{CD40}}\}$ are the means for age and CD40 respectively. This semi-synthetic dataset will be made publicly available.

C.4.2 Quantitative Results

See Table C.1 for additional quantitative comparisons on ACTG dataset.

C.4.3 Qualitative Results

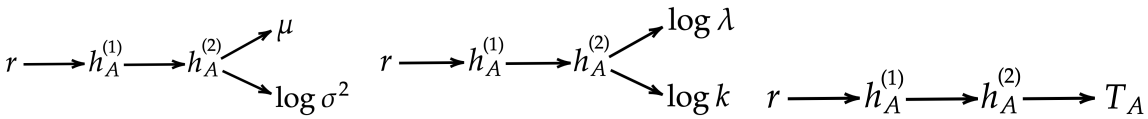
Figure C.3 demonstrates model comparisons across of population hazard, HR(t), on ACTG and FRAMINGHAM datasets. Figures C.1 - C.2, summarizes the positive and negative covariate statistics from the isolated extreme top and bottom quantiles on FRAMINGHAM datasets.

C.4.4 Architecture of the neural network

We detail the architecture of neural-based methods, namely, baselines (AFT-log-Normal, AFT-Weibull, SR) and our proposed methods (CSA and CSA-INFO). All methods are trained using one NVIDIA P100 GPU with 16GB memory. In all experiments we set the minibatch size $M = 200$, Adam optimizer with the following hyper-parameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and epsilon

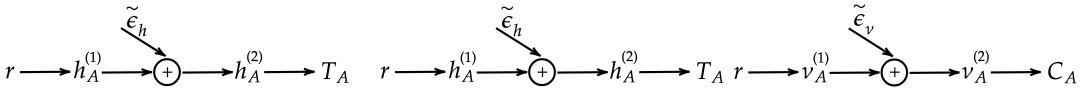
1×10^{-8}. Further, all network weights are initialed according to Uniform$(-0.01, 0.01)$. Datasets are split into training, validation and test sets according to 70%, 15% and 15% partitions, respectively, stratified by event and treatment proportions. The validation set is used for hyperparameter search and early stopping. All hidden units in $\{h_A(\cdot), \nu_A(\cdot)\}$, are characterized by Leaky Rectified Linear Unit (ReLU) activation functions, batch normalization and dropout probability of $p = 0.2$ on all layers. The output layers of predicted times $\{T_A, C_A\}$ have an additional exponential transformation.

Encoder The encoding function $\Phi(\cdot)$ for mapping $r = \Phi(x)$ is shared among all the neural based methods (proposed and baselines) and specified in terms of two-layer MLPs of 100 hidden units.



(a) AFT-log-Normal (b) AFT-Weibull (c) SR

FIGURE C.4: Decoding architecture of baselines.



(a) CSA (b) CSA-INFO

FIGURE C.5: Decoding architecture of proposed methods.

Decoder Figure C.4 shows the architectural details of the baselines, where the decoding function $h_A(\cdot)$ is specified in terms of two-layer MLPs of 100 hidden units. Further, the proposed *planar* flow based methods shown in Figure C.5, are comprised of two-layer MLPS for $\{h_A(\cdot), \nu_A(\cdot)\}$ of dimensions $[100, 200]$. Moreover, the hidden

layers $\{h_A^{(2)}, \nu_A^{(2)}\}$, take as input the concatenated $[h_A^{(1)}, \tilde{\epsilon}_h]$ and $[\nu_A^{(1)}, \tilde{\epsilon}_\nu]$ respectively. Finally, we set the planar flow dimensions for both $\{\tilde{\epsilon}_\nu, \tilde{\epsilon}_h\}$ to 100.

Appendix D

Supplemental Material for "Survival Cluster Analysis"

D.1 Notation

Refer to Table D.1 for summary descriptions of notation used in the SCA formulation.

Table D.1: SCA notations.

Notation	Description
$\boldsymbol{x}_i$	vector of covariates for subject i
t_i	empirical time-to-event (censored or non-censored)
l_i	censoring indicator, where $l_i = 1$ represents a non-censored event
$f(t\|\boldsymbol{x})$	time density function
$F(t\|\boldsymbol{x})$	cumulative density function
$S(t\|\boldsymbol{x})$	survival function, defined as $1 - F(t\|\boldsymbol{x})$
$\lambda(t\|\boldsymbol{x})$	hazard rate (risk score) function
$r_{\boldsymbol{\psi}}(\boldsymbol{x})$	deterministic encoder mapping, parametrized by $\boldsymbol{\psi}$
$g_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{\epsilon})$	stochastic mapping for synthesizing event times, parametrized by $\boldsymbol{\theta}$, where $\boldsymbol{\epsilon}$ is the source of stochasticity
$st(\boldsymbol{c}_{u_n}, \nu)$	Student's t-distribution, with ν degrees of freedom, means $\{\boldsymbol{c}_k\}_{k=1}^{\infty}$ and u_n mixture-component indicator for $\boldsymbol{z}_n$
π_k	mixture proportions
$q(\boldsymbol{\pi}\|\boldsymbol{\xi}, \{\boldsymbol{x}_1\}_{n=1}^{M}) = \mathrm{Dir}(\boldsymbol{\pi}\|\boldsymbol{\xi})$	MLE based distribution for mixture assignments and proportions
$p(\boldsymbol{\pi}\|\boldsymbol{\gamma}, \{\boldsymbol{x}_n\}_{n=1}^{M}) = \mathrm{Dir}(\boldsymbol{\pi}\|\boldsymbol{\gamma})$	DP based distribution for mixture assignments and proportions

D.2 Experimental Setup

In all experiments, SCA, SFM, DATE, DRAFT and S-CRPS are specified in terms of two-layer MLPs of 50 hidden units with Rectified Linear Unit (ReLU) activation functions, batch normalization [IS15] and apply dropout of $p = 0.2$ on all layers.

We set the minibatch size to $M = 350$ and use the Adam [KB14] optimizer with the following hyperparameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and epsilon 1×10^{-8}. We initialize all the network weights according to *Xavier* [GB10]. SFM and DATE inject noise in all layers, see [CTL$^+$18] for more details; while SCA injects noise only in the last layer. Datasets are split into training, validation and test sets as 80%, 10% and 10% partitions, respectively, stratified by non-censored event proportion. The validation set is used for early stopping and learning model hyperparameters. All models are trained using one NVIDIA P100 GPU with 16GB memory.

D.3 C-index, mean CoV and RAE Results

See Table D.2 for additional quantitative evaluations on C-index, mean CoV and RAE.

Table D.2: Performance metrics. SCA is the proposed model.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
RAE (non-censored, censored)						
DATE	(**0.6107**, 0.2148)	(0.5222, 0.1159)	(0.6691, 0.1471)	(0.5289, 0.0294)	(0.5224, 0.2939)	(0.5122, 0.1888)
DRAFT	(0.7099, 0.1195)	(0.6399, 0.0445)	(0.7109, **0.0418**)	(0.6097, **0.0139**)	(0.7465, 0.3317)	(0.6697, 0.1119)
S-CRPS	(0.7240, 0.0773)	(0.6378, **0.0321**)	(**0.4851**, 0.2333)	(0.5323, 0.0427)	(0.7330, **0.0519**)	(0.8369, **0.0108**)
CoxPH	-	-	-	-	-	-
MTLR	-	-	-	-	-	-
SFM	(0.6146, 0.2066)	(**0.5111**, 0.1313)	(0.6398, 0.3274)	(0.5294, 0.0329)	(**0.5162**, 0.2391)	(**0.5074**, 0.2070)
SCA	(0.6186, 0.1897)	(0.5134, 0.1199)	(0.6295, 0.2671)	(**0.5193**, 0.0480)	(0.5424, 0.2244)	(**0.5074**, 0.1819)
Mean Cov						
DATE	**0.2477**	**0.3585**	0.2987	**0.1485**	0.5168	**0.3624**
DRAFT	5.0305	6.2952	3.8689	3.4501	8.4918	3.9911
S-CRPS	0.8585	0.9412	0.7351	0.6036	1.0240	0.6225
CoxPH	-	-	-	-	-	-
MTLR	-	-	-	-	-	-
SFM	0.2953	0.4484	0.3930	0.1993	0.5045	0.3696
SCA	0.2842	0.4458	**0.2891**	0.2154	**0.4619**	0.3870
C-Index						
DATE	0.7756	0.8264	0.8421	**0.8320**	0.7416	0.7048
DRAFT	0.7796	0.8341	0.8560	0.8310	**0.7617**	0.7005
S-CRPS	0.7704	0.8286	**0.8685**	0.8298	0.7529	0.7015
CoxPH	0.7542	**0.8344**	0.8389	0.8223	0.6435	**0.7603**
MTLR	-	-	-	-	-	-
SFM	0.7786	0.8318	0.8319	0.8314	0.7491	0.7009
SCA	**0.7809**	0.8330	0.8465	0.8306	0.7498	0.7072

D.4 Calibration and Survival Function Results

The model calibration and survival plots for datasets SUPPORT, FLCHAIN, SLEEP, SEER, FRAMINGHAM, and EHR are shown in Figures D.1 - D.6.

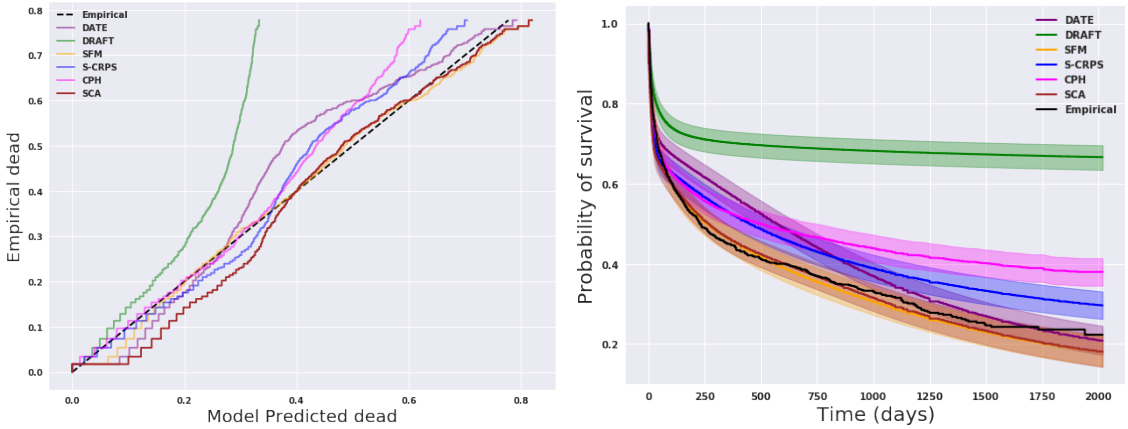


FIGURE D.1: Calibration (left) and Survival function estimates (right) for SUPPORT data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CoxPH).

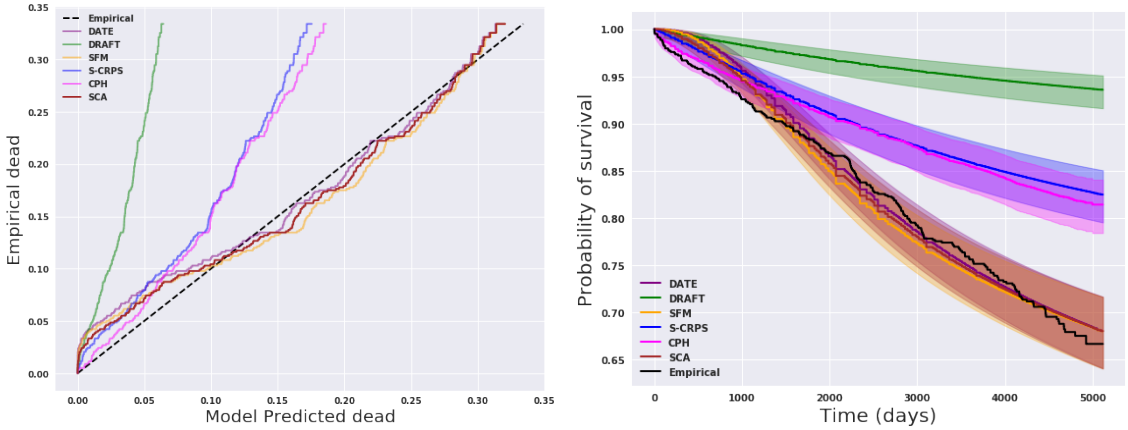


FIGURE D.2: Calibration(left) and Survival function estimates (right) for FLCHAIN data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CPH).

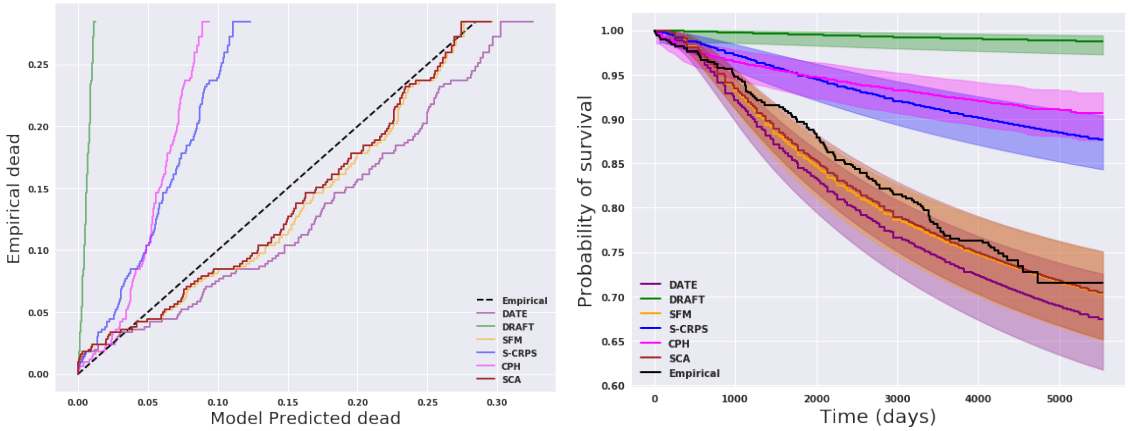


FIGURE D.3: Calibration (left) and Survival function estimates (right) for SLEEP data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CPH).

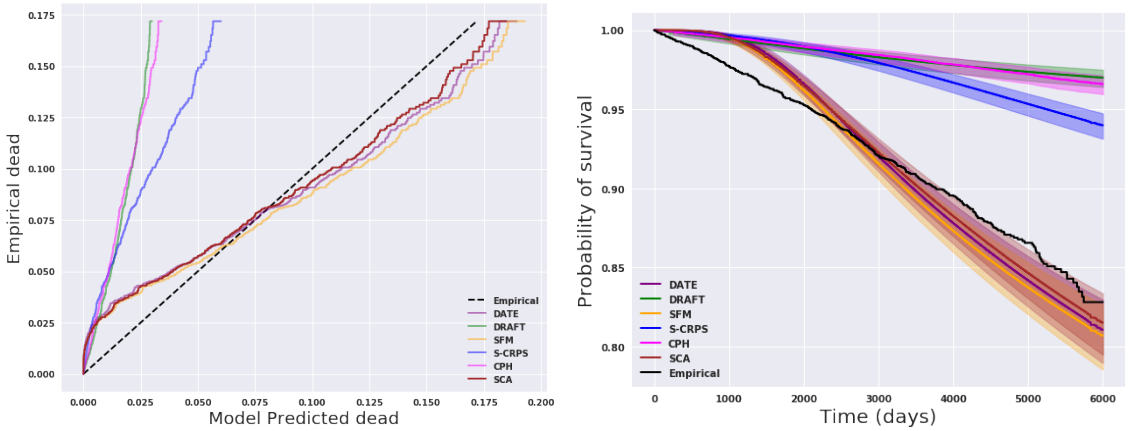


FIGURE D.4: Calibration(left) and Survival function estimates (right) for FRAMINGHAM data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CPH).

D.5 Latent-Space Representation Results

We provide all the qualitative visualization of the latent-space representation, namely, *a*) estimated individualized cluster assignment probability distributions; *b*) *t*-SNE plots of both the centroids $\boldsymbol{c}_k$ with $\boldsymbol{z}$; *c*) cluster-specific Kaplan Meir curves. Refer to Figures D.7 - D.12 for SCA results on FRAMINGHAM, SUPPORT, FLCHAIN, SLEEP,

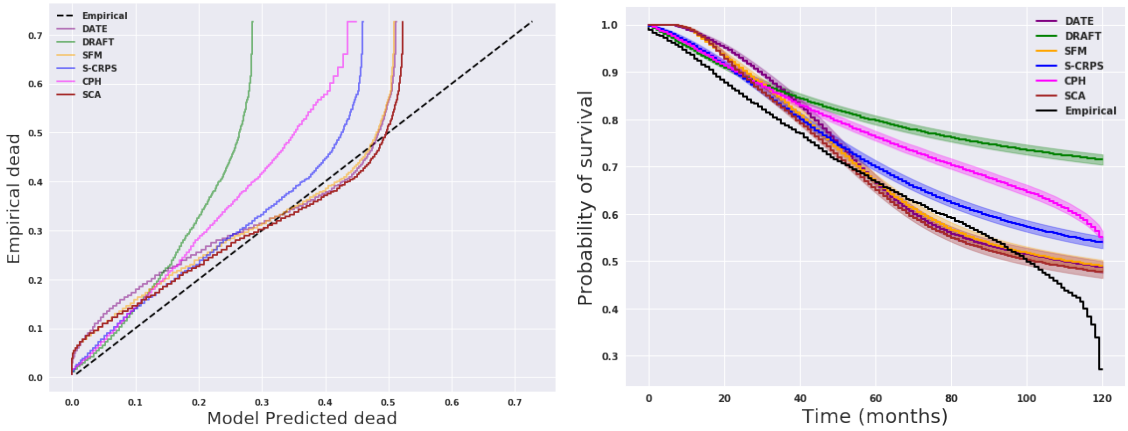


FIGURE D.5: Calibration(left) and Survival function estimates (right) for SEER data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CPH).

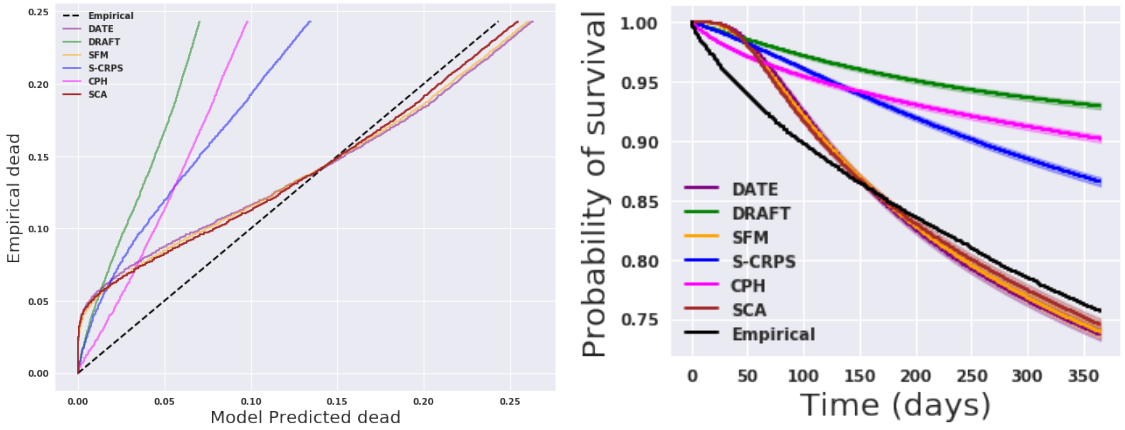
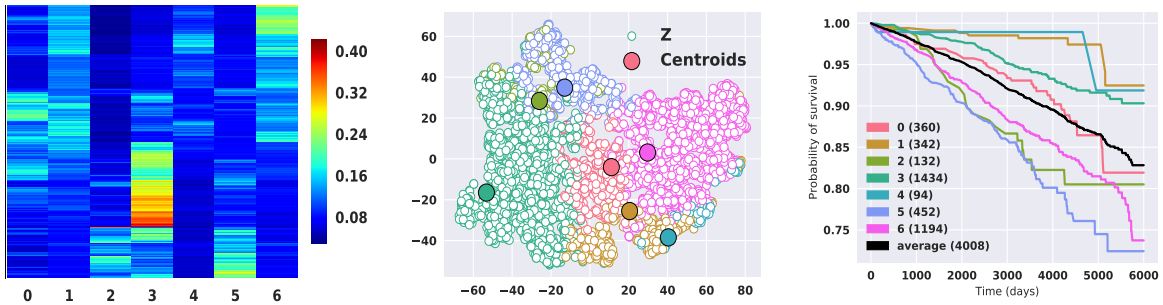


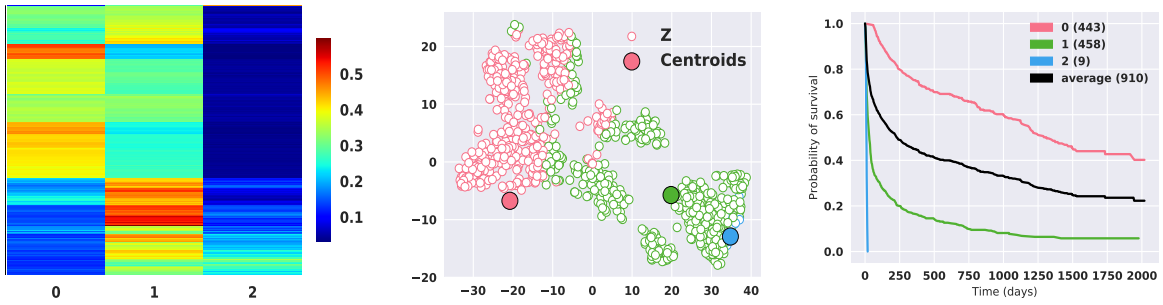
FIGURE D.6: Calibration(left) and Survival function estimates (right) for EHR data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CPH).

SEER and EHR datasets respectively.



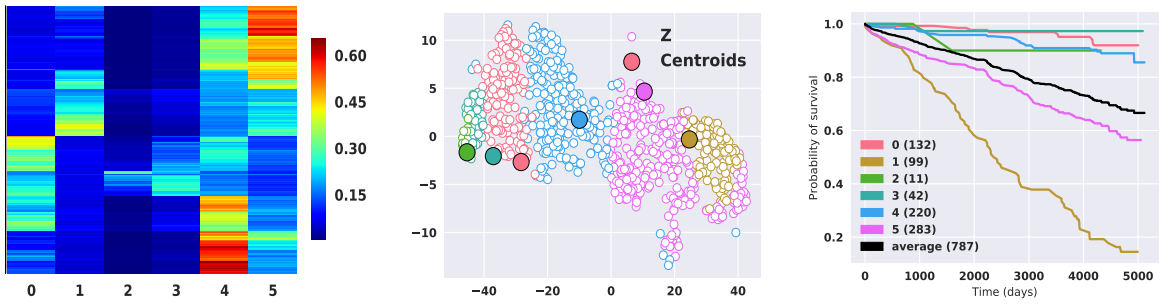
(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.7: Inferred clusters on the testing set of FRAMINGHAM dataset, with $K = 25$ and $\gamma_o = 8$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.



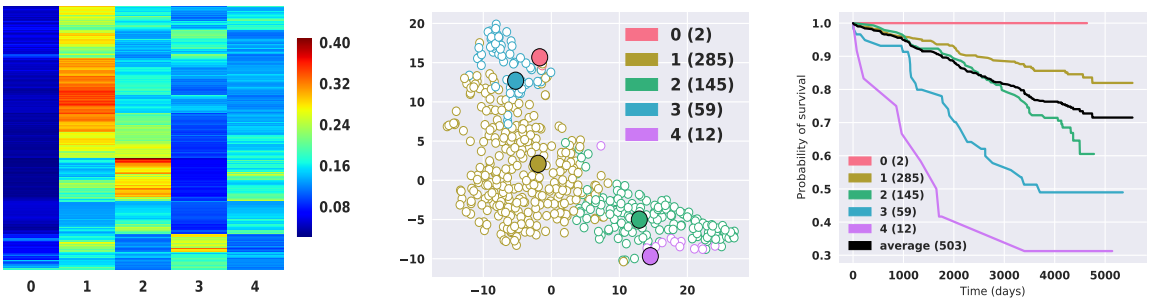
(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.8: Inferred clusters on the testing set of SUPPORT dataset, with $K = 25$ and $\gamma_o = 2$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.



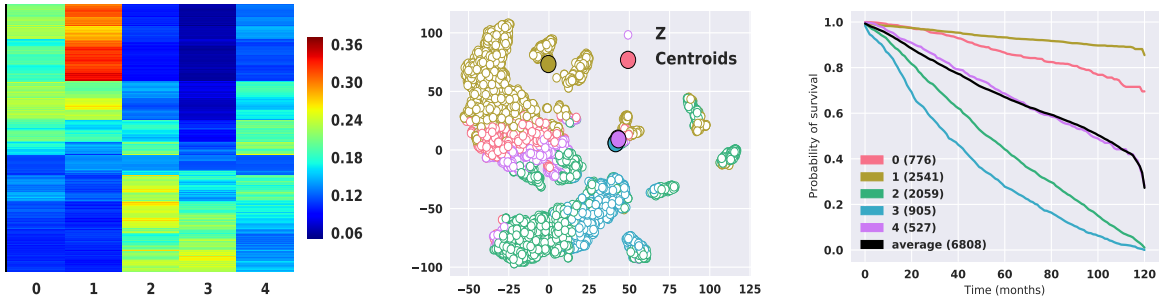
(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.9: Inferred clusters on the testing set of FLCHAIN dataset, with $K = 25$ and $\gamma_o = 4$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.



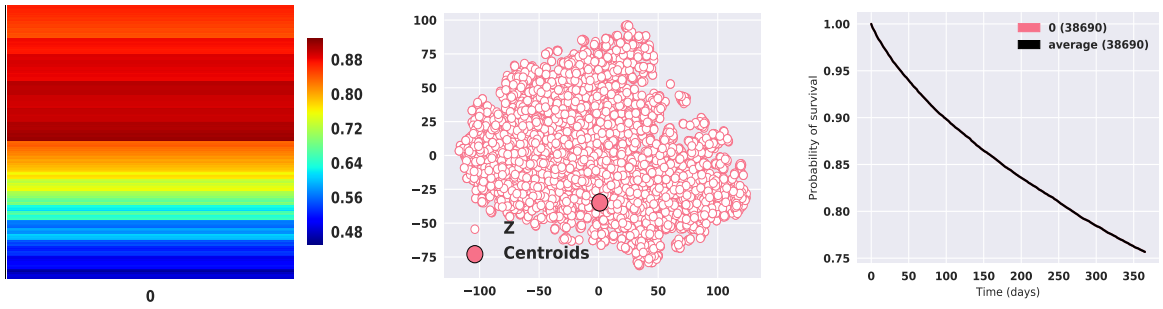
(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.10: Inferred clusters on the testing set of SLEEP dataset, with $K = 25$ and $\gamma_o = 3$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.



(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.11: Inferred clusters on the testing set of SEER dataset, with $K = 25$ and $\gamma_o = 2$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.



(a) Cluster assignments (b) t-SNE plot (c) Cluster Kaplan-Meir

FIGURE D.12: Inferred clusters on the testing set of EHR dataset, with $K = 25$ and $\gamma_o = 2$ with corresponding individual probability distribution $q(\boldsymbol{\pi}|\mathcal{D})$.

Bibliography

[Aal94] Odd O Aalen. Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 1994.

[ADZ+20] Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Y Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, 2020.

[AG+01] Odd O Aalen, Håkon K Gjessing, et al. Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). *Statistical Science*, 2001.

[APS+14] Tariq Ahmad, Michael J Pencina, Phillip J Schulte, Emily O'Brien, David J Whellan, Ileana L Piña, Dalane W Kitzman, Kerry L Lee, Christopher M O'Connor, and G Michael Felker. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *Journal of the American College of Cardiology*, 2014.

[ASK+18] Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology*, 2018.

[ASR88] Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.

[Aus07] Peter C Austin. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, 2007.

[Aus14] Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 2014.

[AvdS17] Ahmed M. Alaa and Mihaela van der Schaar. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In *NeurIPS*, 2017.

[AZT+21] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin Duke. Counterfactual representation learning with balancing weights. In *AISTATS*, 2021.

[BAB05] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 2005.

[BC74] Norman Breslow and John Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 1974.

[BFJ11] Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of statistics*, 2011.

[BHC+14] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Bryan Lau, Adaora A Adimora, and Women's Interagency HIV Study. Worth the weight: using inverse probability weighted cox models in aids research. *AIDS research and human retroviruses*, 2014.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[BJ+06] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 2006.

[BLV+94] Emelia J Benjamin, Daniel Levy, Sonya M Vaziri, Ralph B D'agostino, Albert J Belanger, and Philip A Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. *Jama*, 1994.

[BP12] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *AISTATS*, 2012.

[Bri50] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 1950.

[BT04] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2004.

[CAZ+21] Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2021.

[CDP+18] Liqun Chen, Shuyang Dai, Yunchen Pu, Erjin Zhou, Chunyuan Li, Qinliang Su, Changyou Chen, and Lawrence Carin. Symmetric variational autoencoder and connections to adversarial learning. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[CGM+10] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 2010.

[CH04] Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 2004.

[CHH+59] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 1959.

[CHIM06] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.

[CKWZ20] Yifan Cui, Michael R Kosorok, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv*, 2020.

[CLLK04] Bernard MY Cheung, Ian J Lauder, Chu-Pak Lau, and Cyrus R Kumana. Meta-analysis of large randomized controlled trials to evaluate the impact of statins on cardiovascular outcomes. *British journal of clinical pharmacology*, 2004.

[CLM+20] 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*, 2020.

[CM09] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2009.

[CNC+16] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific reports*, 2016.

[Col15] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[Cox72] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972.

[Cox92] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*. 1992.

[CTD09] Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 2009.

[CTL+18] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. *ICML*, 2018.

[CTL+20] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Irfan Khan, Karen J Chandross, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

[CV15] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 2015.

[CYA13] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Development of a prognostic model for breast cancer survival in an open challenge environment. *Science translational medicine*, 2013.

[CZZ+19] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. *ICLR*, 2019.

[Daw82] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 1982.

[DBP+16] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv*, 2016.

[DDT+16] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *ACM SIGKDD*, 2016.

[DF83] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *The statistician*, 1983.

[Día19] Iván Díaz. Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in Medicine*, 2019.

[DKK+12] 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, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, 2012.

[DZAD17] Ugljesa Djuric, Gelareh Zadeh, Kenneth Aldape, and Phedias Diamandis. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology*, 2017.

[Efr77] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *JASA*, 1977.

[ESBB98] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 1998.

[Fer73] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, 1973.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 2010.

[FLS11] Jennifer Frankovich, Christopher A Longhurst, and Scott M Sutherland. Evidence-based medicine in the emr era. *N Engl J Med*, 2011.

[Fot18] Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv*, 2018.

[FRG+87] Margaret A Fischl, Douglas D Richman, Michael H Grieco, Michael S Gottlieb, Paul A Volberding, Oscar L Laskin, John M Leedom,

Jerome E Groopman, Donna Mildvan, Robert T Schooley, et al. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine*, 1987.

[FRT16] Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. Gaussian processes for survival analysis. In *NeurIPS*, 2016.

[FS95] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 1995.

[G+26] Major Greenwood et al. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, 1926.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[GB13] Sheila Gaynor and Eric Bair. Identification of relevant subtypes via preweighted sparse clustering. *Biostatistics*, 2013.

[GCC+19] Saurabh Gombar, Alison Callahan, Robert Califf, Robert Harrington, and Nigam H Shah. It is time to learn from patients like mine. *NPJ digital medicine*, 2019.

[GCW+18] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *ICLR*, 2018.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[GPC+16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *ICML*, 2017.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

[GSG20] Stefan Groha, Sebastian M Schmon, and Alexander Gusev. Neural odes for multi-state survival analysis. *arXiv*, 2020.

[HBR00] Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 2000.

[HCM+05] Miguel A Hernán, Stephen R Cole, Joseph Margolick, Mardge Cohen, and James M Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 2005.

[HDWF+17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 2017.

[HHDG20] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 2020.

[Hil11] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2011.

[HJL20] Liangyuan Hu, Jiayi Ji, and Fan Li. Estimating heterogeneous survival treatment effect in observational data using machine learning. *arXiv*, 2020.

[HJLC+84] Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 1984.

[HKH+96] Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 1996.

[HLM11] David W Hosmer, Stanley Lemeshow, and Susanne May. *Applied survival analysis*. Wiley Blackwell, 2011.

[HLRV20] Nicholas C Henderson, Thomas A Louis, Gary L Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 2020.

[Hou95] Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1995.

[HR20] Miguel A Hernán and James M Robins. Causal inference: what if. *Boca Raton: Chapman & Hill/CRC*, 2020.

[HSN08] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics*, 2008.

[HT52] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 1952.

[IJ01] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2001.

[IKBL08] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2008.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[JDC+09] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 2009.

[JGJS99] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 1999.

[JZT+17] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *IJCAI*, 2017.

[KA15] D Kinga and J Ba Adam. A method for stochastic optimization. In *ICLR*, 2015.

[KAK97] Niels Keiding, Per Kragh Andersen, and John P Klein. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine*, 1997.

[KALL18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KFE18] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *ICML*, 2018.

[KHL+95] William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 1995.

[KK02] Richard Kay and Nelson Kinnersley. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug information journal*, 2002.

[KK10] David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

[KM58] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *JASA*, 1958.

[KM05] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[KPT+17] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 2017.

[KSC+16] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *stat*, 2016.

[KSC+18] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 2018.

[KW14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[LHR+15] Wenzhao Lian, Ricardo Henao, Vinayak Rao, Joseph Lucas, and Lawrence Carin. A multitask point process predictive model. In *ICML*, 2015.

[LHS14] Christopher A Longhurst, Robert A Harrington, and Nigam H Shah. A 'green button'for using aggregate patient data at the point of care. *Health affairs*, 2014.

[LLC+17] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *NeurIPS*, 2017.

[LMZ18] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 2018.

[LSC+17] Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. *arXiv*, 2017.

[LXZ+18] Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. In *NeurIPS*, 2018.

[LZAvdS19] Changhee Lee, William R Zame, Ahmed M Alaa, and Mihaela van der Schaar. Temporal quilting for survival analysis. In *AISTATS*, 2019.

[LZYvdS18] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, 2018.

[Man66] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 1966.

[ME17] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NeurIPS*, 2017.

[MEB+12] B Mihaylova, J Emberson, L Blackwell, A Keech, J Simes, EH Barnes, M Voysey, 3A Gray, R Collins, and C Baigent. The effects of lowering ldl cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials., 2012.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[MJV+12] Aditya Krishna Menon, Xiaoqian J Jiang, Shankar Vembu, Charles Elkan, and Lucila Ohno-Machado. Predicting accurate probabilities with a ranking loss. In *ICML*, 2012.

[MO14] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial nets. *arXiv*, 2014.

[MPER18] Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, 2018.

[MTRN19] S Chandra Mouli, Leonardo Teixeira, Bruno Ribeiro, and Jennifer Neville. Deep lifetime clustering. *arXiv*, 2019.

[Mül97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

[Mur73] Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 1973.

[NLD21] Chirag Nagpal, Xinyu Rachel 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*, 2021.

[NS17] Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. In *ICLR*, 2017.

[Pit02] Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 2002.

[PJ77] Arthur V Peterson Jr. Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, 1977.

[QHI+97] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 1997.

[Ras00] Carl Edward Rasmussen. The infinite gaussian mixture model. In *NeurIPS*, 2000.

[RAY+16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[RCW15] Alexander M. Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*, 2015.

[RM15] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *ICML*, 2015.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, 2015.

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

[Rob86] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 1986.

[ROC+18] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 2018.

[RPEB16] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, 2016.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 2000.

[Rub05] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2005.

[RYJK+07] Lynn A Gloeckler Ries, John L Young Jr, Gretchen E Keel, Milton P Eisner, Yi Dan Lin, and Marie-Josephe D Horner. Cancer survival among adults: US SEER program, 1988–2001. *Patient and tumor characteristics SEER Survival Monograph Publication*, 2007.

[Set94] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, 1994.

[SFG+12] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 2012.

[SGZ+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

[Sil86] Bernard W Silverman. *Density estimation for statistics and data analysis*. CRC press, 1986.

[Sil18] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[SJS17] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

[SKDo+08] Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *NeurIPS*, 2008.

[SKS+15] Sanjiv J Shah, Daniel H Katz, Senthil Selvaraj, Michael A Burke, Clyde W Yancy, Mihai Gheorghiade, Robert O Bonow, Chiang-Ching Huang, and Rahul C Deo. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 2015.

[SLML16] Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in medicine*, 2016.

[SSC+18] Dinghan Shen, Qinliang Su, Paidamoyo Chapfuwa, Wenlin Wang, Guoyin Wang, Lawrence Carin, and Ricardo Henao. Nash: Toward end-to-end neural architecture for generative semantic hashing. In *ACL*, 2018.

[STS+17] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *ICLR*, 2017.

[SVC+10] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and

Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 2010.

[SWD+18] Jincheng Shen, Lu Wang, Stephanie Daignault, Daniel E Spratt, Todd M Morgan, and Jeremy MG Taylor. Estimating the optimal personalized treatment strategy based on selected variables to prolong survival via random survival forest with weighted bootstrap. *Journal of biopharmaceutical statistics*, 2018.

[SWH09] Michael Schemper, Samo Wakounig, and Georg Heinze. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine*, 2009.

[SZRM18] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *ICLR*, 2018.

[TJCP16] Ludovic Trinquart, Justine Jacot, Sarah C Conner, and Raphaël Porcher. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 2016.

[Tsi07] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[UDR18] Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. Deep reinforcement learning of marked temporal point processes. In *NeurIPS*, 2018.

[vdLR03] Mark J van der Laan and James M Robins. Unified approach for causal inference and censored data. In *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

[VdLR11] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[VEJ20] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *The New England Journal of Medicine*, 2020.

[Vil08] Cédric Villani. *Optimal transport: old and new*. Springer Science & Business Media, 2008.

[VLL+10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning

useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 2010.

[VLR17] Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. Pre-processing censored survival data using inverse covariance matrix based calibration. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[WA18] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2018.

[WBM+04] Timothy J Wilt, Hanna E Bloomfield, Roderick MacDonald, David Nelson, Indulis Rutks, Michael Ho, Gregory Larsen, Anthony McCall, Sandra Pineros, and Anne Sales. Effectiveness of statin therapy in adults with coronary heart disease. *Archives of internal medicine*, 2004.

[Wei92a] L. J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 1992.

[Wei92b] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine*, 1992.

[Wil92] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

[XBK+15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[XDM+19] Eryu Xia, Xin Du, Jing Mei, Wen Sun, Suijun Tong, Zhiqing Kang, Jian Sheng, Jian Li, Changsheng Ma, Jianzeng Dong, et al. Outcome-driven clustering of acute coronary syndrome patients using multi-task neural network with attention. In *MedInfo*, 2019.

[XFY+17] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In *NeurIPS*, 2017.

[XGF16] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

[XTH20] Zidi Xiu, Chenyang Tao, and Ricardo Henao. Variational learning of individual survival distributions. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.

[XXY+18] Shuai Xiao, Hongteng Xu, Junchi Yan, Mehrdad Farajtabar, Xiaokang Yang, Le Song, and Hongyuan Zha. Learning conditional generative models for temporal point processes. In *AAAI*, 2018.

[YBD+16] Salim Yusuf, Jackie Bosch, Gilles Dagenais, Jun Zhu, Denis Xavier, Lisheng Liu, Prem Pais, Patricio López-Jaramillo, Lawrence A Leiter, Antonio Dans, et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *New England Journal of Medicine*, 2016.

[YGLB11] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NeurIPS*, 2011.

[YZWY17] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

[ZBvdS20] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *AISTATS*, 2020.

[ZCM+18] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 2018.

[ZGF+17] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *ICML*, 2017.

[ZTU+12] Lihui Zhao, Lu Tian, Hajime Uno, Scott D Solomon, Marc A Pfeffer, Jerald S Schindler, and Lee Jen Wei. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical trials*, 2012.

[ZYH16] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, 2016.

[ZYW19] Panpan Zheng, Shuhan Yuan, and Xintao Wu. Safe: A neural survival analysis model for fraud early detection. *AAAI*, 2019.

[ZZ18] Quan Zhang and Mingyuan Zhou. Nonparametric Bayesian lomax delegate racing for survival analysis with competing risks. In *NeurIPS*, 2018.

Biography

Paidamoyo Chapfuwa received B.S.E. with distinction, M.S., and Ph.D. degrees in electrical and computer engineering from Duke University, Durham, NC, USA, in 2013, 2018, and 2021, respectively. Paidamoyo has been advised throughout her Ph.D. by Drs. Lawrence Carin and Ricardo Henao. Her research focuses on developing modern machine learning approaches, i.e., representation and deep learning, to characterize individualized survival (event times) from clinical data such as electronic health records and more recently, immunomics. Her work incorporates statistical techniques from causal inference, generative modeling, and Bayesian nonparametrics. Her work has culminated in publications at prestigious venues such as IEEE, ACM, ACL, and ICML. See `https://paidamoyo.github.io` for more information.