Leveraging Heterogeneity in Time-to-Event Predictions

Chirag Nagpal

CMU-LTI-23-004

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee

Artur Dubrawski (Chair)	Carnegie Mellon University
Louis-Philippe Morency	Carnegie Mellon University
Bhiksha Ramakrishnan	Carnegie Mellon University
Russell Greiner	University of Alberta
Katherine Heller	Google Research

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Acknowledgement

I would first and foremost like to thank my advisor, Professor Artur Dubrawski, who instilled more faith and confidence in me than I did in myself. Artur gave me opportunities to pursue summer research even though I came from a relatively unknown undergraduate program. Subsequently, as my doctoral advisor, Artur was extremely patient and supportive and let me pursue an independent research agenda which gave me immense confidence to grow as an independent researcher.

I would also like to convey my gratitude to Prof. Bhiksha Raj and Prof. Rita Singh who first introduced me to Carnegie Mellon during my undergraduate days. During my PhD, Profs. Bhiksha and Rita were extremely warm and welcoming and never let me feel away from home.

I would also like to extend my gratitude to members of my thesis committee, Prof. Russ Greiner who gave me detailed feedback on my manuscript which immensely helped improve the quality of the final thesis, Dr. Katherine Heller for extending multiple opportunities to interact and learn from colleagues at Google and Prof. LP Morency from whose course first exposed me to deep learning.

During my graduate studies, I received immense support from Kush Varshney and Berk Ustun. Both Kush and Berk were very encouraging mentors and made me think about solving real-world problems in a holistic sense beyond just academic publications.

My thesis research involved extensive collaboration, I would especially like to acknowledge members of the Auton Lab including Predrag Punosevac, Xinyu (Rachel) Li, Mingzhu Liu, Keith Dufendach, Vincent Jeanselme, Willa Potosnak, Mononito Goswami, Vedant Sanil, Emma Erickson, Vanessa Le and Sibi Venkatesan as well as my colleagues from my internships, Stephen Pfohl, Negar Rostamzadeh, Dennis Wei, Bhanukiran Vinzamuri and Robert Tillman for being extremely supportive collaborators.

During my time at Carnegie Mellon, I was extremely fortunate to have Rachel Burcin, Stacey Young and Kathleen Schaich supporting me administratively. Graduate School would have been a much more difficult experience had it not been for them helping me out whenever I was in need of their support.

Surviving Pittsburgh would have been impossible had it not been for some of the closest friends I made in Evangelia Spiliopolou, Paul Michel, Raphael Olivier, Ankit Shah, Maria Ryskina, Sid Dalmia, Dhruv Malik, Shruti Rijhwani, Shohin Mukherjee, Laura Simandl, Mihir Mongia, Baljit Singh, Tyler Vuong, Samuel Sokota, Jeremy Cohen, Sumeet Singh, Yash Savani, Afonso Tinoco and Christina Akirtava. I would also like to acknowledge my office mates over the years Mahmoud Alismail, Patrick Fernandez and Shruti Palaskar who kept up with having me as an office-mate.

Finally, I would like to thank my parents and my grandparents whose constant support and appreciation made this thesis possible.

Contents

Preface

Introduction

Real-world decision-making often requires reasoning about **when** an **event** will occur. The overarching goal of such reasoning is to help aid decision-making for optimal triage and subsequent intervention. Such problems involving estimation of Times-to-an-Event frequently arise across multiple application areas, including,

Healthcare and Bio-informatics: More commonly known as 'Survival Analysis' involves prognostication of an adverse physiological event like a stroke, the onset of cancer, re-hospitalization, and mortality. Time-to-event or survival analysis can be used to proactively mitigate adverse outcomes and extend the longevity of patients.

Internet Marketing and e-commerce: Models employed for estimating customer churn and retention in large commercial organizations are essentially time-to-event regression models and help determine best practices to maximize customer retention.

Predictive Maintenance: Reliability engineering and systems safety research involves the use of remaining useful life prediction models to help extend the longevity of machinery and equipment by proactive part and component replacement.

Finance and Actuarial and Sciences: Time-to-Event models are ubiquitous in the estimation of optimal financial strategies for setting insurance premiums, as well as estimating credit defaulting behavior.

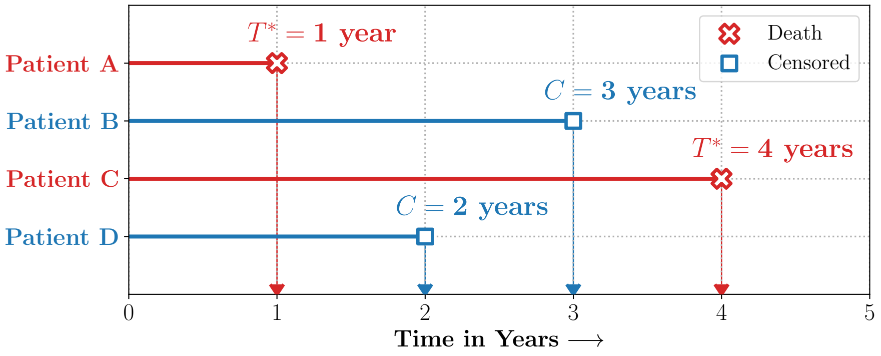


Figure 1: **Censoring and Time-to-Event Predictions: Patients A** and **C** died **1** and **4 years** from entry into the study, whereas **Patients B** and **D** exited the study without experiencing death (were lost to follow up) at **2** and **3 years** from entry in the study. *Time-to-Event* or *Survival Regression* thus involves estimates that are adjusted for individuals whose outcomes were censored.

Figure 1 illustrates a typical example of a Time-to-Event problem in healthcare. The challenge of working with time-to-event data is compounded by the fact that as evidenced in the figure, such data typically includes individuals whose outcomes are unobserved, or 'censored,' either due to a loss of follow-up or end of the study. Discretizing time-to-event outcomes to predict if an event will occur is a common approach in standard machine learning. However, this neglects temporal context, which could result in models that misestimate and lead to poorer generalization

Time-to-Event Regression, often referred to as *Survival Analysis* or *Censored Regression* involves learning of statistical estimators of the survival distribution of an individual given their covariates. As opposed to standard regression, survival analysis is challenging as it involves accounting for outcomes *censored* due to loss of follow up. This circumstance is common in, e.g., bio-statistics, predictive maintenance, and econometrics. With the recent advances in machine learning methodology, especially *deep learning*, it is now possible to exploit expressive representations to help model survival outcomes. My thesis contributes to this new body of work by demonstrating that problems in survival analysis often manifest inherent *heterogeneity* that can be effectively discovered, characterized, and modeled to learn better estimators of survival.

Heterogeneity may arise in a multitude of settings in the context of survival analysis. Some examples include heterogeneity in the form of input features or covariates (for instance, static vs. streaming, *time-varying* data), or multiple outcomes of simultaneous interest (more commonly referred to as *competing risks*). Other sources of heterogeneity involve latent subgroups that manifest different base survival rates or diverse responses to an intervention or treatment.

In this thesis, I aim to demonstrate that carefully modelling the *inherent structure* of heterogeneity can boost predictive power of survival analysis models while improving their specificity and precision of estimated survival at an individual level. An overarching methodological framework of this thesis is the application of graphical models to impose inherent structure in time-to-event problems that explicitly model heterogeneity, while employing advances in deep learning to learn powerful representations of data.

Furthermore, through innovative probabilistic and numerical optimization techniques we explore how the learnt estimators can be made actionable tools for decision support. By enforcing constraints that improve model interpretability, we explore opportunities for enhancing the utility of such models, a requirement that is paramount in critical scenarios such as healthcare.

Our major contributions can be summarized as follows:

✓ **Part I: Estimators for Survival Analysis**

- ✓ **Chapter 1 :** We first introduce a new approach, *Deep Survival Machines* to estimating the survival distribution $\mathbb{P}(T > t|X)$ using a parametric mixture model with representations learnt with neural networks. We demonstrate the superiority of this model type in situations with competing risks where knowledge is shared across outcomes. In the *IEEE Journal of Biomedical and Health Informatics* (Nagpal et al., 2021c).
- ✓ **Chapter 2 :** We extend Deep Survival Machines to settings where input data has -variates that vary over time prior to when the survival curve is estimated. We learn time-varying representations of the covariates with the use of recurrent neural architectures and demonstrate the effectiveness of the result at estimating length of stay and mortality in critically ill Intensive Care Unit patients. This work was published at the *AAAI Spring Symposium on Survival Prediction* (Nagpal et al., 2021b).
- ✓ **Chapter 3 :** We propose a model that expresses the time-to-event distribution using a mixture of Cox models, parameterized with neural representations. We develop an efficient Expectation Maximization

based inference procedure for this model that involves Monte-Carlo sampling to infer latent components. We demonstrate that our approach has improved calibration over existing alternatives, especially on minority demographics. This work was completed during an internship at Google Brain and is accepted for publication at the *Machine Learning for Healthcare Conference* (Nagpal et al., 2021d).

✓ **Part II: Subgroup Discovery for Heterogeneous Treatment Effects**

✓ **Chapter 4 :** We hypothesize that in the observational settings, the individuals' responses to a treatment or intervention manifest heterogeneity conditioned on certain latent characteristics of the subjects. We propose *Heterogenous Effects Mixture Model* to recover the latent subgroups from data if they exist, while estimating confounding effects, using deep learning. Work was carried out during an internship with IBM Research and published at the *ACM Conference on Health, Inference and Learning* (Nagpal et al., 2020).

✓ **Chapter 5 :** We propose to extend our *Heterogenous Effects Mixture Model* to settings with censored outcomes. In such settings, latent group membership may jointly affect base survival rates as well as prognosed treatment effects. By decoupling the base effect of survival with treatment effect heterogeneity, we propose to extend estimators introduced in Part I to reflect counterfactual outcomes and discover actionable phenotypes from the perspective of decision support. This work was presented at the *ACM Conference on Knowledge Discovery and Datamining* 2022 (Nagpal et al., 2022a).

✓ **Part III: Interpretable Approaches for Actionable Time-to-Event Analysis.**

✓ **Chapter 6 : Recovering Sparse and Interpretable Subgroups with Heterogeneous Treatment Effects with Censored Time-to-Event Outcomes**
In this section we propose a statistical approaches to recovering sparse phenogroups (or subtypes) that demonstrate differential treatment effects as compared to the study population. Our approach involves modelling the data as a mixture while enforcing parameter shrinkage through structured sparsity regularization. We propose a novel inference procedure for the proposed model and demonstrate its efficacy in recovering sparse phenotypes across large landmark real world clinical studies in cardiovascular health. Presented at the *Machine Learning for Health Symposium 2022* (Nagpal et al., 2023).

✓ **Chapter 7 : Integer Risk Scoring with Censored Time-to-Event Outcomes**
Integer risk scoring methods are ubiquitous across multiple aspects of healthcare for disease severity estimation and subsequent intervention, management and resource allocation. Historically, integer risk scoring methods have involved heuristic methods driven by domain expertise. Recently there has been an interest in using machine learning to recover optimal solutions to the integer scoring problem for binary outcomes. However, a large number of outcomes are time-to-events. In this chapter, we propose an alternate formulation for the integer risk scoring system involving Mixed Integer Non-Linear Programming that directly models the relative hazard for each individual. We demonstrate that, as opposed to existing methods, our formulation has robust performance across multiple time horizons. We demonstrate the ability of our approach to recover interpretable disease staging models across multiple real world health studies.

Papers of Direct Relevance to Thesis

1. Nagpal, C., Wei, D., Vinzamuri, B., Shekhar, M., Berger, S. E., Das, S., & Varshney, K. R. (2020, April). Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines. *In Proceedings of the ACM Conference on Health, Inference, and Learning* (pp. 19-29).

2. Nagpal, C., Li, X. R., & Dubrawski, A. (2021). Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics* (**Spotlight Presentation in NeurIPS 2019 ML4H Workshop, Top 3% of over 300 submissions**).

3. Nagpal, C., Jeanselme, V., & Dubrawski, A. (2021, May). Deep Parametric Time-to-Event Regression with Time-Varying Covariates. *In Survival Prediction-Algorithms, Challenges and Applications* (pp. 184-193). PMLR.

4. Nagpal, C., Yadlowsky, S., Rostamzadeh, N., & Heller, K. (2021). Deep Cox mixtures for survival regression. *Machine Learning for Health Conference*. PMLR.

5. Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. 2022. Counterfactual Phenotyping with Censored Time-to-Events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. Association for Computing Machinery, New York, NY, USA.

6. Nagpal, C., Potosnak, W. & Dubrawski, A.. (2022). auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data. *Proceedings of the 7th Machine Learning for Healthcare Conference*, in *Proceedings of Machine Learning Research*.

7. Nagpal, C., Sanil, V., & Dubrawski, A. (2023). Recovering Sparse and Interpretable Subgroups with Heterogeneous Treatment Effects with Censored Time-to-Event Outcomes. *arXiv preprint*.

Additional Papers and Clinical Abstracts

1. Nagpal, C., Li, X., Pinsky, M. R., & Dubrawski, A. (2019, October). Dynamically Personalized Detection of Hemorrhage. In Machine Learning for Healthcare Conference (pp. 109-123). PMLR.

2. Nagpal, C., Miller, K., Pellathy, T., Hravnak, M., Clermont, G., Pinsky, M., & Dubrawski, A. (2017, November). Semi-supervised prediction of comorbid rare conditions using medical claims data. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 478-485). IEEE.

3. Nagpal, C., Tillman, R. E., Reddy, P., & Veloso, M. (2019). Bayesian Consensus: Consensus Estimates from Miscalibrated Instruments under Heteroscedastic Noise. In 2019 NeurIPS Workshop on Robust AI in Financial Services.

4. Pellathy, T., Saul, M., Clermont, G., Nagpal, C., Dubrawski, A., Pinsky, M., & Hravnak, M. (2018). Accuracy of Identifying Venous Thromboembolism by Administrative coding compared to Manual Review. Critical Care Medicine, 46(1), 85.

5. Nagpal C & Dubrawski A. Interpretable Treatment Prioritization Rule defines Diabetic patients that benefit from prompt coronary revascularization. J Am Coll Cardiol. 2023 Mar, 81 (8_Supplement) 2263.

Glossary

Time-to-Event typically refers to a quantity of interest that represents the elapsed duration between beginning of the study (or the time when an individual enters a study) till the time at which an event of interest takes place. In the case of health data this could be an event like Death or Stroke or a combination of multiple such events of clinical relevance. Survival analysis typically involve statistical estimation of the distribution of time-to-events.

Survival Probability is the probability that a time-to-event for a subject would occur only beyond a certain period of time. Typically notated as $\boldsymbol{S}(t)$, the survival probability is mathematically equal to unity minus the cumulative density of the time-to-event at a certain time t. Thus,

$$\begin{aligned}\boldsymbol{S}(t) = \mathbb{E}[\mathbf{1}\{T > \boldsymbol{t}\}] &= \mathbb{P}(T > \boldsymbol{t}) \\ &= 1 - \mathbb{P}(T \leq \boldsymbol{t}).\end{aligned} \tag{1}$$

Kaplan-Meier (KM) Estimator It is a non-parametric estimate of the survival function that assumes that the time-to-event and the censoring distribution are independent. The **KM** Estimator can be given as,

$$\widehat{\boldsymbol{S}}(t) = \prod_{i\colon t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i.

Hazard Rate The hazard rate, typically notated as $\boldsymbol{\lambda}(\cdot)$, of a time-to-event is the probability that an event will occur in a given time interval, given that the event has not already occurred. It is a measure of the risk of an event occurring at a particular point in time. The hazard rate can be used to compare the risk of an event occurring between different groups of people or under different conditions.

Suppose that an individual has survived for a time t and we desire the probability that it will not survive for an additional time dt, the hazard rate $\boldsymbol{\lambda}(\cdot)$ over time is,

$$\boldsymbol{\lambda}(t) = \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}(t \leq T < t + \mathrm{d}t)}{\mathrm{d}t \cdot \boldsymbol{S}(t)} = \frac{\mathbb{P}(T = t)}{\boldsymbol{S}(t)}, \quad \text{or} \quad \boldsymbol{\lambda}(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log \boldsymbol{S}(t). \tag{2}$$

Cumulative Hazard Rate The cumulative hazard rate, also known as the cumulative hazard function for a time-to-event distribution, is a function that measures the total amount of hazard (risk) that has been accumulated up to a certain point in time. Typically denoted as $\boldsymbol{\Lambda}(\cdot)$, it is calculated by integrating the hazard function over time. Thus,

$$\boldsymbol{\Lambda}(t) = \int_0^t \boldsymbol{\lambda}(t), \quad \text{and} \quad \boldsymbol{S}(t) = \exp\big(-\boldsymbol{\Lambda}(t)\big). \tag{3}$$

Proportional Hazards (PH) An assumption common to survival and time-to-event analysis that involves assuming that the hazard rates of two individuals or sub-populations are constant over time. The constant value is referred to as the **hazard-ratio** of the study.

Semi-Parametric Models A class of modelling approaches that involve both parametric and non-parametric components. Note that Semi-Parametric models are more general than non-parametric or parametric models.

Cox Model The Cox Proportional Hazards model (Cox, 1972) is a popular semi-parametric regression approach for estimating a survival curve for an individual conditioned on its co-variates. The Cox Model is subject to assumptions of Proportional Hazards, that is, that the conditional hazard rates for any two individuals are constant across time. Mathematically the Cox model assumes that following parametrization,

$$\log \boldsymbol{\lambda}(t|X = \boldsymbol{x}) = \log \boldsymbol{\lambda}_0(t) + f(\boldsymbol{\theta}, \boldsymbol{x}).$$

Typically the function $f(\cdot)$ is assumed to be affine. More recent research has involved modelling $f(\cdot)$ using neural networks. $\boldsymbol{\lambda}_0(\cdot)$ is the non-parametric baseline hazard rate.

Partial Likelihood The parameters of the Cox Proportional Hazards model are estimated by minimizing the partial likelihood. The partial likelihood is so called as it is independent of the non-parametric component $\boldsymbol{\lambda}_0(\cdot)$ representing the base hazard rate. Mathematically, the partial likelihood $\mathcal{PL}(\boldsymbol{\theta})$ is defined as

$$\mathcal{PL}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\boldsymbol{\lambda}(t|\boldsymbol{x}_i)}{\sum\limits_{j\in\mathcal{R}(t_i)} \boldsymbol{\lambda}(t|\boldsymbol{x}_j)} = \prod_{i:\delta_i=1} \frac{\cancel{\boldsymbol{\lambda}_0(t)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathcal{R}(t_i)} \cancel{\boldsymbol{\lambda}_0(t)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)} \tag{4}$$

$$= \prod_{i:\delta_i=1} \frac{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathcal{R}(t_i)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)}. \tag{5}$$

where $\mathcal{R}(t_i)$ is the 'risk set' – the set of individuals that survived beyond time t_i.

Breslow's Estimator The non-parametric component of the Cox Proportional Hazards model, the Base Cumulative Hazard can be estimated using the Breslow's estimator (Breslow, 1972a).

$$\widehat{\boldsymbol{\Lambda}}_0(t) = \sum_{i:t_i<t} \frac{1}{\sum\limits_{j\in\mathcal{R}(t_i)} \exp\big(f(\widehat{\boldsymbol{\theta}};\boldsymbol{x}_j)\big)}, \quad \text{and,} \quad \widehat{\boldsymbol{S}}_0(t) = \exp\big(-\widehat{\boldsymbol{\Lambda}}_0(t)\big) \tag{6}$$

here, $\widehat{\boldsymbol{S}}_0(t)$ is the corresponding base survival rate. The individual survival can then be estimated as,

$$\widehat{\boldsymbol{S}}(t|X = \boldsymbol{x}) = \boldsymbol{S}_0(t)^{\exp(f(\boldsymbol{\theta},\boldsymbol{x}))}.$$

Heteogeneous Treatment Effects (HTE) refers to the variation in the magnitude and direction of the effect of a treatment or an intervention across different subgroups of a population. Not everyone responds to a treatment in the same fashion. Some individuals may benefit more from a treatment than others, and some people may not benefit at all (or worse, are harmed).

Mixed Integer Non-Linear Programming (MINLP) It is a class of numerical optimization techniques involving optimization of a non-linear function (typically a convex function) over a mixed set of integer and continuous variables. MINLP problems are typically NP-hard and involve the use of dynamic programming techniques such as branch-and-bound search to make them tractable to solve.

Preliminaries

Throughout this thesis, unless specified otherwise, we will work with a dataset of right-censored instances $\mathcal{D} := \{(\boldsymbol{x}_i, \boldsymbol{t}_i, \delta_i)\}_{i=1}^{n}$ where $\boldsymbol{x}_i$ is the set of covariates of an individual. $\boldsymbol{t}_i \in \mathbb{R}^{+}$ is the time to event or censoring as indicated by the indicator $\delta_i \in \{0, 1\}$. Time-to-event or survival estimation problem thus reduces to estimating the conditional distribution of survival notated as

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}\{T > \boldsymbol{t}\}|X = \boldsymbol{x}] \\
&\qquad = \mathbb{P}(T > \boldsymbol{t}|X = \boldsymbol{x}) \\
&\qquad = 1 - \mathbb{P}(T \leq \boldsymbol{t}|X = \boldsymbol{x}). \qquad (7)
\end{aligned}$$

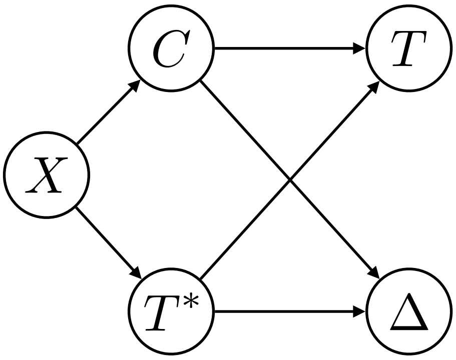


Figure 2: Conditional Independence of the True Time-to-Event T^* and the censoring times C. Only X, T and Δ are observed.

Here, T refers to the distribution of the censored survival time $T = \min(T^*, C)$, where T^* is the distribution of the true time-to-event and C is the distribution of the censoring time (Figure 2). Δ is the distribution of the censoring indicator $\Delta = \mathbf{1}\{T^* < C\}$. Typically we do not observe the true event times for individuals lost to follow up as in Figure 1. Note that we will assume that the event of interest can only take place once. This assumption is natural when modelling time-to-death or time to the first event, a common end-point in clinical settings. Events that can recur along the lifetime of an individual are beyond the scope of the discussion for the purposes of this thesis.

For these individuals, we observe the censored survival time, T and an indicator of if they were censored, $\Delta = T < C$. Assuming conditional independence between T and C ie. $T \perp C|X$ allows identification of the distribution of $\mathbb{P}(T|X)$.

For the censored individuals, we maximize the probability corresponding to the survival function. The likelihood, $\ell(\cdot)$ under censoring is thus given as

$$\boldsymbol{\ell}(\{\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\delta}\}) \propto \mathbb{P}(T = \boldsymbol{t}|X = \boldsymbol{x})^{\delta}\mathbb{P}(T > \boldsymbol{t}|X = \boldsymbol{x})^{1-\delta}. \qquad (8)$$

Note that $\mathbb{P}(T > \boldsymbol{t}|X = \boldsymbol{x})$ is the conditional survival function and typically notated as $\boldsymbol{S}(\boldsymbol{t}|X = \boldsymbol{x})$. Often in survival analysis literature likelihoods are expressed in terms of instantaneous hazard rates $\boldsymbol{\lambda}(t)$. The instantaneous hazard maybe defined as the event rate at a time t, conditional on survival $(T > t)$ till that time. Thus,

$$\boldsymbol{\lambda}(t) = \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}(t \leq T < t + \mathrm{d}t)}{\mathrm{d}t \cdot S(t)} = \frac{\mathbb{P}(T = t)}{\boldsymbol{S}(t)} \quad \text{or,} \quad \boldsymbol{\lambda}(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log \boldsymbol{S}(t). \qquad (9)$$

Now reasoning in terms of hazard rates, we can rewrite Equation 8 equivalently as

$$\boldsymbol{\ell}(\{\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\delta}\}) \propto \boldsymbol{\lambda}(\boldsymbol{t}|X = \boldsymbol{x})^{\delta}\boldsymbol{S}(\boldsymbol{t}|X = \boldsymbol{x}). \tag{10}$$

Broadly the popular approaches for maximizing the likelihood above are classified as,

Parametric Assumes the distribution of the time-to-event $\mathbb{P}(T|X = \boldsymbol{x})$ is parametric like Weibull or Log-Normal. Examples includes the popular Accelerated Failure Time model.

Non-Parametric Involves learning kernels or similarity functions of the input covariates followed by a non-parametric (Kaplan-Meier or Nelson-Aalen) estimation of the survival rate weighted with a kernel obtained either heuristically or learnt via representation learning.

Semi-Parametric the Cox Proportional Hazards model and its extensions arguably, remain the most popular approaches and are classified as *semi-parametric*. The Cox model involves a two step estimation where the feature interactions are learnt through a parametric model followed by non-parametric estimation of the base survival (hazard) rate.

Part I

Machine Learning Estimators for Survival Analysis

Motivation

Survival regression is a field of statistics and machine learning that deals with the estimation of a survival function representing the probability of an event of interest, typically a failure, to occur beyond a certain time in the future. Survival regression models time-to-event by estimating the survival function, $\mathbb{S}(\cdot|X) \triangleq \mathbb{P}(T > t|X)$, conditional on X, the input covariates. Examples include estimating the survival times of patients after certain treatment using clinical variables, or predicting the failure times of machines using their usage histories, etc. Survival regression differs from standard regression due to censoring of data, i.e. observation of some subjects stops before occurrence of an event of interest.

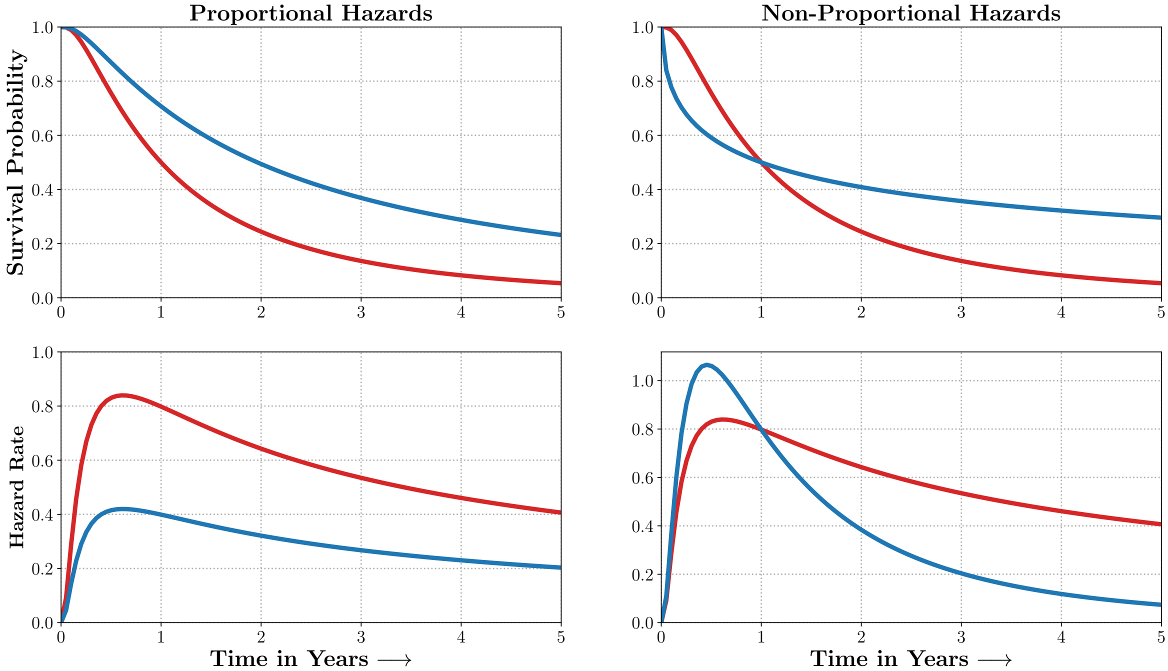


Figure 3: **Non-Proportional Hazards**: When the Proportional Hazards assumptions are satisfied, the Survival Curves and their corresponding Hazard Rates dominate each other and do not intersect. In many real world scenarios however, the survival curves. Part I of this thesis proposes flexible estimators of Time-to-Events in the presence of non-proportional hazards.

Classical statistical machine learning techniques for survival regression rely on non-parametric or semi-parametric methods for survival function estimation, primarily because they make working with censored data relatively straightforward. However, non-parametric methods may suffer from the curse of dimensionality, and semi-parametric approaches usually depend on strong modeling assumptions. In particular, the prevailing assumption of constant proportional hazards over lifetime as proposed by Cox (1972) in the Proportional Hazards model (commonly referred to as CPH), is very likely to be unrealistic in many practical scenarios encountered in healthcare, predictive maintenance, econometrics, or operations research. See Figure 3 for an illustration of a violation of the proportional hazards assumption.

Significant volume of recent research is focused on improving the CPH model. Kraisangka & Druzdzel (2016, 2018) combined Bayesian networks with the CPH model to improve both the model interpretability and predictive power. Researchers have also tried to incorporate structural sparsity, regularization, as well as active and multitask learning when available data is scarce (Vinzamuri et al., 2014; Vinzamuri & Reddy, 2013; Li et al., 2016). Other efforts have involved incorporating non-linear interactions between the covariates in the original model. Rosen & Tanner (1999) proposed using a mixture of linear experts for the original Cox model. Other approaches for incorporating non-linearities considered replacing the linear interaction terms in the CPH model with deep neural networks, first explored in Faraggi & Simon (1995), followed by Xiang et al. (2000), and again recently by Katzman et al. (2018) with their *DeepSurv* approach. Extensions to that work have involved convolutional neural networks and active learning for healthcare applications in oncology (Mobadersany et al., 2018; Nezhad et al., 2019). However, those approaches are still subject to the same strong assumption of proportional hazards as the original CPH formulation.

In the first part of this thesis, we investigate improvements over existing survival analysis methods by proposing approaches that involve modeling the Time-to-Event distribution conditioned on the input data as a fixed size mixture, while learning expressive representations of the input data using deep learning. Our contributions allow for flexible modelling of the Time-to-Event analysis challenges without making strong assumptions on the event time distributions.

The first chapter of this thesis proposes a general parametric approach, *Deep Survival Machines* (DSM), to model static survival data with censored outcomes. We apply the DSM methodology to several real-world datasets and demonstrate its advantages over existing baselines. We further evaluate DSM's capability to model data with multiple competing outcomes or risks.

The subsequent chapter extends the Deep Survival Machines methodology to the longitudinal or dynamic setting involving covariates that vary over time. The performance is benchmarked on Length of Stay and Mortality prediction from critically-ill ICU patients.

Finally, in Chapter 3, we propose a finite size mixture of Cox regressions and an efficient learning algorithm to perform inference under such model. We demonstrate that this approach can be effective at modeling non-proportional hazards and has competitive discriminative capability while outperforming several baselines in terms of calibration.

Chapter 1

Fully Parametric Survival Regression and Representation Learning

1.1 Preliminaries

We assume that the survival data we have access to is *right-censored*. This implies that our data, $\mathcal{D}$ is a set of tuples $\{(\mathbf{x}_i, t_i, \delta_i)\}_{i=1}^{N}$. Where typically, $\mathbf{x}_i \in \mathbb{R}^d$ are features associated with an individual t_i is the time at which an event of interest took place, or the censoring time and δ_i is an indicator that signifies whether t_i is event time or censoring time. For a given individual, we only either observe the actual failure or censoring time but not both. For simplicity, it is assumed that the true data generating process is such that the censoring process is independent of the actual time to failure. We denote the uncensored subset $(\delta = 1)$ of data as $\mathcal{D}_U$ and the censored $(\delta = 0)$ subset as $\mathcal{D}_C$.

1.2 Proposed Model: *Deep Survival Machines*

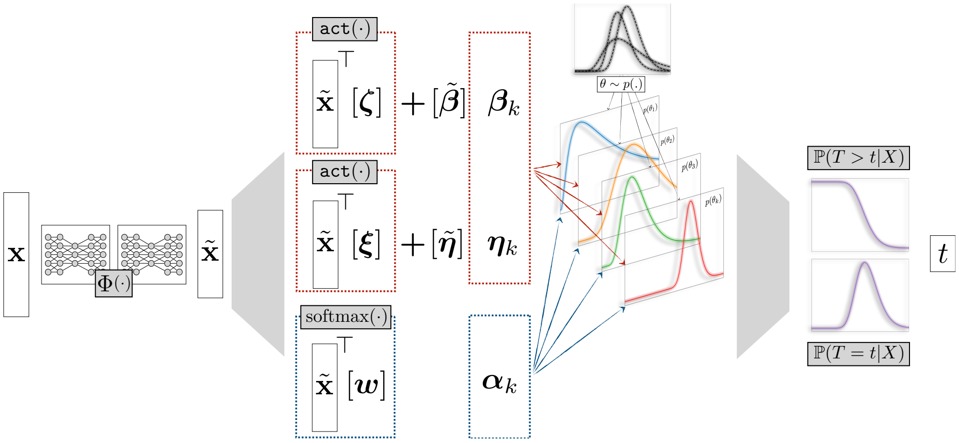


Figure 1.1: The proposed *Deep Survival Machines* pipeline. The input features, $\mathbf{x}$ are passed through a deep multilayer perceptron followed by a softmax over mixture size, K. The Conditional Distribution of $\mathbb{P}(T|X = \mathbf{x})$ is then described as a mixture of K Primitive distributions, drawn from some prior.

In this section we describe our approach, *Deep Survival Machines* (DSM) architecture and inference in further detail. Fig. 1.1 is a visual representation of our approach while Fig. 1.2 describes the model in plate notation.

1.2.1 Primitive Distributions

We choose to model the conditional distribution $\mathbb{P}(T|X = \mathbf{x})$ as a mixture over K well-defined, parametric distributions which we will refer to as PRIMITIVE distributions for the remainder of this chapter. Given that we are modeling survival times, a natural assumption for these PRIMITIVE distributions is to have support only in the space of non-negative reals. Another property of interest is to have a closed form solution for the CDF (cumulative distribution function), as this would enable the use of gradient based optimization for Maximum Likelihood Estimation.

Table 1.1: Choices for the PRIMITIVE distributions.

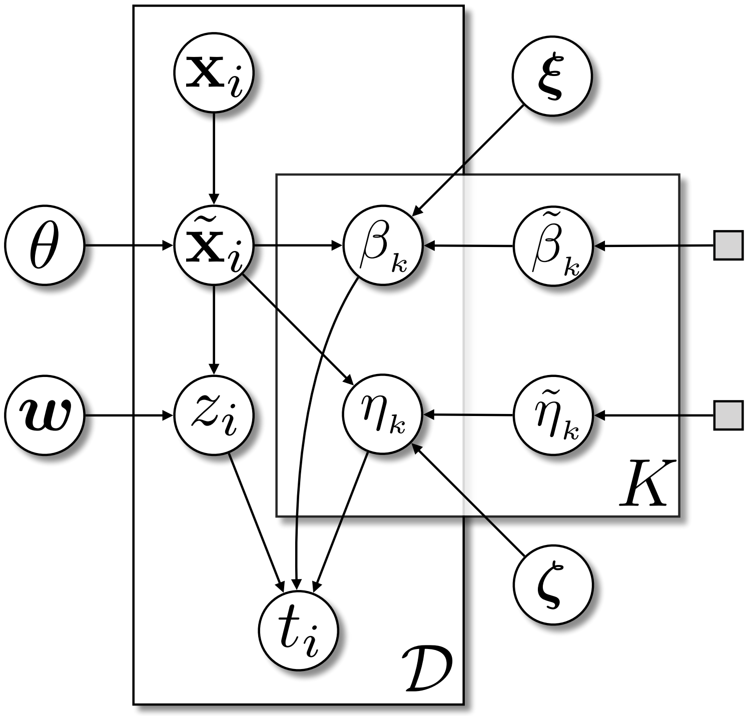
	WEIBULL	LOG-NORMAL
PDF(t)	$\frac{\eta}{\beta}\left(\frac{t}{\beta}\right)^{\eta-1}e^{-\left(\frac{t}{\beta}\right)^{\eta}}$	$\frac{1}{t\beta\sqrt{2\pi}}e^{-\frac{(\ln t-\eta)^2}{2\beta^2}}$
CDF(t)	$e^{-\left(\frac{t}{\beta}\right)^{\eta}}$	$\frac{1}{2}\text{erfc}\left(-\frac{\ln t-\eta}{\sqrt{2}\beta}\right)$

For DSM, we experiment with two types of distributions that satisfy this property, the Weibull and the Log-Normal distribution. The first of them has closed form PDF (probability distribution function) and CDF. For the Log-Normal, we compute the CDF by using the standard approximation of the complementary error function `erfc` as implemented in `PyTorch`. The full functional forms of the distributions are listed in Table 1.1. We parameterize the β_k and η_k as:

$$\beta_k = \tilde{\beta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\zeta}), \quad \eta_k = \tilde{\eta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\xi})$$

Here the $\texttt{act}(\cdot)$ is the SELU and Tanh activation functions for the Weibull and Log-Normal respectively, and $\boldsymbol{\Phi}(.)$ is a Multilayer Perceptron (MLP). $\mathbf{x}_i$ are the input covariates. $\{\theta, \boldsymbol{\xi}, \boldsymbol{\zeta}, \beta$ and $\eta\}$ are all parameters that are learnt during training. Another set of parameters that are learnt are $\boldsymbol{w}$ that determine the mixture weights for each data point.

Figure 1.2 introduces the proposed model in plate notation and the corresponding conditional independence assumptions of the Graphical Model. The input features, $\boldsymbol{x}_i$, are passed through the MLP, $\boldsymbol{\Phi}_\theta$ to determine the representation $\widetilde{\boldsymbol{x}}_i$. This representation then interacts with the additional set of parameters to determine the mixture weights $\boldsymbol{w}$ and the parameters of each of K underlying survival distributions $\{\eta_k, \beta_k\}_{k=1}^{K}$. The final individual survival distribution for the event time T is a weighted average over these K distributions.



The Generative Story at Test Time

1. $\mathbf{x}_i \sim \mathcal{D}$
   We draw the co-variates of the individual, $\mathbf{x}_i$

2. $\boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\xi} \sim \mathcal{N}(0, 1/\lambda)$
   The parameters of the model are drawn from a zero mean Gaussian distribution.

3. $z_i \sim \text{Discrete}\big(\texttt{softmax}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{w})\big)$
   Conditioned on the covariates, $\boldsymbol{x}_i$ and the parameters, $\boldsymbol{w}$ we draw the latent z_i

4. $\log \tilde{\beta}_k \sim \mathcal{N}(\beta_0, 1/\lambda)$
   $\log \tilde{\eta}_k \sim \mathcal{N}(\eta_0, 1/\lambda)$
   The set of parameters $\{\tilde{\beta}_k\}_{k=1}^{K}$ and $\{\tilde{\eta}_k\}_{k=1}^{K}$ are drawn from the prior β_0 and η_0.

5. $\boldsymbol{t}_i \sim \textsc{Primitive}\big(\beta_k, \eta_k\big)$
   where, $\beta_k = \tilde{\beta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\zeta})$
   $\eta_k = \tilde{\eta_k} + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\xi})$
   Finally, the event time t_i is drawn conditioned on β_{z_i} and η_{z_i}.

Figure 1.2: *Deep Survival Machines* in plate notation.

1.2.2 Parameter Estimation

In order to accommodate for heterogeneity arising in the data, we propose to model the survival distribution of each individual as a fixed size mixture of survival distribution primitives. At test time, the survival function corresponding to this held out individual is described as a weighted mixture of the survival distribution primitives. Here, the weights are a softmax of the output of a deep neural network. At training time, the parameters of the network and the survival distribution primitives are learnt jointly.

Uncensored Loss. We consider the maximum likelihood estimator for the uncensored data which can be written as:

$$\begin{aligned}
\ln \mathbb{P}(\mathcal{D}_U|\boldsymbol{\Theta}) &= \ln\Bigg(\prod_{i=1}^{|\mathcal{D}|}\mathbb{P}(T=t_i|X=\mathbf{x}_i,\boldsymbol{\Theta})\Bigg)\\
&= \sum_{i=1}^{|\mathcal{D}|}\ln\Bigg(\sum_{k=1}^{K}\mathbb{P}(T=t_i|Z=k,\beta_k,\eta_k)\mathbb{P}(Z=k|X=\mathbf{x}_i,\boldsymbol{w})\Bigg)\\
&= \sum_{i=1}^{|\mathcal{D}|}\ln\Bigg(\underset{Z\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})}{\mathbb{E}}[\mathbb{P}(T=t_i|Z,\beta_k,\eta_k)]\Bigg)\\
&\text{(Applying Jensen's Inequality)}\\
&\geq \sum_{i=1}^{|\mathcal{D}|}\Bigg(\underset{Z\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})}{\mathbb{E}}[\ln\mathbb{P}(T=t_i|Z,\beta_k,\eta_k)]\Bigg)\\
&\triangleq \mathbf{ELBO}_U(\Theta)
\end{aligned}$$

Censoring Loss. Proceeding as above, we can write the lower bound of loss for the censored observations as:

$$\begin{aligned}
\ln \mathbb{P}(\mathcal{D}_C|\Theta) &= \ln\Bigg(\prod_{i=1}^{|\mathcal{D}|}\mathbb{P}(T>t_i|X=\mathbf{x}_i,\Theta)\Bigg)\\
&\geq \sum_{i=1}^{|\mathcal{D}|}\Bigg(\underset{Z\sim(\cdot|\mathbf{x}_i,w)}{\mathbb{E}}[\ln\mathbb{P}(T>t_i|Z,\beta_k,\eta_k)]\Bigg)\\
&\triangleq \mathbf{ELBO}_C(\Theta)
\end{aligned}$$

Mitigating Long Tail Bias. Survival distributions with positive support typically have long tails, a complication that adds to the bias when performing Maximum Likelihood Estimation. Note that for the censored instances of data, we are maximizing the probability $\mathbb{P}(T > t)$. One conceivable way of adjusting for the long-tail bias is to instead maximize $\mathbb{P}(t_{\text{max}} > T > t) = \mathbb{P}(T > t) - \mathbb{P}(T > t_{\text{max}})$ where t_{max} is some arbitrarily large value that represents the last event or the follow up time in the study. However, for simplicity, we choose to directly discount the censoring loss by multiplying it with a factor $\alpha \in [0, 1]$, which has a similar effect of diminishing bias arising from the long tails.

Prior Loss. We include the strength of the prior on the β_k, η_k as:

$$\begin{aligned}\mathcal{L}_{\text{prior}} &= \ln\left(\prod_{k=1}^{K}\mathbb{P}(\beta_k,\eta_k|\beta,\eta)\right)\\ &= \sum_{k=1}^{K}\ln\mathbb{P}(\beta_k|\beta)+\ln\mathbb{P}(\eta_k|\eta)\\ &= \lambda\sum_{k=1}^{K}||\beta_k-\beta||_{\mathbf{2}}^{2}+||\eta_k-\eta||_{\mathbf{2}}^{2}.\end{aligned}$$

Combined Loss. We finally combine the individual components of loss described above into:

$$\mathcal{L}_{\text{combined}} = \textbf{ELBO}_U(\Theta)+\alpha\cdot\textbf{ELBO}_C(\Theta)+\mathcal{L}_{\text{prior}}.$$

Here, α is a scalar hyperparameter that trades off the contribution of regression loss vis-à-vis the evidence lower bound of the uncensored observations to the combined objective function. For a complete formulation of the loss function, in terms of functions and parameters please refer to Appendix A.1.

Table 1.2: Descriptive statistics of the datasets used in the experiments.

Dataset	Type	Dataset Dim.	Feature Dim.	No. Events		No. Censoring
SUPPORT	Single Risk	9,105	30	6,201 (68.1 %)		2,904 (31.9 %)
METABRIC	Single Risk	1,904	9	1,103 (57.9 %)		801 (42.1 %)
SYNTHETIC	Competing Risks	30,000	12	Event 1 Event 2	7,600 (25.3%) 7,400 (24.7%)	15,000 (50.0 %)
SEER	Competing Risks	65,481	21	BC CVD	13,564 (20.7%) 4,245 (6.5%)	47,672 (72.8 %)

1.2.3 Handling Multiple *Competing Risks*

We adapt *Deep Survival Machines* to scenarios involving multiple competing risks by allowing learning of a common representation for the multiple risks by passing through a single MLP ($\mathbf{\Phi}(.)$ in Fig. 1.1). This representation then interacts with a separate set of $\{\boldsymbol{\xi},\boldsymbol{\zeta},\boldsymbol{w}\}$ in order to describe the event distribution for each competing risk. Maximum Likelihood Estimation is performed by treating the occurrence of a competing event before the other event as a form of independent censoring. This strategy allows the model to leverage knowledge from the two (in general, more than two) competing tasks by allowing parameter sharing through a single intermediate representation.

1.3 Experiments

We evaluate *Deep Survival Machines* on their ability to measure relative risks for a single event of interest in the presence of censoring, and then we further consider ablation experiments where we artificially increase the amount of censoring to demonstrate the robustness of the proposed approach. Finally, we demonstrate DSM's ability to learn representations of the covariates for transferring knowledge across two events in the *competing risks* scenario with censoring.

1.3.1 Datasets

Single Event/Single Risk. We evaluated performance of the proposed method on the following real-world medical datasets with single events: Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) (Knaus et al., 1995), and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012). A brief introduction of each dataset is provided below.

SUPPORT: The SUPPORT dataset resulted from a study conducted to describe a prognostic model to estimate survival over a 180-day period for 9,105 seriously ill hospitalized patients. Of the 9,105 patients, 6,201 patients (68.1%) were followed to death, with a median survival time of 58 days. We used 30 patient covariates, including age, gender, race, education, income, physiological measurements, co-morbidity information, etc. Missing values of certain physiological measurements were imputed using the suggested normal values[1] and other missing values were imputed using the mean value for numerical features and the mode for categoricals.

METABRIC: The METABRIC data came from a study conducted to determine new breast cancer subgroups and facilitate treatment improvement using patients' gene expressions and clinical variables. The dataset consists of 1,904 patients and 9 features. 1,103 patients (57.9%) were followed to death with a median survival time of 115.9 months. The dataset used was preprocessed as in Katzman et al. (2018) and downloaded from the `PySurvival` library.[2]

Competing Risks. We also evaluated the performances on two datasets with competing risks: a synthetic dataset and the Surveillance, Epidemiology, and End Results (SEER) dataset.

SYNTHETIC: In order to demonstrate the effectiveness of DSM as a representation learning framework, we experiment with synthetic data that is generated following the spirit of Alaa & van der Schaar (2017b) & Lee et al. (2018) using the same generative process as they described.

$$\begin{aligned}
\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \mathbf{x}_3^{(i)} &\sim \mathcal{N}(0, \mathbf{I}) \\
T_1^{(i)} &\sim \exp\left((\gamma_3^\top \mathbf{x}_3^{(i)})^2 + \gamma_1^\top \mathbf{x}_1^{(i)}\right) \\
T_2^{(i)} &\sim \exp\left((\gamma_3^\top \mathbf{x}_3^{(i)})^2 + \gamma_2^\top \mathbf{x}_2^{(i)}\right)
\end{aligned}$$

Here $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \mathbf{x}_3^{(i)})$ is a tuple representing the covariates of the individual i. The Event times T_1 and T_2 are exponentially distributed around functions that are both linear and quadratic in X. We generate 30,000 patients from the distribution out of which 50% are subjected to random right censoring by uniformly sampling the censoring times in the interval $[0, \min\{T_1, T_2\}]$. Clearly, the choice of our distributions for the event times are not independent and would allow a model to leverage knowledge of one event to better predict the other, which is what we intend to demonstrate.

[1] http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc

[2] https://square.github.io/pysurvival/

SEER: This dataset[3]provides information on cancer statistics among the United States population. We focused on the breast cancer patients in the registries of Alaska, San Jose-Monterey, Los Angeles and rural Georgia during the years from 1992 to 2007, with the follow-up period restricted to 10 years. Among the 65,481 patients, 13,564 (20.7%) died due to breast cancer (BC) and 4,245 (6.5%) died due to cardiovascular disease (CVD), which were treated as the two competing risks in our experiments. We used 21 patient covariates, including age, race, gender, diagnostic confirmation, morphology information (primary site, laterality, histologic type, etc.), tumor information (size, type, number etc.), and surgery information. Missing values were imputed using the mean value for numerical features and the mode for categorical features.

1.3.2 Baselines

We compare the performance of DSM to the following competing baseline approaches:

Cox Proportional Hazards (CPH): This is the standard semi-parametric model, making the assumption of constant baseline hazard. The features interact with the learnt set of weights in a log-linear fashion in order to determine the hazard for a held-out individual.

Random Survival Forests (RSF): This is a popular non-parametric approach involving learning an ensemble of trees, adapted to censored survival data (Ishwaran et al., 2008).

DeepSurv (DS): Proposed by Katzman et al. (2018), DeepSurv involves learning a non-linear function that describes the relative hazard of a test instance. It makes the familiar assumption of constant baseline hazard, as does CPH.

DeepHit (DH) (Lee et al., 2018): This approach involves learning the joint distribution of all event times by jointly modelling all competing risks and discretizing the output space of event times.

Fine-Gray (FG) (Fine & Gray, 1999): This is a classic approach used for modelling competing risks that focuses on the cumulative incidence function by extending the proportional hazards model to sub-distributions.

For the SYNTHETIC and SEER datasets with competing risks, we compare performance of DSM to cause-specific (**cs-**) versions of CPH and RSF that involve learning separate survival regressions for each competing event by treating the other event as censored.

Performance Metrics

We evaluate DSM by assessing the ordering of pairwise relative risks using Concordance-Index (C-Index) (Harrell, 1982). To demonstrate performance of our approach vs. the methods subject to Coxian assumption, we show the comparison of performances using the time-dependent Concordance-Index C^{td} (Antolini et al., 2005).

$$C^{td}(t) = \mathbb{P}\big(\hat{F}(t|\mathbf{x}_i) > \hat{F}(t|\mathbf{x}_j)|\delta_i = 1, T_i < T_j, T_i \leq t\big)$$

Here, $\hat{F}(t|X)$ is the estimated CDF by the model at the truncation time t, given features X. The probability is estimated by comparing relative risks pair-wise. In order to obtain an unbiased estimate for the quantity, we adjust the estimate with an inverse propensity of censoring estimate, as is common practice in survival analysis literature (Gerds et al., 2013).

[3] https://seer.cancer.gov/

Observing C^{td} by different evaluation time horizons enable us to measure how good the models are at capturing the possible changes in risk over time, thus alleviating the restrictive assumption C-Index makes of constant proportional hazards. For completeness, we report the C^{td} at different truncation event horizon quantiles of 25%, 50%, 75%.

Practical utility of deployed applications and their interpretability requires survival analysis models to be well calibrated. We thus further assess calibration of DSM in comparison to the baselines by computing the Censoring Weighted Brier Score (Gerds & Schumacher, 2006; Graf et al., 1999) at each event's quantile.

Experimental Setup

Hyperparameters: For all the experiments described subsequently we train DSM with the Adam optimizer (Kingma & Ba, 2014) using learning rates of $\{1 \times 10^{-3}, 1 \times 10^{-4}\}$. The number of experts, K for each event is chosen from $\{4, 6, 8\}$ and the discounting factor α is chosen from $\{1/2, 3/4, 1\}$. The prior strength λ is set as 1×10^{-8} for all the experiments and not tuned. We report the C^{td} for the best performing set of parameters over the grid in cross validation for both DSM and the baselines. The representation learning function $\mathbf{\Phi}(.)$ is a fully connected Multi-Layer Perceptron with 1 or 2 hidden layers with the number of nodes in $\{50, 100\}$ and ReLU6 activations. The choice of Log-Normal or Weibull outcome is further tuned as a hyper parameter. All experiments were conducted in `PyTorch` (Paszke et al., 2019).

Evaluation Protocol: For each experiment we report the standard error around the mean of the C^{td} in 5-fold cross validation.[4]

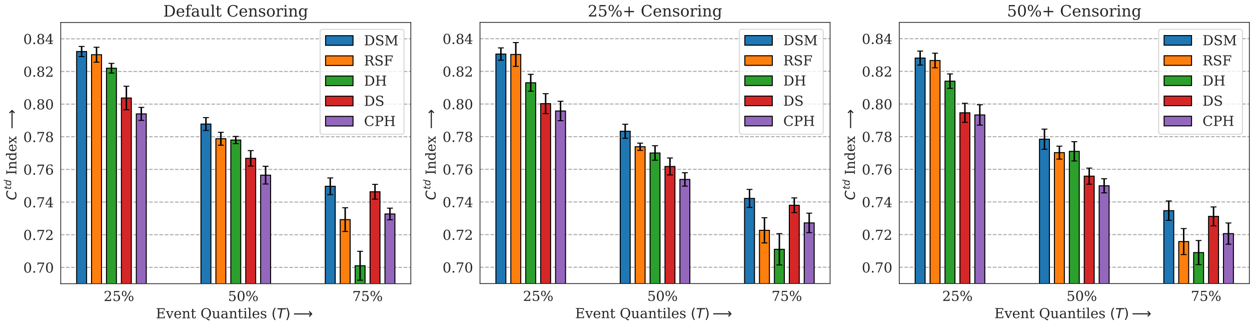


Figure 1.3: Time-Dependent Concordance Index C^{td} for SUPPORT dataset at different quantiles of event times for different levels of censoring.

1.4 Results

1.4.1 Single Event Survival Regression

Parameter inference for DSM involves the exploitation of a closed form of the CDF, which makes DSM amenable to gradient based optimization. Naturally one would expect that a greater amount of censoring will reduce the available information to be modeled, thus adding bias and leading to poorer estimates of the survival function.

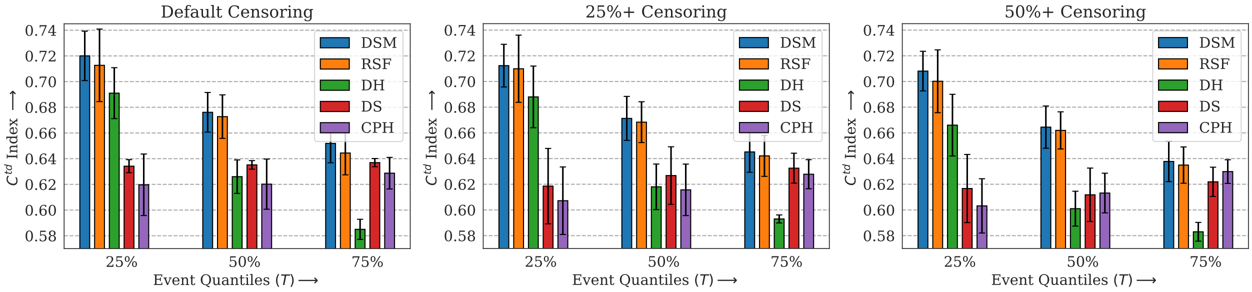


Figure 1.4: Time-Dependent Concordance Index (C^{td}) for METABRIC dataset at different quantiles of event times for different levels of censoring.

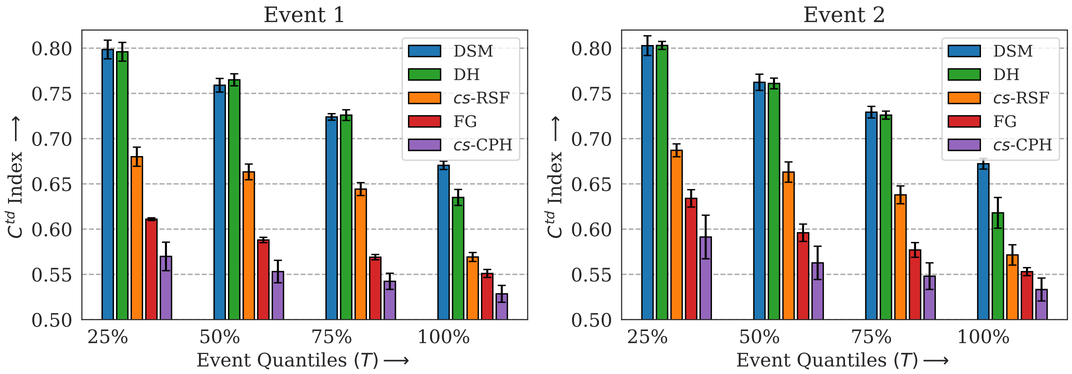


Figure 1.5: C^{td} for competing risks on SYNTHETIC data.

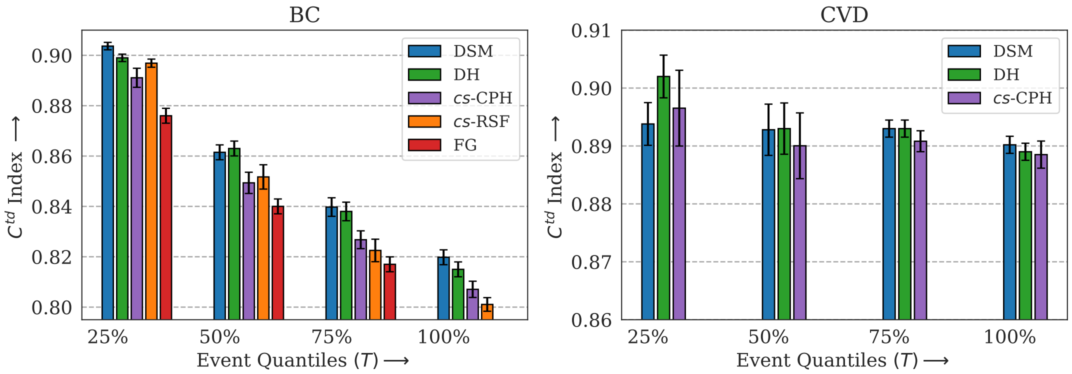


Figure 1.6: C^{td} for competing risks on SEER data.

In this section, we will empirically investigate DSM's robustness to censoring and compare it to the relevant baselines by artificially censoring the event times. We uniformly sample a censoring time between $[0, T)$ for a randomly chosen subset of the uncensored training data. This is only applied to the uncensored instances of the training splits with the same experimental protocol as used in the previous Section 1.3.2. (By not censoring the test splits we are able to better estimate the C^{td}). We perform this artificial censoring on the single event METABRIC and SUPPORT datasets and reduce the uncensored training data to 50% and 25% of its original amount.

Figure 1.3 summarizes the performance of DSM on the SUPPORT dataset in 5-fold cross validation. Notice that RSF is comparable to DSM in the 25% quantile of event time horizons across all levels of censoring, however, DSM significantly outperforms RSF on the longer event quantiles. Similarly, we observed that although DeepSurv was competitive in longer event horizons, DSM significantly outperformed DeepSurv in the shorter horizons, demonstrating superiority.

For METABRIC, we observed that DSM outperformed the Deep Learning baselines significantly. Although RSF was competitive, DSM outperformed RSF on average in 10-fold cross-validation.

Brier Scores and C^{td} obtained for METABRIC and SUPPORT data, can be found in Appendix A.1.1, and they present DSM's calibration ability to be at least on-par and often better when compared to the alternatives.

We close this section with a brief discussion on the scenarios when DSM can possibly achieve performance gain over alternative methods. When the data is sparse during certain time spans, e.g. the longer event horizons, the parametric nature of DSM makes it more robust than the non-parametric or deep learning based methods. On the contrary, the fitting of other competing non-parametric models and deep learning based approaches suffer from this lack of data. Furthermore, since DSM does not make the constant proportional hazard assumption like the CPH model or its variants, DSM is able to capture the flexible patterns of the survival functions when such assumptions do not hold.

1.4.2 Competing Risks Scenario

For the SYNTHETIC dataset, we observe in Fig. 1.5 that DSM is competitive with DeepHit and outperforms all the other baselines in the 25%, 50%, 75% quantiles of event horizons. For comparison, we also report the performance at 100% quantile and observe that DSM is significantly superior to DeepHit for both events, thus confirming its robustness to events at longer horizons.

From Fig. 1.6, on the SEER dataset we observe that for the majority risk, Breast Cancer, DSM significantly outperformed all the other baselines. The results for CVD were less conclusive with DeepHit being competitive at the 25% quantile. We owe this to the class imbalance between the two types of risks. Note that for visual clarity we do not report Fine-Gray and cs-RSF since their performance was poor. We defer the actual numbers and confidence intervals to Appendix A.1.1.

1.4.3 Representation Learning and Knowledge Transfer

We also conducted a set of experiments to evaluate the performance of DSM as a representation learning framework in the competing risks scenario. We compare its ability to transfer knowledge across multiple competing risks to relevant deep learning alternatives.

We divide the SYNTHETIC data into two equal subsets of 15,000 samples each. For the first set, we discard all records that had Event 2 before Event 1. For the second set, we perform similar preprocessing and discard all rows where Event 1 occurred before Event 2. This effectively converts the two subsets into single event censored datasets for Event 1 and Event 2 respectively. We train DSM, *DeepSurv* and *DeepHit* on the first half of the dataset for the prediction of Event 1. The learned model is then used to extract representations for the second subset. The output of the final layer is exploited as an overcomplete representation of the original set of co-variates of the individual observation. In both cases, we tune the models by brute-force over one and two hidden layers and the dimensionality of the hidden layers in $\{25, 50, 100\}$.

Table 1.3: Knowledge transfer across tasks and representation learning capability on the SYNTHETIC dataset. Representations were trained on Event 1 and used to predict relative risks for a held-out set on Event 2 using the CPH model.

Model	C-Index (90%-CI)
NNMF	0.5940 ± 0.0044
VAE	0.6494 ± 0.0044
K-PCA	0.7422 ± 0.0055
DeepSurv	0.6988 ± 0.0038
DeepHit	0.7688 ± 0.0040
DSM	$\mathbf{0.7724 \pm 0.0025}$

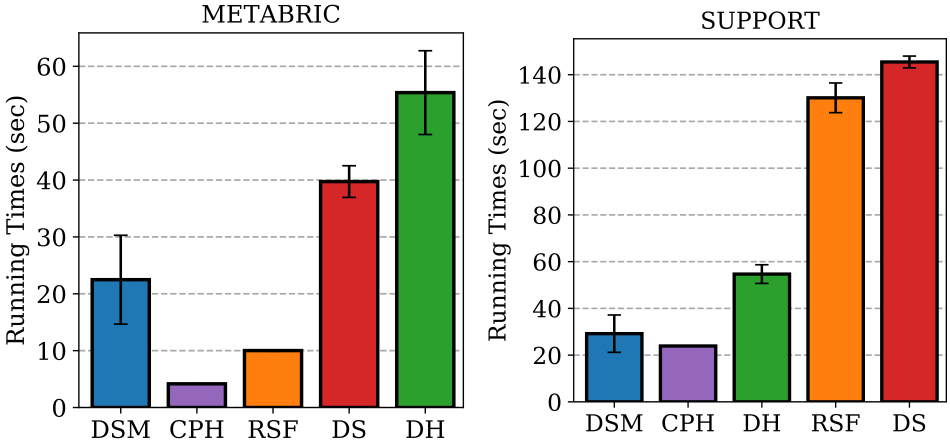


Figure 1.7: Comparison of training times required. Parameter inference with DSM is faster than other deep learning approaches, and it scales better with data size.

For completeness, we also experiment with Kernel-PCA (K-PCA) (Schölkopf et al., 1997), Non-Negative Matrix factorization (NNMF) (Lee & Seung, 2001) and modern Variational Auto Encoders (VAE) to learn latent representations. Note that as compared to *DeepSurv* and DSM, K-PCA, NNMF and VAE are intrinsic methods that do not have access to the label of the original risk (Event 1) at training time and hence are limited in their expressive capability.

Once the representations are extracted for the second subset of the data, a linear Cox Proportional Hazards (CPH) Model is trained on them for the competing risk (Event 2). Table 1.3 presents the result of concordance of the learned CPH model on the extracted embeddings. DSM outperforms the competing baselines.

1.4.4 Model Complexity and Scalability

We would like to stress that the advantage of *Deep Survival Machines* is not only in terms of competitive predictive performance but also in the ability to manage computational and inference complexity. Since DSM involves making reasonable parametric assumptions, inference requires fewer parameters to learn as compared to the considered alternative approaches. In this section, we compare the training time and

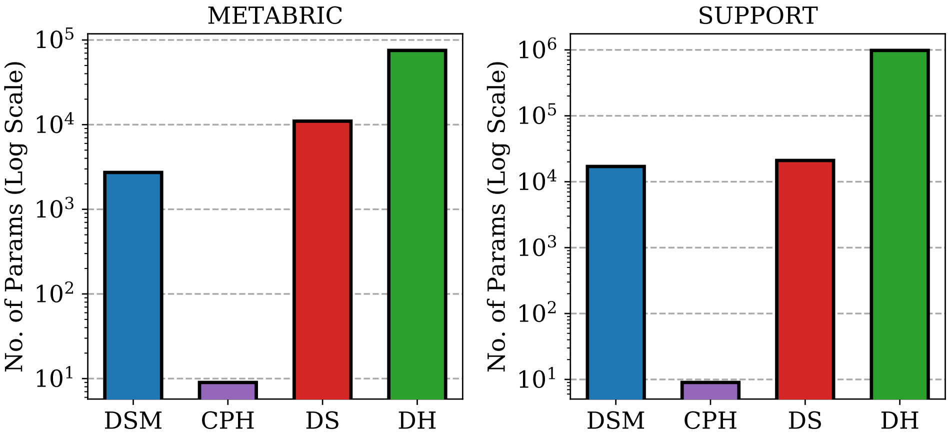


Figure 1.8: Number of learnable parameters in best architectures. DSM requires fewer parameters than deep learning alternatives.

the model complexity in terms of the number of parameters of DSM vis-à-vis the established deep learning baselines, DeepHit and DeepSurv, as well as the linear Cox Proportional Hazards regression CPH.

From Figures 1.7 and 1.8, the advantage of DSM in runtime and space complexity vs. considered deep learning alternative models is clearly visible. Note that while RSF is faster in training on METABRIC, it scales poorly with increasing amounts of data as evidenced by slower runtime on the larger SUPPORT dataset.

Chapter 2

Deep Parametric Time-to-Event Regression with Time-Varying Covariates

2.1 Introduction

Time-to-event regression in healthcare and other domains, such as predictive maintenance, require working with time-series (or time-varying) data such as continuously monitored vital signs, electronic health records, or sensor readings. In such scenarios, the event-time distribution may have temporal dependencies at various time scales that are not easily captured by classical survival models that assume training data points to be independent. In this chapter, we describe a fully parametric approach to model censored time-to-event outcomes with time varying covariates. It involves learning representations of the input temporal data using Recurrent Neural Networks such as LSTMs and GRUs, followed by describing the conditional event distribution as a fixed mixture of parametric distributions. The use of the recurrent neural networks allows the learned representations to model long-term dependencies in the input data while jointly estimating the Time-to-Event. We benchmark our approach on MIMIC III: a large, publicly available dataset collected from Intensive Care Unit (ICU) patients, focusing on predicting duration of their ICU stays and their short-term life expectancy, and we demonstrate competitive performance of the proposed approach compared to established time-to-event regression models.

Several important applications of survival analysis require working with interdependent temporal data such as multiple hemodynamic vital signs. Standard extensions to survival models for longitudinal data involve representing input covariates with aggregate statistics accrued over time in order to make them compatible with standard survival regression approaches. However, some data modalities, such as time series, cannot always be sufficiently captured using statistical featurization of static snapshots of the input features.

Furthermore, in the case of discrete temporal data, such as electronic health records, certain historical events may be more informative and more consequential, with long term effects that require modeling at multiple scales with factors more capacious than simple aggregate statistics such as moving averages and variances over time.

In time-to-event regression literature, such input data are known as time-varying covariates. While there is a large amount of existing work on extensions of classical statistical methods involving such data, modern machine learning approaches to model time-varying covariates are relatively understudied.

In this chapter, we propose Recurrent Deep Survival Machines (RDSM), a fully parametric survival analysis method for modeling time-to-event data in the presence of time-varying covariates. RDSM builds on the original DSM model (Nagpal et al., 2021c) by replacing the learned representation with a Recurrent Neural Network (RNN) architecture, such as a standard RNN or its variants, e.g., GRU (Cho et al., 2014a) or LSTM (Gers et al., 1999). As in the case of the original DSM model, we assume that once the representations are obtained, the event arrival times are distributed as a mixture of underlying parametric distributions. The parameters of these underlying distributions are also assumed to be functions of the obtained representations, and are learned jointly with the recurrent neural architectures. Our key contributions in this chapter are:

- We propose a novel censored Time-to-Event regression model, Recurrent Deep Survival Machines, that allows modeling of time-varying coefficients by using RNN layers with flexible parametric choices on the event-time distribution.
- We demonstrate the utility of RDSM on Mortality and Length-of-Stay prediction in the MIMIC-III dataset of ICU patients, and compare performance of the proposed approach against established censored survival regression baselines.
- Finally, we release the RDSM model as part of the open-source `auton-survival` (Nagpal et al., 2022b) python package for wider dissemination with the survival analysis research community.

2.2 Related Work

The surge of deep learning methods for machine learning have prompted related research in the use of deep learning for augmenting classic survival models. Katzman et al. (2018) propose *DeepSurv*, a proportional hazard model where relative risks are estimated using a neural network. DeepSurv allows to model non-linear proportional hazards however, is still restricted to the strong cox assumptions of proportional hazards.

Lee et al. (2019a) propose *DeepHit* that involves discretizing the event space with fixed intervals and treating the survival analysis problem as binary classification over these horizons. *DeepHit* has competitive performance by alleviating the proportional hazards assumption but scales poorly to large datasets especially at longer horizons. Ren et al. (2019) propose an RNN based model but only in the context of static features and do not discuss time-varying covariates.

Apart from Deep Learning approaches popular approaches for survival analysis also include non-parametric techniques including Random Survival Forests (Ishwaran et al., 2008) and Gaussian Processes (Fernández et al., 2016; Alaa & van der Schaar, 2017b).

Statistical methods for longitudinal data have proposed to model censored survival data by analysing different follow up times, commonly known as landmark times. These approaches involve building separate models for each time (Van Houwelingen, 2007). Two stage landmarking have been explored to model the disease evolution within those intervals and extract meaningful representations which are then used to build standard survival regression models (van Houwelingen & Putter, 2011).

Another approach aims to model longitudinal and survival outcomes simultaneously by sharing a latent representation (Henderson et al., 2000). This joint modeling benefits from shared knowledge between models but suffers from intractability and computational complexity.

Modern ML methods for modeling time-to-event outcomes in the presence of time-varying covariates are relative understudied. Lee et al. (2019a) propose *Dynamic-DeepHit* involving recurrent neural networks but

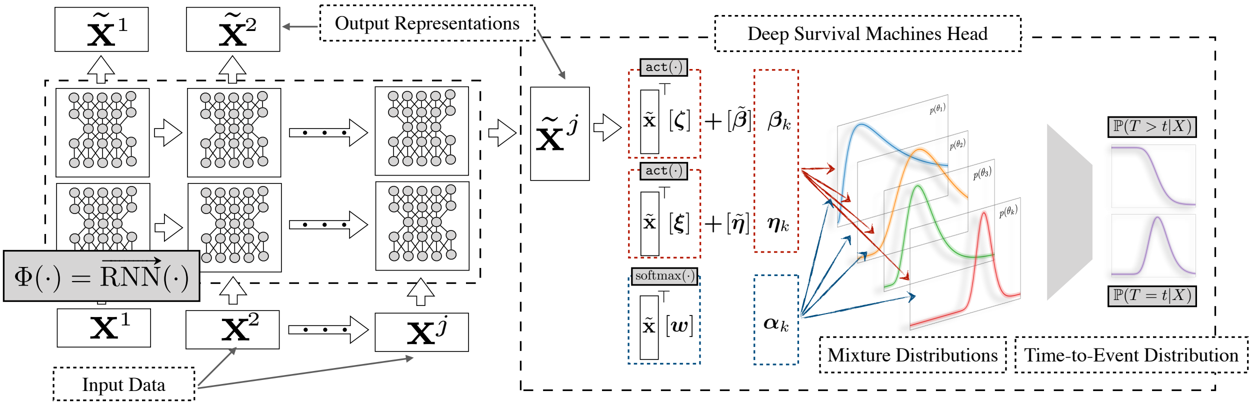


Figure 2.1: **Recurrent Deep Survival Machines:** The model involves first passing the set of input covariates sampled over time $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^j\}$ through the RNN, $\Phi(\cdot)$ to obtain the corresponding set of representations $\{\tilde{\mathbf{x}}^1, \tilde{\mathbf{x}}_i^2, ..., \tilde{\mathbf{x}}^j\}$ at each time step. The remaining Time-to-Event distribution is then modeled as a mixture of parametric distributions where the parameters of the distributions are functions of the input representations. The use of RNNs allows us to learn representations that retain knowledge from previous times steps.

suffers from similar limitations as the original *DeepHit* model. In a similar vein, Jarrett et al. (2019) further propose *MatchNet* involving temporal convolution networks.

2.3 Proposed Model: *Recurrent* Deep Survival Machines

Notation and Setting

We assume that we want to model a *right-censored* dataset, implying that our data, $\mathcal{D}$ is a set of tuples $\{(\mathcal{X}_i, t_i, \delta_i)\}_{i=1}^N$, where t_i is the time at which an event of interest took place, or the censoring time, and δ_i is an indicator that signifies whether t_i is the event time or censoring time. $\mathcal{X}_i$ is the set of features sampled over time along with the corresponding timestamp at which the data was sampled $\mathcal{X}_i := \{(\mathbf{x}_i^1, \tau_i^1), (\mathbf{x}_i^2, \tau_i^2), ...(\mathbf{x}_i^j, \tau_i^j)\}$. We notate $\mathcal{X}_i^j$ as the set of all covariates observed before time-step j and introduce remaining time-to-event at a time-step j as $r^j = (t - \tau^j)$. Thus the learning problem reduces to estimating the distributions of the conditional remaining time-to-event at each time step j, $\widehat{\mathbb{P}}(T(j) \mid \mathcal{X}_i^j)$.

We assume that we either only observe the actual failure or censoring time, but not both, for each individual. Furthermore, for the purposes of identification, it is assumed that the true data generating process is such that the censoring process is independent of the actual time to failure.

Deep Survival Machines

The key idea behind the original Deep Survival Machines model is to assume that the conditional survival distribution of an individual with covariates $\mathbf{x}$ is a mixture of fixed-size parametric distributions like the Weibull or Log-Normal. The shape and scale parameters of the underlying distributions, as well as the mixing weights, are implemented as a function of the input covariates using neural networks. Thus:

$$\mathbb{P}(T = t|X = \mathbf{x}) = \sum_k \mathbb{P}(Z = k|X = \mathbf{x})\mathbb{P}(T = t|X = \mathbf{x}, Z = k). \tag{2.1}$$

Here, $\mathbb{P}(Z = k|X = \mathbf{x}) = \text{softmax}\big(f(\Phi(\mathbf{x}))\big)$ where f is a linear function and $\mathbb{P}(T = t|X = \mathbf{x}, Z = k)$ is Weibull or Log-Normal with shape and scale parameterized as functions of the representation $\Phi(\mathbf{x})$.

$$\begin{aligned} \ln \mathbb{P}(T = t|X = \mathbf{x}) &= \ln \sum_k \mathbb{P}(Z = k|X = \mathbf{x})\mathbb{P}(T = t|X = \mathbf{x}, Z = k) \\ &\geq \sum_k \mathbb{P}(Z = k|X = \mathbf{x}) \ln \mathbb{P}(T = t|X = \mathbf{x}, Z = k). \end{aligned} \tag{2.2}$$

In the original DSM model, the authors proposed to perform inference using a lower bound to the maximum likelihood estimate under the above model as in Equation 2.2.

Structure of Recurrent DSM

Figure 2.1 gives a schematic description of our proposed approach. In order to allow RDSM to incorporate streaming data inputs, we will extend our discussion and consider the remaining time-to-event distribution $T(j)$ at a timestep, j. Instead of treating this distribution as static as in standard survival settings, modeling it as a function of time allows capturing the time varying effects of the input covariates.

Assumption 1 (Independent Censoring) *For an individual with remaining time-to-event distribution, $T(j)$, censoring time C and set of observed covariates, $\mathcal{X}^j$ till time j;*

$$T(j) \perp C \mid \mathcal{X}^j$$

Assumption 1 is analogous to the standard assumptions of random (non-informative) censoring in static survival analysis. This assumption is required for the purpose of identifiability. Assuming independent censoring, we can factorize the log-likelihood of the data from an individual at a time step j with remaining time to event $r^j = t - \tau^j$ over the censored and uncensored likelihood. We thus arrive at the following:

$$\mathbb{P}(\{\mathcal{X}^j, r^j, \delta\}) = \mathbb{P}(T = r^j|\mathcal{X}^j)^{\delta}\mathbb{P}(T > r^j|\mathcal{X}^j)^{(1-\delta)}. \tag{2.3}$$

Assumption 2 (Statefulness) *For an individual with event-time distribution at time j, $T(j)$ and the set of covariates observed till time-step j, $\mathcal{X}^j$;*

$$T(j) \perp (\mathcal{X} \setminus \mathcal{X}^j) \mid \mathcal{X}^j$$

Assumption 2 essentially states that the distribution of the remaining time-to-event at time j, $T(j)$ is completely characterized by the data at time steps preceding j. Although, in practice, later events in the data stream might effect the final event-time t, notice here that we consider the instantaneous event time distribution at time j, $T(j)$ instead, which we allow to evolve as more data is accrued. We require to make this assumption in order to enable inferring event risks dynamically, in a streaming fashion.

Learning

In the case of time-varying covariates, we have access to the full data stream $\mathcal{X}$ and not just a single feature vector $\mathbf{x}$. We thus replace the learning objective with the following objective modified to allow streaming (time-varying) data. For Maximum Likelihood Estimation we can represent the log-likelihood of a single time-step of the data stream as follows:

$$\mathcal{L}(\{\mathcal{X}^j, r^j, \delta\}; \theta) = (1-\delta) \ln \mathbb{P}(T > r^j | \theta, \mathcal{X}^j) + \delta \ln \mathbb{P}(T = r^j | \theta, \mathcal{X}^j). \tag{2.4}$$

Where δ is an indicator of whether the individual experienced the event or was lost to follow-up (censored) and θ is the set of parameters to be inferred. This is a direct consequence of Equation 2.3. Now, leveraging the assumed statefulness, we can rewrite the loss factorized over each time-step as:

$$\mathbb{P}(\{r^1, r^2, ... r^j\} | \theta, \mathcal{X}^j) = \prod_j \mathbb{P}(T = r^j | \theta, \mathcal{X}^j), \text{or} \prod_j \mathbb{P}(T > r^j | \theta, \mathcal{X}^j). \tag{2.5}$$

From Equation 2.5 we can rewrite the loss in Equation 2.4 as a sum over each time-step as

$$\mathcal{L}(\{\mathcal{X}, t, \delta\}; \theta) = \sum_j (1-\delta) \ln \mathbb{P}(T > r^j | \theta, \mathcal{X}^j) + \delta \ln \mathbb{P}(T = r^j | \theta, \mathcal{X}^j).$$

Here, $\mathbb{P}(T > r^j | \theta, \mathcal{X}^j)$ and $\mathbb{P}(T = r^j | \theta, \mathcal{X}^j)$ are functions of the input representation:

$$\mathbb{P}(T = r^j | \theta, \mathcal{X}^j) = \sum_k \text{softmax}\big(f\,(\overrightarrow{\text{RNN}}(\mathcal{X}^j)\big)\mathbb{P}\big(T = r^j | Z = k, \overrightarrow{\text{RNN}}(\mathcal{X}^j)\big).$$

Note that here the term $\overrightarrow{\text{RNN}}(\cdot)$ refers to the output of the Recurrent Neural Network (LSTM or GRU). $\mathbb{P}(T = r^j | Z = k, \overrightarrow{\text{RNN}}(\mathcal{X}^j))$ is obtained by making a parametric choice (like the 'Weibull' distribution) and the shape and scale parameters for each Z modelled as functions of $\overrightarrow{\text{RNN}}(\mathcal{X}^j)$. $\mathbb{P}(T = r^j | \theta, \mathcal{X}^j)$ is defined similarly for the uncensored data.

And finally, the full log-likelihood over the entire dataset, $\mathcal{D} := \{(\mathcal{X}_i, t_i, \delta_i)\}_{i=1}^N$ is:

$$\mathcal{L}(\mathcal{D}; \theta) = \sum_i^{|\mathcal{D}|} \sum_j (1-\delta_i) \ln \mathbb{P}(T > r_i^j | \theta, \mathcal{X}_i^j) + \delta_i \ln \mathbb{P}(T = r_i^j | \theta, \mathcal{X}_i^j).$$

Here we introduce subscript i to refer to a single individual in the dataset. This likelihood can be optimized using a gradient based, first order optimizer. In practice, we optimize a lower bound of the following likelihood similar to Equation 2.2 as was proposed for DSM.

2.4 Experiments

Task and Dataset

We work with the large publicly available dataset of over 50,000 ICU admissions from the Beth Israel Deaconess Medical Center, MIMIC III (Johnson et al., 2016). In critical care scenarios, it is of interest to be able to accurately quantify the Length-of-Stay (LOS) of an admitted patient, and their in-hospital mortality. Such estimation allows decision makers to help allocate adequate healthcare resources including staff and equipment, as well as get a sense of the seriousness of the patient's condition and guide their triage upon admission, and treatment thereafter. For our experiments, we use MIMIC-extract (Wang et al., 2020) to obtain a subset of all patients who were in the ICU for at least 30 hours. Our subset includes all measurements including vital signs and medications administered, sampled every hour. A stay at ICU can be terminated by either of the two competing events: Discharge, or Death. For mortality prediction, we define the target variable to be time in the ICU from admission till death, with discharge being a censoring event. For Length-of-Stay prediction, the target variable is time from admission till discharge, with death being the censoring event.

2.4.1 Evaluation

We use the first 24 hours of data from admission to estimate if the remaining ICU Length of Stay would exceed 1, 3 or 7 days. Simultaneously, we aim to estimate if a patient would experience death within the next 1, 3 or 7 days. Among 24,430 patients spending at least 30 hours in the ICU, 4,886 were used as a test set to evaluate the performance of models, while the rest were left for training and hyper-parameter tuning. For both RDSM and the baselines, the best set of hyperparameters were chosen to maximize the log-likelihood on a development set consisting of 2,443 patients.

We evaluate performance in terms of Area under the Receiver Operating Characteristic curve (AuROC) and Brier score on a left-out test set for the two tasks: 'Length of Stay' and 'Death'. Metrics were measured on the agglomerated risks computed every hour during the first 24 hours after admission.

During evaluation, the AuROC and Brier Scores are computed by considering censoring events as positive (negative) for Length-of-Stay (Mortality) prediction.[1] The metrics could also be adjusted to account for the censoring using Inverse Propensity of Censoring Weighting (IPCW) by employing a Kaplan-Meier estimator of the censoring distribution (Uno et al., 2007; Hung & Chiang, 2010a) as is popular in survival analysis. We however avoid this approach as in ICU or critical-care settings time-to-event (death) and censoring time (discharge) are not independent, leading to biased estimates when adjusted using the Kaplan-Meier estimator.

2.4.2 Baselines

We compare the performance of RDSM to the standard Deep Survival Machines (DSM) model which assumes data at each time-step to be independent. We also benchmark performance against the *DeepSurv* model that assumes proportional hazards and *DeepHit* model that discretizes event times.

2.4.3 Hyperparameter Choices

We performed a simple grid search to coarsely optimize performance of the RDSM model. The choices of hyperparameters include the type of the recurrent neural network cell (selected from vanilla RNNs, LSTMs or GRUs), the batch size (selected from $\{125, 250\}$), the learning rate ($\{1 \times 10^{-3}, 1 \times 10^{-4}\}$), the number of hidden layers ($\{1, 2, 3\}$), the dimensionality of the hidden layer ($\{50, 100, 200\}$), and the number of underlying parametric distributions to use for the mixture ($k \in \{3, 4, 6\}$). For fair comparison the baseline

[1] This is an intuitive choice: a discharged patient is not likely to experience mortality immediately post ICU admission. Similarly, a dead patient most likely had poor physiology and would end up with a longer ICU length of stay. **Note**: This adjustment is only made during evaluation. At training time censored observations are treated as censored.

Model	Death<1	Death<3	Death<7	LOS>1	LOS>3	LOS>7
DeepSurv	0.897 (0.004)	0.859 (0.003)	0.841 (0.002)	0.740 (0.002)	0.715 (0.002)	0.756 (0.003)
DeepHit	0.897 (0.004)	0.859 (0.003)	0.798 (0.003)	0.851 (0.001)	0.724 (0.001)	0.786 (0.002)
DSM	0.898 (0.004)	0.863 (0.002)	0.846 (0.002)	0.819 (0.002)	0.737 (0.001)	**0.801 (0.001)**
RDSM	**0.923 (0.003)**	**0.890 (0.002)**	**0.872 (0.002)**	**0.864 (0.002)**	**0.740 (0.002)**	0.796 (0.003)

Table 2.1: Area under ROC Curves on the MIMIC-III dataset for Length of Stay (LOS) and Mortality Prediction (Death). (95% CIs were generated by bootstrapping the test set, DSM: Deep Survival Machines, RDSM: Recurrent Deep Survival Machines)

Model	Death<1	Death<3	Death<7	LOS>1	LOS>3	LOS>7
DeepSurv	0.005 (<0.001)	0.027 (<0.001)	0.058 (<0.001)	0.115 (<0.001)	0.215 (0.001)	0.096 (0.001)
DeepHit	0.005 (<0.001)	0.028 (<0.001)	0.168 (<0.001)	0.084 (0.001)	0.224 (<0.001)	0.100 (0.001)
DSM	0.005 (<0.001)	0.027 (<0.001)	0.059 (<0.001)	0.080 (0.001)	0.198 (<0.001)	**0.088 (0.001)**
RDSM	0.005 (<0.001)	**0.026 (0.001)**	0.059 (0.002)	**0.073 (<0.001)**	**0.193 (0.001)**	0.091 (0.001)

Table 2.2: Brier score on the MIMIC-III dataset for Length of Stay (LOS) and Mortality Prediction (Death). (95% CIs were generated by bootstrapping the test set, DSM: Deep Survival Machines, RDSM: Recurrent Deep Survival Machines)

models were also optimized over the same grid design as for RDSM. Additionally, all models were trained using the Adam optimizer (Kingma & Ba, 2014). For fair comparison, we employed early stopping by evaluating the likelihood on a 10% subset of the training set.

2.5 Results

As summarized in Tables 2.1 and 2.2, performance of the DSM model is improved by the use of Recurrent Neural Networks across all the tasks, a clear indicator that considering time-varying covariates allows to better model the multivariate longitudinal time series data such as MIMIC-III. Furthermore RDSM demonstrates competitive performance when compared to other published deep learning based survival approaches. As compared to approaches such as DeepHit, RDSM does not involve discretization of the event times leading to risk estimates that are better calibrated as well as making inference scalable.

2.5.1 Discussion

We observe that the proposed approach shows leading performance when making predictions at shorter horizons of time, while still being comparable at longer horizons. We believe that these results can be further improved by making more appropriate parametric choices to model the event time distributions. Further research will involve more flexible parametric assumptions to encourage robust performance across a wide range of time horizons.

We have demonstrated that the use of recurrent neural architectures helps improve representation learning capability for data with time-varying covariates such as time series vital signs. Currently, the MIMIC data we work with were preprocessed to obtain hourly resolution data by aggregating observations at regular intervals and imputing missing entries. There are significant challenges with such real-world healthcare data including missing values and irregular sampling frequencies. Extensions to RDSM could involve integrating handling of missing data into the overall process. The use of RNNs for data imputation as has been demonstrated to have utility in clinical contexts like for example, the GRU-D model (Che et al., 2018). In addition, irregularly or asynchronously sampled time series could be directly handled by time aware RNNs without imputation as shown in Baytas et al. (2017) and Rubanova et al. (2019).

Extensions could also involve the use of attention in order to capture and characterize specific events in the patients history relevant to the outcome of interest enabling studies towards establishing causal relationships between observable factors and outcomes in such data. Future work could also involve the use of generative models or adversarial training to better learn robust representations of the temporal data.

2.6 Conclusion

We proposed an extension of the Deep Survival Machines model to handle temporal data with time-varying coefficients. Our approach uses recurrent neural networks to handle long term temporal dependencies in the input data stream. Our approach obtained favorable results from empirical evaluation of our model on predicting in-hosptial ICU mortality and length of stay. We also made the software implementation of our tool available to the survival analysis research community online as an open source package.

Chapter 3

Deep Cox Mixtures

3.1 Preliminaries

In this section, we first introduce the notation and background for the standard Cox model.

3.1.1 Notation

We consider a dataset of right censored observations $\mathcal{D} = \{(\boldsymbol{x}_i, \delta_i, u_i)\}_{i=1}^{N}$ of three tuples, where $\boldsymbol{x}_i$ are the covariates of an individual i, δ_i is an indicator of whether an event occurred or not and u_i is either the time of event or censoring as indicated by δ_i.

We consider a maximum likelihood (MLE) based approach to learning $S(t|x) = \mathbb{P}(T > t|X = x)$ from the data. Recall that the survival distribution $S(t|x)$ is isomorphic to the cumulative hazard function $\boldsymbol{\Lambda}(t|x)$, and under continuity, this is equivalent to the hazard function $\boldsymbol{\lambda}(t|x)$. As a result, we will refer them in the parameters of the likelihood interchangeably. Lin (2007) shows that the likelihood of the observed data $\mathcal{D}$ is, up to constant factors,

$$\mathcal{L}(\boldsymbol{\Lambda}) = \prod_{i=1}^{|\mathcal{D}|} (\boldsymbol{\lambda}(u_i|\boldsymbol{x}_i))^{\delta_i}\, \boldsymbol{S}(u_i|\boldsymbol{x}_i). \tag{3.1}$$

In the following sections, we show how plugging in specific functional forms for $S(t|x)$ allows us to derive survival function estimators.

3.1.2 MLE for the standard Cox PH model

The key idea behind the Cox model is to assume that the conditional hazard of an individual, is $\boldsymbol{\lambda}(t|x) = \boldsymbol{\lambda}_0(t)\exp\big(f(\boldsymbol{\theta}, x)\big)$, where f is typically a linear function. Under the Cox model, the full likelihood as in equation 3.1 is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_0) = \prod_{i=1}^{|\mathcal{D}|} \Big(\boldsymbol{\lambda}_0(u_i)\exp\big(f(\theta, \boldsymbol{x}_i)\big)\Big)^{\delta_i} \boldsymbol{S}_0(u_i)^{\exp\big(f(\boldsymbol{\theta}; \boldsymbol{x}_i)\big)} \tag{3.2}$$

Cox (1972) and the discussion of his paper by Breslow (1972b), suggest deriving a maximum likelihood estimate of $\boldsymbol{\theta}$ by maximizing the partial likelihood, $\mathcal{PL}(\boldsymbol{\theta})$ defined below, and using the following estimator of the baseline survival function $\mathbf{\Lambda}_0(\cdot)$,

$$\mathcal{PL}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathcal{R}(t_i)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)}, \quad \widehat{\mathbf{\Lambda}}_0(t) = \sum_{i:t_i<t} \frac{1}{\sum\limits_{j\in\mathcal{R}(t_i)} \exp\big(f(\widehat{\boldsymbol{\theta}};\boldsymbol{x}_j)\big)}, \tag{3.3}$$

where $\mathcal{R}(t_i)$ is the 'risk set' – the set of individuals that survived beyond time t_i.

3.2 Proposed Approach

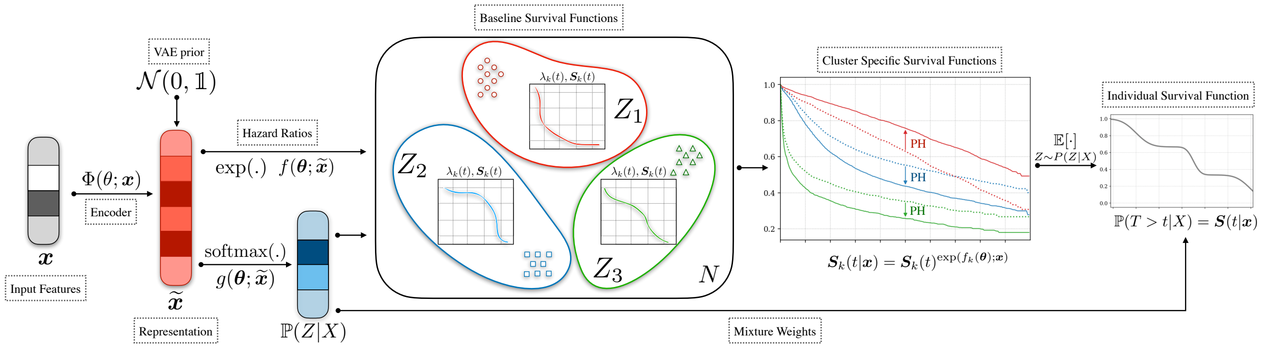

Baseline Survival Functions
VAE prior
$\mathcal{N}(0, \mathbb{1})$
Hazard Ratios
$\exp(.)$ $f(\boldsymbol{\theta};\widetilde{\boldsymbol{x}})$
$\Phi(\theta;\boldsymbol{x})$
Encoder
$\boldsymbol{x}$
Input Features
$\widetilde{\boldsymbol{x}}$
Representation
softmax(.)
$g(\boldsymbol{\theta};\widetilde{\boldsymbol{x}})$
$\mathbb{P}(Z|X)$
$\lambda_k(t), \boldsymbol{S}_k(t)$
Z_1
Z_2
Z_3
N
Cluster Specific Survival Functions
1.0
0.8
0.6
0.4
0.2
PH
PH
PH
$\boldsymbol{S}_k(t|\boldsymbol{x}) = \boldsymbol{S}_k(t)^{\exp(f_k(\boldsymbol{\theta});\boldsymbol{x})}$
$\mathbb{E}[\cdot]$
$Z\sim P(Z|X)$
Individual Survival Function
1.0
0.8
0.6
0.4
0.2
0.0
$\mathbb{P}(T > t|X) = \boldsymbol{S}(t|\boldsymbol{x})$
Mixture Weights

Figure 3.1: **Deep Cox Mixtures**: Representation of the individual covariates $\boldsymbol{x}$ are generated using an encoding neural network. The output representation $\widetilde{\boldsymbol{x}}$ then interacts with linear functions f and g that determine the proportional hazards within each cluster $Z \in \{1, 2, ...K\}$ and the mixing weights $\mathbb{P}(Z|X)$ respectively. For each cluster, baseline survival rates $\boldsymbol{S}_k(t)$ are estimated non-parametrically. The final individual survival curve $S(t|\boldsymbol{x})$ is an average over the cluster specific individual survival curves weighted by the mixing probabilities $\mathbb{P}(Z|X = \boldsymbol{x})$.

3.2.1 Proposed Model

In the case of DCM we propose an extension to the Cox model, modeling an individual's survival function using a finite mixture of K Cox models, with the assignment of an individual i to each latent group mediated by a gating function $g(.)$ The full likelihood for this model is

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{\Lambda}_k) = \prod_{i=1}^{|\mathcal{D}|} \int_Z \big(\boldsymbol{\lambda}(u_i|\boldsymbol{x}_i)\big)^{\delta_i}\, \boldsymbol{S}_k(u_i|\boldsymbol{x}_i)\mathbb{P}(Z = k|\boldsymbol{x}_i).$$

$$\text{where, } \boldsymbol{\lambda}(u_i|\boldsymbol{x}_i) = \boldsymbol{\lambda}_k(u_i)\exp\big(f_k(\boldsymbol{\theta},\boldsymbol{x}_i)\big), \quad \boldsymbol{S}_k(u_i|\boldsymbol{x}_i) = \boldsymbol{S}_k(u_i)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)}$$
$$\text{and, } \mathbb{P}(Z = k|X = \boldsymbol{x}_i) = \operatorname{softmax}\big(g(\boldsymbol{\theta};\boldsymbol{x}_i)\big) \tag{3.4}$$

Architecture: We allow the model to learn representations for the covariates $\boldsymbol{x}_i$ by passing them through a encoding neural network, $\Phi(.) : \mathbb{R}^d \rightarrow \mathbb{R}^h$. This representation then interacts with linear functions f and g defined on $\mathbb{R}^h \rightarrow \mathbb{R}^k$; that determine the log hazard ratios and the mixture weights respectively. The set of parameters for the encoder Φ and the linear functions f and g are jointly notated as $\boldsymbol{\theta}$. We experiment with a simple feed forward MLP and a variational auto-encoder for $\Phi(.)$ The parameters of the MLP and the VAE are learnt jointly during learning. For the VAE variant the encoder and the decoder architecture is kept the same.

We also experiment with a variant that does not involve representation learning and thus the functions f and g are linear and restricted to operate on the original features $\boldsymbol{x}$. Figure 3.1 provides a schematic description of our approach.

3.2.2 Learning

Notice that under the model in Eq. 3.4, the corresponding partial likelihood is not independent of $\boldsymbol{\lambda}(.)$, the hazard rate. We hence cannot directly optimize the partial likelihood to perform parameter learning. This inference complexity is outlined in Appendix B.1.1. Since our model requires inference over the latent assignments Z for learning the Expectation Maximization (Dempster et al., 1977) algorithm is a natural approach to perform inference. The major challenge to applying exact EM lies in the fact that under the our model requires a summation over all possible combinations of latent assignments which is intractable to compute. We propose an approximate, Monte Carlo EM algorithm (Wei & Tanner, 1990; Song et al., 2016) involving the drawing of posterior samples to learn the parameters, $\boldsymbol{\theta}$ and the baseline survival functions $\{\boldsymbol{S}_k(.)\}_{i=1}^K$.

Algorithm 1: Learning for DCM

```
Input   :Training set, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)_{i=1}^N\}$; batches, $B$;
while <not converged> do
    for $b \in \{1, 2, ..., B\}$ do
        $\mathcal{D}_b$ ←sampleMiniBatch($\mathcal{D}$)
        $\{\gamma_i\}_{i=1}^B \leftarrow$ E-Step( $\boldsymbol{\theta}, \{\widetilde{\boldsymbol{S}}_k\}_{i=1}^K$ ) ;
        $\{\zeta_i\}_{i=1}^B \sim$ Categorical($\gamma$) ;
        $\boldsymbol{\theta} \leftarrow$ M-Step($\boldsymbol{\theta}, \{\zeta_i, \gamma_i\}_{i=1}^B$);
        for $k \in \{1, 2, ..., K\}$ do
            $\widehat{\boldsymbol{S}}_k \leftarrow$ breslow($\boldsymbol{\theta}, \{(t_i, \delta_i)\}_{i=1;\zeta_i=k}^{|\mathcal{D}|}$) ;
            $\widetilde{\boldsymbol{S}}_k$ ←splineInterpolate($\widehat{\boldsymbol{S}}_k$) ;
        end
    end
end
Return :learnt parameters, $\boldsymbol{\theta}$;
        baseline survival splines $\{\widetilde{\boldsymbol{S}}_k\}_{i=1}^K$
```

E-Step: Involves estimating the posteriors of Z, $\gamma_i \propto \mathbb{P}(T = t|X, Z)^{\delta_i}\mathbb{P}(T > t|X, Z)^{1-\delta_i}$. The Breslow estimator only gives us the estimates of the survival rates, thus computing the posterior counts, $h_i \propto \mathbb{P}(T = t_i|Z, X)$ for the uncensored instances challenging. We mitigate this by interpolating the Baseline Survival Rate for each latent group, $\boldsymbol{S}_k(.)$ using a polynomial spline. Equation 3.5 provides the interpolated event probability estimates. (Appendix B.1.2 describes this in detail.)

$$\begin{aligned}
&\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}_i, Z = k) = \widetilde{\boldsymbol{S}}_k(t)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)} \text{ and,}\\
&\widehat{\mathbb{P}}(T = t|X = \boldsymbol{x}_i, Z = k) =\\
&\qquad\qquad - \exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)\frac{\widehat{\mathbb{P}}(T > t|\boldsymbol{x}_i, Z = k))}{\widetilde{\boldsymbol{S}}_k(t)}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)
\end{aligned} \tag{3.5}$$

Here, $\widetilde{\boldsymbol{S}}_k(t)$ is the baseline survival rate interpolated with a polynomial spline.

M-Step: Once the posterior counts γ_i are obtained, the M-Step involves learning maximizing the corresponding $Q(.)$ function given as

$$Q(\theta) = \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \gamma_i^k \ln \mathbb{P}(Z|X) + \gamma_i^k \ln \mathbb{P}(t|Z,X);$$
$$\text{where, } \gamma_i \propto \mathbb{P}(T|X,Z) \tag{3.6}$$

Notice that the γ_i^k are soft counts ($\gamma_i \in [0,1]$) making parameter inference for the term $\mathbb{P}(T|Z,X)$ intractable. Motivated from Monte-Carlo EM methods We instead sample hard posterior counts $\zeta_i \sim \text{Categorical}(\gamma_i)$.

We replace this with hard posterior counts for the second term, $\ln \mathbb{P}(t|Z,X)$

$$\overline{Q}(\theta) = \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \gamma_i^k \ln \mathbb{P}(Z|X) + \zeta_i^k \ln \mathbb{P}(t|Z,X);$$
$$\text{where, } \zeta_i \sim \text{Categorical}(\gamma_i) \tag{3.7}$$

Note that $\mathbb{E}[\overline{Q}(\cdot)] = Q(\cdot)$. Thus, $\overline{Q}(\cdot)$ is an unbiased estimate of the exact $Q(\cdot)$

The first term in $\overline{Q}(\cdot)$ can be optimized using gradient based approaches. The second term can be re-written as a sum over k latent groups variables.

$$\begin{aligned}
\overline{Q}(\theta) &= \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \gamma_i^k \ln \mathbb{P}(Z|X) + \mathbb{1}\{\zeta_i = k\} \ln \mathbb{P}(t|Z,X) \\
&= \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \gamma_i^k \ln \mathbb{P}(Z|X) + \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \mathbb{1}\{\zeta_i = k\} \ln \mathbb{P}(t|Z,X) \\
&= \sum_{i=1}^{|\mathcal{D}|} \sum_{k} \gamma_i^k \ln \mathbb{P}(Z|X) + \sum_{k} \sum_{i=1}^{|\mathcal{D}_k|} \ln \mathbb{P}(t|Z,X) \qquad (3.8)\\
&\text{(Here, } \mathcal{D}_k \text{ is the set of all } \mathcal{D} \text{ with } \zeta_i = k\text{)}
\end{aligned}$$

Now using the fact that the Proportional Hazards assumption holds within each group $\mathcal{D}_k$ we arrive at the form of the $Q(\cdot)$ that we optimize in each minibatch as

$$\widehat{Q}(\boldsymbol{\theta}) = \sum_{i}^{|\mathcal{D}_b|} \sum_{k} \gamma_i^k \ln \text{softmax}\big(g(\boldsymbol{\theta}; \boldsymbol{x}_i)\big) + \sum_{k} \ln \mathcal{PL}(\mathcal{D}_b^k; \boldsymbol{\theta})$$

Here, $\mathcal{D}_b^k$ is the subset of all individuals that have $\zeta_i = k$ within the minibatch b and $\mathcal{PL}(.)$ is the partial likelihood as defined in Equation 3.3. Thus, the use of hard counts ζ effectively reduces the problem to learning K separate Cox models allowing us to maximize the partial likelihood independently within each $k \in K$.

The parameters of the encoder are also updated during the **M-Step** by adding the loss corresponding to the VAE. Altogether the loss function for optimization is

$$\text{Loss}(\boldsymbol{\theta}; \mathcal{D}_b) = \widehat{Q}(\boldsymbol{\theta}; \mathcal{D}_b) + \alpha \cdot \text{VAE-Loss}(\boldsymbol{\theta}; \mathcal{D}_b) \tag{3.9}$$

Here, the VAE-Loss is the Evidence Lower Bound for the VAE with representations drawn from a zero mean and identity covariance gaussian prior as in Kingma & Welling (2013).

Algorithm 2 describes the learning procedure for DCM. We sample minibatches $\mathcal{D}_b$ from the data $\mathcal{D}$ and compute the soft and hard posterior counts, $\{\gamma_i, \zeta_i\}_{i\in\mathcal{D}_b}$ for each batch. This is followed by the **M-Step** involving a gradient update the parameter set $\boldsymbol{\theta}$. Finally, we update the Baseline Survival Splines, $\widetilde{\boldsymbol{S}}_k$ computed using the Breslow's estimator (Eq. 3.3) for each cluster. Note that the Breslow's estimator is computed over the full batch, $\mathcal{D}$. This is computed analytically, does not involve gradient computation and so is not expensive.

3.2.3 Inference

Following Equation 3.4, at test time the estimated risk of an individual at time t is given as

$$\begin{aligned}\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}_i) &= \mathbb{E}_{Z\sim\widehat{\mathbb{P}}(Z|X)}[\widehat{\mathbb{P}}(T|X = \boldsymbol{x}_i, Z)]\\ &= \sum_k \widetilde{\boldsymbol{S}}_k(t)^{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)} \times \mathrm{softmax}_k\big(g(\boldsymbol{\theta};\boldsymbol{x}_i)\big) \end{aligned} \tag{3.10}$$

3.3 Experiments

Table 3.1: Summary statistics for the datasets used in the experiments.

Dataset	N	d	Censoring (%)	Minority Class (%)	Event Quantiles		
					$t = 25$th	$t = 50$th	$t = 75$th
SUPPORT	9,105	44	31.89%	Non-White (21.02%)	14	58	252
FLCHAIN	6,524	8	69.93%	Female (44.94%)	903.25	2085	3246
SEER	55,993	168	72.82%	Non-White (23.77%)	25	55	108

In this section we describe the datasets, the survival analysis tasks and baselines we compare DCM against. We also describe the corresponding metrics we employ for evaluation.

3.3.1 Datasets

We experiment with the following real world, publicly available survival analysis datasets:

FLCHAIN (Assay of Serum Free Light Chain): This is a public dataset introduced by Dispenzieri et al. (2012) aiming to study the relationship between serum free light chain and mortality. It includes covariates like age, gender, serum creatinine and presence of monoclonal gammapothy. We removed all the individuals with missing covariates and experiment with the remaining subset of 6,524 individuals. Out of this subset 45% of the participants were coded as female and are considered as 'minority'.

SUPPORT (Study to understand prognoses and preferences for outcomes and risks of treatments (Connors et al., 1995): Dataset from study instituted to understand patient survival for 9,105 terminally ill patients on life support. The median survival time for the patients in the study was 58 days. Out of the $9,105$ patients

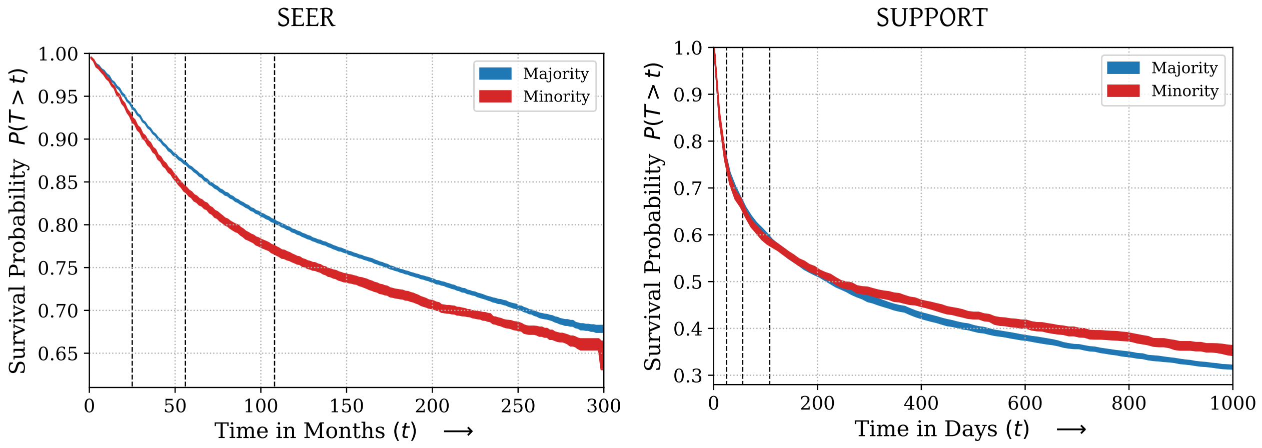


Figure 3.2: Base survival rates for the majority (White) vs. the other demographics in the SEER dataset estimated with a Kaplan-Meier estimator. Notice that the baseline survival rates differ across groups. Dashed lines respresent the 25^{th}, 50^{th} and 75^{th} event quantiles.

79% were coded as 'White', while the rest were coded as 'Black', 'Hispanic' and 'Asian'.

SEER (Surveillance, Epidemiology and End Results Study)[1] This dataset from National Cancer Institute (2019) consists of survival characteristics of oncology patients taken from cancer registries covering about one-third of the US Population. For our study we consider a cohort of patients over a 15 year period from 1992-2007 diagnosed with breast cancer with a median survival time of 55 months. A majority (76%) of the patients were coded as 'White' and the rest were other minorities consisting of 'Blacks', 'American Indians', 'Asians', etc.[2]

Our choice of datasets encompasses varying ranges of dimensionality of covariates, levels of censoring and size vis-a-vis the minority demographics. Table 3.1 describes some summary statistics of the considered datasets. Figure 3.2 compares the baseline survival rates for the majority and minorities in the SEER and SUPPORT dataset. Notice that base survival rates across demographics can vary over time.

3.3.2 Baselines

We compare the proposed DCM against the following baselines.

Accelerated Failure Time (AFT): This is an extension of generalized linear models to the survival setting with censored data. The target variable is assumed to follow a Weibull distribution and the shape and scale parameters are modelled as linear functions of the covariates. Parameter learning is performed using Maximum Likelihood Estimation.

Deep Survival Machines (DSM) (Nagpal et al., 2021c): This is another fully parametric approach and improves on the Accelerated Failure Time model by modeling the event time distribution as a fixed size mixture over Weibull or Log-Normal distributions. The individual mixture distributions are themselves parametrized with

[1] https://seer.cancer.gov/

[2] SEER has a very intricate coding pattern vis-a-vis race. Refer to https://seer.cancer.gov/tools/codingmanuals/race_code_pages.pdf for details.

neural networks allowing to learn complex non-linear representations of the data.

Deep Hit (DHT) (Lee et al., 2018): A discrete time model, DeepHit is a popular Neural Network approach that involves discretizing the event outcome space and treating the survival analysis problem as a multiclass classification problem over the discrete intervals.

Cox Proportional Hazards (CPH): CPH assumes that individuals across the population have constant proportional hazards overtime.

Faraggi-Simon Net (FSN)/DeepSurv (Faraggi & Simon, 1995; Katzman et al., 2018): An extension to the CPH model, FSN involves modeling the proportional hazard ratios over the individuals with Deep Neural Networks allowing the ability to learn non linear hazard ratios.

Random Survival Forest (RSF) (Ishwaran et al., 2008): RSF is an extension of Random Forests to the survival settings where risk scores are computed by creating Nelson-Aalen estimators in the splits induced by the Random Forest.[3] The full set of the model hyper parameters settings on which we perform grid search is deferred to Appendix B.2.

3.3.3 Evaluation Metrics

We compare the performance of DCM against baselines in terms of both discriminative performance and calibration using the following metrics:

Area under ROC Curve (AUC): Involves treating the survival analysis problem as binary classification at different quantiles of event times and computing the corresponding area under the ROC curve.

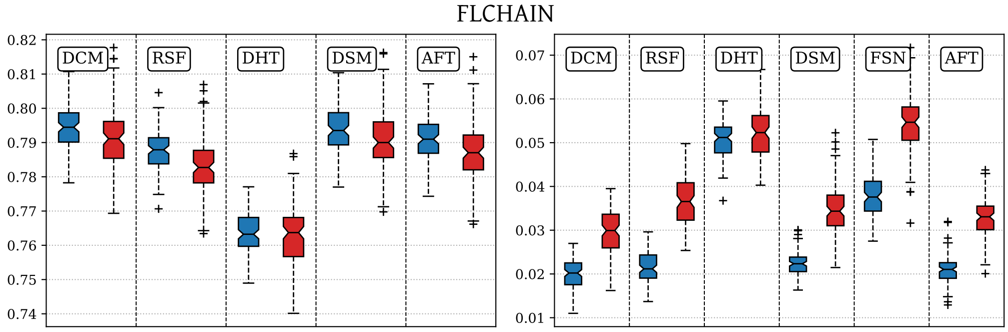
Time Dependent Concordance Index (C^{td}): Concordance Index estimates ranking ability by exhaustively comparing relative risks across all pairs of individuals in the test set. We employ the 'Time Dependent' variant of Concordance Index that truncates the pairwise comparisons to the events occurring within a fixed time horizon.

$$C^{td}(t) = \mathbb{P}\big(\hat{F}(t|\mathbf{x}_i) > \hat{F}(t|\mathbf{x}_j)|\delta_i = 1, T_i < T_j, T_i \leq t\big)$$

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $R(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\text{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|R(t)) - R(t)\big|\big]$$

[3] In practice we observe that performance of the **Random Survival Forest** model, especially in terms of calibration is strongly influenced by the choice of the hyper-parameters, `mtry` (the number of features considered at each split) and `min_node_size` (the minimum number of data samples to continue growing a tree). We thus advise carefully tuning these hyper-parameters while benchmarking **RSF**.



Model	C^{td}		**ECE**	
	Population	Minority	Population	Minority
CPH	0.6621 ± 0.0087	0.6737 ± 0.0124	0.0992 ± 0.0044	0.0878 ± 0.0071
AFT	0.7911 ± 0.0060	0.7875 ± 0.0087	0.0212 ± 0.0034	0.0329 ± 0.0046
RSF	0.7880 ± 0.0059	0.7830 ± 0.0089	0.0215 ± 0.0037	0.0368 ± 0.0053
FSN	0.6608 ± 0.0081	0.6212 ± 0.0131	0.0381 ± 0.0046	0.0545 ± 0.0068
DHT	0.7636 ± 0.0059	0.7631 ± 0.0092	0.0505 ± 0.0041	0.0525 ± 0.0056
DSM	0.7937 ± 0.0061	0.7909 ± 0.0087	0.0223 ± 0.0029	0.0347 ± 0.0056
DCM	0.7943 ± 0.0103	0.7911 ± 0.0091	0.0200 ± 0.0034	0.0294 ± 0.0049

Figure 3.3: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75th event quantile. The columns represent different quantiles at which we evaluate the individual metrics. (Tabulated results are in Appendix B.3.1)

by partitioning the risk scores R into q quantiles $\{[r_j, r_{j+1})\}_{j=1}^{q}$.

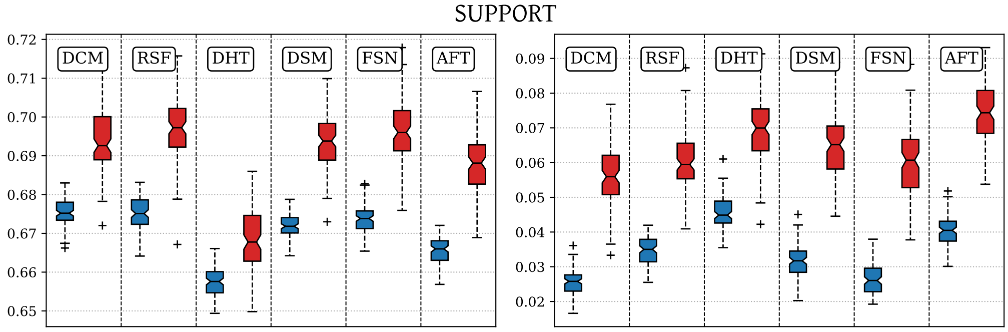
Brier Score (BS): The Brier Score involves computing the Mean Squared Error around the binary forecast of survival at a certain event quantile of interest. Brier Score is a proper scoring rule and can be decomposed into components that measure both discriminative performance and calibration.

$$\text{BS}(t) = \mathbb{E}_{\mathcal{D}}\Big[\big(\mathbb{1}\{T > t\} - \widehat{\mathbb{P}}(T > t|X)\big)^2\Big]$$

Each of the metrics described above are adjusted for censoring by using standard Thompson-Horvitz style Inverse Propensity of Censoring Weights (IPCW) estimates learnt with a Kaplan-Meier estimator over the censoring times.

3.3.4 Experimental Protocol

For the proposed model, DCM and the baselines we perform 5-fold cross validation. The predictions of each fold at the 25th, 50th and 75th quantiles of event times are collapsed together and bootstrapped in order to generate standard errors. For the proposed model and the baselines we report the mean of the evaluation metric and the bootstrapped (100 times) standard errors for the model that has the lowest Brier Score amongst all the competing set of hyper parameter choices. For DCM, the set of hyperparameter choices include the number of hidden layers for Φ tuned from $\{1, 2\}$, units in each hidden layer selected from $\{50, 100\}$, the number of mixture components K which are tuned between $\{3, 4, 6\}$ and the discounting factor for the



Model	C^{td}		**ECE**	
	Population	Minority	Population	Minority
CPH	0.6686 ± 0.0034	0.6905 ± 0.0078	0.0310 ± 0.0041	0.0685 ± 0.0079
AFT	0.6657 ± 0.0034	0.6883 ± 0.0078	0.0402 ± 0.0046	0.0741 ± 0.0085
RSF	0.6751 ± 0.0040	0.6974 ± 0.0084	0.0348 ± 0.0041	0.0603 ± 0.0080
FSN	0.6736 ± 0.0037	0.6961 ± 0.0074	0.0262 ± 0.0040	0.0601 ± 0.0097
DHT	0.6575 ± 0.0038	0.6680 ± 0.0088	0.0457 ± 0.0044	0.0696 ± 0.0089
DSM	0.6718 ± 0.0033	0.6939 ± 0.0079	0.0315 ± 0.0047	0.0650 ± 0.0087
DCM	0.6753 ± 0.0036	0.6939 ± 0.0079	0.0256 ± 0.0037	0.0561 ± 0.0085

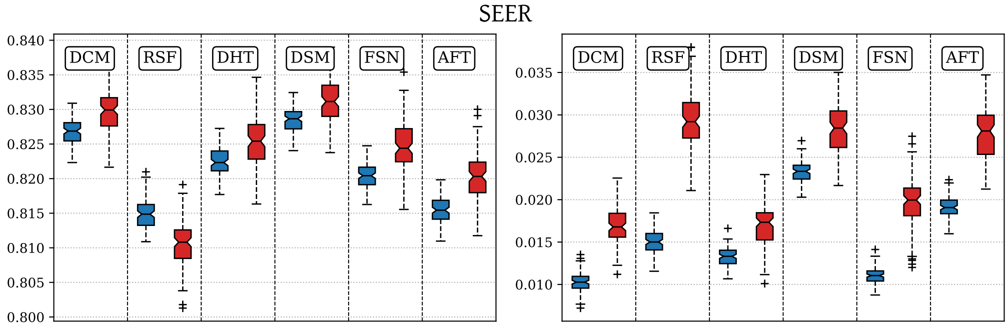
Figure 3.4: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75th event quantile. The columns represents different quantiles at which we evaluate the individual metrics. (Tabulated results are in Appendix B.3.1)

VAE-Loss, α tuned from $\{0, 1\}$. Optimization is performed using the Adam optimizer (Kingma & Ba, 2014) in `tensorflow` with learning rates fixed 1×10^{-3} and mini batch size of 128. The Baseline Survival Splines are fixed to be of degree 3 and fit using the `scipy` python package.

3.4 Results

In this section we describe the results of our various experiments with DCM and the competing baselines. We present the discriminative performance and calibration for DCM against the baselines on the three datasets for the entire population as well as the minority demographic on the 75th quantile of event times in Figures 3.3, 3.4 and 3.5 and the corresponding tables. (For tabulated results including AuROC and Brier Scores, refer to B.3.1.)

FLCHAIN: DCM beat all the other baselines in terms of discriminative performance on the entire population as well as on the minority, ‘Female’ subgroup. In terms of calibration DCM was also consistently better than all the other baselines as evidenced from low ECE scores. Interestingly, both the FSN and the linear Cox model did poorly in terms of concordance and calibration while DCM had good performance suggesting it is not sensistive to proportional hazards (PH).



Model	C^{td}		ECE	
	Population	Minority	Population	Minority
CPH	0.8082 ± 0.0020	0.8121 ± 0.0037	0.0718 ± 0.0015	0.0764 ± 0.0028
AFT	0.8155 ± 0.0020	0.8204 ± 0.0035	0.0192 ± 0.0011	0.0278 ± 0.0029
RSF	0.8153 ± 0.0021	0.8105 ± 0.0035	0.0147 ± 0.0013	0.0270 ± 0.0029
FSN	0.8204 ± 0.0019	0.8248 ± 0.0036	0.0119 ± 0.0011	0.0196 ± 0.0029
DHT	0.8224 ± 0.0020	0.8255 ± 0.0037	0.0133 ± 0.0012	0.0170 ± 0.0024
DSM	0.8281 ± 0.0019	0.8243 ± 0.0036	0.0259 ± 0.0014	0.0311 ± 0.0031
DCM	0.8270 ± 0.0019	0.8296 ± 0.0034	0.0103 ± 0.0011	0.0169 ± 0.0024

Figure 3.5: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75th event quantile. The columns represents different quantiles at which we evaluate the individual metrics. (Tabulated results are in Appendix B.3.1)

SUPPORT: For the SUPPORT dataset, RSF had the best discriminative performance at a population level, and DCM came a close and beat the other deep learning baselines. Interestingly we found that the proposed DCM had the best discriminative performance on the minority demographic beating all other baselines including RSF. While RSF was strong in terms of discriminative performance, it was however poorly calibrated in comparison to other baselines. DCM had the lowest ECE at each quantile amongst all baselines, at both the population level as well as on the minority demographic. The performance of FSN was close to DCM in terms of calibration but did poorly in terms of discrimination, further lending evidence to the fact that DCM is not restricted by PH.

SEER: In terms of calibration DCM beat all the other baselines at all quantiles of interest for the entire population as well as for the minority group. DCM also consistently had good discriminative power. (DSM did slightly better in terms of discrimination at a population level, but this was not significant). We found that DHT was a strong competitor, which is understandable since it is particularly well suited for discrete time datasets like SEER. Note that in the case of SEER we also report results stratified by the four largest minority demographics in the subset of the dataset we work with in Figure B.1. DCM has better discriminative performance across groups especially at longer horizons of event times. In terms of calibration, DCM comes close to or outperforms the semi-parametric approaches like FSN/DeepSurv.

In order to assess the influence of the protected attribute to determine the outcome we conduct additional studies involving removal of the protected attribute (unawareness) and training decoupled classifiers for each demographic. We find that in the case of unawareness we ended up with poorer calibration while decoupled classifiers had poor discriminative performance. These results are deferred to Appendices B.3.2 and B.3.3.

3.4.1 Learnt Latent Groups

For the SUPPORT dataset, we run DCM with the default set of hyperparameters with $k = 3$ latent groups. We compare the subgroup level survival curves for the learnt subgroups using DCM by plotting the mean survival rates within each group estimated with DCM as well as a Kaplan Meier Estimator.

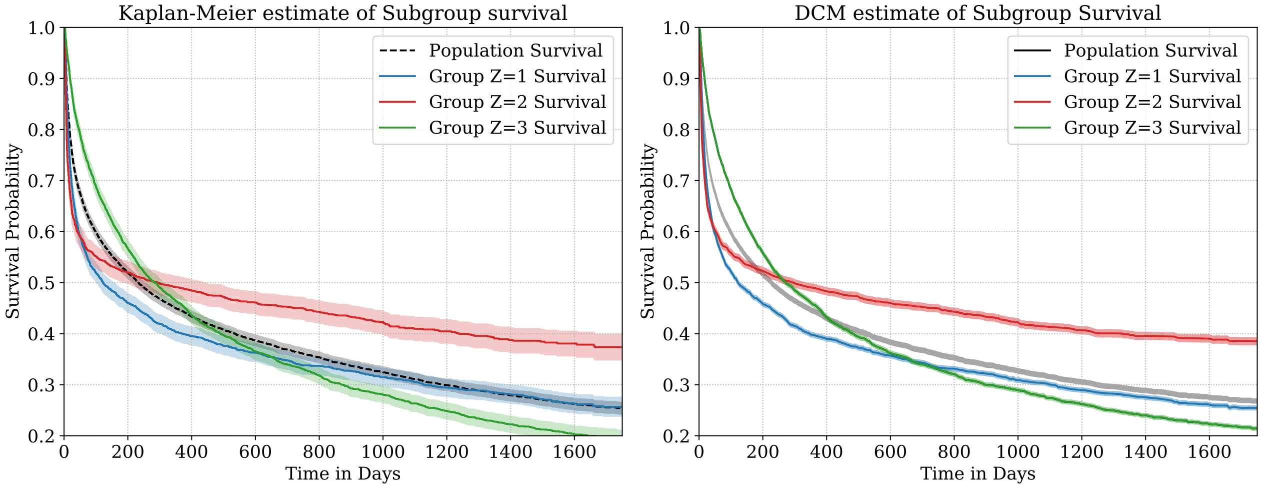


Figure 3.6: Group specific baseline survival rates for the estimated subgroups using DCM with $k = 3$ for a heldout fold of the SUPPORT dataset. The first plot is the group specific Kaplan-Meier plot and the second plot is the survival estimated with DCM.

Figure 3.6 present the estimate survival rates of the discovered subgroups using a Kaplan-Meier curve and DCM respectively. Consider the subgroup level survival curves for groups $Z = 2$ and $Z = 3$ that intersect. Intersecting survival curves indicate non Proportional Hazards which DCM is able to capture.

3.5 Conclusion

We proposed '*Deep Cox Mixtures*' to model censored Time-to-Event data. Our approach involves estimating hazard ratios within latent clusters followed by non-parametric estimation of the baseline survival rates but is not limited by the strong assumptions of constant proportional hazards. We experiment with several real-world health datasets and demonstrate superiority of our approach both in terms of discriminative performance and calibration with an emphasis on improvements especially on the minority demographics.

Part II

Discovery of Subgroups with Heterogeneous Treatment Effects

Motivation

Survival analysis is often conducted to establish the efficacy of an intervention or treatment from the result of a randomized experiment such as a clinical trial. In many scenarios however, conducting a randomized experiment is infeasible and policy makers have to defer to using observational data for decision making. In observational settings however, treatment assignment is subject to factors that might simultaneously affect outcomes, leading to a *confounding* of the treatment effect.

Furthermore, in real world scenarios not all individuals experience the same benefit from a treatment. The response to an intervention is thus *heterogenous* across individuals. Optimal decision making should account for the aforementioned heterogeneity when modelling outcomes at the level of an individual.

In this part of the thesis, we propose to model this form of heterogeneity using latent variables representing subgroups of subjects or phenotypes with similar responses to an intervention. We thus depart from our previous discussion that focussed on just factual regression. We contextualize our proposed models in a *causal* setting and propose approaches to model *counterfactual outcomes*. In a similar vein, we evaluate model performance in terms of causal metrics such as estimation of the Average Treatment Effects and Conditional Average Treatment Effects.

Chapter 4 first introduces the *Heterogenous Effect Mixture Model*, a general approach to model individual treatment effects with latent variables in observational settings. We benchmark the proposed *Heterogenous Effect Mixture Model* on multiple datasets with binary and continuous outcomes at recovering Individual Treatment Effects. We further deploy this model to a large dataset of medical claims to identify actionable subgroups of individuals at high risk of addiction from synthetic opioids such as Fentanyl, a major healthcare concern in the U.S.

Chapter 5 proposes to build on the on the introduced framework for settings in which outcomes are *censored*. Censoring is common in when treatment effect is established by comparing time to adverse effects such as onset of a medical condition, hospitalization or death. In this chapter, we will investigate relevant extensions to quantities such as Individual Treatment Effects in the time-to-event settings, and study approaches to estimate those with censored data under our proposed model.

Chapter 4

Deep Subgroup Discovery in Treatment Effect Estimation

4.1 Introduction

To identify subgroups with different treatment effects, we hypothesize that a latent variable determines the treatment effect of each individual. Moreover, individuals with similar characteristics belong to the same latent subgroup, resulting in similar responses to treatment across the subgroup. Figure 4.1 is an abstract representation of such a phenomenon.

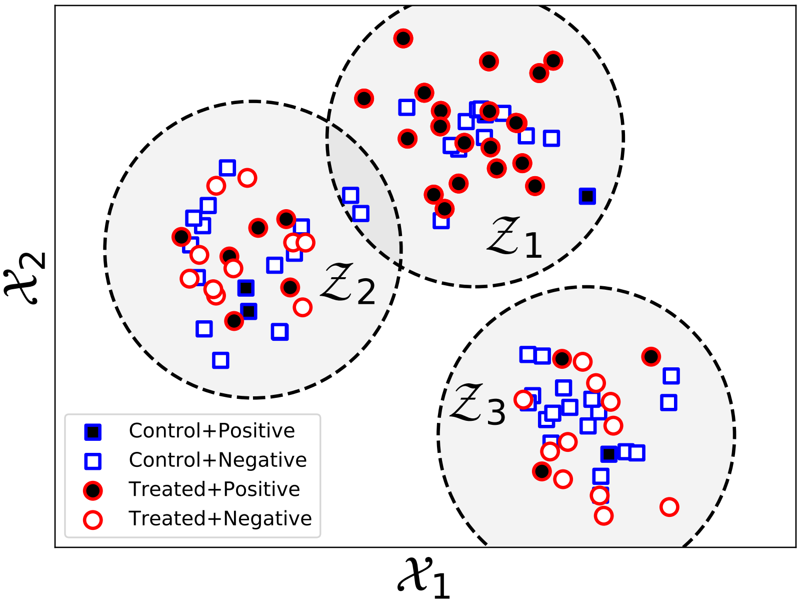


Figure 4.1: The heterogeneous effect subgroup discovery problem. Almost all instances receiving treatment in $\mathcal{Z}_1$ have a positive outcome, while very few in $\mathcal{Z}_3$ do. We are interested in recovering such latent subgroups.

We propose a Bayesian network to model these subgroups, specifically as a mixture model, along with their corresponding treatment effects. Sparsity is induced in the learned mixture component parameters to improve interpretability. Our approach, which we name the heterogeneous effect mixture model (HEMM), is similar in spirit to causal rule sets for identifying subgroups with enhanced treatment effect (Wang & Rudin, 2022) but does not require hard partitions or assignments. Moreover, we incorporate nonlinear as well as linear outcome models, which increases the expressiveness of the model to better adjust for confounding without sacrificing the interpretability of the subgroup definitions. We thus benefit from both the interpretability of sparse mixture models and the representation learning capability of neural networks. In contrast, recent works (Louizos et al., 2017; Shalit et al., 2017; Alaa & van der Schaar, 2017a) that use neural networks or nonparametric methods to estimate heterogeneous treatment effects do not identify subgroups of individuals with similar responses. While our motivating application is opioid use, the proposed approach applies to any problem domain requiring the discovery of subgroups with heterogeneous responses to actions. In this spirit, we also validate our method on synthetic data and the Infant Health and Development Program (IHDP) dataset in terms of its heterogeneous effect estimation and subgroup identification performance.

With respect to opioids, we provide domain expert interpretation of the enhanced treatment effect subgroup discovered using MarketScan data, i.e. patients at higher risk of adverse outcomes after an initial synthetic opioid prescription. Some characteristics of this subgroup are well-known and/or reflected in CDC opioid prescribing guidelines (Dowell et al., 2016): chronic pain conditions, psychological comorbidities, heart disease and obesity.

Overall, our contributions can be summarized as follows:

i) We propose the HEMM for discovering subgroups with enhanced and diminished treatment effects in a potential outcomes causal inference framework, using sparsity priors to enhance identifiability.

ii) We extend the HEMM's outcome model to include neural networks to better adjust for confounding and develop a joint inference procedure for the overall graphical model and the neural networks.

iii) We demonstrate strong performance in estimating heterogeneous effects and identifying subgroups compared to existing approaches. Additionally, we apply the methodology to a large-scale medical claims dataset and discover subgroups at enhanced risk of adverse outcomes with synthetic opioids.

4.2 Related Work

The machine learning literature has traditionally focused on outcome prediction and regression problems. The study of Outcomes along with potential counterfactuals within a causal framework, where treatment has to be modeled explicitly is relatively under studied in machine learning.

The identification of subgroups with heterogeneous or enhanced treatment effects has been addressed in the statistics literature by building separate factual and counterfactual outcome models and then regressing the difference of the two using another method, e.g. a decision tree (Su et al., 2009). This final model can then be deployed to identify subgroups. Within this category of approaches, Lipkovich et al. (2011) propose the subgroup identification based on differential effect search (SIDES) algorithm, Dusseldorp & Mechelen (2014) propose the qualitative interaction trees (QUINT) algorithm, and Foster et al. (2011) propose the virtual twins (VT) method. We consider empirical comparisons to these algorithms in the sequel.

Wang & Rudin (2022) propose causal rule sets for discovering subgroups with enhanced treatment effect. This is the closest to and an inspiration for our work. That work seeks to learn discrete human-interpretable rules predictive of enhanced treatment effect and involves optimization by Monte Carlo methods. We consider instead a mixture of experts approach with soft assignment to groups that retains most of the interpretability but allows greater expressiveness and can be optimized via gradient methods. Our outcome model equation 4.4, equation 4.5 also differs from that of Wang & Rudin (2022). Most importantly, we allow nonlinearity in the form of neural networks whereas Wang & Rudin (2022) considers only linear models. Our model also has a single term representing the main effect of treatment whereas Wang & Rudin (2022) has three such terms: a population average, a subgroup term that is always active, and a subgroup term that is only active under treatment.

Recent papers have proposed estimating heterogeneous/individual treatment effects using neural networks (Louizos et al., 2017; Shalit et al., 2017) or a Bayesian nonparametric method involving Gaussian processes (Alaa & van der Schaar, 2017a). These methods rely on constructing distributional representations of the factual and counterfactual outcomes that are similar in a statistical sense. While these methods perform well on estimating heterogeneous effects, they do not identify subgroups of individuals with similar treatment effects and characteristics and are thus less interpretable. This makes the application of such methods to inform policy decisions more difficult.

4.3 Proposed Approach: Heterogeneous Effect Mixture Model

In this section, we propose a generative mixture model for heterogeneous treatment effects. One way to model heterogeneity, and the one in our proposal, is as a finite mixture of components with a different treatment effect model in each component (some enhanced and some diminished). For tractability, we keep the form of the mixtures to be the simplest possible: Gaussian-distributed for continuous covariates and Bernoulli-distributed for discrete covariates. Section 4.3.3 presents a model for the outcomes as they depend on treatment, covariates, and mixture membership, including nonlinear dependence on the covariates.

4.3.1 Preliminaries

We adopt the Neyman-Rubin potential outcomes framework (Rubin, 1974) for causal inference. Define random variables $\mathbf{X} \in \mathbb{R}^d$ representing covariates and $T \in \{0, 1\}$ as the treatment indicator. The subset of continuous-valued covariates are denoted $\mathbf{X}_{\text{cont}}$ and the discrete covariates (binary-valued or binarized) are denoted $\mathbf{X}_{\text{disc}}$. We will sometimes refer to $T = 1$ as 'treatmed' and $T = 0$ as 'control.'

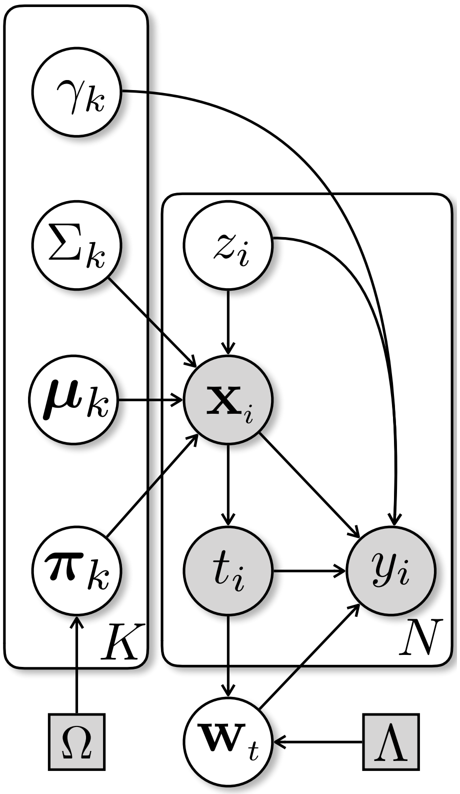


Figure 4.2: The proposed heterogeneous effect mixture model (HEMM) in plate notation. For each instance i, $(\mathbf{x}_i, t_i, y_i)$ are the observed variables and z_i is a latent variable that determines membership in one of the K mixture components. Each component has an associated coefficient γ_k that determines the main treatment effect.

Corresponding to the levels of treatment are two potential outcomes $Y(0)$ and $Y(1)$, which are the outcomes under $T = 0$ and $T = 1$ respectively. These outcomes can be discrete- or continuous-valued. We are given an observational dataset of samples $\mathcal{D} = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^N$ in which only one of the outcomes is observed for each individual: if $t_i = 0$ then $y_i = y(0)_i$, and if $t_i = 1$ then $y_i = y(1)_i$. Our interest lies in estimating the conditional average treatment effect (CATE) conditioned on $\mathbf{X}$, defined as

$$\tau(\mathbf{X}) = \mathbb{E}\big[Y(1) - Y(0) \mid \mathbf{X}\big].$$

In this work, the dependence on $\mathbf{X}$ is mediated primarily through subgroup membership, i.e. members of the same subgroup have similar treatment effects. We make the standard assumptions that allow CATE to be identifiable from observational data, namely exchangeability conditioned on the available covariates, $T \perp (Y(0), Y(1)) \mid \mathbf{X}$, positivity of the treatment propensity, $0 < p(T = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$, and no dependence between individuals i.e. the stable unit treatment value assumption (SUTVA) (Rubin, 1986; Hernán & Robins, 2018). The first two assumptions are collectively known as strong ignorability (SITA).

For the mixture model proposed in this chapter, we additionally define the latent random variable $Z \in \mathcal{Z} = \{1, \ldots, K\}$ to indicate mixture membership. Both the distribution of covariates and the treatment effect are dependent on Z as described next.

4.3.2 Generative Model

The generative model is presented in Figure 4.2 in plate notation. We first give an overview of the distributions and then go into more detail regarding $\mathbf{X}$ and Y in Section 4.3.3.

1. We draw a sample z_i independently for each individual i that determines the latent group membership. The prior distribution for Z is uniform over the K groups,

$$Z \sim \text{Uniform}(K). \tag{4.1}$$

2. Conditioned on the latent group assignment $z_i = k$,

- The $\mathbf{x}_{\text{cont},i}$ are drawn i.i.d. from a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance Σ_k:

$$\mathbf{X}_{\text{cont}} \mid z_i = k \sim \text{Normal}(\boldsymbol{\mu}_k, \Sigma_k). \tag{4.2}$$

  In this chapter, we constrain the off-diagonal elements of Σ_k to be 0 to reduce the number of parameters, although non-diagonal covariances can be easily accommodated.

- The $\mathbf{x}_{\text{disc},i}$ are drawn i.i.d. from a multivariate Bernoulli distribution with mean $\boldsymbol{\pi}_k$:

$$\mathbf{X}_{\text{disc}} \mid z_i = k \sim \text{Bernoulli}(\boldsymbol{\pi}_k). \tag{4.3}$$

3. Conditioned on the covariates $\mathbf{x}_i$, the treatment assignment t_i is drawn from a Bernoulli distribution whose mean is a function of $\mathbf{x}_i$:

$$T \mid \mathbf{x} \sim \text{Bernoulli}(\rho(\mathbf{x})).$$

   This corresponds to a model for treatment *propensity*. Note from Figure 4.2 that the generative model assumes that T is conditionally independent of Z given $\mathbf{X}$. Under this assumption, it will be seen in Section 4.3.4 that inference for the propensity model can be done independently from the other components of the generative model.

4. Finally, an outcome sample y_i is drawn from a distribution whose mean μ_y is a function of the covariates $\mathbf{x}_i$, treatment assignment t_i, and latent group assignment z_i. If Y is binary-valued, the distribution is Bernoulli,

$$Y \mid \mathbf{x}, t, z \sim \text{Bernoulli}\big(\mu_y(\mathbf{x}, t, z)\big),$$

whereas if Y is continuous, the distribution is Gaussian,

$$Y \mid \mathbf{x}, t, z \sim \text{Normal}\big(\mu_y(\mathbf{x}, t, z), \sigma_y^2\big),$$

where σ_y^2 is the variance. The outcome model is discussed further in Section 4.3.3.

Note we are interested in estimating the causal quantity,

$$\mathbb{E}[Y(t)|X] = \mathbb{E}[Y|\mathbf{do}(T = t), X] = p(Y|\mathbf{do}(T = t), X).$$

Here, the first equality is from definition of interventional quantities and the second equality holds due to Y being binary.

Theorem 1 (Identifiability) *Under the Directed Acyclic Graph in Figure. 4.2,*

$$p(Y|\boldsymbol{do}(T = t), X) = \int_Z p(Y|X, Z, T = t)p(Z|X).$$

Theorem 1 confirms that we can estimate the CATE from the observational quantities introduced above. The proof is deferred to the Appendix C.1.

4.3.3 Treatment Outcome Model

We model the enhanced or diminished treatment effect in a subgroup through the following relationships. In the case where Y is binary, its mean is equal to the probability of $Y = 1$. We define the latter using the logistic sigmoid function g to be

$$p(Y = 1 \mid \mathbf{x}, t, Z = k; \mathbf{w}_t, \gamma_k) = g\big(f(\mathbf{x}; \mathbf{w}_t) + \gamma_k t\big), \tag{4.4}$$

where $f(\mathbf{x}; \mathbf{w}_t)$ is a function of $\mathbf{x}$ parametrized by $\mathbf{w}_t$, $t = 0, 1$. The term $\gamma_k t$ represents the main effect due to treatment and the coefficient γ_k, i.e. the size of the effect, depends on the group membership $Z = k$. The parameters $\mathbf{w}_t$ are allowed to be different for $t = 0$ and $t = 1$ to better account for differing covariate distributions $p(\mathbf{x} \mid t)$ between the two treatment groups, a.k.a. selection bias. In the case of continuous Y, we replace g with the identity function as follows:

$$\mathbb{E}[Y = 1 \mid \mathbf{x}, t, Z = k; \mathbf{w}_t, \gamma_k] = f(\mathbf{x}; \mathbf{w}_t) + \gamma_k t. \tag{4.5}$$

The simplest choice for function $f(\cdot)$ is linear, i.e., two linear functions $\mathbf{w}_0^\top \mathbf{x}$ and $\mathbf{w}_1^\top \mathbf{x}$. In practice, however, the outcome may have a highly nonlinear dependence on the covariates. To accommodate nonlinear covariate interactions and thus better adjust for confounding, we also allow f to be a nonlinear function. In this chapter, we experiment with one- and two-hidden-layer feedforward neural networks with ReLU activations. Outcomes under $t = 0$ and $t = 1$ are produced by two different heads of the network, following Shalit et al. (2017); Louizos et al. (2017); Johansson et al. (2016). Even in the nonlinear case, the assignment of an individual to a subgroup is still described by a mixture model and directly interpretable in terms of the original feature representation, thus preserving interpretability of the discovered subgroups.

It is possible to regularize the outcome models equation 4.4, equation 4.5 with ℓ_2 or ℓ_1 regularization $\Lambda(\mathbf{w}_t)$, which is equivalent to adding a normal or Laplace prior on the parameter $\mathbf{w}_t$. In this work however, we use weight decay instead as discussed in Section 4.3.5.

4.3.4 Inference

We would like to fit our proposed model to a given observational dataset $\mathcal{D}$. Denote by $\boldsymbol{\Theta} = (\{\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\pi}_k, \gamma_k\}_{k=1}^K, \mathbf{w})$ the set of all parameters of the model.

We have considered two approaches: maximizing the joint likelihood $p(\mathbf{x}_i, t_i, y_i; \boldsymbol{\Theta})$, and maximizing the conditional likelihood $p(y_i|\mathbf{x}_i, t_i; \boldsymbol{\Theta})$. The joint and conditional likelihoods can be related as follows:

$$\sum_{i=1}^{N} \ln p(\mathbf{x}_i, t_i, y_i) = \sum_{i=1}^{N} \left[\ln p(\mathbf{x}_i) + \ln p(t_i \mid \mathbf{x}_i) + \ln p(y_i \mid \mathbf{x}_i, t_i)\right]. \tag{4.6}$$

The conditional likelihood can be further expanded as

$$\ln p(y_i \mid \mathbf{x}_i, t_i) = \ln \left(\sum_{k=1}^{K} p(z_i = k \mid \mathbf{x}_i) p(y_i \mid \mathbf{x}_i, t_i, z_i = k) \right), \tag{4.7}$$

where we have used the conditional independence of Z and T given $\mathbf{X}$ in the first factor on the right-hand side. The resulting first factor $p(z_i = k \mid \mathbf{x}_i)$ as well as the term $p(\mathbf{x}_i)$ in equation 4.6 depend only on the mixture model equation 4.1–equation 4.3, with $p(\mathbf{x}_i) = \sum_{k=1}^{K} p(z_i = k) p(\mathbf{x}_i \mid z_i = k)$ and $p(z_i = k \mid \mathbf{x}_i) = p(z_i = k) p(\mathbf{x}_i \mid z_i = k)/p(\mathbf{x}_i)$. The second factor on the right-hand side of equation 4.7 depends only on the outcome model equation 4.4, equation 4.5. The remaining term $p(t_i \mid \mathbf{x}_i)$ in equation 4.6 depends on the propensity model. Since this is the only place where the propensity model appears, its inference is separable from the remainder of the problem, as claimed in Section 4.3.2. We do not discuss propensity modeling further as it is not the focus of this work.

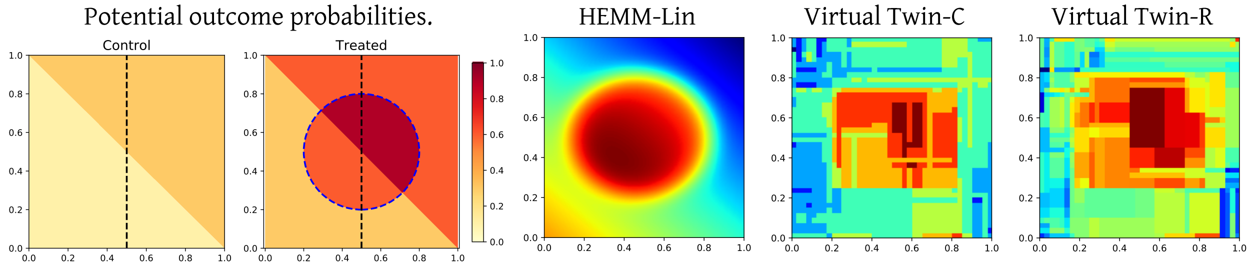
Although maximizing the joint likelihood equation 4.6 results in some closed-form expressions and accordingly easier inference of parameters, we have observed in practice that maximizing the conditional likelihood equation 4.7 has superior performance in estimating the potential outcomes $Y(t)$ and treatment effects. Therefore, we pursue this discriminative approach in this work. The full objective function is therefore

$$\sum_{i=1}^{N} \ln p(y_i \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta}) - \lambda \Omega(\boldsymbol{\pi}), \tag{4.8}$$

where λ controls the strength of the prior, which here is the weight decay.

4.3.5 Evidence Lower Bound (ELBO) Optimization

Instead of optimizing the conditional log-likelihood in equation 4.8 directly using a gradient method, we choose to lower bound the likelihood with a variational approximation, more commonly known as the



$p(Y(0) = 1 \mid \mathbf{X})$ (control) and $p(Y(1) = 1 \mid \mathbf{X})$ (treated). Treatment propensity is greater to the right of the black dashed line, while the blue dashed line denotes the region with enhanced effect.

Figure 4.3: **Synthetic** dataset and enhanced effect subgroups discovered by HEMM and Virtual Twins (VT).

Evidence Lower Bound (ELBO) (Blei et al., 2017). For any variational distribution $q(Z)$ over the latent Z,

$$\begin{aligned}\ln p(y_i \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta}) &= \ln \sum_{k=1}^{K} p(y_i, z_i = k \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta}) \\ &= \ln \left(\mathbb{E}_q \left[\frac{p(y_i, z_i \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta})}{q(Z)} \right] \right) \\ &\geq \mathbb{E}_q \left[\ln \frac{p(y_i, z_i \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta})}{q(Z)} \right] \end{aligned} \tag{4.9}$$

using Jensen's inequality. Now, replacing $q(Z)$ with $p(z_i|\mathbf{x}_i; \boldsymbol{\Theta})$ and using equation 4.7 (and $Z \perp\!\!\!\perp T \mid \mathbf{X}$ from Figure 4.2), we obtain

$$\text{ELBO}(y_i, \mathbf{x}_i, t_i; \boldsymbol{\Theta}) = \sum_{k=1}^{K} p(z_i = k \mid \mathbf{x}_i; \boldsymbol{\Theta}) \ln p(y_i \mid \mathbf{x}_i, t_i, z_i = k; \boldsymbol{\Theta}). \tag{4.10}$$

We hence substitute equation 4.10 in place of $\ln p(y_i \mid \mathbf{x}_i, t_i; \boldsymbol{\Theta})$ in equation 4.8 and proceed to maximize the objective function using the Adam gradient method (Kingma & Ba, 2014), a variant of stochastic gradient descent that is a popular choice for non-convex functions like neural networks. The same method is used for both linear and nonlinear f in equation 4.4, equation 4.5. As noted above, the first factor $p(z_i = k \mid \mathbf{x}_i; \boldsymbol{\Theta})$ in equation 4.10 depends only on the mixture model parameters in equation 4.1–equation 4.3 while the second factor depends only on the outcome model parameters in equation 4.4, equation 4.5. We enable "weight decay" (Krogh & Hertz, 1992) on the parameters $\mathbf{w}_t$ as a form of regularization. For tractability, we compute the ELBO only over a fixed-size mini-batch of the data before each parameter update. Additional details on the algorithm and parameter initialization can be found in Appendices C.3 and C.4 in the supplement.

We also considered an expectation-maximization (EM) algorithm to maximize equation 4.8 as an alternative to ELBO. Our experience however was that ELBO provided better fit in terms of Log-Likelihood and heterogeneous effect estimates in terms of the metric reported in Section 4.5.1. A full description of the EM method and a comparison to the ELBO optimization is deferred to the Appendix.

4.4 Experiments

We demonstrate the performance of HEMM on a synthetic dataset, the semi-synthetic Infant Health and Development Program (**IHDP**) dataset, and a real-world dataset on opioids. These datasets are described

Total Covariate Dimension	**1226**	3 Continuous, 1223 Binary
ICD-9 Diagnostic Codes	1013	Binary
CPT Procedure Codes	171	Binary
Hand-Crafted Comorbidities	41	Binary
Daily Morphine Equivalent, Total Number of Visits, Age	3	Continuous

	Addicted (Y=1)	Not-Addicted (Y=0)	Total
Treated (T=1)	2060	19983	**22043**
Control (T=0)	7269	176156	**183425**
Total	**9329**	**196139**	**205468**

Table 4.1: **Opioid** Dataset Statistics

further below.

Synthetic: We take $\mathbf{X} = (X_0, X_1) \in \mathbb{R}^2$ and sample it from a uniform distribution over $\mathcal{X} = [0,1]^2$. In order to simulate the selection bias inherent in observational studies, the treatment variable depends on $\mathbf{X}$ as $T \sim$ Bernoulli(0.4) for $x_0 < 0.5$ and Bernoulli(0.6) for $x_0 > 0.5$. The potential outcomes $Y(0)$ and $Y(1)$ are also Bernoulli with means given by the functions of $\mathbf{X}$ shown in Figure 4.3 (Outcome). The figure shows that $p(Y(1) = 1 \mid \mathbf{X}) > p(Y(0) = 1 \mid \mathbf{X})$, i.e. treatment increases the probability of positive outcome. Note that under the conditional exchangeability assumption we have $p(Y(t) = 1 \mid \mathbf{X}) = p(Y = 1 \mid T = t, \mathbf{X})$. We model the effect of the confounders $\mathbf{X}$ by assigning higher probability to the upper triangular region of $\mathcal{X}$. This together with the distribution of T imply that individuals who are more likely to have positive outcome regardless of treatment (upper triangle) are also more likely to receive treatment (right half-square). Lastly, we model the enhanced treatment effect group as a circular region $\mathcal{S} = \{x : \|x - c\|_2 < r\}$, where $p(Y(1) = 1 \mid \mathcal{S}) > p(Y(1) = 1 \mid \mathcal{X} \backslash \mathcal{S})$. We set $c = (\frac{1}{2}, \frac{1}{2})$ and $r = \frac{1}{4}$. A total of 1000 samples $(\mathbf{x}_i, t_i, y_i)$ were generated as described above.

IHDP (Semi-Synthetic): The IHDP dataset has gained popularity in the causal inference literature dealing with heterogenous treatment effects (Alaa & van der Schaar, 2017a; Shalit et al., 2017; Louizos et al., 2017; Hill, 2011). The original data includes 25 real covariates and comes from a randomized experiment to evaluate the benefit of IHDP on IQ scores of three-year-old children. A selection bias was introduced by removing some of the treated population, thus resulting in 608 control patients and 139 treated (747 total). The outcomes were simulated using the standard non-linear 'Response Surface B' as described in Hill (2011).

Opioid: We sampled a sub-population consisting of healthcare claims for five million patients from the MarketScan Commercial claims database. These claims describe patients' medical histories, including both inpatient admissions and outpatient services. Diagnoses, procedures, prescriptions and dosages are recorded. We follow the cohort selection procedure outlined in Zhang et al. (2017) to filter patients based on several criteria.

For each patient in our final cohort, we create a feature vector that includes basic demographic information such as age, gender and geographic region. We also included a predefined set of procedures along with diagnostic codes which are associated with opioids and/or addiction, based on input from a physician.

We label all patients who received addiction diagnoses and patients who continued use of opioids for more than one year after the initial prescription as belonging to the positive (adverse outcome) class. Patients who discontinued opioid use within one year of initial treatment were labeled as negative. We use the terms "addicted" and "not addicted" as shorthand for these outcomes. Patients prescribed natural or semi-synthetic opioids are considered the control group, whereas patients administered synthetic opioids are considered the treated group. Table 4.1 summarizes the basic statistics of this dataset.

4.4.1 Algorithms in Comparison

We have considered Virtual Twins (VT) (Foster et al., 2011), QUINT (Dusseldorp & Mechelen, 2014), and SIDES (Lipkovich et al., 2011) among methods that identify subgroups with different treatment effects. We implemented two versions of VT in which the treatment effect is modeled by a decision tree classifier (VT-C) or regressor (VT-R). For VT-C, to better represent the continuous-valued treatment effect (which is a difference in probabilities even if Y is binary), we use a collection of decision tree classifiers obtained by applying different thresholds to the treatment effect.

For QUINT and SIDES, we utilized the standard R implementations and performed extensive hyperparameter tuning. However both QUINT and SIDES failed to recover any subgroups on **Synthetic** and **Opioid** and we thus did not consider them further. For QUINT, the likely reason is that its assumption of a subgroup with diminished effect is not always met, whereas for SIDES, there may be a numerical issue in how it discretizes continuous covariates.

In terms of methods that only predict heterogeneous effects and do not identify subgroups, we also compare our method with some common approaches in Table 4.2. Here Linear-1 corresponds to a single ordinary least squares (for continuous outcomes) or logistic regression (for binary outcomes) for both factual and counterfactual outcomes. In the case of Linear-2, we fit two separate linear models to the control and treated populations to better accommodate selection bias and confounding. The other baselines, k-NN, GP, and CFRF are non-parametric versions of this approach where the estimators of the factual and counterfactual outcomes are k-nearest neighbours, Gaussian processes with a linear kernel and Random Forests respectively.

4.5 Results

We evaluated the proposed HEMM quantitatively on two tasks, prediction of heterogeneous treatment effects and identification of subgroups with enhanced or reduced effect. These results are discussed in Sections 4.5.1 and 4.5.2 respectively in comparison to existing methods, focusing on those that also estimate heterogeneous effects in an interpretable manner. Methods used in the comparison are described in Section 4.4.1 and parameter selection details are in Appendix C.4. In Section 4.5.3, we provide qualitative results for the **Opioid** dataset on the features the model discovers as characteristics of "at-risk" individuals, i.e. those in enhanced effect subgroups.

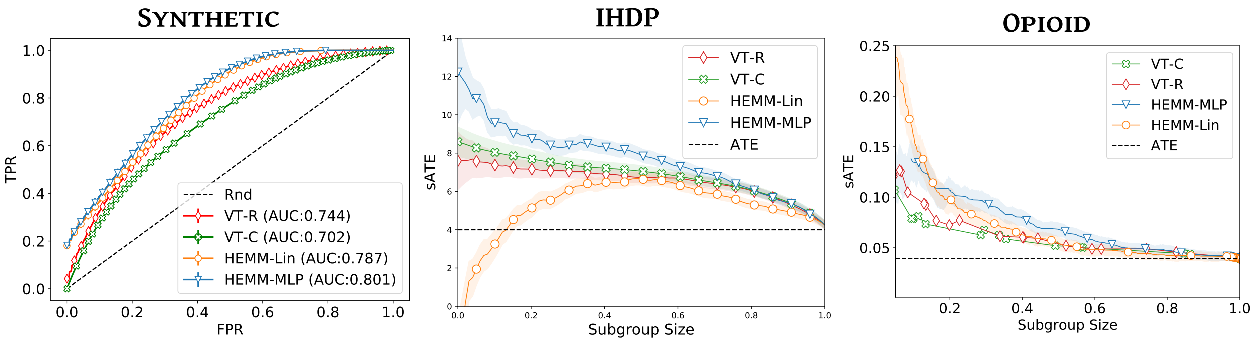


Figure 4.4: Performance of the proposed HEMM and Virtual Twins on the subgroup discovery task. For **SYNTHETIC** data, we have access to ground truth labels for the enhanced treatment effect group and hence compare performance using the ROC. For the **IHDP** and **OPIOID** datasets, we compare average treatment effect (ATE) estimates within the identified subgroup as a function of subgroup size (as a fraction of the population).

4.5.1 Heterogeneous Effect Estimation

We first evaluate our performance on estimation of the CATE ($\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X}]$). A popular metric for this evaluation is the *Precision in Estimating Heterogeneous Effects* (PEHE). The PEHE is defined as

$$\text{PEHE} = \frac{1}{n}\sum_{i=1}^{n}\left(f_1(\mathbf{x_i}) - f_0(\mathbf{x}_i) - \mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}_i]\right)^2 .$$

Here $f_1(\cdot)$ and $f_0(\cdot)$ are the estimated potential outcomes under treatment and control, respectively.

Table 4.2 compares the performance of HEMM against the methods described in Section 4.4.1 on both in-sample PEHE (corresponding to a retrospective study) computed on the training data, and out-of-sample PEHE computed on held-out test data. HEMM-MLP and HEMM-Lin refer to the proposed approach with f in equation 4.4, equation 4.5 as a multilayer perceptron and linear function respectively to model the effect of confounders on the outcome.

HEMM consistently outperforms these standard causal inference baselines. GP and Linear-2 perform close to HEMM on **SYNTHETIC**. We noticed that when a larger sample of data points is available to VT-R, its performance increases dramatically. However, its performance drops in higher-dimensional settings as in the case of **IHDP** and **OPIOID**; this is expected with methods involving non-parametric regression.

4.5.2 Subgroup Identification

SYNTHETIC: In the case of synthetic data, the subgroup with enhanced treatment effect is known. We first visualize the performance of HEMM and VT in identifying this subgroup. For HEMM-Lin, Figure 4.3 shows the estimated probability $p(Z|\mathbf{X})$ of belonging to the enhanced effect subgroup evaluated on the test set. The true circular region is recovered well. Furthermore, in Figure 4.3 we plot the VT-C and VT-R predictions of CATE on the test set. For VT-C, the prediction represents an average over the collection of decision tree classifiers, while for VT-R, it is simply the output of the decision tree regressor. While the difficulty in reproducing the circular shape is expected for decision trees, the enhanced effect estimates are also less uniform as compared to HEMM.

	Synthetic		**IHDP**	
	In-sample	Out-Sample	In-sample	Out-Sample
HEMM-MLP	$\mathbf{0.101 \pm 10^{-3}}$	$\mathbf{0.102 \pm 10^{-3}}$	$\mathbf{1.6 \pm 0.10}$	$\mathbf{1.8 \pm 0.10}$
HEMM-Lin	$\mathbf{0.116 \pm 10^{-3}}$	$\mathbf{0.116 \pm 10^{-3}}$	$\mathbf{2.8 \pm 0.32}$	$\mathbf{2.9 \pm 0.33}$
Linear-1	0.278 ± 10^{-3}	0.278 ± 10^{-3}	7.9 ± 0.46	7.9 ± 0.47
Linear-2	0.106 ± 10^{-3}	0.107 ± 10^{-3}	2.3 ± 0.18	2.4 ± 0.21
k-NN	0.210 ± 10^{-3}	0.210 ± 10^{-3}	3.2 ± 0.12	4.2 ± 0.22
GP	0.106 ± 10^{-3}	0.107 ± 10^{-3}	2.1 ± 0.11	2.3 ± 0.14
CFRF	0.146 ± 10^{-3}	0.142 ± 10^{-3}	2.7 ± 0.31	3.3 ± 0.72
VT-R	0.130 ± 10^{-3}	0.130 ± 10^{-3}	2.5 ± 0.26	2.9 ± 0.51

Table 4.2: $\sqrt{\text{PEHE}}$ values in estimating heterogeneous effects. Error represents 95% confidence interval of multiple Monte Carlo initializations.

We also evaluate subgroup identification more quantitatively by treating it as a problem of classifying whether or not points in the test set belong to the enhanced effect subgroup. ROC curves may then be plotted as in Figure 4.4. For HEMM, the ROC is traced by varying the threshold on the probability $p(Z \mid \mathbf{X})$ of being in the enhanced effect subgroup. Similarly for VT-C and VT-R, the threshold on the CATE estimates is varied. HEMM has higher ROCs than VT on this example, in line with Figure 4.3. There is little difference between HEMM-Lin and HEMM-MLP since the dependence on the covariates $\mathbf{X}$ in Figure 4.3 is simple and complex adjustment is not needed.

IHDP and **Opioid**: For the other two datasets, we conduct a relative comparison with VT since we lack data on ground truth subgroups. The evaluation involves two steps. First, we assign individuals to an enhanced effect subgroup of varying size. (The same procedure can be used for a diminished effect subgroup but we omit the results due to space.) For HEMM, we choose the subgroup k with the largest main effect γ_k and vary the threshold applied to the corresponding membership probability $p(Z = k \mid \mathbf{X})$ returned by the model. For VT-C and VT-R, we vary the threshold applied to the CATE estimates, either the composite estimate of the decision tree classifiers or the regressor estimate, the same quantities as for the synthetic data.

In the second step, we build a propensity score model (an estimator of treatment propensity $p(T = 1|\mathbf{X})$) to estimate the average treatment effect (ATE) conditioned on belonging to the enhanced treatment effect subgroup defined in the first step. For the propensity score model $e(\mathbf{X})$, we fit a random forest, for which parameter tuning is performed on the dev set. We then use the inverse probability of treatment weighting (IPTW) estimator (Imbens, 2004) of the ATE within a subgroup $\mathcal{S}$ as follows:

$$\hat{\tau}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left(\frac{y_i t_i}{e(\mathbf{X}_i)} - \frac{y_i (1 - t_i)}{1 - e(\mathbf{X}_i)} \right). \tag{4.11}$$

IPTW estimation is used for both HEMM- and VT-defined subgroups to be consistent.

Further, Figure 4.4 plot subgroup ATE versus subgroup size (as a fraction of the population) as the threshold for subgroup assignment is varied. When the subgroup is the entire population at size 1.0, all curves meet at the population ATE (dashed line). Since we have selected the enhanced effect subgroup, the curves are then expected to increase as the subgroup is restricted to individuals with larger treatment effects. The fact that

Musculoskeletal System	Nervous System
1.0 spinal curve (kyphosis, lordosis, scoliosis)	1.0 extrapyramidal diseases/movmt. disorders
1.0 ankle fracture	1.0 idiopathic peripheral neuropathies
1.0 sprains/strains of hand and wrist	1.0 headaches
Integumentary System	**Endocrine System**
1.0 cellulitis and abscess of finger and toe	.70 simple and unspecified goiter
1.0 local skin infections	.67 other endocrine disorders
1.0 psoriasis and similar disorders	.65 thyrotoxicosis with or without goiter
Reproductive System	**Digestive and Excretory Systems**
1.0 female infertility	.71 benign neoplasm of intestinal tract
.82 testicular dysfunction	.69 inguinal hernia
.82 disorders of penis	.58 diverticulitis
Circulatory System	**Immune System**
1.0 hypertensive heart disease	.56 immunization
.79 other disorders of circulatory system	.54 strep throat and scarlet fever
.70 cardiac dysrhythmias	.52 bacterial infections in other conditions
Nutrition	**Visual System**
1.0 BMI	.87 keratitis
1.0 b-complex deficiency	.60 other disorders of eye (epi)scleritis
.71 disorder of electrolyte/acid-base balance	.57 visual disturbances
Auditory System	**Psychology**
.53 vertiginous syndrome/vestibular disorder	1.0 suspected mental health condition
.50 otitis media/eustachian tube disorders	.58 adjustment reaction
.44 disorders of pinna and mastoid process	.55 nondependent abuse of drugs
Digestive System (upper/oral)	**Respiratory System**
.77 hernia, abdominal cavity w/o obstruction	1.0 other diseases of respiratory tract
.77 dentofacial anomalies of jaw	.74 deviated nasal septum
.76 diseases of oral soft tissues	.69 influenza

Table 4.3: Top features of the enhanced effect subgroup k discovered by HEMM-MLP on the **Opioid** dataset. The numbers are the ratios $\pi_{jk}/\sum_{k'}\pi_{jk'}$, where $1/2$ represents no increase in prevalence over the other subgroup ($K = 2$).

this increase is nearly monotonic for HEMM-MLP is evidence for the validity of the discovered subgroup, since the IPTW estimator used here is an independent check on the treatment effect model equation 4.4, equation 4.5 used by HEMM. Compared to VT, the subgroups identified by HEMM-MLP have higher ATE. This suggests that for a given subgroup size, HEMM-MLP is better at grouping together individuals with more enhanced effects. HEMM-Lin on the other hand displays contrasting performances. On IHDP in Figure 4.4, the estimated ATE actually decreases for subgroup sizes less than 0.5, likely due to the inadequacy of a linear model to adjust for confounding and accurately estimate CATE. However, the ATE does increase monotonically and faster than for VT.

4.5.3 Interpretation of the Opioid Enhanced Effect Subgroup

We now turn our focus to the motivating application of opioids and analyze key characteristics of the enhanced effect subgroup, i.e. those patients at greater risk of adverse outcomes when treated initially with synthetic opioids. To interpret these features, we collaborated with a subject matter expert (SME) with a PhD in cognitive neuroscience and a clinical research emphasis in chronic pain conditions and treatments, including opioids. Table 4.3 shows the top features of the enhanced effect subgroup as identified by HEMM-MLP. The features are organized by general bodily system and sorted in descending order of prevalence relative to the other subgroup (see table caption); the selection of 3 features in each system was arbitrary and chosen primarily for simplicity and space constraints.

Patients with a history of chronic conditions in general, as well as chronic pain conditions more specifically, are at an increased risk for addiction. Many of the chronic conditions in Table 3, e.g. heart disease (circulatory system), psoriasis (integumentary system), and BMI/obesity (nutrition) also appear in the CDC opioid prescribing guidelines (Dowell et al., 2016) or have extensive literature linking them to increased risk for long-term pain, either intrinsic to the condition or due to needed medical procedures that are more likely to expose patients to opioids (Glanz et al., 2018). For example, numerous papers show a link between increased body-mass index (BMI) and increased pain intensity and duration (with anti-correlations between BMI and pain recovery) (Okifuji & Hare, 2015), and obesity has also been associated with higher initial opioid doses (Kobus et al., 2012). Additionally, the chronic nutritional deficiencies and imbalances shown in Table 4.3 have been linked to acute but intense muscle spasms as well as peripheral polyneuropathies and paresthesias (Mostacci et al., 2018), pain disorders which also show up as increased risk factors (nervous system).

Regarding chronic pain conditions, patients with a history of abnormal spinal curvatures (which can produce low back pain and neuropathy), idiopathic peripheral neuropathies, and headaches (musculoskeletal and nervous systems) are at increased risk for addiction. These are not surprising as they are notoriously difficult to treat using non-opioid therapies such as non-steroidal anti-inflammatory drugs (NSAIDs), steroids, or common procedures and surgeries (e.g. joint replacement or local injections) (Crofford, 2013). They involve pain that may be severely intense or debilitating, sometimes unpredictable or idiopathic, and often non-specific or diffuse (pain is referred, not well localized, or difficult to describe) and thus require a cocktail of prescription medications or invasive procedures, increasing the likelihood of exposure to opioids (Volkow & McLellan, 2016; Rosenblum et al., 2008; Patil et al., 2015). The intensity, duration, and non-specificity of pain may also be a reason why digestive excretory, digestive (upper/oral), and reproductive conditions also show up as moderately strong features in Table 4.3. These diagnoses may either directly result in acute or chronic non-somatic visceral pain (hernia, diverticulitis) (Davis, 2012) or relate to conditions with chronic visceral pain (e.g. female infertility may be secondary to endometriosis or pelvic inflammatory diseases). Opioids are widely utilized for such visceral pain conditions (Gebhart et al., 2000), although often in short duration due to adverse events.

Another expected finding was that individuals with psychological comorbidities (mental health conditions) also have high probability of belonging to the enhanced response group, with individuals participating in psychotherapy having reduced risk of addiction (psychotherapy was among the lowest-scored features and hence not shown in the table). Substantial research has already linked mental illness with opioid misuse (Glanz et al., 2018; Volkow & McLellan, 2016; Rosenblum et al., 2008). Although adjustment reaction (psychology) appears with a lower score in Table 4.3, it encompasses reactions to trauma, episodic emotional disorders, and chronic anxiety that have been shown to be comorbid with many of the chronic diagnoses and pain conditions discussed above (Glanz et al., 2018; Volkow & McLellan, 2016; Rosenblum et al., 2008) and have also been linked with increased opioid dosages (Helmerhorst et al., 2014). Similarly, nondependent abuse of drugs also has a lower score but it is well known that opioid dependence and addiction are associated with polysubstance use and abuse (Soyka, 2015; Pergolizzi et al., 2012). Based on this initial overview, the SME judged the majority of the identified features to be scientifically meaningful with potential clinical utility for future prescribing guidelines. In summary, acute or chronic conditions that put patients at increased risk for initial exposure to opioids, via acute procedures or comorbid prolonged intense pain, both increased a patient's addiction likelihood. Co-morbid mental health disorders, particularly those related to stress or trauma and substance abuse also put individuals at greater risk for future opioid addiction.

4.6 Conclusion

We presented a Heterogeneous Effect Mixture Model (HEMM) for inferring subgroups of individuals that exhibit an enhanced effect caused by treatment. Our work contrasts with existing heterogeneous effect estimation methods as we learn interpretable subgroups using soft assignments while retaining expressiveness in the model. The latter is attributed to the capabilities of neural networks, used here to adjust for confounding. We evaluated the performance of HEMM on a synthetic dataset, the semi-synthetic IHDP dataset, and a large real-world healthcare claims dataset (**Opioid**). We additionally conducted qualitative analysis of the results obtained by HEMM on the **Opioid** dataset and found the derived insights in concordance with existing CDC opioid prescribing guidelines.

Chapter 5

Counterfactual Phenotyping with censored Time-to-Events

5.1 Introduction

Real world studies to estimate the effect of an intervention often involve time-to-event outcomes which are typically followed up only for a fixed period of time. Such studies are commonplace in healthcare and frequently arise when evaluating the effect of a drug or medical intervention on the time to events of interest such as death, re-hospitalization, or a composite physiological outcome. Randomized Control Trials (RCT) aim to eliminate group imbalance through randomizing treatment and control groups. Covariates are evaluated to ensure balanced control and treatment groups so the two groups can be compared without confounding the treatment effect. Hence, in an RCT the population-level event time distributions can be directly compared to obtain estimates of average treatment efficacy.

Indeed, popular population-level metrics for survival and time-to-event prediction involve comparing hazard ratios or summary metrics such as restricted mean time to event by building Proportional Hazard or Kaplan-Meier estimators on the treatment and control arms.

While population-level effect estimation is important as it informs current clinical guidelines and practices, the effect of any intervention is rarely uniform across any population under observation. The advent of precision medicine aims to address these differences in treatment effect by applying individualized treatments designed based on each patient's individuals characteristics. This strategy assumes there are differences in treatment effects that may be explained by varying demographic factors, baseline physiology, or prior medical history. The crux of precision medicine thus entails careful phenotyping of individuals that may receive augmented benefit from a treatment versus those who would not benefit (or worse, suffer adverse effects). These individual patient factors can then be accounted for when framing clinical guidelines based on randomized trials.

Consider the following scenario from an established clinical practice. Landmark studies such as the ALLHAT clinical trial (Davis et al., 1996; Cushman et al., 2002) have established that thiazide diuretic treatment (Chlorthalidone) are not inferior to angiotensin-converting enzyme (ACE) inhibitors (Lisinopril) or calcium channel blockers (Amlodipine) for reducing cardiovascular risk in hypertensive patients. However, there are certain sub-populations, such as those with baseline chronic kidney disease, who may benefit from

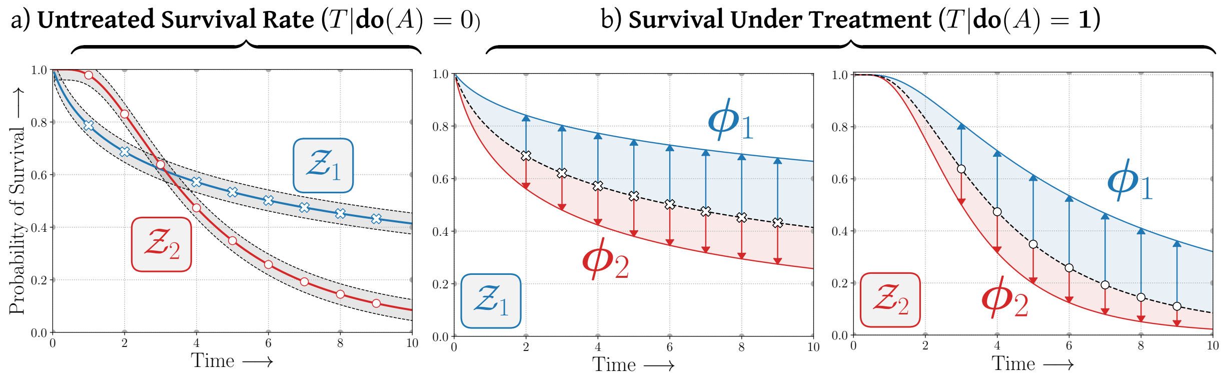


Figure 5.1: COUNTERFACTUAL PHENOTYPING: a) In the untreated population $(T|\mathbf{do}(A)=0)$ there are two subgroups (✖) and (○) that demonstrate differential baseline survival rates represented as the latent group, $\mathcal{Z}$. The intersecting survival curves between the latent groups in $\mathcal{Z}$ suggest that the observed untreated survival rates do not obey the Proportional Hazards assumption. b) Under intervention, we observe that treatment benefit is independent of base survival group, $\mathcal{Z}$. Individuals in ϕ_1 benefit ↑ while individuals in ϕ_2 are harmed ↓. We propose to recover such *counterfactual phenotypes*, ϕ that demonstrate heterogeneous effects to the intervention.

the renal-protective effects of an ACE inhibitor. Additionally, ACE inhibitors are not indicated as an initial treatment for hypertension in black patients, and either a thiazide diuretic or calcium channel blocker is recommended for this group (Whelton et al., 2018).

Such scenarios demonstrate that certain risk groups or phenotypes may not benefit uniformly from a given treatment. It is therefore of immense clinical interest to recover such groups or cohorts of patients to help guide more precise interventions, which leads to personalized medicine and improved patient safety and outcomes. In this chapter, we propose a principled approach, **Cox Mixtures with Heterogeneous Effects** to discover subgroups or cohorts of individuals that demonstrate heterogeneous effects to an intervention in the presence of censored outcomes. The proposed method is not sensitive to strong assumptions of proportional hazards, and it can be applied in situations where such strong assumptions do not generalize uniformly across population.

Our specific contributions can be summarized as follows:

✓ We propose a deep latent variable approach to recover subgroups of patients that respond differentially to an intervention, in the presence of censored outcomes.

✓ We present conditions in which the counterfactuals are identifiable using observational data under the proposed model, along with an efficient approach for learning and inference.

✓ We demonstrate the proposed approach applied to multiple large landmark clinical trials that were originally carried out to assess the efficacy of medical interventions to reduce risk of adverse cardiovascular outcomes among hypertensive and diabetic patients, and we discover clinically actionable counterfactual phenotypes.

Cox Mixtures with Heterogeneous Effects has been released as part of the open-source package, `auton-survival` and is available at `autonlab.github.io/auton-survival/models/cmhe`.

5.2 Related Work

Survival Regression, involving estimation of Time-to-Events in the presence of censored outcomes is a classic problem in statistical estimation, which has recently received attention of the machine learning research community. Arguably, the semi-parametric Cox Proportional Hazards model (Cox, 1972) and its extensions involving multi layer perceptrons (Faraggi & Simon, 1995; Katzman et al., 2018) remain popular, even though they are constrained by strong assumptions on the event time distributions.

Recent research in survival analysis has focused on developing flexible estimators that can ease the restrictive assumptions of the classical Cox model. Non-parametric approaches have been introduced involving Random Forests (Ishwaran et al., 2008) and Gaussian Processes (Fernández et al., 2016; Alaa & van der Schaar, 2017b). (Yu et al., 2011) proposed to treat survival as multitask classification over discrete time horizons. Work of Lee et al. (2018) extends that with deep neural networks in the presence of longitudinal data (Lee et al., 2019a).

Other relevant deep learning approaches include adversarial learning (Chapfuwa et al., 2018) and flexible parametric mixture models (Nagpal et al., 2021c,a; Ranganath et al., 2016b). Further research into survival estimation involves modelling of competing risks, easing restrictive proportional hazard assumptions, and ensuring that models are well calibrated (Chapfuwa et al., 2020b; Goldstein et al., 2020; Yadlowsky et al., 2019). The focus on estimating counterfactuals with censored data has been somewhat limited in current machine learning research literature. Chapfuwa et al. (2021); Curth et al. (2021) propose using integrated probability metric penalties (Shalit et al., 2017) to learn overlapping representations of the treated and control populations in the presence of censored outcomes. More traditional statistical literature in counterfactual survival regression includes propensity score (Linden & Yarnold, 2018; Hassanpour & Greiner, 2019b,a) and doubly robust estimators (Zhao et al., 2015).

In this chapter, we focus specifically on the problem of subgroup identification and phenotyping with censored outcomes. Recent research involving phenotyping of censored time to events are restricted to factual or observational phenotypes (Nagpal et al., 2021d; Chapfuwa et al., 2020a; Manduchi et al., 2021). Our work, however, focuses on simultaneous discovery of latent clusters (or phenotypes) that are *counterfactual*, in that they demonstrate heterogeneous effects to an intervention, while learning effective predictive models capable of capturing the uncovered heterogeneity of response functions.

Figure 5.1 illustrates this '*Counterfactual Phenotyping*' problem. Amongst the untreated population, the latent groups $\boldsymbol{Z}$ mediate the base survival rates. However when an intervention, $\mathbf{do}(A) = \mathbf{1}$ is performed, the treatment effect is mediated by another latent group ϕ, independent of the base survival rate. Identification of such counterfactual phenotypes is of immense utility from the standpoint of clinical decision making as it can be used to administer an optimal treatment strategy to populations that are most likely to benefit.

Machine learning techniques for recovery of subgroups with heterogeneous treatment effects has restricted focus to problems with outcomes that are either continuous or binary in nature. (Foster et al., 2011; Vittinghoff et al., 2010) propose using non-parametric estimators (Decision Trees or Random Forests) to directly regress the difference of the outcomes of the treated and control groups in a framework often called "Virtual Twins" (VT). Wang & Rudin (2022) propose a sampling based approach to recover sparse rule-sets identifying subgroups with enhanced effects. More recently, Nagpal et al. (2020) introduced a deep latent variable approach, Heterogeneous Effect Mixture Model. While close in spirit to our contributions, this approach 1) does not decouple effects of baseline physiology on survival from the treatment effect and 2) is

5.3 Proposed Methodology

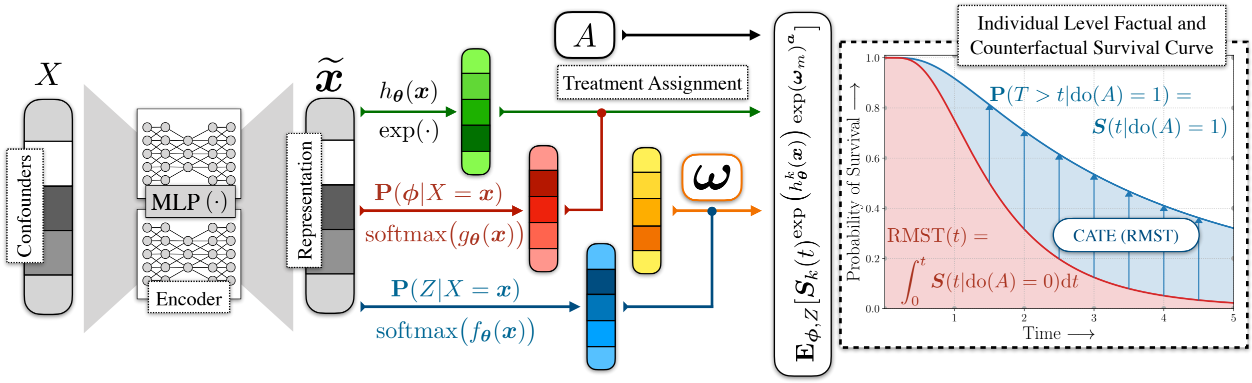

X
Confounders
MLP $(\cdot)$
Encoder
$\widetilde{x}$
Representation
$h_{\theta}(x)$
$\exp(\cdot)$
$\mathbf{P}(\phi|X=x)$
$\text{softmax}(g_{\theta}(x))$
$\mathbf{P}(Z|X=x)$
$\text{softmax}(f_{\theta}(x))$
A
Treatment Assignment
ω
$\mathbf{E}_{\phi,Z}\left[S_k(t)^{\exp\left(h^k_{\theta}(x)\right)\exp(\omega_m)^a}\right]$
Individual Level Factual and Counterfactual Survival Curve
$\mathbf{P}(T>t|\text{do}(A)=1)=$
$S(t|\text{do}(A)=1)$
$\text{RMST}(t)=$
$\int_0^t S(t|\text{do}(A)=0)\text{d}t$
CATE (RMST)
Probability of Survival ⟶
Time ⟶
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5

Figure 5.2: Schematic description of the proposed CMHE. The set of features (confounders) $\boldsymbol{x}$ are passed through an encoder to obtain deep non-linear representations. These representations then describe the latent phenogroups $\mathbf{P}(Z|X=\boldsymbol{x})$ and $\mathbf{P}(\phi|X=\boldsymbol{x})$ that determine the base survival rate and the treatment effect respectively. Finally, the individual level hazard (survival) curve under an intervention $A=\boldsymbol{a}$ is described by marginalizing over Z and ϕ as $S(t|X=x,A=a)=\mathbf{E}_{(Z,\phi)\sim\mathbf{P}(\cdot|X)}\big[S(t|A=\boldsymbol{a},X,Z,\phi)\big]$.

incompatible with censored time-to-events and hence, cannot be applied to many real-world studies in a straightforward manner.

5.3.1 Notation and Setting

We consider a dataset of right censored observations in the form of four tuples, $\mathcal{D}=\{(\boldsymbol{x}_i,\delta_i,t_i,a_i)\}_{i=1}^{N}$, where $t_i\in\mathbb{R}^+$ is either the time to event or censoring as indicated by $\delta_i\in\{0,1\}$, $a_i\in\{0,1\}$ is the indicator of treatment assignment, and $\boldsymbol{x}_i$ are individual covariates that confound the treatment assignment and the outcome.

5.3.2 Cox PH Model and Cox Mixtures

The Cox Proportional Hazards model is arguably the most popular approach to model censored survival outcomes. The Cox model involves assuming that the conditional hazard of an individual is

$$\boldsymbol{\lambda}(t|x)=\boldsymbol{\lambda}_0(t)\exp\big(h_{\boldsymbol{\theta}}(x)\big),\qquad(5.1)$$

where h is typically a linear function. Thus, under the Cox model, the full likelihood in terms

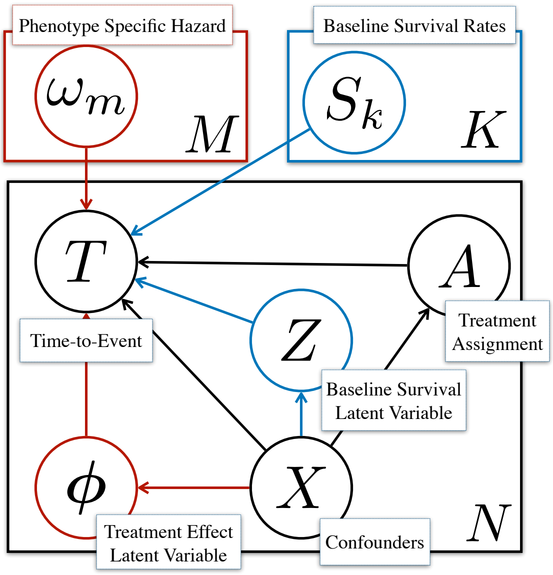


Figure 5.3: The proposed model in Plate Notation. X confounds treatment assignment A and outcome T (Model parameters and censoring distribution have been abstracted out).

of the cumulative hazard[1] $\boldsymbol{\Lambda}_0$ and parameters $\boldsymbol{\theta}$ is as follows:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_0) = \prod_{i=1}^{|\mathcal{D}|} \Big(\boldsymbol{\lambda}_0(t_i) \exp\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)\Big)^{\delta_i} \boldsymbol{S}_0(t_i)^{\exp\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)} \tag{5.2}$$

However, the assumption that the hazard rates remain proportional over time is a strong one and is often violated in practice. One example of such a violation is the presence of intersecting survival curves. Nagpal et al. (2021d) propose to relax the PH assumption by describing the data as belonging to a mixture of fixed size K, such that the PH assumptions hold only conditioned on the latent component assignment, Z. The assignment function of an individual to a latent group can then be learned jointly along with the component-specific hazard ratios. Under this model the hazard rate of an individual belonging to latent $Z = \boldsymbol{k}$ is:

$$\begin{aligned} \boldsymbol{\lambda}_{\boldsymbol{\theta}}(t|X = \boldsymbol{x}, Z = \boldsymbol{k}) &= \boldsymbol{\lambda}_k(t) \exp\left(h_{\boldsymbol{\theta}}^k(\boldsymbol{x})\right) \\ \text{where, } \mathbf{P}(Z = \boldsymbol{k}|X = \boldsymbol{x}) &\propto \exp\left(f_{\boldsymbol{\theta}}^k(\boldsymbol{x})\right) \end{aligned} \tag{5.3}$$

5.3.3 Cox Mixtures with Heterogeneous Effects

In this chapter we show how to model counterfactual outcomes and individual level treatment effects using the proposed approach: **Cox Mixtures with Heterogeneous Effects** (CMHE). To accomplish that, we further extend the model in Equation 5.3 by introducing another latent variable $\boldsymbol{\phi}$ that determines the treatment effect for an individual with confounders, $\boldsymbol{x}$. Thus under this new model,

$$\begin{gathered} \underbrace{\boldsymbol{\lambda}(t|X = \boldsymbol{x}, Z = k, \boldsymbol{\phi} = m, A = \boldsymbol{a})}_{\text{Conditional Hazard Rate}} = \overbrace{\boldsymbol{\lambda}_k(t)}^{\text{Base Survival Rate}} \underbrace{\exp\left(h_{\boldsymbol{\theta}}^k(\boldsymbol{x})\right)}_{\text{Effect of Confounders}} \overbrace{\exp(\omega_m)^{\boldsymbol{a}}}^{\text{Treatment Effect}} \\ \mathbf{P}(Z = k|X = \boldsymbol{x}_i) \propto \exp\left(f_{\boldsymbol{\theta}}(\boldsymbol{x})\right), \mathbf{P}(\boldsymbol{\phi} = m|X = \boldsymbol{x}) \propto \exp\left(g_{\boldsymbol{\theta}}(\boldsymbol{x})\right). \end{gathered} \tag{5.4}$$

Figure 5.3 presents CMHE in plate notation.

Assumption 1 (Independent Censoring) *The distribution of the time-to-event T and the censoring times C are independent conditional on the covariates, X and the treatment assignment A.*

Assumption 2 (Conditional PH) *Conditional on the latent group $\mathcal{Z}$, individual time-to-event distributions obey proportional hazards.*

Assumption 3 (Latent Independence) *The baseline survival rate group, Z and the treatment effect group, ϕ are independent given the confounders X, i.e., $Z \perp \phi \mid X$.*

Assumption 4 (Ignorability) *The treatment assignment, A is independent of the potential time-to-event outcomes, $T(A)$ conditioned on the set of confounders X, ie. $A \perp T(A) \mid X$.*

[1] The cumulative hazard is defined as $\boldsymbol{\Lambda}_0(t) = \int_0^t \boldsymbol{\lambda}_0(t)$. It can equivalently be described in terms of the base survival rate as $\boldsymbol{\Lambda}_0(t) = -\ln \boldsymbol{S}_0(t)$.

Note that Assumption 1 is stronger compared to standard regression as it requires conditioning on both the covariates and the treatment assignment. Assumption 4 essentially states that the treatment assignment A is completely characterized by the available confounders X. In other words, there are no exogenous factors (unobserved confounders) that may affect treatment assignments.

Remark 1 (Identifiability) *Under Assumptions 1 - 3, the counterfactual Time-to-Event distribution is identifiable with observables as follows,* $\mathbf{P}(T|\mathbf{do}(A) = \boldsymbol{a}, X) = \mathbf{E}_{(Z,\boldsymbol{\phi})\sim\mathbf{P}(\cdot|X)}\big[\mathbf{P}(T|A = \boldsymbol{a}, X, Z, \boldsymbol{\phi})\big]$.

Remark 1 allows us to make phenogroup level counterfactual inference in terms of observables. It stems directly from the standard application of Pearl's **do**-calculus and Assumptions 3 and 4 (proof available in Appendix D.1). As a consequence of the Assumptions 1 - 3 under this new model, the full likelihood is:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_k) = \prod_{i=1}^{|\mathcal{D}|} \underbrace{\int_{k\in Z} \overbrace{\int_{m\in\boldsymbol{\phi}} \Big(\boldsymbol{\lambda}_k^m(t_i|\boldsymbol{x}_i, a_i)^{\delta_i} \boldsymbol{S}_k^m(t_i|\boldsymbol{x}_i, a_i)\Big) \mathrm{d}\mathbf{P}(\boldsymbol{\phi}|\boldsymbol{x}_i)}^{\text{Marginzalization over Treatment Effect Phenotypes, } \boldsymbol{\phi}} \mathrm{d}\mathbf{P}(Z|\boldsymbol{x}_i)}_{\text{Marginzalization over Baseline Survival Clusters, } Z}$$

$$\text{where, } \boldsymbol{\lambda}_k^m(t|\boldsymbol{x}, a) = \boldsymbol{\lambda}_k(t) \exp\big(h_{\boldsymbol{\theta}}^k(\boldsymbol{x})\big) \exp\big(\boldsymbol{\omega}_m\big)^{\boldsymbol{a}},$$

$$\boldsymbol{S}_k^m(t|\boldsymbol{x}, a) = \boldsymbol{S}_k^m(t)^{\exp\big(h_{\boldsymbol{\theta}}^k(\boldsymbol{x})\big) \exp\big(\boldsymbol{\omega}_m\big)^{\boldsymbol{a}}}. \tag{5.5}$$

5.3.4 Architecture

A non-linear representation $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{d^{‘}}$ of the input $\boldsymbol{x}$ is obtained using a multilayer perceptron with parameters $\boldsymbol{\theta}$. This representation reflects the baseline survival specific log-hazard ratios, for the each of the k base survival phenotypes through the function $h : \mathbb{R}^{d^{‘}} \to \mathbb{R}^k$ and the non-normalized probability of belonging to one of the $\boldsymbol{k}$ base survival clusters, ie. $\mathbf{P}(Z|X = \boldsymbol{x}) \propto f(\boldsymbol{x})$ as $f : \mathbb{R}^{d^{‘}} \to \mathbb{R}^k$. Further, the non-normalized probability of assignment to counterfactual phenotype is defined as $g : \mathbb{R}^{d^{‘}} \to \mathbb{R}^m$, i.e., $\mathbf{P}(\boldsymbol{\phi}|X = \boldsymbol{x}) \propto g(\boldsymbol{x})$. Figure 5.2 provides a schematic diagram of the proposed architecture for CMHE.

5.3.5 Learning

Parameter inference in semi-parametric latent variable models such as CMHE is hard as estimation of the baseline hazard rates $(\{\boldsymbol{\Lambda}_k\}_{k=1}^K)$ is carried out non-parametrically. Naive application of the Expectation Maximization (EM) algorithm requires inference over all possible $(M \times K)^{|\mathcal{D}|}$ latent assignments for the entire dataset which is intractable. As described in Nagpal et al. (2021d) we propose a stochastic EM algorithm involving Monte-Carlo sampling to make inference tractable. The M-step of our proposed EM algorithm involves the following $Q(\cdot)$ function,

$$\widehat{Q}(\boldsymbol{\theta}) = \sum_k \ln \mathcal{PL}_k(\mathcal{D}_b, \boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\theta}) + \sum_k \sum_i^{|\mathcal{D}_b|} \gamma_i^k \ln \mathrm{softmax}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\big) + \sum_m \sum_i^{|\mathcal{D}_b|} \zeta_i^m \ln \mathrm{softmax}\big(g_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\big), \tag{5.6}$$

which we arrive at using the assumption that the Proportional Hazards hold within each baseline survival rate group $Z = \boldsymbol{k}$. Here γ and $\boldsymbol{\zeta}$ are the respective soft posterior counts of Z and $\boldsymbol{\phi}$. $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ corresponds

Algorithm 2: Learning for CMHE with Stochastic EM

```
Input  :Training set, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, a_i, \delta_i)_{i=1}^{N}\}$; batches, $B$;

while <not converged> do
    for $b \in \{1, 2, ..., B\}$ do
        $\mathcal{D}_b \sim \mathcal{D}$                                    ▷ Draw a minibatch from the full dataset.
        ........................................................................ E-STEP
        for $i \in |\mathcal{D}_b|$ do
            $\gamma_i \sim \mathbf{P}(Z = k|\boldsymbol{x}_i)$;    $\boldsymbol{\zeta}_i \sim \mathbf{P}(\boldsymbol{\phi} = m|\boldsymbol{x}_i)$
                                                                        ▷ Sample soft posteriors.
            $\boldsymbol{\psi}_i \sim \text{Categorical}(\boldsymbol{\gamma}_i)$;    $\boldsymbol{\xi}_i \sim \text{Categorical}(\boldsymbol{\zeta}_i)$
                                                                        ▷ Draw hard posteriors.
        end
        ........................................................................ M-STEP
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \cdot \nabla_\theta \widehat{Q}(\boldsymbol{\theta}; \mathcal{D}_b)$
                                                        ▷ Update $\boldsymbol{\theta}$ with gradient of $\widehat{Q}$.
        for $k \in \{1, 2, ..., K\}$ do
            $\widehat{\boldsymbol{\Lambda}}_k(t) \leftarrow \sum_{i:t_i<t} \left( \sum_{j \in \mathcal{R}(t_i)} \exp\left(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}_j) + \boldsymbol{a} \cdot \omega_{\zeta_j}\right) \right)^{-1}$
                                                                ▷ Breslow (1972b)'s estimator.
            $\widehat{\boldsymbol{S}}_k(t) \leftarrow \exp\left(-\widehat{\boldsymbol{\Lambda}}_k(t)\right)$
            $\widetilde{\boldsymbol{S}}_k \leftarrow \arg\min_s \sum_{i=1}^{n} \left(\widehat{\boldsymbol{S}}_k(t) - s(t)\right)^2 + \lambda \int_{t_{\min}}^{t_{\max}} s''(t)$
                                                                ▷ Cubic-spline interpolation.
        end
        ........................................................................
    end
end

Return :learnt parameters, $\boldsymbol{\theta}$; baseline survival splines $\{\widetilde{\boldsymbol{S}}_k\}_{i=1}^{K}$
```

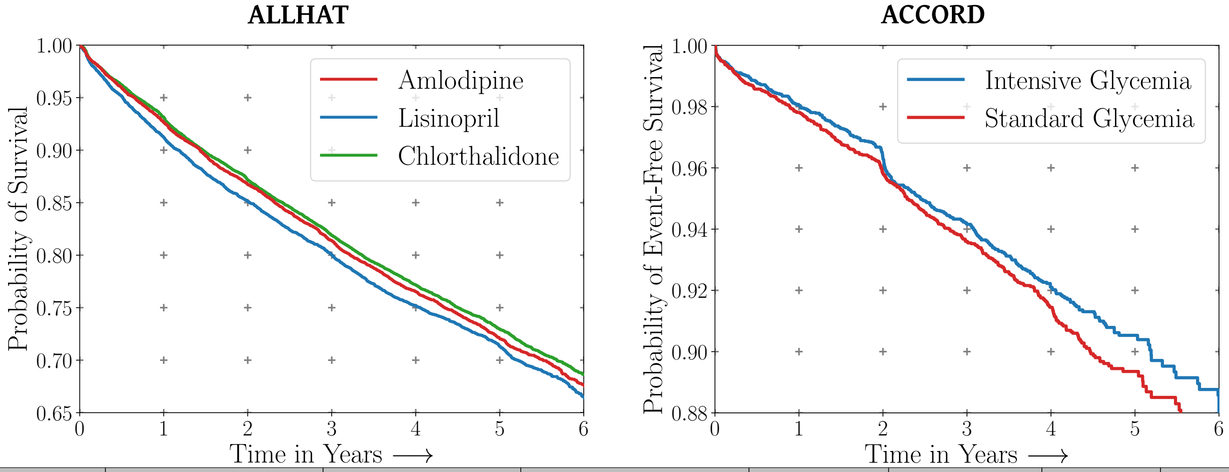
to the hard posterior counts sampled as $\boldsymbol{\psi} \sim \text{Categorical}(\boldsymbol{\gamma})$, $\boldsymbol{\xi} \sim \text{Categorical}(\boldsymbol{\zeta})$ and $\mathcal{PL}_k(\cdot)$ is the *partial likelihood*. Algorithm 2 describes the stochastic EM learning algorithm for CMHE. The proposed stochastic EM makes inference is tractable with a complexity of $(|\mathcal{D}| \times M \times K)$. A complete discussion of our formulation along with the functional form of the $Q(\cdot)$ and $\mathcal{PL}_k(\cdot)$ functions are deferred to Appendix D.2

5.3.6 Inference

Following Remark 1, the estimated risk of an individual with confounders $\boldsymbol{x}$ under an intervention $\mathbf{do}(A) = a$ at a time t is

$$\begin{aligned}\widehat{\mathbf{P}}(T > t|X = \boldsymbol{x}, \mathbf{do}(A) = \boldsymbol{a}) &= \mathbf{E}_{(Z,\boldsymbol{\phi}) \sim \widehat{\mathbf{P}}(\cdot|X)}[\widehat{\mathbf{P}}(T|X = \boldsymbol{x}, A = \boldsymbol{a}, Z, \boldsymbol{\phi})] \\ &= \sum_k \sum_m \text{softmax}_k\big(f_{\boldsymbol{\theta}}(\boldsymbol{x})\big) \cdot \text{softmax}_m\big(g_{\boldsymbol{\theta}}(\boldsymbol{x})\big) \cdot \widetilde{\boldsymbol{S}}_k^m(t)^{\exp\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{a} \cdot \omega_m\right)}.\end{aligned}$$

5.4 Experiments



Dataset	Outcome	Treatment	Control	Hazard Ratio	$\text{ATE}_{\text{RMST}}(t)$	Event Rate	N
ALLHAT-A	Cardiovascular Death	Chlorothalidone	Amlodipine/Lisinopril	0.94 ± 0.04	21.23 ± 12.34	26.84%	$33,357$
ALLHAT-B	Cardiovascular Death	Amlodipine	Lisinopril	0.95 ± 0.05	23.87 ± 15.56	27.47%	$18,102$
ACCORD	Primary End Point	Intensive Glycemia	Standard Glycemia	0.89 ± 0.12	8.28 ± 14.42	8.50%	$9,822$

Figure 5.4: Kaplan-Meier estimates and summary statstics of the datasets used in the chapter. (For ACCORD, the primary endpoint was the time to first instance of Non-Fatal Myocardial Infarction, Stroke, or Death.)

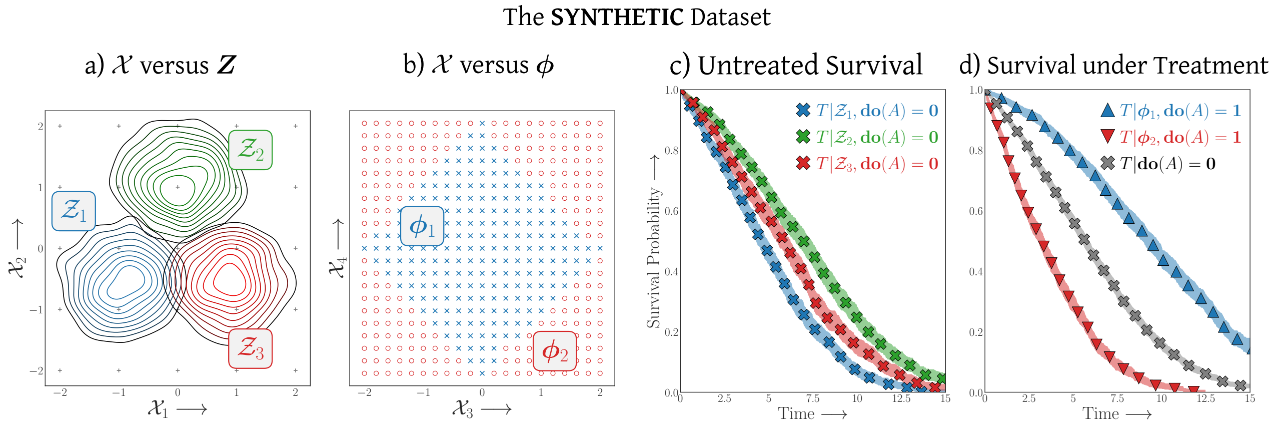
In our experiments, we consider data from the landmark ALLHAT and ACCORD clinical trials originally conducted to determine the optimal treatment for reducing risk from cardiovascular disease.

5.4.1 Datasets

▶ **Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack (allhat)**: The ALLHAT clinical trial (Furberg et al., 2002) was constituted to establish the appropriate intervention between chlorthalidone (a diuretic), amlodipine (calcium channel blocker) and lisinopril (angiotensin converting enzyme (ACE) inhibitor) for hypertensive patients to reduce adverse cardiovascular events. Patients were enrolled over a four year period with a mean follow up time of 4.9 years. The complete study involved 33,357 participants older than 55 years of age, all with hypertension. 15,255 ($\approx 50\%$) of participants were assigned to the Chlorthalidone arm, while 9,048 ($\approx 25\%$) assigned to amlodipine and 9,054 ($\approx 25\%$) to lisinopril. For the purposes of this chapter, we consider two separate experiments on the ALLHAT dataset. **ALLHAT-A**: We consider all patients in the trial assigned to chlorthalidone as 'Treated' and the rest as 'Controls'. **ALLHAT-B**: We only consider the 18,102 assigned to lisinopril or amlodipine. Amlodipine is considered the 'Treatment'. For both ALLHAT A and B we consider the time to death from cardiovascular events as the outcome of interest.

▶ **Action to Control CVD Risk in Diabetes (accord)** : The ACCORD study (Ismail-Beigi et al., 2010) involved 10,251 patients over the age of 40 with Type-2 Diabetes Mellitus with a median follow up time of 3.7 years. 5,128 ($\approx 50\%$) patients were randomized to intensive glycemic control. (`HbA1C`[2]: $< 6\%$) and the rest to standard glycemic control (`HbA1C`: $7 - 7.9\%$). In this chapter we consider the patients assigned to the standard

[2] Glycated Hemoglobin (A1c)



The **SYNTHETIC** Dataset

a) $\mathcal{X}$ versus $\boldsymbol{Z}$

b) $\mathcal{X}$ versus ϕ

c) Untreated Survival

d) Survival under Treatment

Figure 5.5: a) The distribution of the latent baseline survival groups, Z in the space of features $\mathcal{X}$. b) The distribution of the latent treatment effect phenogroups ϕ in the space of features $\mathcal{X}$. c) Kaplan-Meier Estimators conditioned on the latent baseline groups in the untreated population ie. $\widehat{\mathbf{P}}(T > t|Z, \text{do}(A) = 0)$. d) Kaplan-Meier estimators of the treated population conditioned on the latent treated effect groups ie. $\widehat{\mathbf{P}}(T > t|\phi, \text{do}(A) = 1)$. Notice that individuals in ϕ_1, benefit ▲ from the intervention but ϕ_2 are harmed ▼.

glycemic control arm as 'Treated' and compare the performance in terms of time to the primary endpoint of the study, a composite endpoint including death, myocardial infarction or stroke. The **ACCORD** trial is of particular interest, as the results from it demonstrate that although intensive hyperglycemia treatment strategy for patients with diabetes reduces rate of adverse cardiovascular events, however it may increase the patient's risk for overall mortality. This is most likely due to adverse effects of the treatment itself.

Figure 5.4 presents the Kaplan-Meier estimates of the overall population level event-free survival for the two studies along with the summary statistics including Hazard Ratio and Average Treatment Effect in Restricted Mean Survival Time. For both the **ALLHAT** and the **ACCORD** trials we consider a set of confounding features measured during the patients baseline visit at the time of randomization. This includes basic demographic information including sex and race, age at entry into the study, previous history of adverse cardiovascular events, etc. (Full list of confounders in Appendix D.7).

▶ **SYNTHETIC**: We further benchmark the proposed model on a synthetic dataset presented in Figure 5.5. This dataset is designed such that the latent treatment effect phenotype ϕ is not linearly separable in $\boldsymbol{x}$. The time-to-event T conditioned on $\boldsymbol{x}$, latent Z and latent effect group ϕ are generated from a Gompertz distribution. (Complete details of this design are deferred to Appendix D.5).

5.4.2 Counterfactual Phenotyping

We evaluate the performance of CMHE in its ability to identify phenotypes of varying sizes that have a more pronounced treatment effect when compared to the baselines. For all the datasets we keep 75% of the dataset as training and the rest as testing to evaluate each model's performance. For CMHE as well as the baselines we identify the phenogroup with the most enhanced and the most diminished treatment effect on the training set and report the estimated Conditional Average Treatment Effect in Restricted Mean Survival Time on the held out sample.

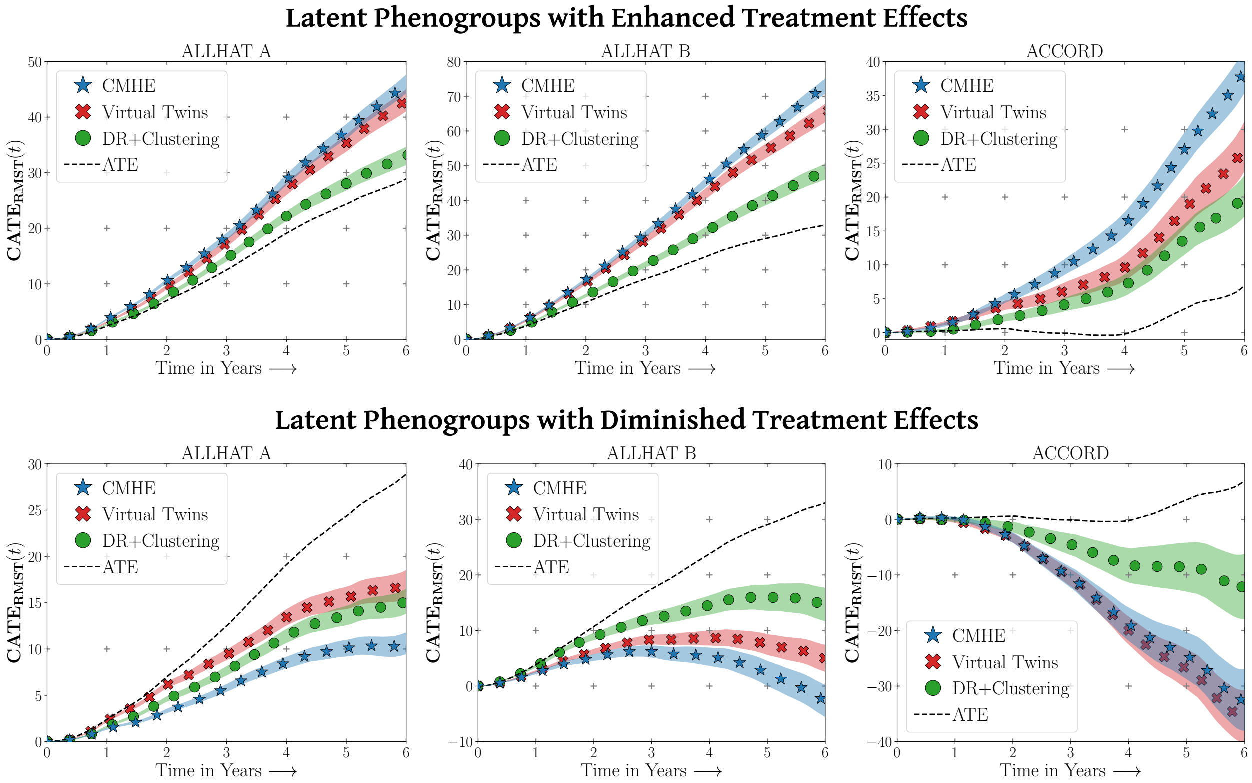


Figure 5.6: **Conditional Average Treatment Effect in Restricted Mean Survival Time** (95% Confidence Bands) over Time for counterfactual phenotypes recovered by CMHE and Baselines in comparison to the Average Treatment Effect (ATE). For each of the datasets we identify phenogroups with enhaced (diminished) treatment effect based on RMST on the training split and report the corresponding RMST on the out-of-sample testing split. Phenogroups of different sizes are generated by varying the threshold, $\alpha > \widehat{\mathbf{P}}(\phi|X = \boldsymbol{x})$ at which an individual $\boldsymbol{x}$ is assigned to the latent phenogroup, $\mathcal{X}$. (Here we report phenogroups that are of the size 15% of the total study population.) Notice how CMHE consistently recovers phenogroups with larger $\textbf{CATE}_{\text{RMST}}(t)$ (Tabulated results in Table D.6).

Definition 1 (RMST) *The Restricted Mean Survival Time at time* $\boldsymbol{t}$ *under an intervention* $\boldsymbol{a}$ *for individual with confounders* $\boldsymbol{x}$ *is the expected conditional Time-to-Event,* $\mathbf{E}[\min\{T, t\}|\text{do}(A) = \boldsymbol{a}, X = \boldsymbol{x}]$.

In the case of time-to-event outcomes, Restricted Mean Survival Time (**RMST**) is the truncated area under the survival curve,

$$\mathbf{E}[\min\{T, t\}|\mathbf{do}(A) = \boldsymbol{a}, X = \boldsymbol{x}] = \int_0^t \boldsymbol{S}(t|\mathbf{do}(A) = \boldsymbol{a}, X = \mathbf{x})\mathrm{d}t.$$

Following from Chen & Tsiatis (2001); Royston & Parmar (2013) we define the Conditional Average Treatment Effect in terms of the difference in Restricted Mean Survival Time under treatment and control.

Definition 2 ($\textbf{CATE}_{\textbf{RMST}}$) *The Conditional Average Treatment Effect at time* t *is expected difference between the treated and control Restricted Mean Survival Time conditioned on the phenogroup,* $\mathcal{X}$.

$$\textbf{CATE}_{\text{RMST}}(\mathcal{X}; t) = \underset{\boldsymbol{x} \in \mathcal{X}}{\mathbf{E}} \big[\mathbf{E}[\min\{T, t\}|\mathbf{do}(A) = 1, X = \boldsymbol{x}] - \mathbf{E}[\min\{T, t\}|\mathbf{do}(A) = 0, X = \boldsymbol{x}]\big].$$

This can now be estimated as:

$$\widehat{\text{CATE}}_{\text{RMST}}(\mathcal{X};t) = \frac{1}{n}\sum_{x\in\mathcal{X}}\left[\int_0^t \widehat{\boldsymbol{S}}_1(t|\mathbf{x})\mathrm{d}t - \int_0^t \widehat{\boldsymbol{S}}_0(t|\mathbf{x})\mathrm{d}t\right]. \tag{5.7}$$

Here $\mathcal{X}$ is the set of all individuals in the phenotype (we control the size of the phenotype by varying the threshold, $\alpha > \widehat{\mathbf{P}}(\boldsymbol{\phi}|\boldsymbol{x})$ and t is the time horizon at which RMST is computed. $\boldsymbol{S}_1(\cdot)$ and $\boldsymbol{S}_0(\cdot)$ are the survival distributions under treatment and control.

Note that we cannot directly compare the survival curves conditioned on the recovered phenotype as within the phenotype treatment assignment is not random. In order to mitigate this problem, we fit separate Random Survival Forests (RSFs) (Ishwaran et al., 2008) on the treated and control populations in the training set in a 5-fold Cross Validation to minimize the Integrated Brier Score. The fitted estimators are then employed to estimate the individual counterfactual survival curves, $\widehat{\boldsymbol{S}}_1(t|X)$ and $\widehat{\boldsymbol{S}}_0(t|X)$ on the test data for evaluation.

BASELINES: We compare the ability of CMHE against the following baseline strategies for counterfactual phenotyping:

▶ **Dimensionality Reduction + Clustering**: Involves first performing dimensionality reduction of the input confounders, $\boldsymbol{x}$, followed by clustering. For the experiments in the chapter we consider Linear-PCA and Kernel-PCA with an Radial Basis Function Kernel for dimensionality reduction, followed by K-Means and Gaussian Mixture Models (GMMs) for clustering. The number of reduced dimensions for the confounders is tuned from {8,16} and the number of clusters from {2,3}. For the GMMs we enforce the learned covariance matrix of the components to be diagonal.

▶ **Virtual Twins** (Foster et al., 2011; Vittinghoff et al., 2010): Involves first building regression estimators of the outcome conditioned on the confounders separately for the treated and control populations, followed by regressing the difference between counterfactual estimates on the confounders using a simpler model such as a Decision Tree. In the experiments, we estimate the virtual twin counterfactual survival models using a Linear Cox model and Cox model parameterized with a 2 hidden layer Multi-Layer Perceptron (MLP) with (Faraggi & Simon, 1995; Katzman et al., 2018) in a 5 fold cross-validation fashion on the training dataset. For both the Linear and MLP Cox Model we tune the batch size from {128,256} the learning rate {2,3} in $\{\mathbf{10^{-3}}, \mathbf{10^{-4}}\}$. For the MLP, the hidden layer was fixed to have `Tanh` activations with a dimensionality of 50. The models were optimized using `Adam` (Kingma & Ba, 2014). Once the counterfactual models are estimated, the difference in there estimates in terms of **RMST@5-Year** is modelled using a Random Forest with 25 trees whose depth is tuned from {4,5}. The trained Random Forest is then employed to recover the counterfactual phenotypes.

RESULTS: CMHE consistently recovered phenogroups that demonstrated higher CATE as compared to the the Virtual Twins and Dimensionality Reduction + Clustering baselines as in Figure 5.6. In the case of **ALLHAT A** and **B**, CMHE recovered a sub-population of 15% of the test data that had a diminished **CATE**$_{\text{RMST}}$**@5-Years** of $\mathbf{10.09 \pm 0.86}$ and $\mathbf{2.28 \pm 2.18}$ Days respectively as compared to the population ATE of $\mathbf{24.30}$ and $\mathbf{28.97}$ Days (Table D.6).

In the ACCORD trial, CMHE recovered a phenogroup involving 15% of the test set that had a much more dramatic treatment effect (**CATE$_{\text{RMST}}$@5-Years** of $\mathbf{27.02 \pm 2.43}$ Days, as compared to the population average treatment effect of $\mathbf{8.28 \pm 14.42}$ Days). These results demonstrate that CMHE can discover compelling phenogroups with substantially different treatment responses. In the case of synthetic data, we directly compare the performance of in-recovery of ϕ by treating it as a binary classification problem. CMHE has higher discriminative performance (Area under Receiver Operating Characteristic: $\mathbf{0.924 \pm 0.01}$) versus the clustering ($\mathbf{0.505 \pm 0.02}$) and Virtual Twins ($\mathbf{0.900 \pm 0.001}$) baselines, as shown in Figure 5.7. Notice that Clustering is not better than random, which is expected due to the nonlinear nature of ϕ as in Figure 5.5.

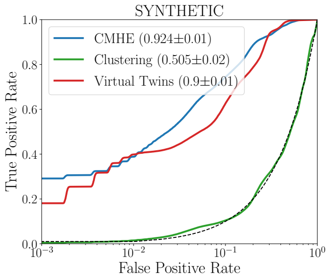


Figure 5.7: ROC curves for recovery of ϕ by CMHE and proposed models on the synthetic dataset.

Rule explanation of CMHE Phenotype				Size	5-Year RMST	
					CATE	ATE
ALLHAT-A						
Height<68.51	Dia.BP<88.5	GFR<65.7	Aspirin=+ve	**5.05%**	**42.87**	**24.30**
Dia.BP<82.50	Prior[T-inv.]=+ve	LLT=-ve	Race[Black]=+ve	**13.00%**	**20.32**	**24.30**
ALLHAT-B						
Sys.BP>144.5	Dia.BP>87.5	GFR>55.7	Race[Black]=+ve	**8.71%**	**53.20**	**28.97**
Dia.BP<=78.5	GFR<=86.1	Race[Black]=+ve	Aspirin=-ve	**6.86%**	**7.56**	**28.97**
ACCORD						
GFR<79.9	UACR<42.9	FPG>194.6	Prior[CVD]=-ve	**5.58%**	**28.94**	**4.79**
potassium<=4.8	GFR>90.2	Prior[MI]=-ve	Prior[CVD]=+ve	**7.56%**	**-23.26**	**4.79**

Table 5.1: We employ a tree ensemble based rule learning algorithm (Details in Appendix D.4) to explain the phenotypes extracted by CMHE. Enhanced (Diminished) Treatment Effect Rules in Blue (Red). We report explanations that maximize F_1 score on the heldout dataset. Extensive discussion on physiologic interpretation of the extracted explanations are in Section 5.4.2.

Interpretation: We employed a tree ensemble based rule learning (Friedman & Popescu, 2008) to interpret the phenotypes discovered with CMHE in terms of parsimonious conjunctions. The extracted rules are presented in Table 5.1. Additional implementation details are provided in Appendix D.4. The extracted rules were subjected to qualitative evaluation by an expert clinician.

► **ALLHAT-A**: Conditions associated with increased protective effect from chlorthalidone treatment include patients who are older, shorter, have decreased renal function evidenced from lower glomerular filtration rate (GFR), and exhibit less baseline cardiac disease (absence of coronary heart disease). Additionally, several conditions favor patients with lower baseline systolic or diastolic blood pressure. In clinical practice, the decision for first-line antihypertensive therapeutic agent continues to be debated (Whelton et al., 2018). However, ACE inhibitors are commonly used for treatment of patients with hypertension and cardiac disease (such as coronary heart disease). The conditions stated above are consistent with this, indicating patients with absence of coronary heart disease are more likely to benefit from diuretic therapy. Additionally, patients treated with calcium channel blockers may be prone to edema and fluid retention. Fluid retention is a risk for patients with lower baseline GFR, indicative of kidney dysfunction, thus diuretic therapy may be preferred in these patients. On the other hand, ACE inhibitors are believed to slow the progression of mild kidney disease, making them a reasonable treatment for these patients.

Metric	Concordance Index			IBS
Time Horizon	1 Year	3 Year	5 Year	
ALLHAT-A				
Cox-Linear	0.6822	0.6716	0.6688	0.1362
Cox-MLP	0.6797	0.6693	0.6677	0.1361
CMHE-Linear	0.6830	0.6722	0.6692	0.1360
CMHE-MLP	**0.6832**	**0.6734**	**0.6705**	**0.1357**
ACCORD				
Cox-Linear	0.6615	0.6737	0.6713	0.0591
Cox-MLP	0.6564	0.6723	0.6706	0.0590
CMHE-Linear	0.6606	0.6720	0.6697	0.0591
CMHE-MLP	**0.6755**	**0.6881**	**0.6850**	**0.0587**

Metric	Concordance Index			IBS
Time Horizon	1 Year	3 Year	5 Year	
ALLHAT-B				
Cox-Linear	0.6741	0.6641	0.6629	0.1402
Cox-MLP	0.6717	0.6605	0.6602	0.1404
CMHE-Linear	0.6753	0.6651	0.6638	0.1401
CMHE-MLP	**0.6760**	**0.6655**	**0.6640**	**0.1399**
SYNTHETIC				
Cox-Linear	0.6224	0.6205	0.6158	0.1723
Cox-MLP	0.6623	0.6727	0.6740	0.1619
CMHE-Linear	0.6356	0.6365	0.6337	0.1698
CMHE-MLP	**0.6676**	**0.6758**	**0.6786**	**0.1604**

Table 5.2: Time Dependent Concordance Index and Integrated Brier Score (**IBS**) for CMHE (Linear and MLP) and baseline Cox models.

► **ALLHAT-B**: Patients treated with amlodipine who achieved the greatest benefit tended to have higher baseline blood pressure (systolic and diastolic), weigh less (lower weight and body mass index), have higher baseline kidney function (higher GFR), and were more likely to be of Black, non-Hispanic ethnicity. These characteristics are useful and actionable clinical parameters to guide clinicians in choosing a first antihypertensive agent. In fact, the American College of Cardiology and American Heart Association guidelines for hypertension management suggest that Black patients should be prescribed either a thiazide diuretic or calcium channel blocker for first-line agent (Whelton et al., 2018). In this case, the rules discovered with CMHE demonstrate that the calcium channel blocker amlodipine is more effective than lisinopril for Black, non-Hispanic patients. Conditions corresponding with diminished treatment effect are the inverse of those described for the group receiving the most benefit. ACE inhibitors are indicated treatment to slow the progression of renal insufficiency for those with mild renal dysfunction and hypertension, and the CMHE conditions support this, indicating amlodipine is better in patients with higher baseline kidney function, whereas lisinopril may be better for patients with kidney disease.

► **ACCORD**: CMHE revealed multiple conditions describing the phenogroup with decreased treatment effect from intense hyperglycemic control. In this group, criteria tended to favor patients with less severe renal disease (higher baseline GFR, lower baseline creatinine, lower baseline potassium) or decreased evidence of cardiovascular disease (lower diastolic blood pressure, less bradycardia), though this was not true in all cases. In contrast, the phenogroup with increased treatment effect from intense hyperglycemic control has a worse baseline kidney function (GFR < 79.9 mL/min, urine creatinine < 42.9 mg/dL), higher baseline fasted glucose levels (fasting blood glucose > 195), and lacked a documented history of cerebrovascular disease. These results suggest that patients at with decreased renal function and poor baseline glucose control, which is commonly seen in advanced diabetics, stand to gain the most from an intensive treatment regimen. Patients with a history of cerebrovascular disease (e.g., stroke), on the other hand, may be at greater risk from hypoglycemic complications related to intensive glucose control. This is consistent with clinical intuition that aggressive treatment is required for the most severe forms of a disease, whereas early stage disease may not derive as much benefit, increasing risk for net harm due to unintended side effects.

5.4.3 Factual Regression

For completeness, we further evaluate the performance of the proposed approach in estimating factual risk over multiple time horizons using the standard survival analysis metrics, including:
Brier Score $\big(\mathrm{BS}(t)\big)$: Defined as the Mean Squared Error (MSE) around the probabilistic prediction at a certain time horizon.

$$\mathrm{BS}(t) = \underset{x\sim\mathcal{D}}{\mathbf{E}}\left[||\mathbf{1}\{T > t\} - \widehat{\mathbf{P}}(T > t|X))||_2^2\right] \tag{5.8}$$

Time Dependent Concordance Index (C^{td}): A rank order statistic that computes model performance in ranking patients based on their estimated risk at a specfic time horizon.

$$C^{td}(t) = \mathbf{P}\big(\hat{F}(t|\mathbf{x}_i) > \hat{F}(t|\mathbf{x}_j)|\delta_i = 1, T_i < T_j, T_i \leq t\big) \tag{5.9}$$

We compute the censoring adjusted estimates of the Time Dependent Concordance Index (Antolini et al., 2005; Gerds et al., 2013) and the Integrated Brier Score[3] (Gerds & Schumacher, 2006; Graf et al., 1999) to assess both discriminative performance and model calibration at multiple time horizons. For each of the datasets we perform 5-fold cross-validation over the hyperparameter grid as described in Appendix D.3, and report the performance of the hyperparameter setting with the lowest Brier Score averaged over all folds. Table 5.2 presents the discriminative performance and calibration of CMHE compared to Cox PH models in factual regression. We find that CMHE had similar or better discriminative performance than a simple Cox Model with a linear and MLP hazard functions. CMHE was also better calibrated as evidenced by overall lower Integrated Brier Score, suggesting utility for factual risk estimation.

5.5 Discussion and Conclusion

We proposed a novel deep learning approach able to discover latent phenogroups that respond differentially to an intervention in the presence of censored time-to-event outcomes. It provides a valuable adjunct to traditional statistical techniques in healthcare survival research and can aid determination of treatment efficacy. This new technique, CMHE, provides the opportunity to gain individualized patient insights by identifying subjects who would benefit from a treatment as well as those who are at highest risk for harm. This can be particularly useful in clinical practice when these personalized insights differ from population-level expectations of the efficacy of the established as well as novel treatment protocols.

[3] Brier Score integrated over 1, 3 and 5 years. IBS $= \sum_t {}^{t}/_{t_{\max}} \cdot \mathrm{BS}(t)$

Part III

Interpretable Approaches for Actionable Time-to-Event Analysis

Motivation

The motivations for this chapter are drawn from my interactions with clinicians in cardiovascular surgery and medicine and our subsequent attempts to derive actionable insights helping improve patient care. Estimators and methods included in Part I and II rely on the use of deep representation learning techniques in order to model complex clinical covariates. While powerful, practical deployment and their use in real world clinical scenarios is severely limited by the inability of these models to produce intepretable hypotheses which could help provide diagnoses and guide personalized therapy.

In this part of the thesis, we thus take a tangential view to modelling time-to-event and survival problems keeping interpretability as a paramount requirement. In light of this we propose two approaches involving 1) estimation of heterogeneous treatment effects using sparsity presevering regularization and, 2) disease stratification and staging using integer risk scoring methods.

Chapter 6 extends the estimator introduced in Part II with the ability to specify sparse feature selection approaches involving structured regularization. As opposed the previous model, we do not require stochastic sampling and instead propose an exact inference algorithm compatible with such sparsity regularization. We demonstrate the ability of our approach in recovering parsimonious subgroups in real world studies of cardio-vascular health helping improve understanding of heterogeneity in these studies.

Clinicians often rely on simplistic stratification of patients into stages and categories that guide subsequent treatment and therapy. The ability for clinicians to quickly use an integer calculator that assigns scores to each patient based on their risk level thus makes for an impactful contribution. Chapter 7 proposes, CoxSLIM, a methodology for data-driven recovery of an integer risk scoring model with censored time-to-event outcomes. We model the problem as a Mixed Integer Non-Linear program involving constraints that impose model structure in terms of maximum number of features employed and propose an efficient algorithm to solve the aforementioned optimization problem.

Chapter 6

Recovering Sparse and Interpretable Subgroups with Heterogeneous Treatment Effects with Censored Time-to-Event Outcomes

6.1 Introduction

Data driven decision making across multiple disciplines including healthcare, epidemiology, econometrics and prognostics often involves establishing efficacy of an intervention when outcomes are measured in terms of the time to an adverse event, such as death, failure or onset of a critical condition. Typically the analysis of such studies involves assigning a patient population to two or more different treatment arms often called the 'treated' (or 'exposed') group and the 'control' (or 'placebo') group and observing whether the populations experience an adverse event (for instance death or onset of a disease) over the study period at a rate that is higher (or lower) than for the control group. Efficacy of a treatment is thus established by comparing the relative difference in the rate of event incidence between the two arms called the hazard ratio. However, not all individuals benefit equally from an intervention. Thus, very often potentially beneficial interventions are discarded even though there might exist individuals who benefit, as the population level estimates of treatment efficacy are inconclusive.

In this chapter we assume that patient responses to an intervention are typically heterogeneous and there exists patient subgroups that are unaffected by (or worse, **harmed**) by the intervention being assessed. The ability to discover or phenotype these patients is thus clinically useful as it would allow for more precise clinical decision making by identifying individuals that actually benefit from the intervention being assessed. Towards this end our contributions in this chapter are as follows:

- We propose **Sparse Cox Subgrouping**, (SCS) a latent variable approach to model patient subgroups that demonstrate heterogeneous effects to an intervention.

- As opposed to existing literature in modeling heterogeneous treatment effects with censored time-to-event outcomes our approach involves structured regularization of the covariates that assign individuals to subgroups leading to parsimonious models resulting in phenogroups that are interpretable.
- We release a `python` implementation of the proposed SCS approach as part of the `auton-survival` package (Nagpal et al., 2022b) for survival analysis: `https://autonlab.github.io/auton-survival/`

6.2 Related Work

Large studies especially in clinical medicine and epidemiology involve outcomes that are time-to-events such as death, or an adverse clinical condition like stroke or cancer. Treatment efficacy is typically estimated by comparing event rates between the treated and control arms using the Proportional Hazards (Cox, 1972) model and its extensions.

Identification of subgroups in such scenarios has been the subject of a large amount of traditional statistical literature. Large number of such approaches involve estimation of the factual and counterfactual outcomes using separate regression models (T-learner) followed by regressing the difference between these estimated potential outcomes. Within this category of approaches, Lipkovich et al. (2011) propose the subgroup identification based on differential effect search (SIDES) algorithm, Su et al. (2009) propose a recursive partitioning method for subgroup discovery, Dusseldorp & Mechelen (2014) propose the qualitative interaction trees (QUINT) algorithm, and Foster et al. (2011) propose the virtual twins (VT) method for subgroup discovery involving decision tree ensembles. We include a parametric version of such an approach as a competing baseline.

Identification of heterogeneous treatment effects (HTE) is also of growing interest to the machine learning community with multiple approaches involving deep neural networks with balanced representations (Song et al., 2016; Johansson et al., 2020), generative models Louizos et al. (2017) as well as Non-Parametric methods involving random-forests (Wager & Athey, 2018) and Gaussian Processes (Alaa & van der Schaar, 2017a). There is a growing interest in estimating HTEs from an interpretable and trustworthy standpoint (Lee et al., 2020; Nagpal et al., 2020; Morucci et al., 2020; Wu et al., 2022; Crabbé et al., 2022). Wang & Rudin (2022) propose a sampling based approach to discovering interpretable rule sets demonstrating HTEs.

However large part of this work has focused extensively on outcomes that are binary or continuous. The estimation of HTEs in the presence of censored time-to-events has been limited. Xu et al. (2022) explore the problem and describe standard approaches to estimate treatment effect heterogeneity with survival outcomes. They also describe challenges associated with existing risk models when assessing treatment effect heterogeneity in the case of cardiovascular health.

There has been some initial attempt to use neural network for causal inference with censored time-to-event outcomes. Curth et al. (2021) propose a discrete time method along with regularization to match the treated and control representations. Chapfuwa et al. (2021)'s approach is related and involves the use of normalizing flows to estimate the potential time-to-event distributions under treatment and control. While contributions are similar to Chapter 5 (Nagpal et al., 2022a), in that we assume treatment effect heterogeneity through a latent variable model, the chapter differs in that
1) Our approach is free of the expensive Monte-Carlo sampling procedure and,
2) Our generalized EM inference procedure allows us to naturally incorporate structured sparsity regular-

ization, which helps recovers phenogroups that are parsimonious in the features they recover that define subgroups.

Survival and time-to-event outcomes occur pre-eminently in areas of cardiovascular health. One such area is reducing combined risk of adverse outcomes from atherosclerotic disease (Herrington et al., 2016; Furberg et al., 2002; Group, 2009; Buse et al., 2007) (a class of related clinical conditions from increasing deposits of plaque in the arteries, leading to Stroke, Myocardial Infarction and other Coronary Heart Diseases.) The ability of recovering groups with differential benefits to interventions can thus lead to improved patient care through framing of optimal clinical guidelines.

6.3 Proposed Model: Sparse Cox Subgrouping

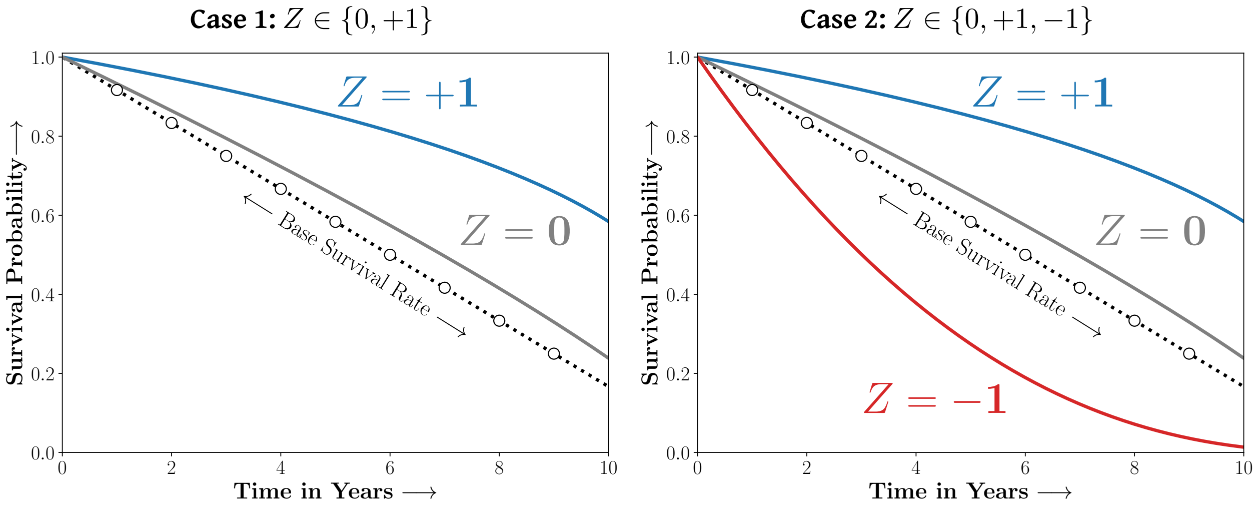


Figure 6.1: Potential outcome distributions under the assumptions of treatment effect heterogeneity. **Case 1**: Amongst the treated population, conditioned on the latent Z, there are two subgroups that **benefit** and are **unaffected** by the intervention. **Case 2**: There is an additional latent subgroup conditioned on which, the treated population is **harmed** with a worse survival rate.

Notation As is standard in survival analysis, we assume that we either observe the true time-to-event or the time of censoring $U = \min\{T, C\}$ indicated by the censoring indicator defined as $\Delta = \mathbf{1}\{T < C\}$. We thus work with a dataset of right censored observations in the form of 4-tuples, $\mathcal{D} = \{(\boldsymbol{x}_i, \delta_i, \boldsymbol{u}_i, \boldsymbol{a}_i)\}_{i=1}^{n}$, where $\boldsymbol{u}_i \in \mathbb{R}^{+}$ is the time-to-event or censoring as indicated by $\delta_i \in \{0, 1\}$, $\boldsymbol{a}_i \in \{0, 1\}$ is the indicator of treatment assignment, and $\boldsymbol{x}_i$ are individual covariates that confound the treatment assignment and the outcome.

Assumption 5 (Independent Censoring) *The time-to-event T and the censoring distribution C are independent conditional on the covariates X and the intervention A.*

Model Consider a maximum likelihood approach to model the data $\mathcal{D}$ the set of parameters $\boldsymbol{\Omega}$. Under Assumption 5 the likelihood of the data $\mathcal{D}$ can be given as,

$$\mathcal{L}(\boldsymbol{\Omega};\mathcal{D}) \propto \prod_{i=1}^{|\mathcal{D}|} \boldsymbol{\lambda}(u_i|X=\boldsymbol{x}_i, A=\boldsymbol{a}_i)^{\delta_i}\boldsymbol{S}(u_i|X=\boldsymbol{x}_i, A=\boldsymbol{a}_i), \tag{6.1}$$

here $\boldsymbol{\lambda}(t) = \lim\limits_{\Delta t\to 0} \frac{\mathbb{P}(t\leq T<t+\Delta t|T\geq t)}{\Delta t}$ is the hazard and $\boldsymbol{S}(t) = \mathbb{P}(T>t)$ is the survival rate.

Assumption 6 (PH) *The distribution of the time-to-event T conditional on the covariates and the treatment assignment obeys proportional hazards.*

From Assumption 6 (Proportional Hazards), an individual with covariates $(X = \boldsymbol{x})$ under intervention $(A = \boldsymbol{a})$ under a Cox model with parameters β and treatment effect ω is given as

$$\boldsymbol{\lambda}(\boldsymbol{t}|A=\boldsymbol{a}, X=\boldsymbol{x}) = \boldsymbol{\lambda}_0(t)\exp\left(\boldsymbol{\beta}^\top\boldsymbol{x} + \boldsymbol{\omega}\cdot\boldsymbol{a}\right), \tag{6.2}$$

Here, $\boldsymbol{\lambda}_0(\cdot)$ is an infinite dimensional parameter known as the base survival rate. In practice in the Cox's model the base survival rate is a nuisance parameter and is estimated non-parametrically. In order to model the heterogeneity of treatment response. We will now introduce a latent variable $Z \in \{0, 1, -1\}$ that mediates treatment response to the model,

$$\begin{aligned}\boldsymbol{\lambda}(\boldsymbol{t}|A=\boldsymbol{a}, X=\boldsymbol{x}, Z=\boldsymbol{k}) &= \boldsymbol{\lambda}_0(t)\exp(\beta^\top\boldsymbol{x})\exp(\boldsymbol{\omega})^{\boldsymbol{k}\boldsymbol{a}},\\ \text{and,}\quad \mathbb{P}(Z=\boldsymbol{k}|X=\boldsymbol{x}) &= \frac{\exp(\boldsymbol{\theta}_k^\top\boldsymbol{x})}{\sum_j \exp(\boldsymbol{\theta}_j^\top\boldsymbol{x})}.\end{aligned} \tag{6.3}$$

Here, $\boldsymbol{\omega} \in \mathbb{R}$ is the treatment effect, and $\boldsymbol{\theta} \in \mathbb{R}^{k\times d}$ is the set of parameters that mediate assignment to the latent group Z conditioned on the confounding features $\boldsymbol{x}$. Note that the above choice of parameterization naturally enforces the requirements from the model as in Figure 6.1. Consider the following scenarios,

Case 1: The study population consists of two sub-strata ie. $Z \in \{0, +1\}$, that are benefit and are unaffected by treatment respectively.

Case 2: The study population consists of three sub-strata ie. $Z \in \{0, +1, -1\}$, that benefit, are harmed or unaffected by treatment respectively.

Following from Equations 6.1 & 6.3, the complete likelihood of the data $\mathcal{D}$ under this model is,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Omega};\mathcal{D}) &= \prod_{i=1}^{|\mathcal{D}|}\sum_{k\in Z}\Big(\boldsymbol{\lambda}_0(u_i)\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})\Big)^{\delta_i}\boldsymbol{S}_0(u_i)^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})}\mathbb{P}(Z=k|X=\boldsymbol{x}_i)\\ \text{where, } \ln\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k}) &= \boldsymbol{\beta}^\top\boldsymbol{x} + \boldsymbol{k}\cdot\boldsymbol{a}\cdot\boldsymbol{w} \text{ and } \ln\boldsymbol{S}_0(\cdot) = -\boldsymbol{\Lambda}_0(\cdot),\end{aligned} \tag{6.4}$$

Note that $\boldsymbol{\Lambda}_0(\cdot) = \int_0^t \boldsymbol{\lambda}_0(\cdot)$ is the infinite dimensional cumulative hazard and is inferred when learning the model. We will notate the set of all learnable parameters as $\boldsymbol{\Omega} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\Lambda}_0\}$.

Shrinkage In retrospective analysis to recover treatment effect heterogeneity a natural requirement is parsimony of the recovered subgroups in terms of the covariates to promote model interpretability. Such parsimony can be naturally enforced through appropriate shrinkage on the coefficients that promote sparsity. We want to recover phenogroups that are 'sparse' in $\boldsymbol{\theta}$. We enforce sparsity in the parameters of the latent Z gating function via a group ℓ_1 (Lasso) penalty. The final loss function to be optimized including the group sparsity regularization term is,

$$\mathcal{L}(\boldsymbol{\Omega};\mathcal{D}) + \boldsymbol{\epsilon} \cdot \mathcal{R}(\boldsymbol{\theta}) \text{ where, } \mathcal{R}(\boldsymbol{\theta}) = \sum_{d} \sqrt{\sum_{k\in\mathcal{Z}} \left(\boldsymbol{\theta}_d^k\right)^2}$$
$$\text{and } \boldsymbol{\epsilon} > 0 \text{ is the strength of the shrinkage parameter.} \tag{6.5}$$

Identifiability is imposed by restricting the gating parameters for the $(Z = 0)$ to be 0. Thus $\boldsymbol{\theta}_1 = 0$.

Inference We will present a variation of the **Expectation Maximization** algorithm to infer the parameters in Equation 6.3. Our approach differs from Nagpal et al. (2022a, 2021d) in that it does not require stochastic Monte-Carlo sampling. Further, our generalized EM inference allows for the incorporation of the structured sparsity in the **M-Step**.

A Semi-Parametric $Q(\cdot)$ Note that the likelihood in Equation 6.3 is semi-parametric and consists of parametric components and the infinite dimensional base hazard $\boldsymbol{\Lambda}(\cdot)$. We define the $Q(\cdot)$ as:

$$Q(\boldsymbol{\Omega};\mathcal{D}) = \sum_{i=1}^{n}\sum_{k\in\mathcal{Z}} \gamma_i^k \Big(\ln \boldsymbol{p_\theta}(Z = k|X = \boldsymbol{x}_i) + \ln \boldsymbol{p}_{\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{\Lambda}}(T|Z = k, X = \boldsymbol{x}_i)\Big) + \mathcal{R}(\boldsymbol{\theta})$$

The E-Step Requires computation of the posteriors counts $\boldsymbol{\gamma} := \boldsymbol{p}(Z = k|T, X = \boldsymbol{x}, A = \boldsymbol{a})$.

Result 1 (Posterior Counts) *The posterior counts $\boldsymbol{\gamma}$ for the latent Z are estimated as,*

$$\begin{aligned}\boldsymbol{\gamma}^k &= \widehat{\mathbb{P}}(Z = k|X = \boldsymbol{x}, A = \boldsymbol{a}, \boldsymbol{u}) \\ &= \frac{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})^{\delta_i}\,\widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})}\exp(\boldsymbol{\theta}_{\boldsymbol{k}}^\top \boldsymbol{x})}{\sum_{j\in\mathcal{Z}} \boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{j})^{\delta_i}\,\widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{j})}\exp(\boldsymbol{\theta}_{\boldsymbol{j}}^\top \boldsymbol{x})}.\end{aligned} \tag{6.6}$$

Result 1 follows directly from the For a full discussion on derivation of the $Q(\cdot)$ and the posterior counts please refer to Appendix E.1

The M-Step Involves maximizing the $Q(\cdot)$ function. Rewriting the $Q(\cdot)$ as a sum of two terms,

$$Q(\boldsymbol{\Omega}) = \underbrace{\sum_{i=1}^{n}\sum_{k\in\mathcal{Z}} \gamma_i^k \ln \boldsymbol{p}_{\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{\Lambda}_0}(T|Z = k, X = \boldsymbol{x}_i, A = \boldsymbol{a}_i)}_{\boldsymbol{A}(\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{\Lambda}_0)} + \underbrace{\sum_{i=1}^{n}\sum_{k\in\mathcal{Z}} \gamma_i^k \ln \boldsymbol{p_\theta}(Z = k|X = \boldsymbol{x}_i) + \mathcal{R}(\boldsymbol{\theta})}_{\boldsymbol{B}(\boldsymbol{\theta})}$$

Result 2 (Weighted Cox model) *The term $\boldsymbol{A}$ can be rewritten as a weighted Cox model and thus optimized using the corresponding 'partial likelihood',*

Updates for $\{\boldsymbol{\beta}, \boldsymbol{\omega}\}$: The partial-likelihood, $\mathcal{PL}(\cdot)$ under sampling weights (Binder, 1992) is

$$\mathcal{PL}(\boldsymbol{\Omega}; \mathcal{D}) = \sum_{i=1, \delta_i=1}^{n} \sum_{k \in \mathcal{Z}} \gamma_i^k \Bigg(\ln \boldsymbol{h}_k(\boldsymbol{x}_i, \boldsymbol{a}_i, \boldsymbol{k}) - \ln \sum_{j \in \mathsf{RiskSet}(u_i)} \sum_{k \in \mathcal{Z}} \gamma_j^k \boldsymbol{h}_k(\boldsymbol{x}_j, \boldsymbol{a}_j, \boldsymbol{k}) \Bigg) \Bigg] \tag{6.7}$$

Here RiskSet$(\cdot)$ is the *'risk set'* or the set of all individuals who haven't experienced the event till the corresponding time, i.e. $\mathsf{RS}(t) := \{i : u_i > t\}$. $\mathcal{PL}(\cdot)$ is convex in $\{\boldsymbol{\beta}, \boldsymbol{\omega}\}$ and we update these with a gradient step.

Updates for $\boldsymbol{\Lambda}_0$: The base hazard $\boldsymbol{\Lambda}_0$ are updated using a weighted Breslow's estimate (Breslow, 1972b; Lin, 2007) assuming the posterior counts γ to be sampling weights (Chen, 2009),

$$\widehat{\boldsymbol{\Lambda}}_0(t)^+ = \sum_{i=1}^{n} \sum_{k \in \mathcal{Z}} \mathbf{1}\{u_i < t\} \frac{\gamma_i^k \cdot \delta_i}{\sum\limits_{j \in \mathsf{RS}(u_i)} \sum\limits_{k \in \mathcal{Z}} \gamma_j^k \boldsymbol{h}_k(\boldsymbol{x}_j, \boldsymbol{a}_j, \boldsymbol{k})} \tag{6.8}$$

Term $\boldsymbol{B}$ is a function of the gating parameters $\boldsymbol{\theta}$ that determine the latent assignment Z along with sparsity regularization. We update $\boldsymbol{B}$ using a Proximal Gradient update as is the case with Iterative Soft Thresholding (ISTA) for group sparse ℓ_1 regression.

Updates for $\boldsymbol{\theta}$: The update for $\boldsymbol{\theta}$ including the group regularization (Friedman et al., 2010) term $\mathcal{R}(\cdot)$,

$$\widehat{\boldsymbol{\theta}}^+ = \mathsf{prox}_{\eta\epsilon}\left(\boldsymbol{\theta} - \frac{d}{d\boldsymbol{\theta}} \boldsymbol{B}(\boldsymbol{\theta})\right), \quad \text{where } \mathsf{prox}_{\eta\epsilon}(\boldsymbol{y}) = \frac{\boldsymbol{y}}{||\boldsymbol{y}||_2} \max\{0, ||\boldsymbol{y}||_2 - \eta\boldsymbol{\epsilon}\}. \tag{6.9}$$

All together the inference procedure is described in Algorithm 3.

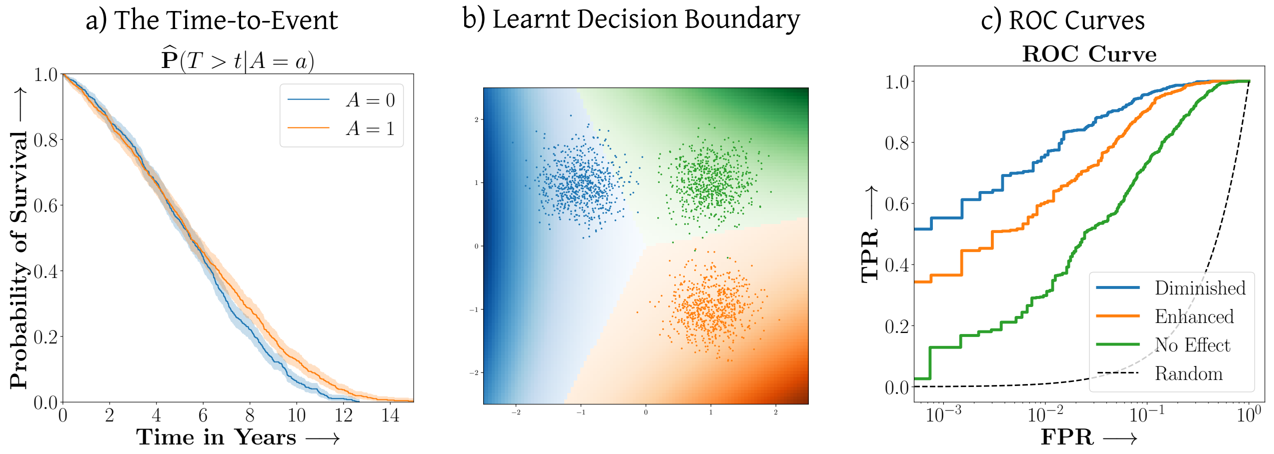
Algorithm 3: Parameter Learning for SCS with a Generalized EM

```
Input  :Training set, $\mathcal{D} = \{(\boldsymbol{x}_i, u_i, a_i, \delta_i)_{i=1}^n\}$; maximum EM iterations, $B$, step size $\eta$

while <not converged> do
    for $b \in \{1, 2, ..., B\}$ do
        E-STEP ..........................................................
        $\gamma_i^k = \frac{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})^{\delta_i} \widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})} \exp(\boldsymbol{\theta}_{\boldsymbol{k}}^\top \boldsymbol{x})}{\sum_{j \in \mathcal{Z}} \boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{j})^{\delta_i} \widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{j})} \exp(\boldsymbol{\theta}_{\boldsymbol{j}}^\top \boldsymbol{x})}$    ▷ Compute posterior counts (Equation 6.6).
        M-STEP ..........................................................
        $\widehat{\boldsymbol{\beta}}^+ \leftarrow \widehat{\boldsymbol{\beta}} - \eta \cdot \nabla_{\boldsymbol{\beta}} \mathcal{PL}(\boldsymbol{\beta}, \boldsymbol{w}; \mathcal{D})$
        $\widehat{\boldsymbol{w}}^+ \leftarrow \widehat{\boldsymbol{w}} - \eta \cdot \nabla_{\boldsymbol{w}} \mathcal{PL}(\boldsymbol{\beta}, \boldsymbol{w}; \mathcal{D})$    ▷ Gradient descent update.
        $\widehat{\boldsymbol{\Lambda}}_0(t)^+ \leftarrow \sum_{i=1}^n \sum_{k \in \mathcal{Z}} \mathbf{1}\{u_i < t\} \frac{\gamma_i^k \cdot \delta_i}{\sum\limits_{j \in \mathsf{RiskSet}(u_i)} \sum\limits_{k \in \mathcal{Z}} \gamma_j^k \boldsymbol{h}_k(\boldsymbol{x}_j, \boldsymbol{a}_j, \boldsymbol{k})}$    ▷Breslow (1972b)'s estimator.
        $\widehat{\boldsymbol{\theta}}^+ \leftarrow \mathsf{prox}_{\eta\epsilon}(\widehat{\boldsymbol{\theta}} - \eta \cdot \nabla_\theta \boldsymbol{B}(\theta))$    ▷ Update $\boldsymbol{\theta}$ with gradient of $\widehat{Q}$.
    end
end

Return :learnt parameters $\boldsymbol{\Omega}$;
```

6.4 Experiments

In this section we describe the experiments conducted to benchmark the performance of SCS against alternative models for heterogenous treatment effect estimation on multiple studies including a synthetic dataset and multiple large landmark clinical trials for cardiovascular diseases.

6.4.1 Simulation


a) The Time-to-Event

b) Learnt Decision Boundary

c) ROC Curves

Figure 6.2: a) Population level Kaplan-Meier Estimates of the Survival Distribution stratified by the treatment assignment. b) Distribution of the Latent Z in X and the recovered decision boundary by SCS. c) Receiver Operator Characteristics of SCS in recovering the true phenotype.

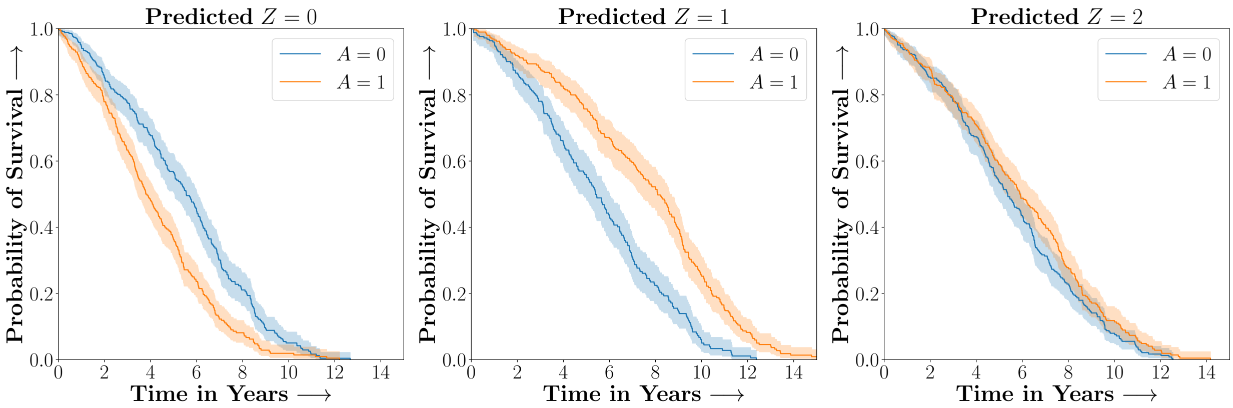


Figure 6.3: The phenotypes recovered with Sparse Cox Subgrouping on the Synthetic Data. As expected, the recovered phenotypes conform to the modelling assumptions as in Figure 6.4.

.

In this section we first describe the performance of the proposed Sparse Cox Subgrouping approach on a synthetic dataset designed to demonstrate heterogeneous treatment effects. We randomly assign individuals to the treated or control group. The latent variable Z is drawn from a uniform categorical distribution that determines the subgroup,

$$A \sim \text{Bernoulli}(1/2), \quad Z \sim \text{Categorical}(1/3)$$

Conditioned on Z we sample $X_{1:2} \sim \text{Normal}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$ as in Figure 6.2 that determine the conditional Hazard Ratios $\text{HR}(k)$, and randomly sample noisy covariates $X_{3:6} \sim \text{Uniform}(-1, 1)$. The true time-to-event T and censoring times C are then sampled as,

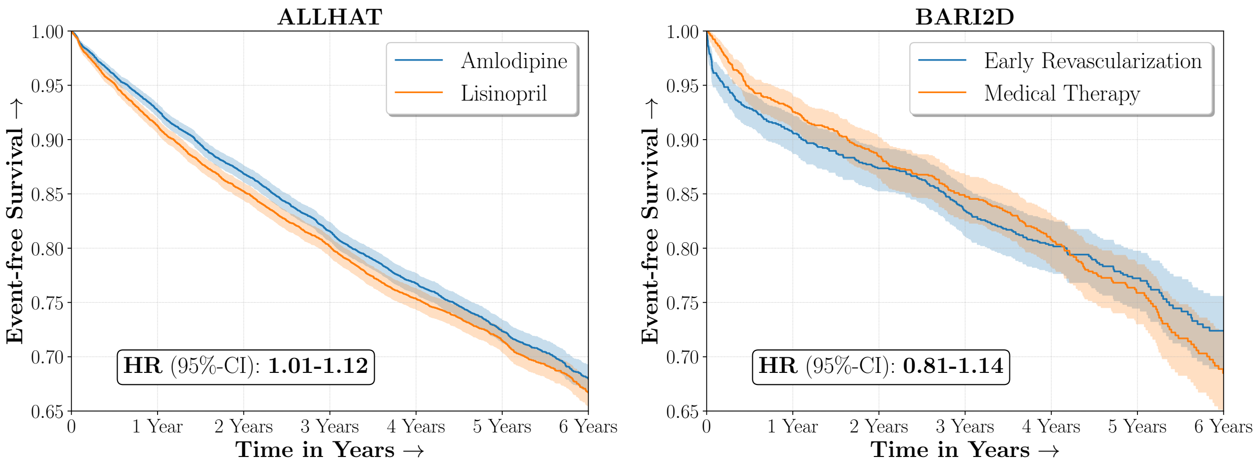
$$T|(X = \boldsymbol{x}, A = \boldsymbol{a}, Z = \boldsymbol{k}) \sim \text{Gompertz}(\beta = 1, \eta = 0.25 \cdot \text{HR}(k)^{\boldsymbol{a}}), \quad C|T \sim \text{Uniform}(0, T)$$

Finally we sample the censoring indicator $\Delta \sim \text{Bernoulli}(0.8)$ and set the observed time-to-event,

$$U = T \text{ if } \Delta = 1, \text{else we set } U = C.$$

Figure 6.2 presents the ROC curves for SCS's ability to identify the groups with enhanced and diminished treatment effects respectively. In Figure 6.3 we present Kaplan-Meier estimators of the Time-to-Event distributions conditioned on the predicted Z by SCS. Clearly, SCS is able to identify the phenogroups corresponding to differential benefits.

6.4.2 Recovering subgroups demonstrating Heterogeneous Treatment Effects from Landmark studies of Cardiovascular Health



ALLHAT	
Size	18,102
Outcome	Combined CVD
Intervention	Lisinopril
Control	Amlodipine
Hazard Ratio	1.06, (1.01, 1.12)
5-year RMST	24.86, (8.89, 37.35)

BARI2D	
Size	2,368
Outcome	Death, MI or Stroke
Intervention	Medical Therapy
Control	Early Revascularization
Hazard Ratio	1.02, (0.81, 1.14)
5-year RMST	23.26, (-27.01, 64.84)

Figure 6.4: Event-free Kaplan-Meier survival curves stratified by the treatment assignment and summary statistics for the **ALLHAT** and **BARI2D** studies. (Combined CVD: Coronary Heart Disease, Stroke, other treated angina, fatal or non-fatal Heart Failure, and Peripheral Arterial Disease.)

Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack (Furberg et al., 2002) The ALLHAT study was a large randomized experiment conducted to assess the efficacy of multiple classes of blood pressure lowering medicines for patients with hypertension in reducing risk of adverse cardiovascular conditions. We considered a subset of patients from the original **ALLHAT** study who were randomized to receive either Amlodipine (a calcium channel blocker) or Lisinopril (an Angiotensin-converting enzyme inhibitor). Overall, Amlodipine was found to be more efficacious than Lisinopril in reducing combined risk of cardio-vascular disease.

Bypass Angioplasty Revascularization Investigation in Type II Diabetes (Group, 2009)

Diabetic patients have been traditionally known to be at higher risk of cardiovascular disease however appropriate intervention for diabetics with ischemic heart disease between surgical coronary revascularization or management with medical therapy is widely debated. The **BARI2D** was a large landmark experiment conducted to assess efficacy between these two possible medical interventions. Overall **BARI2D** was inconclusive in establishing the appropriate therapy between Coronary Revascularization or medical management for patients with Type-II Diabetes.

Figure 6.4 presents the event-free survival rates as well as the summary statistics for the studies. In our experiments, we included a large set of confounders collected at baseline visit of the patients which we utilize to train the proposed model. A full list of these features are in Appendix E.2.

6.4.3 Baselines

Cox PH with ℓ_1 Regularized Treatment Interaction (COX-INT)

We include treatment effect heterogeneity via interaction terms that model the time-to-event distribution using a proportional hazards model as in Kehl & Ulm (2006). Thus,

$$\boldsymbol{\lambda}(t|X = \boldsymbol{x}, A = \boldsymbol{a}) = \boldsymbol{\lambda}_0(t) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x} + \boldsymbol{a} \cdot \boldsymbol{\theta}^\top \boldsymbol{x}\right) \tag{6.10}$$

The interaction effects $\boldsymbol{\theta}$ are regularized with a lasso penalty in order to recover a sparse phenotyping rule defined as $G(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$.

Binary Classifier with ℓ_1 Regularized Treatment Interaction (BIN-INT)

Instead of modelling the time-to-event distribution we directly model the thresholded survival outcomes $Y = \mathbf{1}\{T < t\}$ at a five-year time horizon using a log-linear parameterization with a logit link function. As compared to COX-INT, this model ignores the data-points that were right-censored prior to the thresholded time-to-event, however it is not sensitive to the strong assumption of Proportional Hazards.

$$\begin{aligned} \mathbb{E}[T > t|X = \boldsymbol{x}, A = \boldsymbol{a}] = \sigma(\boldsymbol{\beta}^\top \boldsymbol{x} + \boldsymbol{\beta}_0 + \boldsymbol{a} \cdot \boldsymbol{\theta}^\top \boldsymbol{x}), \\ \text{and, } \sigma(\cdot) \text{ is the logistic link function.} \end{aligned} \tag{6.11}$$

Cox PH T-Learner with ℓ_1 Regularized Logistic Regression (COX-TLR)

We train two separate Cox Regression models on the treated and control arms (T-Learner) to estimate the potential outcomes under treatment $(A = 1)$ and control $(A = 0)$. Motivated from the '*Virtual Twins*' approach as in Foster et al. (2011), a logistic regression with an ℓ_1 penalty is trained to estimate if the risk of the potential outcome under treatment is higher than under control. This logistic regression is then employed as the phenotyping function $G(\cdot)$ and is given as,

$$\begin{aligned} G(\boldsymbol{x}) &= \mathbb{E}[\mathbf{1}\{f_1(\boldsymbol{x},t) > f_0(\boldsymbol{x},t)\}|X=\boldsymbol{x}] \\ \text{where, } f_{\boldsymbol{a}}(\boldsymbol{x},t) &= \mathbb{P}(T>t|\text{do}(A=\boldsymbol{a}), X=\boldsymbol{x}). \end{aligned} \tag{6.12}$$

The above models involving sparse ℓ_1 regularization were trained with the `glmnet` (Friedman et al., 2009) package in R.

The ACC/AHA Long term Atheroscleoratic Cardiovascular Risk Estimate

The American College of Cardiology and the American Heart Association model for estimation of risk of Atherosclerotic disease risk (Goff Jr et al., 2014) involves pooling data from multiple observational cohorts of patients followed by modelling the 10-year risk of an adverse cardiovascular condition including death from coronary heart disease, Non-Fatal Myocardial Infarction or Non-fatal Stroke. While the risk model was originally developed to assess factual risk in the observational sense, in practice it is also employed to assess risk when making counterfactual decisions.

6.4.4 Protocol

We compare the performance of SCS and the corresponding competing methods in recovery of subgroups with enhanced (or diminished treatment effects). For each of these studies we stratify the study population into equal sized sets for training and validation while persevering the proportion of individuals that were assigned to treatment and experienced the outcome in the follow up period. The models were trained on the training set and validated on the held-out test set. For each of the methods we experiment with models that do not enforce any sparsity ($\epsilon = 0$) as well as tune the level of sparsity to recover phenotyping functions that involve 5 and 10 features. The subgroup size are varied by controlling the threshold at which the individual is assigned to a group. Finally, the treatment effect is compared in terms of Hazard Ratios, Risk Differences as well as Restricted Mean Survival Time over a 5 Year event period.

6.4.5 Results and Discussion

We present the results of SCS versus the baselines in terms of Hazard Ratios on the **ALLHAT** and **BARI2D** datasets in Figures 6.5 and 6.6. In the case of **ALLHAT**, SCS consistently recovered phenogroups with more pronounced (or diminished) treatment effects. On external validation on the heldout dataset, we found a subgroup of patients that had similar outcomes whether assigned to Lisinopril or Amlodipine, whereas the other subgroup clearly identified patients that were harmed with Lisinopril. The group harmed with Lisinopril had higher Diastolic BP. On the other hand, patients with Lower kidney function did not seem to benefit from Amlodipine.

In the case of **BARI2D**, SCS recovered phenogroups that were both harmed as well as benefitted from just medical therapy without revascularization. The patients who were harmed from Medical therapy were typically older, on the other hand the patients who benefitted primarily included patients who were otherwise assigned to receive PCI instead of CABG revascularization, suggesting PCI to be harmful for diabetic patients. Tables 6.1 and 6.2 present the features that were selected by the proposed model for the studies.

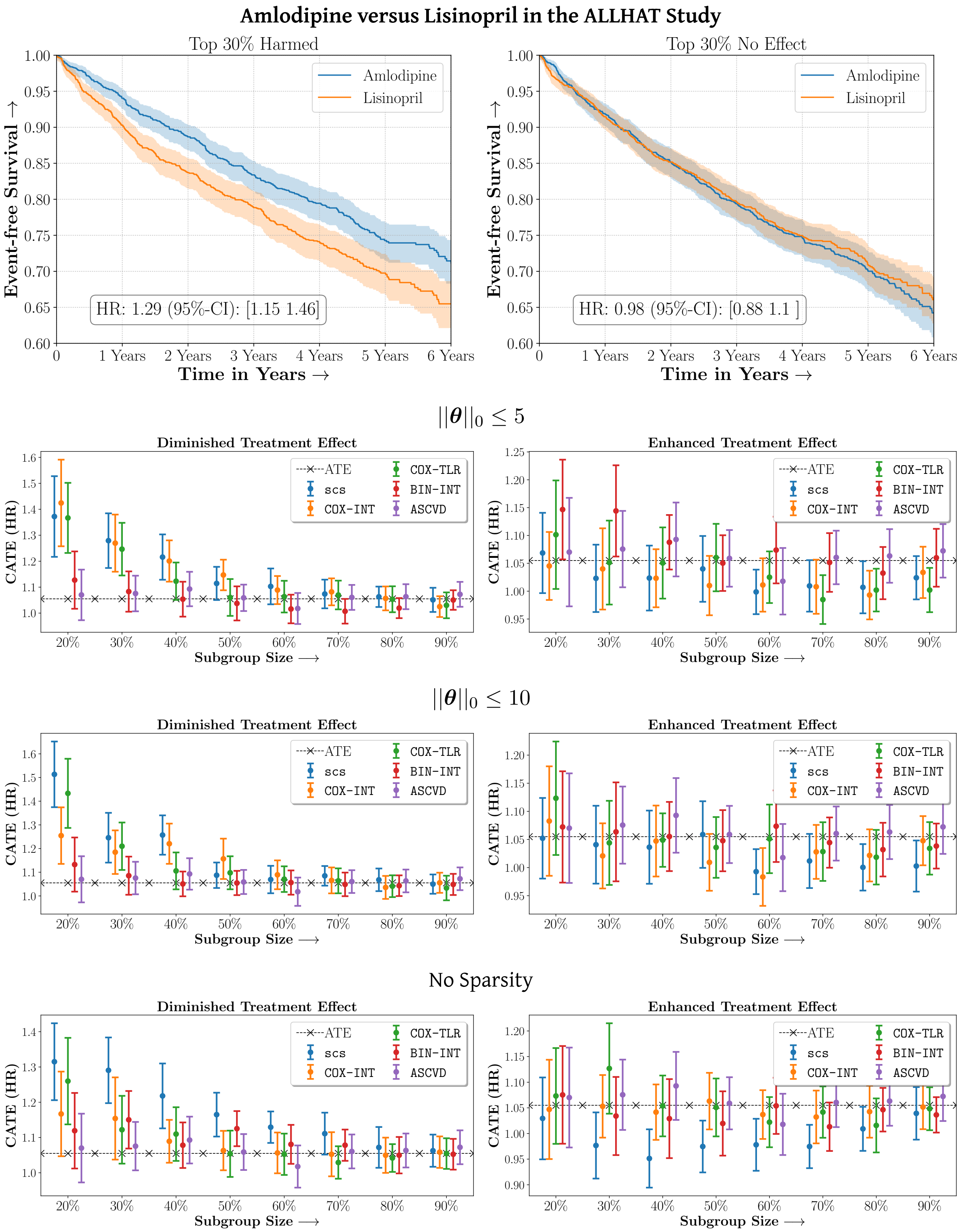


Figure 6.5: Conditional Average Treatment Effect in Hazard Ratio versus subgroup size for the latent phenogroups extracted from the **ALLHAT** study.

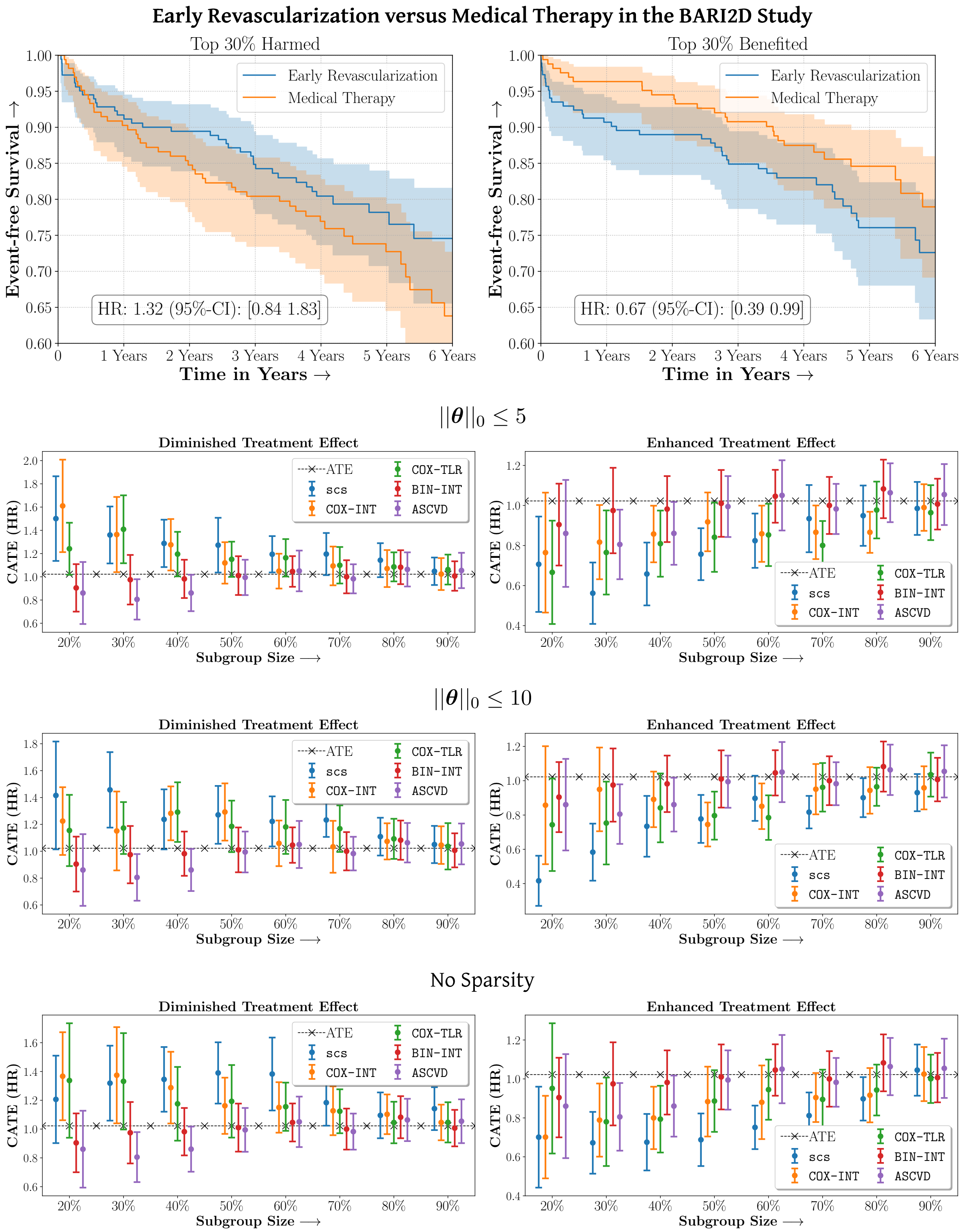


Figure 6.6: Conditional Average Treatment Effect in Hazard Ratio versus subgroup size for the latent phenogroups extracted from the **BARI2D** study.

ALLHAT

Name	Description
BV2SBP	Baseline Seated Diastolic Pressure
BLGFR	Baseline est Glomerular Filteration Rate
BLMEDS	Antihypertensive Treatment
CURSMOKE	Current Smoking Status
SEX	Sex of Participant

BARI2D

Name	Description
age	Age upon entry
asp	Aspirin use
hxhtn	History of hypertension requiring tx
hxchl	Hypercholesterolemia req tx
priorstent	Prior stent

Table 6.1: List of selected features with sparsity level: $||\boldsymbol{\theta}||_0 \leq 5$

ALLHAT

Name	Description
BV2SBP	Baseline Seated Diastolic Pressure
BLGFR	Baseline est Glomerular Filtration Rate
BLMEDS	Antihypertensive Treatment
CURSMOKE	Current Smoking Status
SEX	Sex of Participant
ASPIRIN	Aspirin Use
ACHOL	Total Cholesterol
BLWGT	Weight upon entry
BMI	Body mass index upon entry
OASCVD	Other atherosclerotic cardiovascular disease

BARI2D

Name	Description
age	Age upon entry
asp	Aspirin use
hxhtn	History of hypertension requiring tx
hxchl	Hypercholesterolemia req tx
priorstent	Prior stent
acei	ACE Inhibitor
acr	Urine albumin/creatinine ratio mg/g
insul_circ	Circulating insulin
screat	Serum creatinine (mg/dl)
tchol	Total Cholesterol

Table 6.2: List of selected features with sparsity level: $||\boldsymbol{\theta}||_0 \leq 10$

6.5 Concluding Remarks

We presented Sparse Cox Subgrouping (SCS) a latent variable approach to recover subgroups of patients that respond differentially to an intervention in the presence of censored time-to-event outcomes. As compared to alternative approaches to learning parsimonious hypotheses in such settings, our proposed model recovered hypotheses with more pronounced treatment effects which we validated on multiple studies for cardiovascular health. While powerful in its ability to recover parsimonious subgroups there exists limitations in SCS in its current form. The model is sensitive to proportional hazards and may be ill-specified when the proportional hazards assumptions are violated as is evident in many real world clinical studies (Maron et al., 2018; Bretthauer et al., 2022). Another limitation is that SCS in its current form looks at only a single endpoint (typically death, or a composite of multiple adverse outcome). In practice however real world studies typically involve multiple end-points. We envision that extensions of SCS would allow patient subgrouping across multiple endpoints, leading to discovery of actionable sub-populations that similarly benefit from the intervention under assessment.

Chapter 7

Learning Integer Risk Scores for Disease Staging with Censored Time-to-Event Outcomes

7.1 Introduction

Clinicians routinely rely on simple heuristically learnt risk scoring models for patient risk assessment and for subsequent prioritization to certain therapies or intervention. Such models are typically built on intuition and heuristic judgement followed by extensive validation on an observational cohort of patients to verify performance in terms of both discriminate performance and calibration.

1.	**C**ongestive Heart Failure	1 point		. . .
2.	**H**ypertension	1 point	+	. . .
3.	**A**ge $\geq$ 75	1 point	+	. . .
4.	**D**iabetes Mellitus	1 point	+	. . .
5.	Prior **S**troke or Transient Ischemic Attack	**2** points	+	
		SCORE	=	

SCORE	0	1	2	3	4	5	6
RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Figure 7.1: The popular integer risk scoring model CHADS, with calibrated risk for Atrial Fibrillation risk estimation. Current approaches involving risk estimation rely on heuristics to determine coefficients or optimization are limited to binary outcomes. We propose extensions that naturally allow modelling of time-to-event outcomes as are common in healthcare settings.

Figure 7.1 demonstrates the popular CHADS risk scoring model (Gage et al., 2004). Such scoring systems are ubiquitous in healthcare and can help clinicians assess patient risk and subsequent therapy. Outcomes such as Death, or occurrence of an event are typically Time-to-Event outcomes and treating them as binary often results is mis-estimated risk estimates especially at longer time horizons. Further, large number of patients might be censored or lost to follow up from the study something that cannot be modelled naturally with binary outcomes.

Contributions The main contributions of this chapter include:

1. We propose a mixed integer non-linear program to recover risk scoring systems with time-to-event outcomes and an efficient algorithm to optimize the same that involves a cutting plane approximation of the proposed non-linear program.
2. We demonstrate that the formulation allows for addition of several constraints that help improve usability of the risk scoring system including, level of sparsity and range of possible coefficient values.
3. We demonstrate, through several real world examples that our approach is able to recover patient strata with high discriminative capability as well as calibration. Further, we demonstrate that our approach has better performance than similar approaches that are limited to binary classification.

7.2 Methodology

In this section we present **CoxSLIM** an integer scoring system for censored time-to-event outcomes.

7.2.1 Setup

We start with a dataset of n examples $\{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{n}$, where example i consists of a features $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^d$, a time-to-event $t_i \in \mathbb{R}^+$, and a binary indicator $\delta_i \in \{0, 1\}$ that indicates if t_i represents a time-to-event or censoring time. We consider a maximum likelihood approach to model the time-to-event data. Thus the likelihood L is given as

$$\mathcal{L}(\mathcal{D}) \propto \prod_{i=1}^{n} \boldsymbol{\lambda}(t|X = x_i)^{\delta_i} \boldsymbol{S}(t|X = x_i) \text{ where,} \tag{7.1}$$

Here $\boldsymbol{\lambda}(t) = \lim_{\Delta t \to 0} \frac{P(t<T<t+\Delta t|T>t)}{\Delta t}$ is the hazard rate of the event and $\boldsymbol{S}(t)$ is the corresponding survival function. We assume a proportional hazards (PH) approach to model the above likelihood. Thus the individual hazard rate for an individual with covariates $\boldsymbol{x}$ is modelled as

$$\boldsymbol{\lambda}(t|X = \boldsymbol{x}) = \boldsymbol{\lambda}_0(t) \exp\left(\boldsymbol{w}^\top \boldsymbol{x}/c\right). \tag{7.2}$$

Here $\boldsymbol{\lambda}_0(t)$ is the infinite dimensional baseline hazard rate of failure and $c \in \mathbb{R}^+$ is a constant hyperparameter that sets the base of the exponent. In practice the above model is learnt by minimizing the *partial likelihood* equation given as follows,

$$P\mathcal{L}(\mathcal{D}; \boldsymbol{w}, c) = \sum_{i=1}^{n} \mathbf{1}_{\delta_i \neq 0} \bigg(\boldsymbol{w}^\top \boldsymbol{x}_i/c - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\boldsymbol{w}^\top \boldsymbol{x}_j/c\right) \bigg) \tag{7.3}$$

Here $\mathcal{R}(t)$ is the *'risk set'* or the set of all individuals with $\{t_i : t_i > t\}$. Notice that the partial likelihood is independent of $\boldsymbol{\lambda}(\cdot)$.

Model Requirement	Example
Feature Selection	Choose between 5 to 10 total features
Group Sparsity	Include either $male$ or $female$ in the model but not both
Optimal Thresholding	Use at most 3 thresholds for a set of indicator variables: $\sum_{k=1}^{100} 1\{age \leq k\} \leq 3$
Logical Structure	If $male$ is in model, then include $hypertension$ or $bmi \geq 30$
Side Information	Predict Stage ≥ 5 when $male =$ TRUE and $hypertension =$ TRUE

Table 7.1: Model requirements that can be addressed by adding operational constraints

7.2.2 Learning Integer Hazard Scoring Systmes with Time-to-Event Outcomes

We determine the values of the coefficients by solving the following mixed integer nonlinear program (MINLP):

Definition 3 (Hazard Scoring Problem) *The hazard scoring problem is a discrete optimization problem of the form:*

$$\min_{\boldsymbol{w}} \quad \mathcal{PL}(\mathcal{D}; \boldsymbol{w}, c) \quad \text{s.t.} \quad \boldsymbol{w} \in \mathcal{W} \quad \textit{and} \quad \|\boldsymbol{w}\|_0 \leq R^{\max}, \tag{7.4}$$

where:

- $\mathcal{PL}(\mathcal{D}; \boldsymbol{w}, c) = \sum_{i=1}^{n} \mathbf{1}_{\delta_i \neq 0} \left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{c} - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\right) \right)$ *is the partial likelihood;*
- $\|\boldsymbol{w}\|_0 = \sum_{j=1}^{d} \mathbf{1}\{w_j \neq 0\}$ *is the ℓ_0-seminorm;*
- $\mathcal{W} \subset \mathcal{Z}^{d+1}$ *is a set of feasible coefficient vectors, e.g., $\mathcal{W} = \{-5, 5\}^{d+1}$;*
- $R^{\max} \in \mathcal{Z}$ *is a user-specified parameter to impose sparsity in the learnt coefficient set.*

This problem captures what we desire in a scoring system. The objective minimizes the *partial likelihood* over the event rate which inturn helps recovering models that are well calibrated with good discriminative performance. Further it penalizes the ℓ_0-seminorm (the count of non-zero coefficients) for sparsity. The trade-off parameter ϵ controls the balance between these competing objectives, and represents the maximum log-likelihood that is sacrificed to remove a feature from the optimal model. The constraints restrict coefficients to a set of small integers such as $\mathcal{W} := \{-5, \ldots, 5\}^{d+1}$, and may be customized to encode other model requirements such as those in Table 7.1.

We optimize the problem in Equation 7.4 by solving the following MINLP:

$$\begin{aligned}
\min_{\boldsymbol{w}} \quad & L + \epsilon R \\
\text{s.t.} \quad & L = \sum_{i=1}^{n} \mathbf{1}_{\delta_i \neq 0} \left(\boldsymbol{w}^\top \boldsymbol{x}_i / c - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\boldsymbol{w}^\top \boldsymbol{x}_j / c\right) \right) && \textit{partial likelihood} && (7.5a) \\
& R = \sum_{j \in [d]} \alpha_j && \textit{model size} && (7.5b) \\
& W_j^{\max} \alpha_j \geq w_j && j \in [d] \quad w_j > 0 \implies \alpha_j = 1 && (7.5c) \\
& -W_j^{\min} \alpha_j \geq -w_j && j \in [d] \quad w_j < 0 \implies \alpha_j = 1 && (7.5d) \\
& L \in [0, \ldots, L^{\max}] && \textit{loss} && (7.5e)
\end{aligned}$$

$$
\begin{aligned}
& R \in \{0, \ldots, R^{\text{max}}\} && && \textit{model size} && (7.5\text{f}) \\
& w_j \in \{W_j^{\text{min}} \ldots, W_j^{\text{max}}\} && j \in [d] && \textit{coef for variable } j && (7.5\text{g}) \\
& \alpha_j \in \{0, 1\} && j \in [d] && \alpha_j := 1[w_j \neq 0] && (7.5\text{h})
\end{aligned}
$$

- L and R are “auxiliary" variables that represent the overall loss and the model size, respectively. In theory, these variables are redundant in that they could be replace by the quantities in equation 7.5a and equation 7.5b. In practice, we include them because they allow us to set upper and lower bounds on feasible models via “variable definition constraints" in equation 7.5e and equation 7.5f.

- The formulation accounts for model size using the indicator variables $\alpha_j := 1[w_j \neq 0]$. These variables are set to 1 whenever $w_j \neq 0$ through the constraints in equation 7.5c and equation 7.5d.

- The coefficient for each variable is constrained to small integer values in Constraints . These constraints restrict each w_j to integers from W_j^{min} to W_j^{max}. By default, we set these values to $W_j^{\text{min}} = -5$ and $W_j^{\text{max}} = +5$.

Cutting-Plane Formulation We recover an optimal solution to the MINLP in equation 7.5 with the lattice cutting-plane algorithm in (Ustun & Rudin, 2019). The cutting-plane algorithm solves a surrogate problem that replaces loss function with a linearized “cutting-plane" approximation. This problem is a MINLP equation 7.5 with the following form:

$$
\begin{aligned}
\min_{\boldsymbol{w}} \quad & L + \epsilon R \\
\text{s.t.} \quad & L \geq \mathcal{PL}(\boldsymbol{w}^t) + \nabla\mathcal{PL}(\boldsymbol{w}^t)(\boldsymbol{w} - \boldsymbol{w}^t) && t \in [T] && \textit{loss cuts} && (7.6\text{a}) \\
& R = \sum_{j \in [d]} \alpha_j && && \textit{model size} && (7.6\text{b}) \\
& W_j^{\text{max}} \alpha_j \geq w_j && j \in [d] && w_j > 0 \implies \alpha_j = 1 && (7.6\text{c}) \\
& -W_j^{\text{min}} \alpha_j \geq -w_j && j \in [d] && w_j < 0 \implies \alpha_j = 1 && (7.6\text{d}) \\
& L \in [0, \ldots, L^{\text{max}}] && && \textit{loss} && (7.6\text{e}) \\
& R \in \{0, \ldots, R^{\text{max}}\} && && \textit{model size} && (7.6\text{f}) \\
& w_j \in \{W_j^{\text{min}} \ldots, W_j^{\text{max}}\} && j \in [d] && \textit{coef for variable } j && (7.6\text{g}) \\
& \alpha_j \in \{0, 1\} && j \in [d] && \alpha_j := 1[w_j \neq 0] && (7.6\text{h})
\end{aligned}
$$

The main difference with the MIP formulation in equation 7.6 and the MINLP formulation in equation 7.5 is that we now compute the loss using a cutting-plane approximation of the loss function. The cutting-plane approximation is captured through T cuts equation 7.6a. Each cut is a supporting hyperplane to the loss function at a specific point $\boldsymbol{w}^t$ – where the values of $\boldsymbol{w}^t$ represent integer-feasible solutions. Since we with the Cox partial likelihood (i.e, a convex loss function), the cutting-plane approximation is an under-approximation.

7.2.3 Estimation the Conditional Survival Function

Once the parameters of the model in $\boldsymbol{w}$ are learnt using the cutting plane formulation of MINLP formulation (Equation 7.6), the individualized survival at a time horizon t, $\widehat{\mathbb{P}}(T > t | X = \boldsymbol{x}, \boldsymbol{w})$ can be estimated using a non-parametric maximum likelihood estimation procedure, more commonly known as the Breslow's Estimator (Lin, 2007; Breslow, 1972b).

$$\widehat{\mathbb{P}}(T > t | X = \boldsymbol{x}, \boldsymbol{w}) = \exp\big(- \widehat{\boldsymbol{\Lambda}}_0(t) \big)^{\exp(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{c})} \text{ and, } \quad \widehat{\boldsymbol{\Lambda}}_0(t) = \sum_{t_i < t} \frac{1}{\sum_{j \in \mathcal{R}(t_i)} \exp\big(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\big)} \tag{7.7}$$

Here, $\widehat{\boldsymbol{\Lambda}}_0(t)$ is the estimated baseline cumulative hazard. We also consider an alternate approach to model the conditional survival function that involves building a conditional Kaplan-Meier for each estimate level using the integer scoring output.

7.3 Experiments

Datasets We compare and evaluate the systems using real-world datasets for clinical decision making. Each of the datasets includes demographic information such as sex, age, as well as clinical variables specific to the study or the results of a medical procedure.

`flchain` (Assay of Serum Free Light Chain): This is a public dataset introduced by Dispenzieri et al. (2012) aiming to study the relationship between serum free light chain and mortality. It includes covariates like age, gender, serum creatinine and presence of monoclonal gammapothy. We removed all the individuals with missing covariates and experiment with the remaining subset of 6,524 individuals.

`support` The `support` dataset is derived from a study of the survival risk of critically-ill patients who were discharged from the ICU conducted by Connors et al. (1995). Here, we have records of 9,105 patients. The outcome variable indicates that a patient has died within six months of discharge. The features cover chronic health conditions (e.g., diabetic status, number of comorbidities), vital signs (e.g., mean blood pressure), and results of lab tests (e.g., white blood cell count). The dataset is publically available for research here: `https://hbiostat.org/data/`.

`seer-lymphoma` (Surveillance, Epidemiology and End Results Study)[1]: We consider a cohort of 60,486 patients who were diagnosed with lymphoma or leukemia cancer between 2000-2004 and monitored as part of the National Cancer Institute SEER study (National Cancer Institute, 2019). Here, the outcome variable indicates if a patient dies within five years from any cause, and 45.83% of patients die within the first five years from diagnosis. The cohort includes patients from New Jersey, Greater California, Kentucky, Lousisiana and Georgia. The features reflect the morphology and histology of the tumor (e.g., size, metastasis, stage, node count and location, number and location of notes) as well as interventions that were administered at the time of diagnosis (e.g., surgery, chemo, radiology).

RiskSLIM

Feature	Score
flc.grp=5	1
age<73.0	-1
kappa<1.70	-1
kappa<2.27	-1
creatinine<0.80	1
creatinine>1.40	1

CoxSLIM

Feature	Score
flc.grp_10	2
age<56.0	-2
age<63.5	-2
age<73.0	-2
age>80.0	3
creatinine>1.4	2

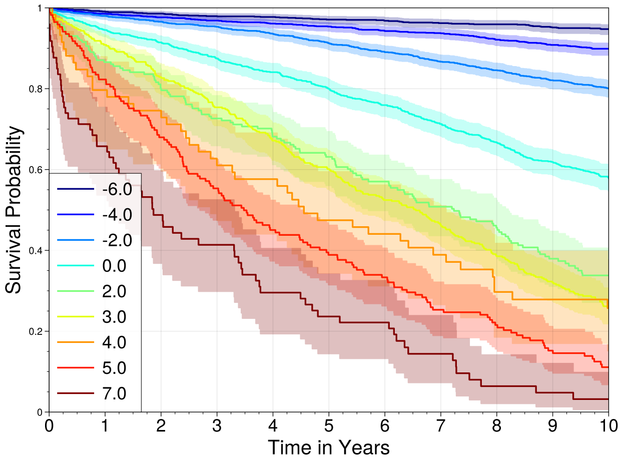
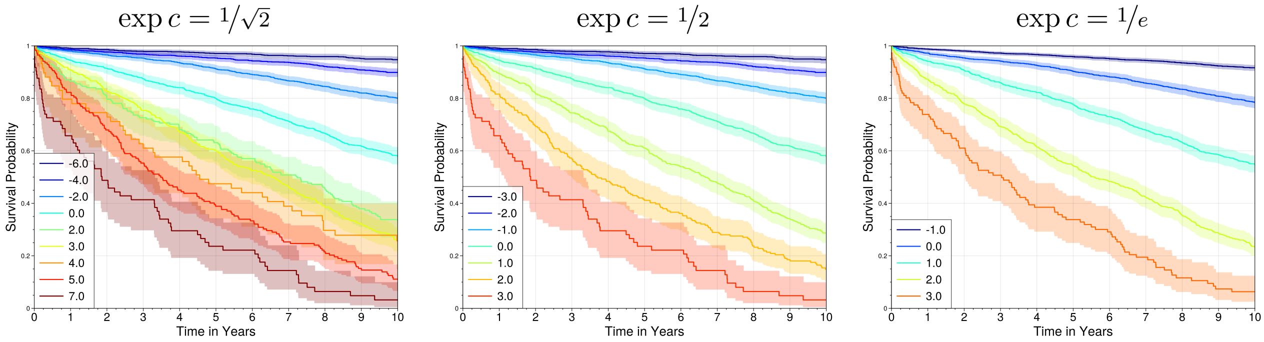


Figure 7.2: **Left**: The learnt integer risk scoring system on `flchain`. **Right**: KM-curves stratified by scores assigned by CoxSLIM. Each curve represents an increase of hazard by a factor of $\sqrt{2}$.


$\exp c = 1/\sqrt{2}$

$\exp c = 1/2$

$\exp c = 1/e$

`flchain`									
	Brier Score			**Area Under ROC Curve**			**ECE**		
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year
				No Censoring					
RiskSLIM	0.034	0.113	0.223	0.815	0.79	0.778	0.010	0.096	0.223
RiskSLIM-KM	0.034	0.095	0.146	0.815	0.79	0.778	0.008	0.021	0.021
CoxSLIM	0.034	0.091	0.131	0.830	0.810	0.829	0.009	0.014	0.027
CoxSLIM-KM	0.034	0.091	0.130	0.833	0.810	0.829	0.013	0.012	0.021
				25% + Censoring					
RiskSLIM	0.033	0.114	0.223	0.821	0.803	0.791	0.011	0.104	0.230
RiskSLIM-KM	0.033	0.093	0.141	0.821	0.799	0.796	0.012	0.027	0.045
CoxSLIM	0.033	0.090	0.133	0.830	0.809	0.829	0.011	0.024	0.046
CoxSLIM-KM	0.033	0.091	0.131	0.833	0.809	0.828	0.014	0.024	0.042
				50% + Censoring					
RiskSLIM	0.033	0.114	0.225	0.817	0.807	0.799	0.015	0.112	0.238
RiskSLIM-KM	0.033	0.095	0.147	0.817	0.807	0.799	0.018	0.053	0.094
CoxSLIM	0.033	0.093	0.143	0.828	0.807	0.827	0.018	0.055	0.095
CoxSLIM-KM	0.034	0.096	0.141	0.821	0.804	0.821	0.020	0.056	0.092

Table 7.2: Discriminative performance and Calibration of CoxSLIM vs. RiskSLIM on the `flchain` data. All of the metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring Weighting.

RiskSLIM	
Feature	Score
Comorbidities>=1	1
Comatose	2
ARF/MOSF	1
Cancer: Metastized	1
-	-
-	-

CoxSLIM	
Feature	Score
bilirubin>2.6	1
Age>=65	1
No Comorbidity	-1
Cancer	1
Comatose	3
Cancer: Metastized	1

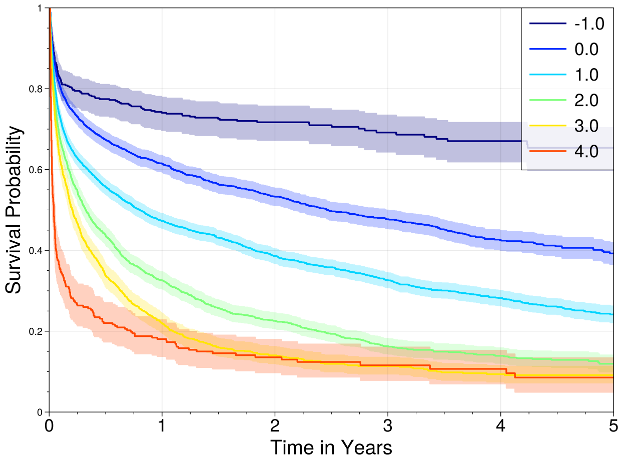
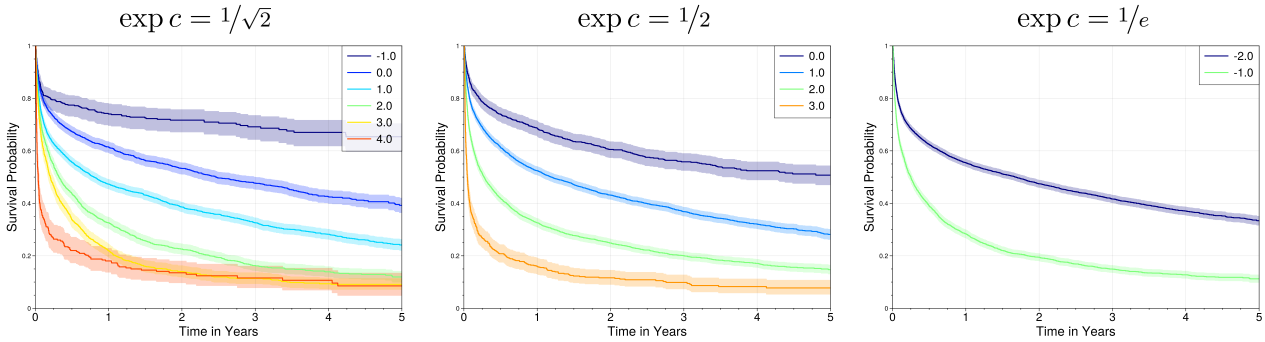


Figure 7.3: **Left**: The learnt integer risk scoring system on `support`. **Right**: KM-curves stratified by scores assigned by CoxSLIM. Each curve represents an increase of hazard by a factor of $\sqrt{2}$.



	support								
	Brier Score			**Area Under ROC Curve**			**ECE**		
	6-Month	1-Year	5-Year	6-Month	1-Year	5-Year	6-Month	1-Year	5-Year
				No Censoring					
RiskSLIM	0.229	0.236	0.257	0.647	0.644	0.67	0.019	0.088	0.29
RiskSLIM-KM	0.228	0.228	0.181	0.646	0.644	0.67	0.009	0.011	0.010
CoxSLIM	0.232	0.224	0.174	0.644	0.670	0.701	0.019	0.014	0.026
CoxSLIM-KM	0.232	0.225	0.175	0.644	0.670	0.701	0.024	0.021	0.033
				25% + Censoring					
RiskSLIM	0.229	0.244	0.284	0.644	0.641	0.663	0.060	0.139	0.341
RiskSLIM-KM	0.231	0.233	0.195	0.644	0.641	0.663	0.081	0.081	0.088
CoxSLIM	0.230	0.228	0.185	0.663	0.678	0.716	0.081	0.079	0.090
CoxSLIM-KM	0.229	0.227	0.185	0.663	0.678	0.712	0.079	0.077	0.078
				50% + Censoring					
RiskSLIM	0.240	0.260	0.329	0.645	0.654	0.674	0.131	0.208	0.408
RiskSLIM-KM	0.245	0.249	0.218	0.645	0.653	0.676	0.175	0.187	0.204
CoxSLIM	0.247	0.248	0.212	0.642	0.669	0.693	0.172	0.184	0.191
CoxSLIM-KM	0.246	0.246	0.214	0.642	0.669	0.692	0.171	0.182	0.188

Table 7.3: Discriminative performance and Calibration of CoxSLIM vs. RiskSLIM on the `support` data. All of the metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring Weighting.

RiskSLIM		CoxSLIM	
Feature	Score	Feature	Score
Primary:head/neck	-1	Primary: bone marrow	1
Laterality: Right	-1	Surgery performed +ve	-1
Surgery performed +ve	-1	AGE<=50	-1
No Benign Tumors	5	AGE>60	1
AGE>70	1	AGE>70	1
AGE>85	1	AGE>80	2

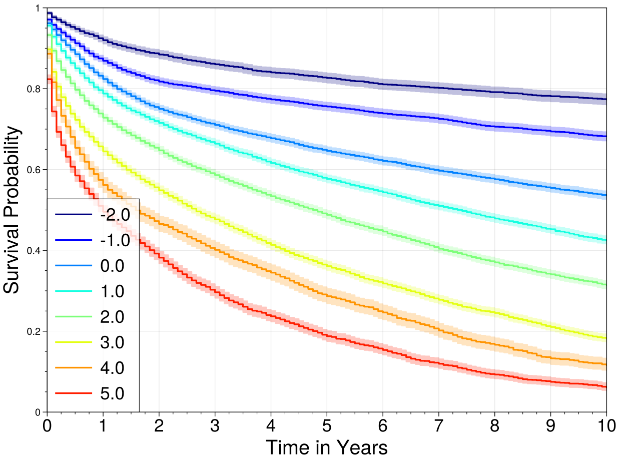
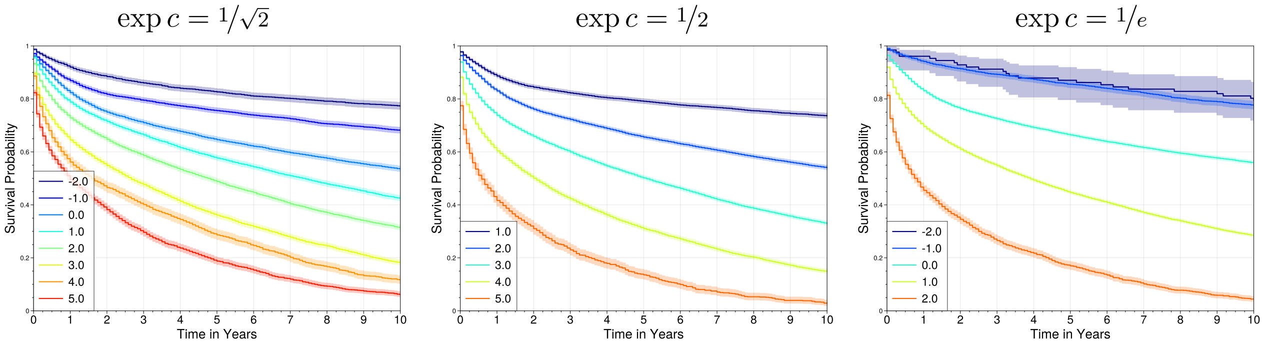


Figure 7.4: **Left**: The learnt integer risk scoring system on `seer-lymphoma`. **Right**: KM-curves stratified by scores assigned by Cox-SLIM. Each curve represents hazard increase of $\sqrt{2}$.


$\exp c = {}^1\!/\!\sqrt{2}$ $\exp c = {}^1\!/\!_2$ $\exp c = {}^1\!/\!_e$

seer-lymphoma									
	Brier Score			**Area Under ROC Curve**			**ECE**		
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year
				No Censoring					
RiskSLIM	0.176	0.262	0.319	0.665	0.687	0.712	0.034	0.204	0.333
RiskSLIM-KM	0.175	0.218	0.205	0.665	0.687	0.712	0.003	0.007	0.004
CoxSLIM	0.175	0.213	0.192	0.680	0.717	0.765	0.022	0.027	0.02
CoxSLIM-KM	0.175	0.212	0.192	0.680	0.717	0.765	0.009	0.007	0.003
				25% + Censoring					
RiskSLIM	0.173	0.270	0.336	0.664	0.685	0.709	0.036	0.244	0.374
RiskSLIM-KM	0.174	0.221	0.212	0.664	0.685	0.709	0.051	0.077	0.074
CoxSLIM	0.174	0.216	0.200	0.678	0.715	0.763	0.053	0.082	0.081
CoxSLIM-KM	0.174	0.215	0.198	0.678	0.715	0.763	0.051	0.076	0.070
				50% + Censoring					
RiskSLIM	0.172	0.283	0.359	0.662	0.682	0.705	0.079	0.295	0.425
RiskSLIM-KM	0.177	0.233	0.231	0.662	0.682	0.705	0.108	0.170	0.178
CoxSLIM	0.176	0.228	0.219	0.675	0.711	0.76	0.108	0.170	0.181
CoxSLIM-KM	0.177	0.228	0.216	0.675	0.711	0.759	0.108	0.167	0.171

Table 7.4: Discriminative performance and Calibration of CoxSLIM vs. RiskSLIM on the `seer-lymphoma` data. All of the metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring Weighting.

Dataset	Description	n	d	5-Yr Survival
`flchain`	effect of light chains on survival	6,524	39	86.54%
`support`	survival post ICU discharge	9,105	90	24.55%
`seer-lymphoma`	lymphoma/leukemia survival	60,486	55	54.16%

Table 7.5: Datasets used in Section 7.3. n and d denote the number of examples and features in each dataset, respectively. All datasets are publicly available.

Methods For all the datasets we experiment with 70% of the data as the training set and test the performance of the learnt scoring system on the remaining 30% held out set. The set of possible model coefficients $\mathcal{W}$ are restricted to be between $\{-5, ..., 5\}$ and we fix the maximum size of the model R^{max} to be 6. We compare the performance of our proposed CoxSLIM to RiskSLIM (Ustun & Rudin, 2017, 2019) involving learning of integer scoring systems with binary outcomes.

For RiskSLIM, our horizon of a positive outcome is determined using best judgment for each dataset. For `flchain` and `seer-lymphoma` we consider an outcome to be positive if a patient survived the first 1 year from entry into the study. For `support` we consider survival post the first 6-months from discharge as a positive outcome. In order to demonstrate the superiority of CoxSLIM in the presence of higher amounts of censoring we also experiment by synthetically augmenting the amount of censoring in the above datasets by randomly sampling a certain percentage of the uncensored individuals and censoring there event times drawn from a uniform distribution.

Evaluation Our goal is to train a risk score that is sparse, has small integer coefficients, and performs well in terms of the following measures:

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $s(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\text{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|s(t)) - s(t)\big|\big]$$

by conditioning on the set of all possible estimated integer risk scores $\{\boldsymbol{w}^\top \boldsymbol{x} | \boldsymbol{x} \in \mathcal{D}\}$.

Brier Score (BS): The Brier Score involves computing the Mean Squared Error around the binary forecast of survival at a certain event quantile of interest. Brier Score is a proper scoring rule and can be decomposed into components that measure both discriminative performance and calibration.

$$\text{BS}(t) = \mathbb{E}_{\mathcal{D}}\big[\big(\mathbb{1}\{T > t\} - \widehat{\mathbb{P}}(T > t|X)\big)^2\big]$$

Area under ROC Curve (AUC): Involves treating the survival analysis problem as binary classification at different horizons of event times and computing the corresponding area under the curve.

Each of the metrics described above are adjusted for censoring by using standard Thompson-Horvitz style Inverse Propensity of Censoring Weights (IPCW) estimates learnt with a Kaplan-Meier estimator over the censoring times.

Results In this section we describe the results of the proposed CoxSLIM approach vs RiskSLIM in terms of both Calibratation and Discriminative performance. We found that for all three datasets we were able to solve the CoxSLIM problem to optimality within a running time of < 10 minutes with 8 parallel threads using the IBM CPLEX solver, while RiskSLIM took longer to converge. CoxSLIM consistently had better discrimination performance as evidenced from the higher area under ROC scores at different horizons of time Tables 7.2, 7.3 and 7.4. Further While RiskSLIM was calibrated at the horizon it was trained on, calibration deteriorated significantly at longer time horizons. We also report the Kaplan-Meier adjusted survival curves for both CoxSLIM and RiskSLIM. We find that KM adjustment greatly improves calibration, but the general trend still favours CoxSLIM. For completeness, we also present the sparse integer scoring system outputs from CoxSLIM and RiskSLIM in Figures 7.2, 7.3 and 7.4 as well as the corresponding survival curves stratified by the estimated risk score by CoxSLIM.

In order to better present the qualitative differences between various different bases for the risk scoring models, we also experiment with different values of the scaling base c coefficient and find that smaller values lead to better stratification with more granular scoring stages, however this comes at a cost of calibration.

7.4 Conclusion

We proposed **CoxSLIM** an integer risk scoring method that allows learning highly interpretable scoring systems involving censored time-to-event outcomes in a data driven manner. Our formulation involves a mixed integer program and allows for specification of several operational constraints helping improve utility of the learnt scoring systems. We benchmark the performance of CoxSLIM to existing solutions and found that across multiple real world risk estimation studies, CoxSLIM recovered highly calibrated risk scoring systems with improved discriminative power.

In the future we aim to extend CoxSLIM to situations where we are interested in stratifying patients with heterogeneous treatment response. Such a model could involve interaction effects with treatment assignment as well as additional constraints that ensure conditions for faithful recovery in counterfactual inference such as positivity are respected.

Discussion and Future Work

In this thesis, I presented multiple modeling approaches for time-to-event data. Part I of the thesis focused on flexible and accurate survival modeling involving deep representation learning and probabilistic inference. Part II on the other hand focused on making survival predictions **actionable** as tools for decision support by recovering subgroups that respond differentially to interventions. Finally, in Part III I proposed using advanced numerical optimization techniques to extract sparse and frugal hypotheses from time-to-event data. Overall through my thesis contributions, I hope to engender discussion both amongst the Machine Learning and Healthcare community on the complex challenges that are involved when modeling time-to-event data which appears ubiquitously across multiple areas of application.

This thesis, however, has only scratched the surface of the multitude of challenges that stem from time-to-event data, and multiple avenues are open for furthering academic research along these directions. Some potentially promising future directions include:

Heterogeneous Treatment Effects and Subgroup Discovery in the Presence of Competing Events Clinical studies often involve multiple outcomes that may respond differently to an intervention. Such outcomes might potentially be non-independent competing risks and subgroup discovery and heterogeneous treatment effect estimation requires careful assumptions in order to correctly estimate counterfactual phenotypes.

Integer Risk Scoring for Heterogeneous Treatment Effects Typical integer risk scoring methods as presented in Chapter 7 typically involve observational outcomes. In our contributions we extend this to the setting where outcomes are times-to-event, however for actionable clinical decision support it would be imperative for these calculators to represent differential treatment responses. In such settings, one must ensure careful covariate balance, especially when models are trained on observational data.

Uncertainty Estimation with Time-to-Events In a large number of applications, input covariates may be unavailable, subject to measurement error, or may not be reflective of the outcome of interest. In such scenarios, data-driven decision-making could benefit from individual level confidence intervals for a specific prediction. A large amount of recent deep learning research has focused on uncertainty estimation with ensembling and bayesian approaches but has restricted focus to scalar or temporal outcomes. In the case of time-to-event models such intervals are functions of time making analysis harder.

Time-to-Recurrent Events Standard practice in clinical medicine and reliability analysis is to perform inference over the time-to-first major adverse event, such as Death or Stroke or failure of a component. In as much, the focus of this thesis was in Time-to-the-First event. While there are extensions of the models proposed in this thesis in the situation with multiple recurrent events, academic literature on this topic is sparse, especially when estimating treatment effects. Potential extensions could involve counting process based semi-parametric extensions of the Cox model or parametric models defined on distributions that have a circular support.

Other than the methodological extensions listed above there are open opportunities for applications of the thesis research across multiple real world risk estimation problems, including:

Healthcare Disease staging with complex multi-modal data, including genetic markers, histopathology, natural language, imaging such as CT scans and Chest X-rays. Another promising research direction is the discovery of treatment effect heterogeneity across a wide spectrum of randomized trials especially but not limited to epidemiology, psychology, and oncology. Finally, there is ample opportunity to validate existing standard clinical risk scoring methods temporally across multiple time horizons of clinical interest and improve upon them especially within minority groups utilizing methodologies presented in this thesis.

Discourse and Conversational Analysis Computation social scientists are often interested in modeling Discourse and Human-Machine Interaction from the perspective of improving the longevity of discourse by keeping participants engaged. Such problems are typically modeled as binary but can be re-framed as modeling time-to-events of potential interest such as the user dropping out of conversation or time till the conversation involves the onset of a certain topic or level of sentiment.

Policy Impact Evaluation and Development Economics Randomized experiments and observational studies in policy impact evaluation and development economics often involve outcomes that are temporal. Estimators proposed in this thesis can be directly applied to such studies with higher statistical power.

Finance and Econometrics Large number of problems arising in mathematical finance such as the time till a financial instrument achieves a certain value or time before an analyst re-adjusts there earnings expectation for a certain financial entity are inherently temporal and maybe better modeled as time-to-event.

Predictive Maintenance and Reliability Engineering In the case of reliability and predictive maintenance presents an opportunity to extend proposed methods to incorporate econometric notions of reward or utility beyond just survival time.

Appendices

Appendix A

Appendix to Chapter 1

A.1 Loss Function Formulation

At test time, DSM describes the survival function of the test individual as a weighted mixture of K survival distribution primitives, and the K weights are a softmax over the output of a neural network. The loss function of DSM is designed to handle both the censored and uncensored data.

Uncensored Loss. The maximum likelihood estimator for the uncensored data can be written as

$$\begin{aligned}
\ln \mathbb{P}(\mathcal{D}_U|\mathbf{\Theta}) &= \ln \left(\prod_{i=1}^{|\mathcal{D}|} \mathbb{P}(T = t_i | X = \mathbf{x}_i, \mathbf{\Theta}) \right) \\
&= \sum_{i=1}^{|\mathcal{D}|} \ln \left(\sum_{k=1}^{K} \mathbb{P}(T = t_i | Z, \beta_k, \eta_k) \mathbb{P}(Z | X = \mathbf{x}_i, \boldsymbol{w}) \right) \\
&= \sum_{i=1}^{|\mathcal{D}|} \ln \left(\underset{Z \sim (\cdot|\mathbf{x}_i, \boldsymbol{w})}{\mathbb{E}} [\mathbb{P}(T = t_i | Z, \beta_k, \eta_k)] \right) \\
&\text{(Applying Jensen's Inequality)} \\
&\geq \sum_{i=1}^{|\mathcal{D}|} \left(\underset{Z \sim (\cdot|\mathbf{x}_i, \boldsymbol{w})}{\mathbb{E}} [\ln \mathbb{P}(T = t_i | Z, \beta_k, \eta_k)] \right) \\
&= \sum_{i=1}^{|\mathcal{D}|} \left(\textsc{Softmax}_{(K)} \big(\ln f(t_i | \beta_{k_i}, \eta_{k_i}) \big) \right) \\
&\triangleq \textbf{ELBO}_U(\Theta)
\end{aligned}$$

Here $\mathbf{x}_i$ are the input covariates of the i-th observation, and $f(t)$ is the probability density function (PDF) of the primitive distribution. β_{k_i} and η_{k_i} for the i-th observation are parameterized as

$$\beta_{k_i} = \tilde{\beta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\zeta}),$$

$$\eta_{k_i} = \tilde{\eta}_k + \texttt{act}(\Phi_\theta(\boldsymbol{x}_i)^\top \boldsymbol{\xi})$$

where $\texttt{act}(\cdot)$ is the SELU activation function if Weibull is used as the primitive distribution and the Tanh activation function if Log-Normal is used as the primitive distribution. $\mathbf{\Phi}(.)$ is a Multilayer Perceptron.

Censoring Loss. As above, the lower bound of the censored observations can be written as

$$\begin{aligned}\ln \mathbb{P}(\mathcal{D}_C|\Theta) &= \ln \left(\prod_{i=1}^{|\mathcal{D}|} \mathbb{P}(T > t_i | X = \mathbf{x}_i, \Theta) \right) \\ &\geq \sum_{i=1}^{|\mathcal{D}|} \left(\underset{Z \sim (\cdot|\mathbf{x}_i, w)}{\mathbb{E}} [\ln \mathbb{P}(T > t_i | Z, \beta_k, \eta_k)] \right) \\ &= \sum_{i=1}^{|\mathcal{D}|} \left(\textsc{Softmax}_{(K)} \big(\ln S(t_i | \beta_{k_i}, \eta_{k_i}) \big) \right) \\ &\triangleq \textbf{ELBO}_C(\Theta)\end{aligned}$$

$S(t)$ is the survival function of the primitive distribution.

For the scenario of M *competing risks*, $\textbf{ELBO}_{U_m}(\Theta)$ and $\textbf{ELBO}_{C_m}(\Theta)$ are computed for the m-th competing risk by treating other events as censoring. The total loss can be written as

$$\mathcal{L} = \sum_{m=1}^{M} \textbf{ELBO}_{U_m}(\Theta) + \alpha \cdot \textbf{ELBO}_{C_m}(\Theta) + \mathcal{L}_{\text{prior}_n}$$

A.1.1 Results in Tabular Format

In this section, we provide the comparison of the performances of DSM with the baseline approaches using C^{td} at different event time horizons. The C^{td} was evaluated at the 25%, 50%, 75% quantiles of event times. The mean and the 90% confidence interval of the C^{td} were computed using 5-fold cross validation.

The results of two single-risk datasets, SUPPORT and METABRIC, are respectively shown in Table A.2 and Table A.4. To investigate the models' robustness to censoring, we also artificially increased the amount of censoring in training set by censoring a randomly chosen subset which included 25% or 50% of the originally uncensored observations in the training data, on both SUPPORT and METABRIC. The results of added censoring are also shown.

The results of two datasets with competing risks, SYNTHETIC and SEER, are shown respectively in Table A.5 and Table A.6. cs-CPH and cs-RSF stand for the cause-specific versions of CPH and RSF models.

A.2 Hyperparameter Tuning for the Baselines

We compared the performance of *Deep Survival Machines* (DSM) to several competing baseline approaches. In this section, we provide details of the hyperparameter tuning for each baseline approach. The hyperparameters tuned for Random Survival Forests (RSF) (Ishwaran et al., 2008) and *DeepHit* (Lee et al., 2018) are described as below, and the best set of hyperparameters was chosen based on the time-dependent Concordance-Index C^{td} (Antolini et al., 2005) on the validation set. For Cox Proportional Hazards (CPH) model (Cox, 1972), we used the default settings in the python `PySurvival` library.[1] For *DeepSurv* (Katzman et al., 2018), We directly used the hyperparameters provided in the *DeepSurv* GitHub repository.[2] For Fine-Gray (FG) model (Fine & Gray, 1999), we used the default settings in the R `cmprsk` package.[3]

[1] https://square.github.io/pysurvival/

[2] https://github.com/jaredleekatzman/DeepSurv/tree/master/experiments/deepsurv

[3] https://cran.r-project.org/web/packages/cmprsk/cmprsk.pdf

Table A.1: Time-dependent Concordance-Index on SUPPORT dataset at different quantiles of event times for different levels of censoring.

Default Censoring.

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.794 ± 0.002	0.756 ± 0.004	0.733 ± 0.003
DeepSurv	0.804 ± 0.004	0.767 ± 0.003	0.746 ± 0.003
DeepHit	0.822 ± 0.003	0.778 ± 0.002	0.701 ± 0.007
RSF	0.830 ± 0.005	0.779 ± 0.004	0.729 ± 0.007
DSM	0.832 ± 0.002	0.788 ± 0.003	0.750 ± 0.003

25%+ Censoring.

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.796 ± 0.003	0.754 ± 0.002	0.727 ± 0.003
DeepSurv	0.800 ± 0.004	0.762 ± 0.002	0.738 ± 0.003
DeepHit	0.813 ± 0.004	0.770 ± 0.004	0.711 ± 0.008
RSF	0.830 ± 0.004	0.774 ± 0.001	0.723 ± 0.005
DSM	0.831 ± 0.002	0.783 ± 0.003	0.742 ± 0.003

50%+ Censoring

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.793 ± 0.006	0.750 ± 0.004	0.721 ± 0.006
DeepSurv	0.795 ± 0.004	0.756 ± 0.004	0.731 ± 0.003
DeepHit	0.814 ± 0.004	0.771 ± 0.005	0.709 ± 0.006
RSF	0.827 ± 0.002	0.770 ± 0.003	0.716 ± 0.005
DSM	0.828 ± 0.002	0.778 ± 0.004	0.735 ± 0.004

Random Survival Forests (RSF): The number of trees in the forest was selected from $[10, 20, 50, 100]$ and the maximum depth of the trees was set to 4.

DeepHit (DH): We followed the experiment settings provided in the *DeepHit* GitHub repository.[4] The number of layers in the shared sub-network and in each cause-specific (CS) sub-network was selected from

[4] https://github.com/chl8856/DeepHit/

Table A.2: Brier Score on SUPPORT dataset at different quantiles of event times for different levels of censoring.

Default Censoring.

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.115 ± 0.002	0.171 ± 0.002	0.193 ± 0.001
DeepSurv	0.110 ± 0.002	0.165 ± 0.002	0.186 ± 0.001
DeepHit	0.164 ± 0.003	0.285 ± 0.009	0.342 ± 0.020
RSF	0.118 ± 0.003	0.181 ± 0.002	0.203 ± 0.002
DSM	0.107 ± 0.002	0.159 ± 0.002	0.188 ± 0.002

25%+ Censoring.

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.125 ± 0.007	0.190 ± 0.003	0.224 ± 0.001
DeepSurv	0.118 ± 0.003	0.181 ± 0.003	0.213 ± 0.001
DeepHit	0.178 ± 0.004	0.336 ± 0.006	0.436 ± 0.006
RSF	0.128 ± 0.003	0.204 ± 0.003	0.243 ± 0.001
DSM	0.115 ± 0.002	0.176 ± 0.003	0.211 ± 0.002

50%+ Censoring

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.140 ± 0.004	0.225 ± 0.005	0.278 ± 0.005
DeepSurv	0.131 ± 0.004	0.210 ± 0.004	0.259 ± 0.002
DeepHit	0.192 ± 0.005	0.365 ± 0.010	0.481 ± 0.017
RSF	0.143 ± 0.004	0.240 ± 0.004	0.301 ± 0.002
DSM	0.126 ± 0.003	0.202 ± 0.004	0.244 ± 0.002

$[1, 2, 3, 5]$; the number of nodes in each layer was selected from $[50, 100, 200, 300]$; the activation function was selected from [RELU, ELU, Tanh]; and the coefficients α_k for trading off the ranking losses of the k competing risks were chosen from $[0.1, 0.5, 1.0, 3.0, 5.0]$. We generated 10 settings by randomly sampling each hyperparameter from the given lists of candidates 10 times, and selected the best set of hyperparameters which had the highest validation C^{td}. The hyperparameters for each dataset are shown in Table A.7.

Table A.3: Time-dependent Concordance-Index on METABRIC dataset at different quantiles of event times for different levels of censoring.

Default Censoring.

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.620 ± 0.016	0.620 ± 0.013	0.629 ± 0.010
DeepSurv	0.634 ± 0.018	0.635 ± 0.011	0.637 ± 0.0010
DeepHit	0.691 ± 0.016	0.626 ± 0.011	0.585 ± 0.006
RSF	0.713 ± 0.017	0.673 ± 0.010	0.644 ± 0.010
DSM	0.720 ± 0.0116	0.676 ± 0.009	0.652 ± 0.009

25%+ Censoring.

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.607 ± 0.015	0.616 ± 0.012	0.628 ± 0.010
DeepSurv	0.619 ± 0.018	0.627 ± 0.012	0.633 ± 0.011
DeepHit	0.688 ± 0.020	0.618 ± 0.015	0.593 ± 0.002
RSF	0.710 ± 0.016	0.668 ± 0.009	0.642 ± 0.010
DSM	0.712 ± 0.010	0.671 ± 0.010	0.645 ± 0.010

50%+ Censoring

Models	Time-dependent Concordance-Index (C^{td})		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.603 ± 0.014	0.613 ± 0.012	0.630 ± 0.010
DeepSurv	0.617 ± 0.018	0.612 ± 0.014	0.622 ± 0.012
DeepHit	0.666 ± 0.020	0.601 ± 0.011	0.583 ± 0.006
RSF	0.700 ± 0.016	0.662 ± 0.010	0.635 ± 0.010
DSM	0.708 ± 0.009	0.664 ± 0.010	0.638 ± 0.010

A.3 Data Preprocessing

Both SUPPORT (single risk) and SEER (competing risks) datasets have missing values. The number and percentage of instances with missing values in each feature of the two datasets are provided in Table A.8 and Table A.9.

Table A.4: Brier Score on METABRIC dataset at different quantiles of event times for different levels of censoring.

Default Censoring.

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.127 ± 0.006	0.209 ± 0.003	0.249 ± 0.002
DeepSurv	0.124 ± 0.006	0.195 ± 0.004	0.226 ± 0.007
DeepHit	0.137 ± 0.003	0.239 ± 0.002	0.284 ± 0.004
RSF	0.119 ± 0.006	0.193 ± 0.003	0.227 ± 0.004
DSM	0.116 ± 0.006	0.187 ± 0.003	0.222 ± 0.005

25%+ Censoring.

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.135 ± 0.007	0.230 ± 0.003	0.288 ± 0.003
DeepSurv	0.131 ± 0.006	0.216 ± 0.004	0.260 ± 0.007
DeepHit	0.149 ± 0.003	0.271 ± 0.002	0.335 ± 0.002
RSF	0.127 ± 0.007	0.214 ± 0.003	0.263 ± 0.004
DSM	0.125 ± 0.007	0.210 ± 0.003	0.264 ± 0.005

50%+ Censoring

Models	Brier Score		
	Quantiles of Event Times		
	25%	50%	75%
CPH	0.148 ± 0.009	0.264 ± 0.004	0.352 ± 0.005
DeepSurv	0.145 ± 0.008	0.251 ± 0.006	0.322 ± 0.009
DeepHit	0.166 ± 0.003	0.321 ± 0.004	0.418 ± 0.010
RSF	0.141 ± 0.008	0.248 ± 0.004	0.324 ± 0.006
DSM	0.137 ± 0.008	0.238 ± 0.003	0.305 ± 0.007

21 out of the 30 features in SUPPORT have missing values. For the following 7 features, we used the suggested normal values[5]for imputation, which are *Serum Albumin (Day 3)*: 3.5, *PaO2/(.01*FiO2) (Day 3)*: 333.3, *Bilirubin (Day 3)*: 1.01, *Serum Creatinine (Day 3)*: 1.01, *BUN (Day 3)*: 6.51, *White Blood Cell Count (Day 3)*: 9, *Urine Output (Day 3)*: $2,502$, as these values are found to be working well in baseline physiologic data imputation.

Table A.5: C^{td} for competing risks on SYNTHETIC data.

(a) Event 1.

Models	Quantiles of Event Times			
	25%	50%	75%	100%
cs-CPH	0.570 ± 0.016	0.553 ± 0.012	0.542 ± 0.009	0.528 ± 0.009
FG	0.610 ± 0.001	0.587 ± 0.002	0.568 ± 0.003	0.548 ± 0.004
cs-RSF	0.680 ± 0.011	0.663 ± 0.009	0.644 ± 0.007	0.569 ± 0.005
DeepHit	0.796 ± 0.009	0.765 ± 0.005	0.726 ± 0.005	0.635 ± 0.007
DSM	0.798 ± 0.010	0.759 ± 0.008	0.724 ± 0.004	0.670 ± 0.005

(b) Event 2.

Models	Quantiles of Event Times			
	25%	50%	75%	100%
cs-CPH	0.591 ± 0.024	0.563 ± 0.018	0.548 ± 0.015	0.533 ± 0.013
FG	0.634 ± 0.008	0.595 ± 0.008	0.575 ± 0.007	0.552 ± 0.005
cs-RSF	0.687 ± 0.007	0.663 ± 0.011	0.638 ± 0.010	0.571 ± 0.011
DeepHit	0.803 ± 0.004	0.761 ± 0.005	0.726 ± 0.004	0.618 ± 0.014
DSM	0.803 ± 0.011	0.762 ± 0.009	0.729 ± 0.006	0.672 ± 0.006

5 out of the 21 patient covariates in SEER have missing data. For these 5 features in SEER, as well as the remaining 14 features in SUPPORT, we followed the data imputation practice used in Lee et al. (2018): missing data was imputed by the mean for numeric features and the mode for categorical features. We first divided the data into train/validation subsets, and the missing values in each subset were imputed with the mean/mode values of the train set.

A.4 Benchmarking Machine Specifications

All experiments except the experiments for *DeepHit* were run on a Linux version 3.10.0-1062.9.1.el7.x86_64 machine with an Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz (8-core CPU) and RAM 32 GB. The experiments for *DeepHit* were run on a TITAN X (Pascal) GPU cluster (1 GPU) with an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz (32-core CPU), NVIDIA driver version 418.74 and CUDA 10.1.

Table A.6: C^{td} for competing risks on SEER data.

(a) Breast Cancer (BC).

Models	Quantiles of Event Times			
	25%	50%	75%	100%
cs-CPH	0.891 ± 0.004	0.849 ± 0.004	0.827 ± 0.004	0.807 ± 0.003
FG	0.876 ± 0.002	0.840 ± 0.002	0.816 ± 0.002	0.798 ± 0.002
cs-RSF	0.897 ± 0.002	0.852 ± 0.005	0.823 ± 0.004	0.801 ± 0.003
DeepHit	0.899 ± 0.001	0.863 ± 0.003	0.838 ± 0.003	0.815 ± 0.002
DSM	0.904 ± 0.001	0.861 ± 0.003	0.840 ± 0.004	0.820 ± 0.003

(b) Cardiovascular Disease (CVD).

Models	Quantiles of Event Times			
	25%	50%	75%	100%
cs-CPH	0.897 ± 0.007	0.890 ± 0.006	0.891 ± 0.002	0.889 ± 0.002
FG	0.842 ± 0.008	0.844 ± 0.006	0.854 ± 0.004	0.857 ± 0.004
cs-RSF	0.845 ± 0.006	0.832 ± 0.008	0.823 ± 0.008	0.816 ± 0.009
DeepHit	0.902 ± 0.003	0.893 ± 0.003	0.893 ± 0.001	0.889 ± 0.001
DSM	0.894 ± 0.004	0.893 ± 0.004	0.893 ± 0.001	0.890 ± 0.001

Table A.7: The hyperparameters of *DeepHit* for each dataset.

Dataset	Type	Shared Sub-network		CS Sub-network		Activation	α
		No. Layers	No. Nodes	No. Layers	No. Nodes		
SUPPORT	Single Risk	3	100	3	300	eLU	0.1
METABRIC	Single Risk	3	100	1	100	Tanh	5.0
SYNTHETIC	Competing Risks	3	300	2	50	eLU	0.1
SEER	Competing Risks	1	100	2	50	eLU	0.5

Table A.8: Statistics of the missing values in each feature of SUPPORT dataset.

Feature Name	No. Instances	Feature Name	No. Instances
Years of Education	1, 634 (17.9%)	Income	2, 982 (32.8%)
SUPPORT Coma Score	1 (0.0%)	Average TISS (Days 3-25)	82 (0.9%)
Race	42 (0.5%)	Mean Arterial Blood Pressure (Day 3)	1 (0.0%)
White Blood Cell Count (Day 3)	212 (2.3%)	Heart Rate (Day 3)	1 (0.0%)
Respiration Rate (Day 3)	1 (0.0%)	Temperature (Day 3)	1 (0.0%)
PaO2/(.01*FiO2) (Day 3)	2, 325 (25.5%)	Serum Albumin (Day 3)	3, 372 (37.0%)
Bilirubin (Day 3)	2, 601 (28.6%)	Serum Creatinine (Day 3)	67 (0.7%)
Serum Sodium (Day 3)	1 (0.0%)	Serum pH Arterial (Day 3)	2, 284 (25.1%)
Glucose (Day 3)	4, 500 (49.4%)	BUN (Day 3)	4, 352 (47.8%)
Urine Output (Day 3)	4, 862 (53.4%)	ADL Patient (Day 3)	5, 641 (62.0%)
ADL Surrogate (Day 3)	2, 867 (31.5%)		

Table A.9: Statistics of the missing values in each feature of SEER dataset.

Feature Name	No. Instances	Feature Name	No. Instances
Surgery Type	36, 524 (55.8%)	Surgery-Beyond Primary Site	59, 295 (90.6%)
Surgery-Primary Site	28, 957 (44.2%)	Surgery-Distant Lymph Nodes/ Other Tissues	35, 143 (53.7%)
No. Lymph Nodes Examined	35, 143 (53.7%)		

Appendix B

Appendix to Chapter 3

B.1 Additional details on DCM implementation

B.1.1 Non Applicability of the Partial Likelihood for the Proposed Model

In this section, we demonstrate that we cannot directly maximize the partial likelihood to learn our model. In the case of the Cox model, the hazard rate for an individual with covariates $\boldsymbol{x}_i$ at time t, $\boldsymbol{\lambda}(t|\boldsymbol{x}_i)$ is given as

$$\boldsymbol{\lambda}(t|\boldsymbol{x}_i) = \boldsymbol{\lambda}_0(t)\exp(f(\beta, \boldsymbol{x}_i)).$$

Here, $\boldsymbol{\lambda}_0(t)$ is the baseline hazard. Now the partial likelihood $\mathcal{PL}(\boldsymbol{\theta})$ is defined as

$$\begin{aligned}\mathcal{PL}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\boldsymbol{\lambda}(t|\boldsymbol{x}_i)}{\sum\limits_{j\in\mathcal{R}(t_i)} \boldsymbol{\lambda}(t|\boldsymbol{x}_j)} &= \prod_{i:\delta_i=1} \frac{\cancel{\boldsymbol{\lambda}_0(t)}\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathcal{R}(t_i)} \cancel{\boldsymbol{\lambda}_0(t)}\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)} \qquad \text{(B.1)}\\ &= \prod_{i:\delta_i=1} \frac{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathcal{R}(t_i)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)}. \qquad \text{(B.2)}\end{aligned}$$

Under our model, the hazard rate for an individual with covariates $\boldsymbol{x}_i$ at time t, $\boldsymbol{\lambda}(t|\boldsymbol{x}_i)$ is given as

$$\boldsymbol{\lambda}(\cdot|\boldsymbol{x}_i) = \frac{\mathbb{P}(t|\boldsymbol{x}_i)}{\boldsymbol{S}(t|\boldsymbol{x}_i)} = \frac{\sum\limits_k \mathbb{P}(t|\boldsymbol{x}_i, Z=k)\mathbb{P}(Z=k|\boldsymbol{x}_i)}{\sum\limits_k \boldsymbol{S}(t|\boldsymbol{x}_i, Z=k)\mathbb{P}(Z=k|\boldsymbol{x}_i)}$$

Clearly, we do not have the proportional hazards form for DCM and so cannot directly optimize the Partial Likelihood independent of the baseline hazard rate.

B.1.2 Spline Estimates

We want to extract the probabilities estimates $\mathbb{P}(T|Z,X)$ in order to compute the posterior $\mathbb{P}(Z|T,X) \propto \mathbb{P}(T|Z,X)$ for the uncensored observations. We only have access to the estimated survival function from the

Breslow's estimate, $\widehat{\boldsymbol{S}}(T > t|X = \boldsymbol{x}_i)$.

$$\begin{aligned}
\mathbb{P}(T > t|X = \boldsymbol{x}_i, Z = k) &= 1 - \mathbb{P}(T \leq t|X = \boldsymbol{x}_i, Z = k)\\
&= 1 - \text{cdf}(T \leq t|X = \boldsymbol{x}_i, Z = k)\\
\text{Now, } \text{cdf}(T \leq t|X = \boldsymbol{x}_i, Z = k) &= 1 - \mathbb{P}(T > t|X = \boldsymbol{x}_i, Z = k)\\
\frac{\partial}{\partial t}\text{cdf}(T \leq t|X = \boldsymbol{x}_i, Z = k) &= \frac{\partial}{\partial t}\Big(1 - \mathbb{P}(T > t|X = \boldsymbol{x}_i, Z = k)\Big) \qquad \text{[taking derivative wrt. } t\text{]}\\
\implies \text{pdf}(T = t|X = \boldsymbol{x}_i, Z = k) &= -\frac{\partial}{\partial t}\mathbb{P}(T > t|X = \boldsymbol{x}_i, Z = k)\\
&= -\frac{\partial}{\partial t}\boldsymbol{S}_k(t)^{\exp(f_k(\boldsymbol{\theta};\boldsymbol{x}_i))}
\end{aligned}$$

Here pdf$(\cdot)$ and cdf$(\cdot)$ are the probability density and the cumulative density functions respectively. Now replacing the baseline survival function $\boldsymbol{S}_k(.)$ with the interpolated spline estimate, $\widetilde{\boldsymbol{S}}_k(.)$ we get the spline estimate of $\mathbb{P}(T = t|Z, X)$ as

$$\begin{aligned}
\widehat{\mathbb{P}}(T = t|Z, X) &= -\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)^{\exp(f_k(\boldsymbol{\theta};\boldsymbol{x}_i))}\\
&= -\exp\big(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\big)\widetilde{\boldsymbol{S}}_k(t)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)-1}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)\\
&= -\exp\big(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\big)\frac{\widehat{\mathbb{P}}(T > t|\boldsymbol{x}_i, Z = k))}{\widetilde{\boldsymbol{S}}_k(t)}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)
\end{aligned}$$

Here, $\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)$ is the derivative of the baseline survival rate interpolated with a polynomial spline.

B.2 Hyper-Parameter tuning for the Baselines

In this section we specify the hyper parameter choices along with a short description over which we perform grid search for the baselines.

Table B.1: DSM Hyper-parameter Grid

Hyper-parameter	Grid
Outcome Distribution	{ 'Weibull' }
No. Clusters (k)	{ '3', '4' }
No. of Hidden Layers	{ '0', '1', '2' }
Hidden Layer Dim.	{ '50', '100' }
Batch Size	{ '128', '256' }
Learning Rate	{ '1e-4', '1e-3' }
Activation	{ 'SeLU' }

Deep Survival Machines (DSM): The choice of hyper parameters for DSM include the number of underlying survival distributions (k) the choice of each outcome survival distribution, the number of hidden layers and neurons for the representation learning network and the activations. We also tune the learning rate and batch size. The choices of hyperparam values is given in Table B.1.

Table B.2: DHT and FSN Hyper-parameter Grid

Hyper-parameter	Grid
No. of Hidden Layers	{ ‘1’, ‘2’ }
Hidden Layer Dim.	{ ‘50’, ‘100’ }
Batch Size	{ ‘128’, ‘256’ }
Learning Rate	{ ‘1e-4’, ‘1e-3’ }
Activation	{ ‘ReLU’ }

Deep Hit (DHT): For Deep Hit, we tune the the Number of Hidden Layers, dimensionality of the hidden layers and the activation function. We also tune the learning rate and minibatch size. Note that Deep Hit requires grid discretization of the output event time space. For the SUPPORT and FLCHAIN datasets we discretize the output grid by dividing it into bins of $\max(T)$ bins. Since, the SEER is a discrete event time dataset we divide the output grid for Deep Hit into $^{\max(T)}/_{10}$ bins.

Faraggi-Simon Net (FSN)/DeepSurv: Similar to Deep Hit, for FSN we tune the the Number of Hidden Layers, dimensionality of the hidden layers and the activation function. We also tune the learning rate and minibatch size.

Both FSN and DHT were implemented using the pycox (Kvamme et al., 2019) python package. Table B.2 describes the hyper-parameter choices for both DHT and FSN.

Table B.3: RSF Hyper-parameter Grid

Hyper-parameter	Grid
Max Depth	{ ‘5’ }
No. of Trees	{‘50’ }
mtry	{‘sqrt’ ,50 , 75, ‘all’ }
min_node_split	{‘150’ , ‘200’, ‘250’ }

Random Survival Forest (RSF): For the RSF model we tune the number of trees and the maximum depth of each tree using the implementation as paart of the pysurvival Python package (Fotso et al., 2019–). Table B.3 presents the chosen grid parameters.

Table B.4: AFT and CPH Hyper-parameter Grid

Hyper-parameter	Grid
$\ell 2$ Penalty	{ ‘1e-3’, ‘1e-2’, ‘1e-1’ }

For **Cox Proportional Hazards (CPH)** and **Accelerated Failure Time (AFT)** the only hyperamaeter is the $\ell 2$ penalty on the parameters. The grid choice is presented in Table B.4.

B.3 Additional Results

B.3.1 Tabulated Results

In this section we present tabulated results for our experiments for the entire population and the minority demogrpahic on the three datasets.

Tables B.5 and B.6 present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the FLCHAIN dataset, respectively.

Tables B.7 and B.8 present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the SUPPORT dataset, respectively.

Tables B.9 and B.10 and present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the SEER dataset, respectively.

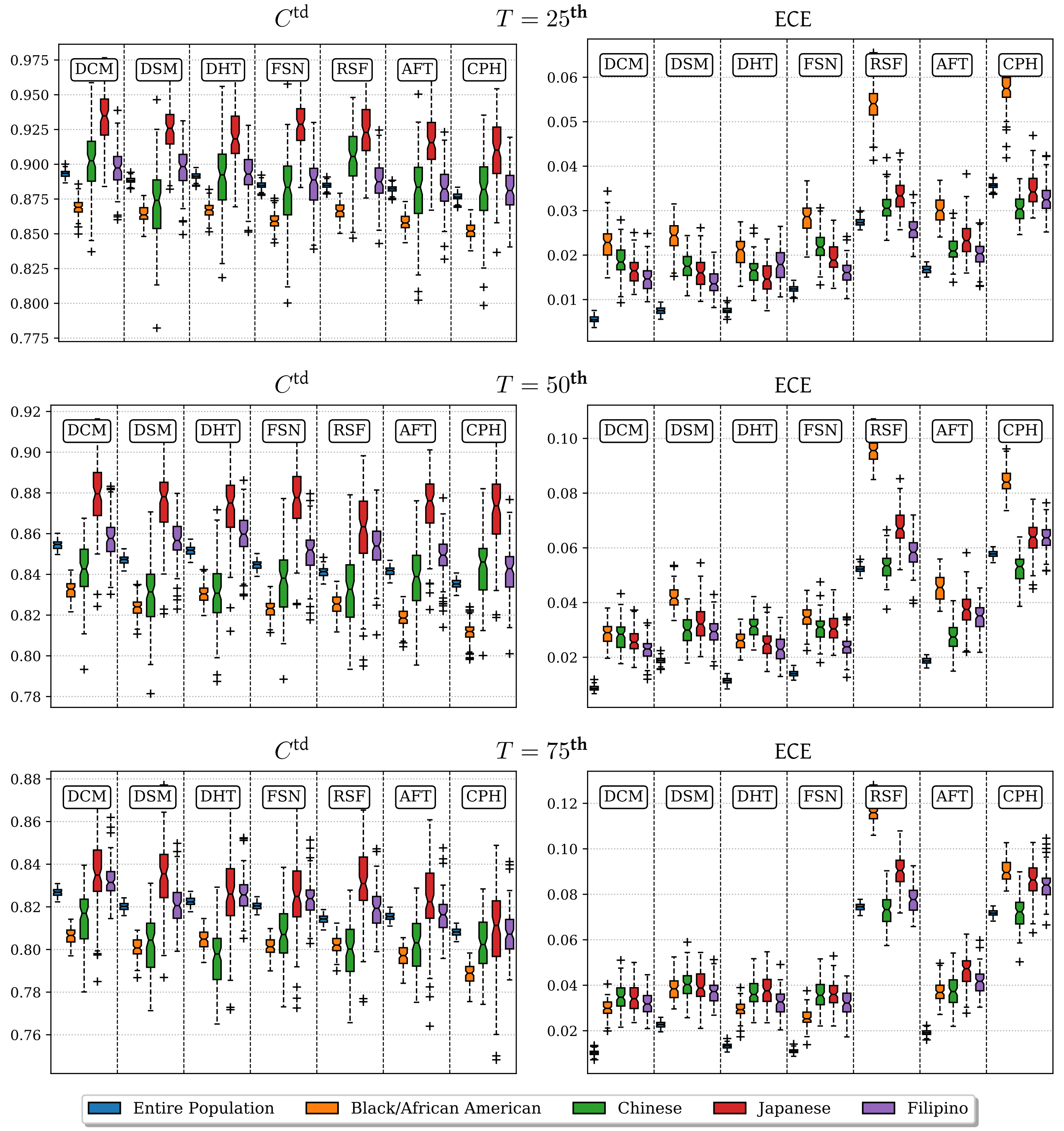


Figure B.1: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at different quantiles of event times for the minority demographics. The rows represents different quantiles at which we evaluate the individual metrics. (Minorities in the dataset are denoted by different colors in the legend)

SEER has multiple minority classes. In Figure B.1 we break results down by the top four largest minorities in the subset of SEER we are working with, 'Black/African American', 'Chinese', 'Japanese' and 'Filipino'.

B.3.2 Unawareness to Group Membership

In the case of CPH and FSN the baseline survival rate is estimated non-parametrically. In Table B.11 we attempt to see how unawareness to the demographic compares to Deep Cox Mixtures. Note that we report Brier Score here as it gives sense of both discrimination and calibration.

B.3.3 Decoupled Survival Models

In the case of CPH and FSN the baseline survival rate is estimated non-parametrically. In Table B.12 we attempt to see how training separate models for each demographic compares to Deep Cox Mixtures by reporting Brier Scores. (Note that we report Brier Score here as it gives sense of both discrimination and calibration.)

B.3.4 Dynamics of the Proposed MCMC EM Algorithm

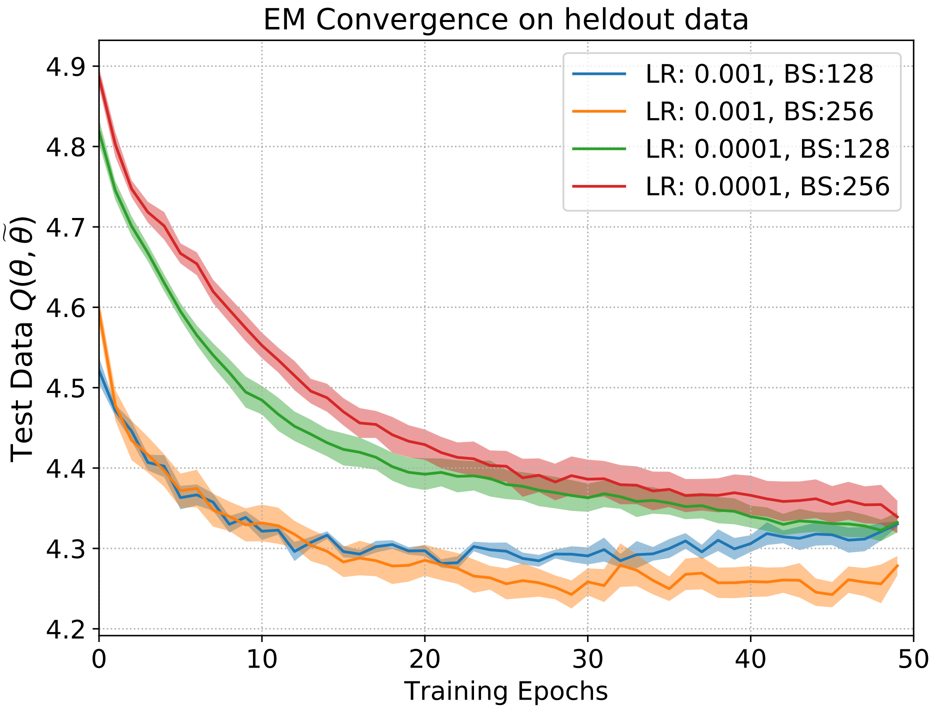


Figure B.2: The estimated $Q(\theta, \widetilde{\theta})$ on a heldout set from the SUPPORT dataset for different hyper-parameters (LR: Learning Rate, BS: Batch size).

Figure B.2 presents the estimated $Q(\theta, \widetilde{\theta})$ function for heldout dataset for the SUPPORT dataset. Empirically our proposed monte carlo EM monotonically decreases the $Q(\theta, \widetilde{\theta})$ suggesting good learning dynamics.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.6621 ± 0.0143	0.6696 ± 0.0110	0.6621 ± 0.0087
AFT	0.7914 ± 0.0107	0.7938 ± 0.0080	0.7911 ± 0.0060
RSF	0.7898 ± 0.0102	0.7908 ± 0.0078	0.7880 ± 0.0059
FSN	0.6353 ± 0.0146	0.6519 ± 0.0104	0.6608 ± 0.0081
DSM	0.8008 ± 0.0100	0.7988 ± 0.0078	0.7937 ± 0.0061
DHT	0.7669 ± 0.0104	0.7666 ± 0.0078	0.7636 ± 0.0059
DCM	0.7991 ± 0.0103	0.7988 ± 0.0077	0.7943 ± 0.0060

AUC(t) (↑)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.6680 ± 0.0149	0.6827 ± 0.0120	0.6821 ± 0.0094
AFT	0.8032 ± 0.0110	0.8170 ± 0.0085	0.8257 ± 0.0063
RSF	0.8015 ± 0.0105	0.8142 ± 0.0083	0.8235 ± 0.0064
FSN	0.6416 ± 0.0150	0.6673 ± 0.0109	0.6904 ± 0.0090
DSM	0.8124 ± 0.0102	0.8218 ± 0.0083	0.8283 ± 0.0066
DHT	0.7771 ± 0.0106	0.7878 ± 0.0082	0.7936 ± 0.0064
DCM	0.8107 ± 0.0106	0.8219 ± 0.0082	0.8291 ± 0.0065

ECE(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0386 ± 0.0031	0.0699 ± 0.0042	0.0992 ± 0.0044
AFT	0.0141 ± 0.0024	0.0216 ± 0.0034	0.0212 ± 0.0034
RSF	0.0155 ± 0.0022	0.0198 ± 0.0027	0.0215 ± 0.0037
FSN	0.0214 ± 0.0027	0.0334 ± 0.0035	0.0381 ± 0.0046
DSM	0.0144 ± 0.0025	0.0214 ± 0.0030	0.0223 ± 0.0029
DHT	0.0283 ± 0.0029	0.0410 ± 0.0036	0.0505 ± 0.0041
DCM	0.0122 ± 0.0024	0.0169 ± 0.0033	0.0200 ± 0.0034

BS(t)(↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0671 ± 0.0027	0.1211 ± 0.0035	0.1665 ± 0.0037
AFT	0.0584 ± 0.0023	0.0991 ± 0.0028	0.1244 ± 0.0025
RSF	0.0603 ± 0.0023	0.1004 ± 0.0027	0.1250 ± 0.0026
FSN	0.0672 ± 0.0026	0.1199 ± 0.0029	0.1589 ± 0.0027
DSM	0.0578 ± 0.0022	0.0975 ± 0.0028	0.1224 ± 0.0026
DHT	0.0631 ± 0.0022	0.1086 ± 0.0026	0.1399 ± 0.0024
DCM	0.0582 ± 0.0023	0.0979 ± 0.0028	0.1228 ± 0.0026

Table B.5: Results for various performance metrics on FLCHAIN (entire population) along with bootstrapped std errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6444 ± 0.0193	0.6692 ± 0.0160	0.6737 ± 0.0124
AFT	0.7822 ± 0.0158	0.7838 ± 0.0112	0.7875 ± 0.0087
RSF	0.7796 ± 0.0147	0.7799 ± 0.0113	0.7830 ± 0.0089
FSN	0.5746 ± 0.0211	0.6014 ± 0.0156	0.6212 ± 0.0131
DSM	0.7849 ± 0.0153	0.7886 ± 0.0113	0.7909 ± 0.0087
DHT	0.7607 ± 0.0153	0.7610 ± 0.0116	0.7631 ± 0.0092
DCM	0.7873 ± 0.0164	0.7893 ± 0.0116	0.7911 ± 0.0091

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6492 ± 0.0202	0.6842 ± 0.0175	0.6983 ± 0.0136
AFT	0.7944 ± 0.0163	0.8069 ± 0.0122	0.8230 ± 0.0095
RSF	0.7918 ± 0.0152	0.8028 ± 0.0124	0.8189 ± 0.0099
FSN	0.5774 ± 0.0219	0.6115 ± 0.0164	0.6477 ± 0.0148
DSM	0.7966 ± 0.0158	0.8118 ± 0.0123	0.8259 ± 0.0095
DHT	0.7710 ± 0.0157	0.7822 ± 0.0127	0.7938 ± 0.0104
DCM	0.7991 ± 0.0169	0.8122 ± 0.0126	0.8265 ± 0.0100

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0378 ± 0.0044	0.0642 ± 0.0056	0.0878 ± 0.0071
AFT	0.0221 ± 0.0035	0.0289 ± 0.0045	0.0329 ± 0.0046
RSF	0.0220 ± 0.0036	0.0330 ± 0.0046	0.0368 ± 0.0053
FSN	0.0325 ± 0.0043	0.0416 ± 0.0059	0.0545 ± 0.0068
DSM	0.0243 ± 0.0038	0.0323 ± 0.0048	0.0347 ± 0.0056
DHT	0.0328 ± 0.0037	0.0411 ± 0.0051	0.0525 ± 0.0056
DCM	0.0209 ± 0.0035	0.0298 ± 0.0054	0.0294 ± 0.0049

BS(t)(↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0693 ± 0.0043	0.1223 ± 0.0053	0.1626 ± 0.0057
AFT	0.0613 ± 0.0035	0.1031 ± 0.0040	0.1262 ± 0.0037
RSF	0.0624 ± 0.0036	0.1041 ± 0.0041	0.1273 ± 0.0038
FSN	0.0715 ± 0.0043	0.1278 ± 0.0051	0.1673 ± 0.0050
DSM	0.0609 ± 0.0035	0.1015 ± 0.0041	0.1244 ± 0.0037
DHT	0.0648 ± 0.0035	0.1107 ± 0.0038	0.1394 ± 0.0039
DCM	0.0607 ± 0.0035	0.1021 ± 0.0042	0.1249 ± 0.0039

Table B.6: Results for various performance metrics on FLCHAIN (minority) along with bootstrapped std errors.

$C^{\text{td}}(t)$ (↑)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.6899 ± 0.0057	0.6713 ± 0.0040	0.6686 ± 0.0034
AFT	0.6826 ± 0.0057	0.6662 ± 0.0040	0.6657 ± 0.0034
RSF	0.7513 ± 0.0063	0.7104 ± 0.0045	0.6751 ± 0.0040
FSN	0.6988 ± 0.0059	0.6779 ± 0.0044	0.6736 ± 0.0037
DSM	0.7459 ± 0.0059	0.7042 ± 0.0038	0.6718 ± 0.0033
DHT	0.7302 ± 0.0067	0.6871 ± 0.0043	0.6575 ± 0.0038
DCM	0.7425 ± 0.0059	0.7057 ± 0.0042	0.6753 ± 0.0036

AUC(t) (↑)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.7011 ± 0.0061	0.6990 ± 0.0049	0.7214 ± 0.0049
AFT	0.6936 ± 0.0061	0.6943 ± 0.0049	0.7209 ± 0.0049
RSF	0.7663 ± 0.0066	0.7379 ± 0.0054	0.7273 ± 0.0054
FSN	0.7091 ± 0.0062	0.7050 ± 0.0052	0.7249 ± 0.0050
DSM	0.7606 ± 0.0063	0.7337 ± 0.0047	0.7236 ± 0.0050
DHT	0.7421 ± 0.0070	0.7123 ± 0.0052	0.7042 ± 0.0052
DCM	0.7576 ± 0.0065	0.7347 ± 0.0049	0.7256 ± 0.0054

ECE(t) (↓)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0201 ± 0.0029	0.0265 ± 0.0038	0.0310 ± 0.0041
AFT	0.0281 ± 0.0031	0.0617 ± 0.0048	0.0402 ± 0.0046
RSF	0.0241 ± 0.0032	0.0368 ± 0.0044	0.0348 ± 0.0041
FSN	0.0220 ± 0.0029	0.0267 ± 0.0036	0.0262 ± 0.0040
DSM	0.0341 ± 0.0033	0.0621 ± 0.0043	0.0315 ± 0.0047
DHT	0.0220 ± 0.0026	0.0351 ± 0.0037	0.0457 ± 0.0044
DCM	0.0179 ± 0.0030	0.0268 ± 0.0038	0.0256 ± 0.0037

BS(t) (↓)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.1334 ± 0.0023	0.1995 ± 0.0019	0.2136 ± 0.0016
AFT	0.1354 ± 0.0025	0.2051 ± 0.0023	0.2147 ± 0.0016
RSF	0.1240 ± 0.0023	0.1899 ± 0.0018	0.2109 ± 0.0017
FSN	0.1315 ± 0.0023	0.1981 ± 0.0020	0.2122 ± 0.0018
DSM	0.1271 ± 0.0024	0.1955 ± 0.0022	0.2130 ± 0.0017
DHT	0.1271 ± 0.0024	0.1971 ± 0.0016	0.2206 ± 0.0014
DCM	0.1258 ± 0.0024	0.1905 ± 0.0020	0.2118 ± 0.0019

Table B.7: Results for various performance metrics on SUPPORT (entire population) along with bootstrapped std. errors.

$C^{\text{td}}(t)$ (↑)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.7161 ± 0.0126	0.6982 ± 0.0089	0.6905 ± 0.0078
AFT	0.7101 ± 0.0126	0.6941 ± 0.0089	0.6883 ± 0.0078
RSF	0.7503 ± 0.0120	0.7198 ± 0.0084	0.6974 ± 0.0084
FSN	0.7203 ± 0.0129	0.7025 ± 0.0090	0.6961 ± 0.0074
DSM	0.7548 ± 0.0132	0.7220 ± 0.0093	0.6939 ± 0.0079
DHT	0.7321 ± 0.0145	0.6943 ± 0.0099	0.6680 ± 0.0088
DCM	0.7570 ± 0.0130	0.7234 ± 0.0089	0.6939 ± 0.0079

AUC(t) (↑)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.7261 ± 0.0127	0.7348 ± 0.0109	0.7446 ± 0.0107
AFT	0.7199 ± 0.0127	0.7311 ± 0.0109	0.7446 ± 0.0108
RSF	0.7667 ± 0.0121	0.7536 ± 0.0106	0.7522 ± 0.0122
FSN	0.7283 ± 0.0128	0.7375 ± 0.0110	0.7518 ± 0.0101
DSM	0.7690 ± 0.0130	0.7594 ± 0.0113	0.7478 ± 0.0109
DHT	0.7400 ± 0.0143	0.7265 ± 0.0123	0.7129 ± 0.0120
DCM	0.7701 ± 0.0129	0.7588 ± 0.0109	0.7424 ± 0.0113

ECE(t) (↓)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0473 ± 0.0071	0.0610 ± 0.0084	0.0685 ± 0.0079
AFT	0.0530 ± 0.0075	0.0891 ± 0.0091	0.0741 ± 0.0085
RSF	0.0401 ± 0.0064	0.0608 ± 0.0077	0.0603 ± 0.0080
FSN	0.0418 ± 0.0067	0.0579 ± 0.0090	0.0601 ± 0.0097
DSM	0.0506 ± 0.0070	0.0818 ± 0.0094	0.0650 ± 0.0087
DHT	0.0483 ± 0.0070	0.0635 ± 0.0087	0.0696 ± 0.0089
DCM	0.0397 ± 0.0059	0.0550 ± 0.0080	0.0561 ± 0.0085

BS(t) (↓)

Model	**Quantiles**		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.1340 ± 0.0050	0.1943 ± 0.0042	0.2069 ± 0.0037
AFT	0.1363 ± 0.0054	0.2026 ± 0.0049	0.2090 ± 0.0039
RSF	0.1263 ± 0.0048	0.1870 ± 0.0039	0.2031 ± 0.0039
FSN	0.1319 ± 0.0051	0.1934 ± 0.0044	0.2037 ± 0.0040
DSM	0.1275 ± 0.0050	0.1919 ± 0.0047	0.2056 ± 0.0040
DHT	0.1298 ± 0.0049	0.1963 ± 0.0039	0.2186 ± 0.0036
DCM	0.1261 ± 0.0048	0.1868 ± 0.0044	0.2073 ± 0.0044

Table B.8: Results for various performance metrics on SUPPORT (minority) along with bootstrapped std. errors.

$C^{\mathrm{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8766 ± 0.0027	0.8354 ± 0.0024	0.8082 ± 0.0020
AFT	0.8823 ± 0.0026	0.8416 ± 0.0024	0.8155 ± 0.0020
RSF	0.8838 ± 0.0025	0.8421 ± 0.0025	0.8153 ± 0.0021
FSN	0.8850 ± 0.0025	0.8447 ± 0.0023	0.8204 ± 0.0019
DHT	0.8915 ± 0.0024	0.8517 ± 0.0024	0.8224 ± 0.0020
DSM	0.8949 ± 0.0022	0.8559 ± 0.0022	0.8281 ± 0.0019
DCM	0.8933 ± 0.0024	0.8550 ± 0.0022	0.8270 ± 0.0019

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8828 ± 0.0028	0.8526 ± 0.0025	0.8337 ± 0.0022
AFT	0.8893 ± 0.0026	0.8596 ± 0.0025	0.8424 ± 0.0021
RSF	0.8899 ± 0.0026	0.8594 ± 0.0026	0.8416 ± 0.0023
FSN	0.8921 ± 0.0026	0.8632 ± 0.0024	0.8477 ± 0.0021
DHT	0.8983 ± 0.0025	0.8701 ± 0.0025	0.8495 ± 0.0022
DSM	0.9022 ± 0.0023	0.8748 ± 0.0023	0.8566 ± 0.0020
DCM	0.9002 ± 0.0025	0.87350 ± 0.0023	0.8552 ± 0.0020

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0356 ± 0.0008	0.0577 ± 0.0012	0.0718 ± 0.0015
AFT	0.0168 ± 0.0008	0.0187 ± 0.0011	0.0192 ± 0.0011
RSF	0.0052 ± 0.0007	0.0092 ± 0.0010	0.0147 ± 0.0013
FSN	0.0124 ± 0.0008	0.0140 ± 0.0011	0.0111 ± 0.0011
DHT	0.0076 ± 0.0008	0.0115 ± 0.0011	0.0133 ± 0.0012
DSM	0.0067 ± 0.0007	0.0211 ± 0.0012	0.0259 ± 0.0014
DCM	0.0055 ± 0.0008	0.0087 ± 0.0010	0.0103 ± 0.0011

BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0501 ± 0.0007	0.0887 ± 0.0009	0.1206 ± 0.0009
AFT	0.0470 ± 0.0006	0.0827 ± 0.0009	0.1107 ± 0.0009
RSF	0.0447 ± 0.0006	0.0802 ± 0.0008	0.1095 ± 0.0010
FSN	0.0462 ± 0.0006	0.0800 ± 0.0008	0.1075 ± 0.0009
DHT	0.0450 ± 0.0006	0.0788 ± 0.0008	0.1074 ± 0.0010
DSM	0.0451 ± 0.0006	0.0797 ± 0.0008	0.1073 ± 0.0009
DCM	0.0450 ± 0.0006	0.0785 ± 0.0008	0.1064 ± 0.0010

Table B.9: Results for various performance metrics on SEER (entire population) along with bootstrapped errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.8804 ± 0.0043	0.8405 ± 0.0039	0.8121 ± 0.0037
AFT	0.8865 ± 0.0042	0.8466 ± 0.0036	0.8204 ± 0.0035
RSF	0.8797 ± 0.0048	0.8379 ± 0.0038	0.8105 ± 0.0035
FSN	0.8870 ± 0.0043	0.8490 ± 0.0038	0.8248 ± 0.0036
DHT	0.8920 ± 0.0039	0.8540 ± 0.0038	0.8255 ± 0.0037
DSM	0.8908 ± 0.0038	0.8506 ± 0.0038	0.8243 ± 0.0036
DCM	0.8933 ± 0.0037	0.8558 ± 0.0036	0.8296 ± 0.0034

AUC(t) (↑)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.8888 ± 0.0043	0.8604 ± 0.0042	0.8398 ± 0.0042
AFT	0.8952 ± 0.0042	0.8676 ± 0.0039	0.8491 ± 0.0040
RSF	0.8867 ± 0.0048	0.8571 ± 0.0041	0.8373 ± 0.0039
FSN	0.8963 ± 0.0043	0.8702 ± 0.0040	0.8538 ± 0.0041
DHT	0.9002 ± 0.0039	0.8754 ± 0.0041	0.8540 ± 0.0041
DSM	0.9033 ± 0.0036	0.8770 ± 0.0039	0.8591 ± 0.0037
DCM	0.9020 ± 0.0037	0.8775 ± 0.0038	0.8595 ± 0.0038

ECE(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0399 ± 0.0018	0.0642 ± 0.0021	0.0764 ± 0.0028
AFT	0.0173 ± 0.0016	0.0271 ± 0.0022	0.0278 ± 0.0029
RSF	0.0112 ± 0.0016	0.0219 ± 0.0023	0.0270 ± 0.0029
FSN	0.0152 ± 0.0016	0.0198 ± 0.0025	0.0196 ± 0.0029
DHT	0.0107 ± 0.0015	0.0134 ± 0.0020	0.0170 ± 0.0024
DSM	0.0125 ± 0.0016	0.0292 ± 0.0023	0.0311 ± 0.0031
DCM	0.0105 ± 0.0016	0.0145 ± 0.0024	0.0169 ± 0.0024

BS(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
CPH	0.0563 ± 0.0014	0.0989 ± 0.0019	0.1285 ± 0.0020
AFT	0.0522 ± 0.0013	0.0907 ± 0.0019	0.1168 ± 0.0021
RSF	0.0508 ± 0.0014	0.0899 ± 0.0018	0.1190 ± 0.0021
FSN	0.0515 ± 0.0013	0.0877 ± 0.0017	0.1133 ± 0.0020
DHT	0.0509 ± 0.0012	0.0861 ± 0.0017	0.1135 ± 0.0021
DSM	0.0509 ± 0.0013	0.0882 ± 0.0018	0.1140 ± 0.0020
DCM	0.0508 ± 0.0012	0.0862 ± 0.0017	0.1127 ± 0.0020

Table B.10: Results for various performance metrics on SEER (minority) along with bootstrapped standard errors.

FLCHAIN BS(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
AFT	0.05847 ± 0.00225	0.09935 ± 0.00273	0.12486 ± 0.00248
CPH	0.07009 ± 0.00266	0.12779 ± 0.00303	0.17540 ± 0.00278
FSN	0.06604 ± 0.00250	0.11902 ± 0.00280	0.15870 ± 0.00251
DCM	0.05803 ± 0.00226	0.09754 ± 0.00281	0.12222 ± 0.00257

SUPPORT BS(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
AFT	0.13538 ± 0.00252	0.20508 ± 0.00235	0.21490 ± 0.00165
CPH	0.13345 ± 0.00236	0.19951 ± 0.00192	0.21363 ± 0.00161
FSN	0.13206 ± 0.00244	0.19777 ± 0.00201	0.21309 ± 0.00188
DCM	0.11684 ± 0.00851	0.18516 ± 0.00882	0.21641 ± 0.00706

SEER BS(t) (↓)

Model	Quantiles		
	$t = 25\text{th}$	$t = 50\text{th}$	$t = 75\text{th}$
AFT	0.04728 ± 0.00061	0.08327 ± 0.00085	0.11150 ± 0.00092
CPH	0.05031 ± 0.00068	0.08912 ± 0.00085	0.12113 ± 0.00090
FSN	0.04634 ± 0.00060	0.08038 ± 0.00080	0.10797 ± 0.00095
DCM	0.04503 ± 0.00058	0.07854 ± 0.00080	0.10641 ± 0.00095

Table B.11: Brier Scores (Lower is better) for various performance metrics along with bootstrapped standard errors on the three datasets compared with baselines unaware of the protected group membership

FLCHAIN BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
AFT	0.05847 ± 0.00225	0.09935 ± 0.00273	0.12486 ± 0.00248
CPH	0.07009 ± 0.00266	0.12779 ± 0.00303	0.17540 ± 0.00278
FSN	0.06604 ± 0.00250	0.11902 ± 0.00280	0.15870 ± 0.00251
DCM	0.05803 ± 0.00226	0.09754 ± 0.00281	0.12222 ± 0.00257

SUPPORT BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
AFT	0.13538 ± 0.00252	0.20508 ± 0.00235	0.21490 ± 0.00165
CPH	0.13345 ± 0.00236	0.19951 ± 0.00192	0.21363 ± 0.00161
FSN	0.13206 ± 0.00244	0.19777 ± 0.00201	0.21309 ± 0.00188
DCM	0.11684 ± 0.00851	0.18516 ± 0.00882	0.21641 ± 0.00706

SEER BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
AFT	0.04728 ± 0.00061	0.08327 ± 0.00085	0.11150 ± 0.00092
CPH	0.05031 ± 0.00068	0.08912 ± 0.00085	0.12113 ± 0.00090
FSN	0.04634 ± 0.00060	0.08038 ± 0.00080	0.10797 ± 0.00095
DCM	0.04503 ± 0.00058	0.07854 ± 0.00080	0.10641 ± 0.00095

Table B.12: Brier Scores (Lower is better) for various performance metrics along with bootstrapped standard errors on the three datasets compared with baselines unaware of the protected group membership

Appendix C

Appendix to Chapter 4

C.1 Identifiability

Theorem 1 (Identifiability) *Under the Directed Acyclic Graph in Figure. 4.2,*

$$p(Y|\mathbf{do}(T=t),X) = \int_Z p(Y|X,Z,T=t)p(Z|X)$$

. Proof.

$$p(Y|\mathbf{do}(T=t),X) = \int_Z p(Y|\mathbf{do}(T=t),Z,X)p(Z|\mathbf{do}(T=t),X)$$

(conditioning on and marginalizing out Z)

Now, $p(Y|\mathbf{do}(T=t),Z,X) = p(Y|T=t,Z,X)$

and, $p(Z|\mathbf{do}(T=t),Z) = p(Z|T=t,X)$

(From Pearl (2009)'s Backdoor Adjustment Formula)

$$p(Y|\mathbf{do}(T=t),X) = \int_Z p(Y|\mathbf{do}(T=t),Z,X)p(Z|T=t,X)$$

(But, under the DAG assumptions, $Z \perp T|X$)

Thus, $p(Y|\mathbf{do}(T=t),X) = \int_Z p(Y|X,Z,T=t)p(Z|X)$ ■

C.2 Parameter Inference with EM

In this section we provide an alternate approach to perform parameter inference using Expectation Maximization and compare it to the proposed ELBO optimization.

C.2.1 Inference

The complete-data log-likelihood used in EM is given by

$$\mathcal{L}_c(\mathbf{\Theta},\mathcal{D}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\{z_i=k\} \ln(P_k^m(\mathbf{x}_i)P_k^t(y_i)), \tag{C.1}$$

here, $P_k^m(\mathbf{x}_i) = p(z_i = k|\mathbf{x}_i)$,$P_k^t(\mathbf{x}_i) = p(y_i|\mathbf{x}_i, t_i, z_i = k)$ and $\mathbb{1}$ is the indicator function.

E-Step As is standard in EM, let us define $Q(\mathbf{\Theta}, \mathbf{\Theta}^l)$ as the expected value of the complete-data log-likelihood equation C.1 respect to the conditional distribution of the latents given the current parameters $\mathbf{\Theta}^l$:

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^l) = \mathbb{E}\left[\mathcal{L}_c(\mathbf{\Theta}, \mathcal{D}) \mid \{y_i, \mathbf{x}_i, t_i\}_{i=1}^N ; \mathbf{\Theta}^l\right].$$

Since the only quantity in equation C.1 that depends explicitly on z_i is the indicator $\mathbb{1}\{z_i = k\}$, we can compute Q by replacing these indicators with the posterior probability of $z_i = k$:

$$\begin{aligned} &\mathbb{E}[\mathbb{1}\{z_i = k\}] \equiv h_i^{(k)} = p(z_i = k|y_i, \mathbf{x}_i, t_i; \mathbf{\Theta}^l) \\ &= \frac{p(y_i|z_i = k, \mathbf{x}_i, t_i; \mathbf{\Theta}^l)p(z_i = k|\mathbf{x}_i, \mathbf{\Theta}^l)}{p(y_i|\mathbf{x}_i, t_i, \mathbf{\Theta}^l)} \\ &= \frac{p(y_i|z_i = k, \mathbf{x}_i, t_i; \mathbf{\Theta}^l)}{p(y_i|\mathbf{x}_i, t_i, \mathbf{\Theta}^l)} \frac{p(\mathbf{x}_i|z_i = k; \mathbf{\Theta}^l)p(z_i = k; \mathbf{\Theta}^l)}{p(\mathbf{x}_i; \mathbf{\Theta}^l)}. \end{aligned} \tag{C.2}$$

The terms in the numerator can be evaluated using equation 4.1–equation 4.3, equation 4.4 from the model. The terms in the denominator are normalization constants that ensure the probabilities sum to one.

M-Step We use a gradient ascent method in the M-step to maximize Q with respect to the parameters:

$$\mathbf{\Theta}^{l+1} = \arg\max_{\mathbf{\Theta}} \left(\sum_{i=1}^{N} \sum_{k=1}^{K} h_i^{(k)} \ln[P_k^m(\mathbf{x}_i) P_k^t(y_i)] - \lambda\Omega(\boldsymbol{\pi}) \right). \tag{C.3}$$

The posterior probabilities $h_i^{(k)}$ are fixed from the E-step. Using Bayes' rule as in equation C.2, $P_k^m(\mathbf{x}_i) = p(z_i = k \,|\, \mathbf{x}_i)$ can be expressed in terms of model parameters $\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\pi}_k$ defined by equation 4.1–equation 4.3. Similarly, $P_k^t(y_i)$ depends on parameters $\mathbf{w}$ and γ_k according to the outcome model equation 4.4. The use of gradient ascent allows for any differentiable nonlinear function $f(\cdot)$ in equation 4.4.

Instead of computing Q over the entire dataset, we sample a mini-batch from the dataset and perform the E-step and M-step over just the mini-batch in each iteration. We observe that this mini-batch procedure is faster than regular EM over the entire dataset.

C.2.2 Comparison to ELBO Optimization

In order to compare the performance of EM vis-à-vis the variational inference-motivated ELBO optimization, we compare the train and test negative log-likelihood for both approaches on 100 realizations of the **IHDP** dataset, with the number of latent components, $K = 3$ and stochastic gradient descent learning rate of 1×10^{-3}. We then average over the resulting 100 curves. Figure C.1 presents the results; it is clear from the figure that the ELBO approach has less tendency to overfit and results in an overall better fit compared to the EM approach. This motivates our choice to directly optimize the ELBO.

C.3 Parameter Initialization

Gradient based optimization strategies can be subject to local minima and hence their performance is dependent on parameter initialization. To initialize the model with 'good' values ensuring better convergence,

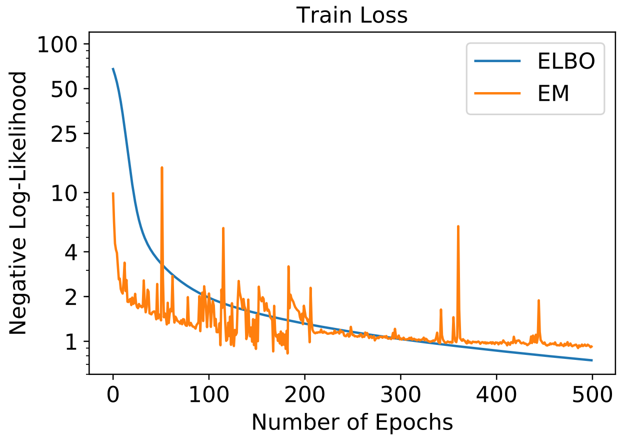


Figure C.1: The negative log-likelihood (NLL) versus the number of optimization epochs for EM and ELBO. Notice how the Test NLL continues to decrease for ELBO vs. EM, suggesting the ELBO approach is less sensitive to overfitting.

we set the mean for each component, $\boldsymbol{\mu}_k$ and $\boldsymbol{\pi}_k$, equal to the sample mean of the entire data, i.e. $\boldsymbol{\mu}_k^0 = \frac{1}{N}\sum_i \mathbf{x}_{\text{cont},i}$, $\boldsymbol{\pi}_k^0 = \frac{1}{N}\sum_i \mathbf{x}_{\text{disc},i}$, and the covariance of every component to $\Sigma_k^0 = \text{diag}(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a vector consisting of the sample variances of the continuous covariates $\mathbf{X}_{\text{cont}}$. We pre-train the parameters $\mathbf{w}_t$ in the outcome model equation 4.4 using standard cross-entropy loss without the subgroup and treatment assignment term $\gamma_k t$. Finally, we initialize the treatment coefficients, γ_k, randomly with positive values for all k.

C.4 Model Fitting

Our implementation of HEMM has two free parameters, the number of groups K and the strength of the sparsity prior, λ. For the **Opioid** dataset we divide the dataset into 3 parts with 70% as TRAIN for model training, 10% as DEV for parameter tuning, and 20% as TEST for evaluation. The partition is done so that the joint distribution of outcome and treatment is approximately the same in the 3 sets: $p_{\text{TRAIN}}(Y,T) \approx p_{\text{TEST}}(Y,T) \approx p_{\text{DEV}}(Y,T)$. For **IHDP** we use the stnadard 80/20 TRAIN/TEST split as is popular in literature.

We perform a grid search over $K \in \{2,3,4\}$ and $\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. For each (K,λ) pair, we perform 5 runs with randomly initialized values of the treatment coefficients γ_k. All other parameters are initialized as described in Section C.3. For Adam, we use a step size of 10^{-4} and mini-batch sizes of 10, 20 & 1000 for **Synthetic**, **IHDP**, and **Opioid** respectively, and stop parameter update if the ELBO on the DEV is lower at the end of an epoch. We also search over the space of models where the outcome and counterfactual have the same or different parameterisation based on treatment assignment. From all the (K,λ) pairs and random initializations above, we select the model that has the best performance on the DEV set in predicting the outcome y_i, in terms of the Area Under the Receiver Operating Characteristic (AU-ROC). For the **Synthetic** dataset, we simply set $K=2$ and $\lambda=0$. In this case there is no need for a DEV set and the data is split 50/50 between TRAIN and TEST.

Appendix D

Appendix to Chapter 5

D.1 Identifiability

Proof of Remark 1. $\mathbf{P}(T|\mathbf{do}(A)=\boldsymbol{a},X)=$

$$\int_Z\int_{\boldsymbol{\phi}}\mathbf{P}(T|\mathbf{do}(A)=\boldsymbol{a},X,Z,\boldsymbol{\phi})\mathbf{P}(Z,\boldsymbol{\phi}|\mathbf{do}(A)=\boldsymbol{a},X)$$

$$=\int_Z\int_{\boldsymbol{\phi}}\mathbf{P}(T|\mathbf{do}(A)=\boldsymbol{a},X,Z,\boldsymbol{\phi})\mathbf{P}(Z|\mathbf{do}(A)=\boldsymbol{a},X)\mathbf{P}(\boldsymbol{\phi}|\mathbf{do}(A)=\boldsymbol{a},X)$$

Because, $Z\perp\phi\,|\,X$. But, $\mathbf{P}(Z|\mathbf{do}(A)=\boldsymbol{a},X)=\mathbf{P}(Z|X)$ and,

$\mathbf{P}(\boldsymbol{\phi}|\mathbf{do}(A)=\boldsymbol{a},X)=\mathbf{P}(\boldsymbol{\phi}|X)$ (From Pearl's 3rd Rule of **do**-Calculus.)

$$=\int_{\boldsymbol{\phi}}\int_Z\mathbf{P}(T|\mathbf{do}(A)=\boldsymbol{a},X,Z,\boldsymbol{\phi})\mathbf{P}(Z|X)\mathbf{P}(\boldsymbol{\phi}|X)$$

$$=\int_{\boldsymbol{\phi}}\int_Z\mathbf{P}(T|A=\boldsymbol{a},X,Z,\boldsymbol{\phi})\mathbf{P}(Z|X)\mathbf{P}(\boldsymbol{\phi}|X)\text{ (Pearl's 2}^{\text{nd}}\text{ Rule.)}$$

$$=\mathbf{E}_{(Z,\boldsymbol{\phi})\sim\mathbf{P}(\cdot|X)}\big[\mathbf{P}(T|A=\boldsymbol{a},X,Z,\boldsymbol{\phi})\big].\qquad\blacksquare$$

D.2 Learning

We propose to maximize the likelihood in Equation 5.5 using a stochastic Expectation Maximization algorithm (Algorithm 2).

E-Step: Involves first computing the posterior counts of the joint of the latent $\boldsymbol{Z}$ and $\boldsymbol{\phi}$ as follows:

$$\mathbb{E}[\mathbf{1}\{Z=k,\boldsymbol{\phi}=m\}|\{\boldsymbol{t},\boldsymbol{x},\boldsymbol{a}\}]=\mathbf{P}(Z=k,\boldsymbol{\phi}=m|\{\boldsymbol{t},\boldsymbol{x},\boldsymbol{a}\})=\frac{\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})\cdot\mathbf{P}(Z=k,\boldsymbol{\phi}=m|\boldsymbol{x},\boldsymbol{a})}{\sum_k\sum_m\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})\cdot\mathbf{P}(Z=k,\boldsymbol{\phi}=m|\boldsymbol{x},\boldsymbol{a})}\tag{D.1}$$

Note that for the censored individuals $\mathbf{P}(t|\cdot)$ is $\mathbf{P}(T>t|\cdot)$. In practice the $\mathbf{P}(t|\cdot)$ are obtained through spline interpolation. Now the latent variable specific soft posterior counts can be computed by marginalizing out the complementary Latent Variable. Thus, soft posterior counts of Z are $\gamma=\sum_{\boldsymbol{\phi}\in m}\mathbf{P}(Z=k,\boldsymbol{\phi}=m|\{\boldsymbol{t},\boldsymbol{x},\boldsymbol{a}\})$, and of $\boldsymbol{\phi}$ are, $\zeta=\sum_{\boldsymbol{Z}\in k}\mathbf{P}(Z=k,\boldsymbol{\phi}=m|\{\boldsymbol{t},\boldsymbol{x},\boldsymbol{a}\})$.

M-Step: as in standard EM, involves maximizing the $Q(\cdot)$ function on the data, $\mathcal{D}$ defined as:

$$Q(\boldsymbol{\theta}) = \underbrace{\sum_{i=1}^{|\mathcal{D}|}\sum_{k}\sum_{m}\gamma^k\cdot\zeta^m\cdot\ln\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})+}_{\textcircled{A}}$$
$$\underbrace{\sum_{i=1}^{|\mathcal{D}|}\sum_{k}\gamma^k\ln\mathbf{P}(Z=k|\boldsymbol{x},\boldsymbol{a})}_{\textcircled{B}}+\underbrace{\sum_{i=1}^{|\mathcal{D}|}\sum_{m}\zeta^m\ln\mathbf{P}(\boldsymbol{\phi}=m|\boldsymbol{x},\boldsymbol{a})}_{\textcircled{C}}. \tag{D.2}$$

Note that $\textcircled{B}$ and $\textcircled{C}$ can be directly optimized with a gradient based approach. However, $\textcircled{A}$ is the semi-parametric Cox event rate which is hard to optimize in the presence of soft posterior weights. We instead replace the soft weights in $\textcircled{A}$ with the hard posterior counts sampled as follows: $\psi\sim$ Categorical(γ), $\xi\sim$ Categorical(ζ).

We thus arrive at $\widehat{Q}(\cdot)$ given as,

$$\widehat{Q}(\boldsymbol{\theta}) = \underbrace{\sum_{i=1}^{|\mathcal{D}|}\sum_{k}\sum_{m}\mathbf{1}_{\{\psi_i=k\}}\mathbf{1}_{\{\xi_i=m\}}\cdot\ln\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})}_{\textcircled{A'}}+\textcircled{B}+\textcircled{C} \tag{D.3}$$

Remark 2 *$\widehat{Q}(\cdot)$ is an unbiased estimate of the $Q(\cdot)$ in Equation D.2.*

Proof. Follows immediately from the fact that $\mathbf{E}[\widehat{Q}]=Q(\cdot)$.

We can now rewrite $\textcircled{A'}$ as,

$$\begin{aligned}\textcircled{A'} &= \sum_{i=1}^{|\mathcal{D}|}\sum_{k}\mathbf{1}_{\{\psi_i=k\}}\sum_{m}\mathbf{1}_{\{\xi_i=m\}}\cdot\ln\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})\\ &= \sum_{k}\underbrace{\sum_{i=1}^{|\mathcal{D}|}\mathbf{1}_{\{\psi_i=k\}}\sum_{m}\mathbf{1}_{\{\xi_i=m\}}\cdot\ln\mathbf{P}(\boldsymbol{t}|Z=k,\boldsymbol{\phi}=m,\boldsymbol{x},\boldsymbol{a})}_{\text{Proportional Hazards, Partial Likelihood}}\end{aligned} \tag{D.4}$$

Maximizing $\textcircled{A'}$ is equal to maximizing $\mathcal{PL}_k(\cdot)$ over each k where,

$$\ln\mathcal{PL}_k(\mathcal{D},\psi,\boldsymbol{\xi};\boldsymbol{\theta}) = \sum_{i:\delta_i=1}^{|\mathcal{D}|}\mathbf{1}_{\{\psi_i=k\}}\bigg(h^k(\boldsymbol{x}_i)+a\omega_{\xi_i}-\ln\sum_{j\in\mathcal{R}(t_i)}\exp\big(h^k(\boldsymbol{x}_j)+a\omega_{\xi_j}\big)\bigg).$$

Combining $\mathcal{PL}_k(\cdot)$ with $\textcircled{B}$ and $\textcircled{C}$ we arrive at the $\widehat{Q}$ in Equation 5.6.

D.3 Factual Regression Experiments

We compare the performance of CMHE in 5 Fold CV to a Linear Cox model and a Deep Cox Model in 5 fold cross validation with a `2 Hidden Layer MLP` with dimensionality of 50 and `Tanh` activations. Each model was trained with `Adam` with learning rates tuned from $\{\mathbf{10e^{-3}},\mathbf{10e^{-4}}\}$ and minibatch size of $\{128,256\}$. For CMHE we tuned the number of treatment effect phenotypes from $\boldsymbol{\phi}$ from $\{2,3\}$ and the base survival rate phenogroups from `{1,2,3}`.

D.4 Rule Learning

We used the python package, `scope-rules`[1] to explain the learnt phenotypes with parsimonius rules. The rules were restricted to have a maximum length of 4 and a precision of `0.8`. Explanations with the highest F_1 score on the train set are reported in Table 5.1.

D.5 Synthetic Dataset

We employ the python package `sklearn`[2] to generate the confounders $\boldsymbol{x}$.

$$
\begin{aligned}
[\boldsymbol{x}_1, \boldsymbol{x}_2], Z &\sim \texttt{sklearn.datasets.make\_blobs}(K = 3) \\
[\boldsymbol{x}_3, \boldsymbol{x}_4] &\sim \text{Uniform}(-2, 2) \\
\boldsymbol{\phi} &\triangleq \mathbf{1}\{|\boldsymbol{x}_3| + |\boldsymbol{x}_4| > 2\} \\
A &\sim \text{Bernoulli}(1/2) \\
\boldsymbol{T}^*|(Z = k, \boldsymbol{\phi} = m, A = \boldsymbol{a}) &\sim \text{Gompertz}\big(\boldsymbol{\beta}_k^\top \boldsymbol{x} + (-\boldsymbol{a}^m)\big) \\
\delta \sim \text{Bernoulli}(3/4), \quad C &\sim \text{Uniform}(0, \boldsymbol{t}^*) \\
\text{if } \delta = 1 : \boldsymbol{T} = \boldsymbol{T}^* &\text{else if, } \delta = 0 : \boldsymbol{T} = C.
\end{aligned}
$$

D.6 Tabulated Results

Latent Phenogroups with Enhanced Treatment Effects

Model	CATE (RMST) in Days		
	1 Year	3 Year	5 Year
		ALLHAT-A	
DR-C	2.68 ± 0.08	14.61 ± 0.52	28.03 ± 1.19
VT	3.08 ± 0.11	17.36 ± 0.65	35.31 ± 1.44
CMHE	3.58 ± 0.10	18.52 ± 0.56	37.43 ± 1.23
		ALLHAT-B	
DR-C	4.23 ± 0.15	21.57 ± 0.79	40.37 ± 1.68
VT	5.34 ± 0.17	28.68 ± 0.93	54.16 ± 1.96
CMHE	5.64 ± 0.13	30.44 ± 0.74	59.54 ± 1.59
		ACCORD	
DR-C	0.36 ± 0.18	4.15 ± 0.94	13.76 ± 2.21
VT	1.40 ± 0.24	6.19 ± 1.16	18.11 ± 2.65
CMHE	1.22 ± 0.24	9.71 ± 1.11	27.02 ± 2.43

Latent Phenogroups with Diminished Treatment Effects

Model	CATE (RMST) in Days		
	1 Year	3 Year	5 Year
		ALLHAT-A	
DR-C	1.60 ± 0.08	7.64 ± 0.46	13.62 ± 1.0
VT	2.24 ± 0.10	9.36 ± 0.54	15.52 ± 1.16
CMHE	1.38 ± 0.06	5.80 ± 0.38	10.09 ± 0.86
		ALLHAT-B	
DR-C	3.57 ± 0.20	12.16 ± 1.08	15.89 ± 2.29
VT	2.64 ± 0.14	8.34 ± 0.80	7.49 ± 1.80
CMHE	2.39 ± 0.18	6.20 ± 1.02	2.28 ± 2.18
		ACCORD	
DR-C	-0.06 ± 0.32	-4.51 ± 1.86	-8.59 ± 4.34
VT	-0.33 ± 0.24	-10.53 ± 1.47	-26.92 ± 3.50
CMHE	0.12 ± 0.29	-10.45 ± 1.79	-24.52 ± 4.15

Table D.1: Tabulated results of the proposed method versus baselines in counterfactual phenotyping. We report the Conditional Average Treatment Effect in Restricted Mean Survival Time over multiple time horizons.

[1] `https://github.com/scikit-learn-contrib/skope-rules`

[2] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

D.7 List of Confounding features

Name	Description
RACE	Race of Participant
HISPANIC	If Participant was Hispanic
ETHNIC	Ethnicity
SEX	Sex of Participant
ESTROGEN	Estrogen supplementation
BLMEDS	Antihypertensive treatment
MISTROKE	History of Stroke
HXCABG	History of coronary artery bypass
STDEPR	Prior ST depression/T-wave inversion
OASCVD	Other atherosclerotic cardiovascular disease
DIABETES	Prior history of Diabetes
HDLLT35	HDL cholesterol < 35mg/dl; 2x in past 5 years
LVHECG	LVH by ECG in past 2 years
WALL25	LVH by ECG in past 2 years
LCHD	History of CHD at baseline
CURSMOKE	Current smoking status.
ASPIRIN	Aspirin use
LLT	Lipid-lowering trial
RACE2	Race (2 groups)
BLMEDS2	Antihypertensive treatment
GEOREGN	Geographic Region
AGE	Age upon entry
BLWGT	Weight upon entry
BLHGT	Height upon entry
BLBMI	Body Mass Index upon entry
BV2SBP	Baseline SBP
BV2DBP	Baseline DBP
EDUCAT	Education
APOTAS	Baseline serum potassium
BLGFR	Baseline est glomerular filtration rate

Table D.2: List of confounding variables used for experiments involving the ACCORD dataset.

Name	Description
female	Indicator if sex is Female
bl_age	Age in years
cvd_hx_bl	CVD History at Baseline: 0=No, 1=Yes
raceclass	Race Class: White, Black, Hispanic, Other
sbp	Systolic Blood Pressure (mmHg)
dbp	Diastolic Blood Pressure (mmHg)
hr	Heart Rate (bpm)
x1diab	Diagnosis of type 2 diabetes of >3 months duration
x2mi	Myocardial infarction
x2stroke	Stroke
x2angina	Angina/Ischemic changes (Graded Exercise/Imaging)
cabg	CABG
ptci	PTCI/PTCA/Atherectomy
cvdhist	Participant has history of clinical CVD events
orevasc	Other revascularization procedure
x2hbac11	HbA1c between 7.5% and 11.0% inclusive
x2hbac9	HbA1c between 7.5% and 9.0% inclusive
x3malb	Micro or macro albuminuria within past 2 years
x3lvh	LVH by ECG or Echocardiogram within past 2 years
x3sten	Low ABI (<0.9)/>= 50% stenosis of coronary, carotid **or**, lower extremity artery within past 2 years
x4llmeds	On lipid lowering medication currently **or**, untreated LDL-C > 130 mg/dL within past 2 years
x4gender	Gender for low HDL-C within past 2 years
x4hdlf	HDL-c < 50 mg/dL within past 2 years, female
x4hdlm	HDL-c < 40 mg/dL within past 2 years, male
x4bpmeds	Participant currently on BP medications
x4notmed	Participant not on BP medication **and**, most recent BP within past 2 years
x4smoke	Current cigarette smoker
x4bmi	BMI > 32 kg/m2 within past 2 years
chol	Total Cholesterol (mg/dL)
trig	Triglycerides (mg/dL)
vldl	Very low density lipoprotein (mg/dL)
ldl	Low density lipoprotein (mg/dL)
hdl	High density lipoprotein (mg/dL)
fpg	Fasting plasma glucose (mg/dL)
alt	ALT (mg/dL)
cpk	CPK (mg/dL)
potassium	Potassium (mmol/L)
screat	Serum creatinine (mg/dL)
gfr	eGFR from 4 var. MDRD eq. (ml/min/1.73 m2)
ualb	Urinary albumin (mg/dL)
ucreat	Urinary creatinine (mg/dL)
uacr	Creatine to albumin ratio

Table D.3: List of confounding variables used for experiments involving the ALLHAT dataset.

Appendix E

Appendix to Chapter 6

E.1 Derivation of the Inference Algorithm

Censored Instances: Note that in the case of the censored instances we will condition on the thresholded survival $(T > \boldsymbol{u})$. The the posterior counts thus reduce to:

$$\begin{aligned}\boldsymbol{\gamma}^k &= \mathbb{P}(Z = k|X = \boldsymbol{x}, A = \boldsymbol{a}, T > \boldsymbol{u}) \\ &= \frac{\mathbb{P}(T > \boldsymbol{t}|Z = \boldsymbol{k}, X = \boldsymbol{x}, A = \boldsymbol{a})\boldsymbol{p}(Z = \boldsymbol{k}|X = \boldsymbol{x})}{\sum_k \mathbb{P}(T > \boldsymbol{t}|Z = \boldsymbol{k}, X = \boldsymbol{x}, A = \boldsymbol{a})\mathbb{P}(Z = \boldsymbol{k}|X = \boldsymbol{x})}\end{aligned} \tag{E.1}$$

$$\text{Here, } \mathbb{P}(T > \boldsymbol{t}|Z = \boldsymbol{k}, X = \boldsymbol{x}, A = \boldsymbol{a}) = \exp\big(-\boldsymbol{\Lambda}(t)\big)^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},k)}$$

Uncensored Instances The posteriors are $\boldsymbol{\gamma}^k = \boldsymbol{p_\theta}(Z = k|X = \boldsymbol{x}, T = \boldsymbol{u})$,

Posteriors for the uncensored data are more involved and involve the base hazard $\boldsymbol{\lambda}_0(\cdot)$. Posteriors for uncensored data are independent of the base hazard function, $\boldsymbol{\lambda}_0(\cdot)$ as,

$$\boldsymbol{\gamma}^k = \frac{\cancel{\boldsymbol{\lambda}_0(u)}\boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{a})\boldsymbol{S}_0(u_i)^{\boldsymbol{h}_k(\boldsymbol{x},\boldsymbol{a})}}{\sum\limits_k \cancel{\boldsymbol{\lambda}_0(u)}\boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{a})\boldsymbol{S}_0(u)^{\boldsymbol{h}_k(\boldsymbol{x},\boldsymbol{a})}} = \frac{\boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{a})\boldsymbol{S}_0(u_i)^{\boldsymbol{h}_k(\boldsymbol{x},\boldsymbol{a})}}{\sum\limits_k \boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{a})\boldsymbol{S}_0(u_i)^{\boldsymbol{h}_k(\boldsymbol{x},\boldsymbol{a})}}$$

Combining Equations E.1 and E.2 we arrive at the following estimate for the posterior counts

$$\begin{aligned}\boldsymbol{\gamma}^k &= \widehat{\mathbb{P}}(Z = k|X = \boldsymbol{x}, A = \boldsymbol{a}, \boldsymbol{u}) \\ &= \frac{\mathbb{P}(\boldsymbol{u}|Z = \boldsymbol{k}, X = \boldsymbol{x}, A = \boldsymbol{a})\mathbb{P}(Z = \boldsymbol{k}|X = \boldsymbol{x})}{\sum_k \mathbb{P}(\boldsymbol{u}|Z = \boldsymbol{k}, X = \boldsymbol{x}, A = \boldsymbol{a})\mathbb{P}(Z = \boldsymbol{k}|X = \boldsymbol{x})} \\ &= \frac{\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{k})^{\delta_i}\widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{k})}\exp(\boldsymbol{\theta}_{\boldsymbol{k}}^\top \boldsymbol{x})}{\sum_{j\in\mathcal{Z}} \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{j})^{\delta_i}\widehat{\boldsymbol{S}}_0(\boldsymbol{u})^{\boldsymbol{h}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{j})}\exp(\boldsymbol{\theta}_{\boldsymbol{j}}^\top \boldsymbol{x})}.\end{aligned} \tag{E.2}$$

E.2 Additional Details on the ALLHAT and BARI 2D Case Studies

Tables E.1 and E.2 represent additional confounding variables found in the **ALLHAT** and **BARI2D** trials respectively.

Name	Description
`ETHNIC`	Ethnicity
`SEX`	Sex of Participant
`ESTROGEN`	Estrogen supplementation
`BLMEDS`	Antihypertensive treatment
`MISTROKE`	History of Stroke
`HXCABG`	History of coronary artery bypass
`STDEPR`	Prior ST depression/T-wave inversion
`OASCVD`	Other atherosclerotic cardiovascular disease
`DIABETES`	Prior history of Diabetes
`HDLLT35`	HDL cholesterol <35mg/dl; 2x in past 5 years
`LVHECG`	LVH by ECG in past 2 years
`WALL25`	LVH by ECG in past 2 years
`LCHD`	History of CHD at baseline
`CURSMOKE`	Current smoking status.
`ASPIRIN`	Aspirin use
`LLT`	Lipid-lowering trial
`AGE`	Age upon entry
`BLWGT`	Weight upon entry
`BLHGT`	Height upon entry
`BLBMI`	Body Mass Index upon entry
`BV2SBP`	Baseline SBP
`BV2DBP`	Baseline DBP
`APOTAS`	Baseline serum potassium
`BLGFR`	Baseline est glomerular filtration rate
`ACHOL`	Total Cholesterol
`AHDL`	Baseline HDL Cholesterol
`AFGLUC`	Baseline fasting serum glucose

Table E.1: List of confounding variables used for experiments involving the **ALLHAT** dataset.

Name	Description
`hxmi`	History of MI
`age`	Age upon entry
`dbp_stand`	Standing diastolic BP
`sbp_stand`	Standing systolic BP
`sex`	Sex
`asp`	Aspirin use
`smkcat`	Cigarette smoking category
`betab`	Beta blocker use
`ccb`	Calcium blocker use
`hxhtn`	History of hypertension requiring tx
`insulin`	Insulin use
`weight`	Weight (kg) upon entry
`bmi`	BMI upon entry
`qabn`	Abnormal Q-Wave
`trig`	Triglycerides (mg/dl) upon entry
`dmdur`	Duration of diabetes mellitus
`ablvef`	Left ventricular ejection fraction <50%
`race`	Race
`priorrev`	Prior revascularization
`hxcva`	Cerebrovascular accident
`screat`	Serum creatinine (mg/dl)
`hmg`	Statin
`hxhypo`	History of hypoglycemic episode
`hba1c`	Hemoglobin A1c(%)
`priorstent`	Prior stent
`spotass`	Serum Potassium(mEq/L)
`hispanic`	Hispanic ethnicity
`tchol`	Total Cholesterol
`hdl`	HDL Cholesterol
`insul_circ`	Circulating insulin (IU/ml)
`tzd`	Thiazolidinedione
`ldl`	LDL Cholesterol
`tabn`	Abnormal T-waves
`nsgn`	Nonsublingual nitrate
`sulf`	Sulfonylurea
`hxchf`	Histoty of congestive heart failure req tx
`arb`	Angiotensin receptor blocker
`acr`	Urine albumin/creatinine ratio mg/g
`diur`	Diuretic
`apa`	Anti-platelet
`hxchl`	Hypercholesterolemia req tx
`acei`	ACE inhibitor
`abilow`	Low ABI (<= 0.9)
`biguanide`	Biguanide
`stabn`	Abnormal ST depression

Table E.2: List of confounding variables used for experiments involving the **BARI2D** dataset.

Bibliography

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017a.

Alaa, A. M. and van der Schaar, M. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2326–2334. Curran Associates Inc., 2017b.

Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Austin, P. C., Harrell Jr, F. E., and van Klaveren, D. Graphical calibration curves and the integrated calibration index (ici) for survival models. *Statistics in Medicine*, 2020.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

Beil, M., Proft, I., van Heerden, D., Sviri, S., and van Heerden, P. V. Ethical considerations about artificial intelligence for prognostication in intensive care. *Intensive Care Medicine Experimental*, 7(1):70, 2019.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Binder, D. A. Fitting cox's proportional hazards models from survey data. *Biometrika*, 79(1):139–147, 1992.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bosco Sabuhoro, J., Larue, B., and Gervais, Y. Factors determining the success or failure of canadian establishments on foreign markets: A survival analysis approach. *The International Trade Journal*, 20(1):33–73, 2006.

Brat, G. A., Agniel, D., Beam, A., Yorkgitis, B., Bicket, M., Homer, M., Fox, K. P., Knecht, D. B., McMahill-Walraven, C. N., Palmer, N., and Kohane, I. Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: Retrospective cohort study. *BMJ*, 360:j5790, January 2018.

Breslow, N. E. Discussion of the paper by D.R. Cox. *J R Statist Soc B*, 34:216–217, 1972a.

Breslow, N. E. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34:216–217, 1972b.

Breslow, N. E. and Chatterjee, N. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

Bretthauer, M., Løberg, M., Wieszczy, P., Kalager, M., Emilsson, L., Garborg, K., Rupinski, M., Dekker, E., Spaander, M., Bugajski, M., et al. Effect of colonoscopy screening on risks of colorectal cancer and related death. *New England Journal of Medicine*, 2022.

Brummett, C. M., Waljee, J. F., Goesling, J., Moser, S., Lin, P., Englesbe, M. J., Bohnert, A. S. B., Kheterpal, S., and Nallamothu, B. K. New persistent opioid use after minor and major surgical procedures in US adults. *JAMA Surg.*, 152(6):e170504, 2017.

Buse, J. B., Group, A. S., et al. Action to control cardiovascular risk in diabetes (accord) trial: design and methods. *The American journal of cardiology*, 99(12):S21–S33, 2007.

Califf, R. M., Woodcock, J., and Ostroff, S. A proactive response to prescription opioid abuse. *New England J. Med.*, 374(15):1480–1485, 2016.

Chapfuwa, P., Tao, C., Li, C., Page, C., Goldstein, B., Carin, L., and Henao, R. Adversarial time-to-event modeling. *arXiv preprint arXiv:1804.03184*, 2018.

Chapfuwa, P., Li, C., Mehta, N., Carin, L., and Henao, R. Survival Cluster Analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020a.

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., and Henao, R. Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 2020b.

Chapfuwa, P., Assaad, S., Zeng, S., Pencina, M. J., Carin, L., and Henao, R. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145, 2021.

Che, Z., St. Sauver, J., Liu, H., and Liu, Y. Deep learning solutions for classifying patients on opioid use. In *AMIA Annu. Symp. Proc.*, pp. 525–534, 2017.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

Chen, G. H. Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates. *arXiv preprint arXiv:1905.05285*, 2019.

Chen, P.-Y. and Tsiatis, A. A. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

Chen, Y.-H. Weighted breslow-type and maximum likelihood estimation in semiparametric transformation models. *Biometrika*, 96(3):591–600, 2009.

Chiang, C. L. *The life table and its applications.* Krieger Malabar, FL, 1984.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014a.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014b.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pp. 301–318, 2016a.

Choi, Y., Chiu, C. Y.-I., and Sontag, D. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016b.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Connors, A. F., Dawson, N. V., Desbiens, N. A., Fulkerson, W. J., Goldman, L., Knaus, W. A., Lynn, J., Oye, R. K., Bergner, M., Damiano, A., et al. A controlled trial to improve care for seriously iii hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (support). *Jama*, 274 (20):1591–1598, 1995.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34 (2):187–202, 1972.

Crabbé, J., Curth, A., Bica, I., and van der Schaar, M. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *arXiv preprint arXiv:2206.08363*, 2022.

Crofford, L. J. Use of NSAIDs in treating patients with arthritis. *Arthritis Res. Ther.*, 15(Suppl. 3):S2, July 2013.

Curth, A., Lee, C., and van der Schaar, M. Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 2021.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., and Yuan, Y. e. a. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983.

Cushman, W. C., Ford, C. E., Cutler, J. A., Margolis, K. L., Davis, B. R., Grimm, R. H., Black, H. R., Hamilton, B. P., Holland, J., Nwachuku, C., et al. Original papers. success and predictors of blood pressure control in diverse north american settings: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (allhat). *The Journal of Clinical Hypertension*, 4(6):393–404, 2002.

Czado, C. and Rudolph, F. Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics*, 31(3):395–413, 2002.

Davis, B. R., Cutler, J. A., Gordon, D. J., Furberg, C. D., Wright Jr, J. T., Cushman, W. C., Grimm, R. H., LaRosa, J., Whelton, P. K., Perry, H. M., et al. Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (allhat). *American Journal of Hypertension*, 9(4):342–360, 1996.

Davis, M. P. Drug management of visceral pain: Concepts from basic research. *Pain Res. Treat.*, 2012:265605, 2012.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Dispenzieri, A., Katzmann, J. A., Kyle, R. A., Larson, D. R., Therneau, T. M., Colby, C. L., Clark, R. J., Mead, G. P., Kumar, S., Melton III, L. J., et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pp. 517–523. Elsevier, 2012.

Dowell, D., Haegerich, T. M., and Chou, R. Cdc guideline for prescribing opioids for chronic pain—united states, 2016. *JAMA*, 315(15):1624–1645, 04 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.1464. URL https://dx.doi.org/10.1001/jama.2016.1464.

Dusseldorp, E. and Mechelen, I. Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. *Statistics in medicine*, 33, 01 2014. doi: 10.1002/sim.5933.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Edlund, M. J., Steffick, D., Hudson, T., Harris, K. M., and Sullivan, M. Risk factors for clinically recognized opioid abuse and dependence among veterans using opioids for chronic non-cancer pain. *Pain*, 129(3): 355–362, June 2007.

Faraggi, D. and Simon, R. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Fernández, T., Rivera, N., and Teh, Y. W. Gaussian processes for survival analysis. *Advances in Neural Information Processing Systems*, 29:5021–5029, 2016.

Fine, J. P. and Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Stat. Med.*, 30(24):2867–2880, 2011.

Fotso, S. et al. PySurvival: Open source package for survival analysis modeling, 2019–. URL https://www.pysurvival.io/.

Friedman, J. and Popescu, B. E. Gradient directed regularization for linear regression and classification. Technical report, Citeseer, 2003.

Friedman, J., Hastie, T., Tibshirani, R., et al. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4):1–24, 2009.

Friedman, J., Hastie, T., and Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Friedman, J. H. and Popescu, B. E. Predictive learning via rule ensembles. *The annals of applied statistics*, 2008.

Furberg, C. D., Wright, J. T., Davis, B. R., Cutler, J. A., Alderman, M., Black, H., Cushman, W., Grimm, R., Haywood, L. J., Leenen, F., et al. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (allhat). *Journal of the American Medical Association*, 288(23):2981–2997, 2002.

Gage, B. F., Van Walraven, C., Pearce, L., Hart, R. G., Koudstaal, P. J., Boode, B., and Petersen, P. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation*, 110(16):2287–2292, 2004.

Gaille, M., Araneda, M., Dubost, C., Guillermain, C., Kaakai, S., Ricadat, E., Todd, N., and Rera, M. Ethical and social implications of approaching death prediction in humans-when the biology of ageing meets existential issues. *BMC Medical Ethics*, 21(1):1–13, 2020.

Gebhart, G. F., Su, X., Joshi, S., Ozaki, N., and Sengupta, J. N. Peripheral opioid modulation of visceral pain. *Ann. N. Y. Acad. Sci.*, 909(1):41–50, January 2000.

George, P., Chandwani, S., Gabel, M., Ambrosone, C. B., Rhoads, G., Bandera, E. V., and Demissie, K. Diagnosis and surgical delays in african american and white women with early-stage breast cancer. *Journal of Women's Health*, 24(3):209–217, 2015.

Gerds, T. A. and Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

Gerds, T. A., Kattan, M. W., Schumacher, M., and Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013.

Gers, F. A., Schmidhuber, J., and Cummins, F. Learning to forget: Continual prediction with lstm. 1999.

Ghasemi, A., Yacout, S., and Ouali, M. S. Optimal condition based maintenance with imperfect information and the proportional hazards model. *International Journal of Production Research*, 45(4):989–1012, 2007. doi: 10.1080/00207540600596882. URL https://doi.org/10.1080/00207540600596882.

Glanz, J. M., Narwaney, K. J., Mueller, S. R., Gardner, E. M., Calcaterra, S. L., Xu, S., Breslin, K., and Binswanger, I. A. Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy. *J. Gen. Intern. Med.*, 33(10):1646–1653, October 2018.

Goff Jr, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'donnell, C. J., et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation*, 129(25_suppl_2):S49–S73, 2014.

Goldstein, M., Han, X., Puli, A., Perotte, A., and Ranganath, R. X-cal: Explicit calibration for survival analysis. *Advances in Neural Information Processing Systems*, 33:18296–18307, 2020.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Group, B. D. S. A randomized trial of therapies for type 2 diabetes and coronary artery disease. *New England Journal of Medicine*, 360(24):2503–2515, 2009.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. Effective ways to build and evaluate individual survival distributions. *The Journal of Machine Learning Research*, 21(1):3289–3351, 2020.

Harbaugh, C. M., Lee, J. S., Hu, H. M., McCabe, S. E., Voepel-Lewis, T., Englesbe, M. J., Brummett, C. M., and Waljee, J. F. Persistent opioid use among pediatric patients after surgery. *Pediatrics*, 144(1):e20172439, January 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Harrell, F. E. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, 247(18): 2543, 1982. doi: 10.1001/jama.1982.03320430047030.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *International Joint Conference on Artificial Intelligence*, 2019a.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

Helmerhorst, G. T. T., Vranceanu, A.-M., Vrahas, M., Smith, M., and Ring, D. Risk factors for continued opioid use one to two months after surgery for musculoskeletal trauma. *J. Bone & Joint Surgery*, 96(6):495–499, March 2014.

Henderson, R., Diggle, P., and Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

Hernán, M. A. and Robins, J. M. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2018. Forthcoming.

Herrington, W., Lacey, B., Sherliker, P., Armitage, J., and Lewington, S. Epidemiology of atherosclerosis and the potential to reduce the global burden of atherothrombotic disease. *Circulation research*, 118(4):535–546, 2016.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hochstein, A., Ahn, H.-I., Leung, Y. T., and Denesuk, M. Survival analysis for hdlss data with time dependent variables: Lessons from predictive maintenance at a mining service provider. In *Proceedings of 2013 IEEE International Conference on Service Operations and Logistics, and Informatics*, pp. 372–381. IEEE, 2013.

Hung, H. and Chiang, C.-T. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010a.

Hung, H. and Chiang, C.-t. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679, 2010b.

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Ishwaran, H. and Kogalur, U. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2019. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.9.1.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Ismail-Beigi, F., Craven, T., Banerji, M. A., Basile, J., Calles, J., Cohen, R. M., Cuddihy, R., Cushman, W. C., Genuth, S., Grimm Jr, R. H., et al. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the accord randomised trial. *The Lancet*, 376(9739):419–430, 2010.

Jarrett, D., Yoon, J., and van der Schaar, M. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE Journal of Biomedical and Health Informatics*, 24(2):424–436, 2019.

Jena, A. B., Goldman, D., Schaeffer, L. D., Weaver, L., and Karaca-Mandic, P. Opioid prescribing by multiple providers in medicare: Retrospective observational study of insurance claims. *BMJ*, 348:g1393, February 2014.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *Proc. Int. Conf. Mach. Learn.*, pp. 3020–3029, 2016.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jones, A. M. and O'Donnell, O. *Econometric analysis of health data*. John Wiley & Sons, 2002.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):53, 2017.

Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

Kehl, V. and Ulm, K. Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis*, 50(5):1338–1355, 2006.

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I.-H., and Kim, H. J. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, 9(1):1–10, 2019.

Kim, N., Matzon, J. L., Abboudi, J., Jones, C., Kirkpatrick, W., Leinberry, C. F., Liss, F. E., Lutsky, K. F., Wang, M. L., Maltenfort, M., and Ilyas, A. M. A prospective evaluation of opioid utilization after upper-extremity surgical procedures: Identifying consumption patterns and determining prescribing guidelines. *J. Bone Joint Surg.*, 98(20):e89, October 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Klueh, M. P., Hu, H. M., Howard, R. A., Vu, J. V., Harbaugh, C. M., Lagisetty, P. A., Brummett, C. M., Englesbe, M. J., Waljee, J. F., and Lee, J. S. Transitions of care for postoperative opioid prescribing in previously opioid-naïve patients in the USA: a retrospective review. *J. Gen. Intern. Med.*, in press, 2018.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., et al. The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

Kobus, A. M., Smith, D. H., Morasco, B. J., Johnson, E. S., Yang, X., Petrik, A. F., and Deyo, R. A. Correlates of higher-dose opioid medication use for low back pain in primary care. *J. Pain*, 13(11):1131–1138, November 2012.

Kouw, W. M. and Loog, M. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019.

Kraisangka, J. and Druzdzel, M. J. Making large Cox's proportional hazard models tractable in Bayesian networks. volume 52 of *Proceedings of Machine Learning Research*, pp. 252–263, Lugano, Switzerland, 06–09 Sep 2016. PMLR.

Kraisangka, J. and Druzdzel, M. J. A bayesian network interpretation of the Cox's proportional hazard model, Oct 2018. URL https://www.sciencedirect.com/science/article/pii/S0888613X17306308.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814, 2018.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3792–3803, 2019.

Kvamme, H., Borgan, Ø., and Scheel, I. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

Le, V., Nagpal, C., and Dubrawski, A. Identification of patients with stable coronary artery disease who benefit from ace inhibitors using cox mixture model for heterogeneous treatment effects. *Journal of Critical Care*, 74:154208, 2023.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Lee, C., Yoon, J., and Van Der Schaar, M. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 2019a.

Lee, C., Zame, W., Alaa, A., and Schaar, M. Temporal quilting for survival analysis. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 596–605, 2019b.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

Lee, K., Bargagli-Stoffi, F. J., and Dominici, F. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

Li, Y., Wang, J., Ye, J., and Reddy, C. K. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1715–1724. ACM, 2016.

Lin, D. On the breslow estimator. *Lifetime data analysis*, 13(4):471–480, 2007.

Linden, A. and Yarnold, P. R. Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, 2018.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. Subgroup identification based on differential effect search —– a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.*, 30(21):2601–2621, 2011.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Adv. Neur. Inf. Process. Syst.*, pp. 6446–6456, 2017.

Makary, M. A., Overton, H. N., and Wang, P. Overprescribing is major contributor to opioid crisis. *BMJ*, 359: j4792, 2017.

Manduchi, L., Marcinkevičs, R., Massi, M. C., Weikert, T., Sauter, A., Gotta, V., Müller, T., Vasella, F., Neidert, M. C., Pfister, M., et al. A deep variational approach to clustering survival data. *arXiv preprint arXiv:2106.05763*, 2021.

Marlin, B. M., Schmidt, M., , and Murphy, K. P. Group sparse priors for covariance estimation. In *Proc. Conf. Uncertainty Artif. Intell.*, pp. 383–392, Montreal, Canada, June 2009.

Maron, D. J., Hochman, J. S., O'Brien, S. M., Reynolds, H. R., Boden, W. E., Stone, G. W., Bangalore, S., Spertus, J. A., Mark, D. B., Alexander, K. P., et al. International study of comparative health effectiveness with medical and invasive approaches (ischemia) trial: rationale and design. *American heart journal*, 201:124–135, 2018.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Mishra, M., Martinsson, J., Rantatalo, M., and Goebel, K. Bayesian hierarchical model-based prognostics for lithium-ion batteries. *Reliability Engineering & System Safety*, 172, 12 2017. doi: 10.1016/j.ress.2017.11.020.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Vega, J. E. V., Brat, D. J., and Cooper, L. A. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

Morucci, M., Orlandi, V., Roy, S., Rudin, C., and Volfovsky, A. Adaptive hyper-box matching for interpretable individualized treatment effect estimation. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1089–1098. PMLR, 2020.

Mostacci, B., Liguori, R., and Cicero, A. F. Nutraceutical approach to peripheral neuropathies: Evidence from clinical trials. *Current Drug Metabolism*, 19(5):460–468, 2018.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, pp. 2901. NIH Public Access, 2015.

Nagpal, C. and Dubrawski, A. Interpretable treatment prioritization rule defines diabetic patients that benefit from prompt coronary revascularization. *Journal of the American College of Cardiology*, 81(8_Supplement): 2263–2263, 2023. doi: 10.1016/S0735-1097(23)02707-9. URL https://www.jacc.org/doi/abs/10.1016/S0735-1097%2823%2902707-9.

Nagpal, C., Li, X., Pinsky, M. R., and Dubrawski, A. Dynamically personalized detection of hemorrhage. In *Machine Learning for Healthcare Conference*, pp. 109–123. PMLR, 2019a.

Nagpal, C., Sangave, R., Chahar, A., Shah, P., Dubrawski, A., and Raj, B. Nonlinear semi-parametric models for survival analysis. *arXiv preprint arXiv:1905.05865*, 2019b.

Nagpal, C., Wei, D., Vinzamuri, B., Shekhar, M., Berger, S. E., Das, S., and Varshney, K. R. Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 19–29, 2020.

Nagpal, C., Jeanselme, V., and Dubrawski, A. Deep parametric time-to-event regression with time-varying covariates. In *Survival Prediction-Algorithms, Challenges and Applications*, pp. 184–193. PMLR, 2021a.

Nagpal, C., Jeanselme, V., and Dubrawski, A. Deep parametric time-to-event regression with time-varying covariates. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, volume 146 of *Proceedings of Machine Learning Research*, pp. 184–193. PMLR, 22–24 Mar 2021b. URL http://proceedings.mlr.press/v146/nagpal21a.html.

Nagpal, C., Li, X., and Dubrawski, A. Deep Survival Machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021c.

Nagpal, C., Yadlowsky, S., Rostamzadeh, N., and Heller, K. Deep cox mixtures for survival regression. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 674–708. PMLR, 06–07 Aug 2021d. URL https://proceedings.mlr.press/v149/nagpal21a.html.

Nagpal, C., Goswami, M., Dufendach, K., and Dubrawski, A. Counterfactual phenotyping with censored time-to-events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 3634–3644, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539110. URL https://doi.org/10.1145/3534678.3539110.

Nagpal, C., Potosnak, W., and Dubrawski, A. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 585–608. PMLR, 05–06 Aug 2022b. URL https://proceedings.mlr.press/v182/nagpal22a.html.

Nagpal, C., Sanil, V., and Dubrawski, A. Recovering sparse and interpretable subgroups with heterogeneous treatment effects with censored time-to-event outcomes. *arXiv preprint arXiv:2302.12504*, 2023.

National Cancer Institute, DCCPS, S. R. P. Surveillance, epidemiology, and end results (seer) program research data (1975-2016), 2019. URL www.seer.cancer.gov.

Neill, D. B. and Herlands, W. Machine learning for drug overdose surveillance. *J. Tech. Human Serv.*, 36(1):8–14, January 2018.

Nezhad, M. Z., Sadati, N., Yang, K., and Zhu, D. A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. *Expert Systems with Applications*, 115:16–26, 2019.

Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, pp. 38–41, 2019.

Okifuji, A. and Hare, B. D. The association between chronic pain and obesity. *J. Pain Res.*, 8:399–408, 2015.

Palmer, B. F. and Clegg, D. J. Electrolyte disturbances in patients with chronic alcohol-use disorder. *New England J. Med.*, 377(14):1368–1377, 2017.

Parente, S. T., Kim, S. S., Finch, M. D., Schloff, L. A., Rector, T. S., Seifeldin, R., and Haddox, J. D. Identifying controlled substance patterns of utilization requiring evaluation using administrative claims data. *Am. J. Managed Care*, 10(11):783–790, November 2004.

Park, S., Bastani, O., Weimer, J., and Lee, I. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3219–3229. PMLR, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Patil, P. R., Wolfe, J., Said, Q., Thomas, J., and Martin, B. C. Opioid use in the management of diabetic peripheral neuropathy (DPN) in a large commercially insured population. *Clin. J. Pain*, 31(5):414–424, May 2015.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pergolizzi, J. V., Gharibo, C., Passik, S., Labhsetwar, S., Taylor, R., Pergolizzi, J. S., and Müller-Schwefe, G. Dynamic risk factors in the misuse of opioid analgesics. *J. Psychosomatic Res.*, 72(6):443–451, 2012.

Pfohl, S., Duan, T., Ding, D. Y., and Shah, N. H. Counterfactual reasoning for fair clinical risk prediction. *arXiv preprint arXiv:1907.06260*, 2019.

Pfohl, S. R., Foryciarz, A., and Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *arXiv preprint arXiv:2007.10306*, 2020.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. Deep survival analysis. *Proceedings of the 1st Machine Learning for Healthcare Conference, PMLR*, 56:101–114, 2016a.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016b.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., and Yu, Y. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4798–4805, 2019.

Rosen, O. and Tanner, M. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9): 1119–1131, 1999.

Rosenblum, A., Marsch, L. A., Joseph, H., and Portenoy, R. K. Opioids and the treatment of chronic pain: controversies, current status, and future directions. *Exp. Clin. Psychopharmacol.*, 16(5):405–416, October 2008.

Royston, P. and Altman, D. G. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1), 2013. doi: 10.1186/1471-2288-13-33.

Royston, P. and Parmar, M. K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(1): 1–15, 2013.

Rubanova, Y., Chen, R. T., and Duvenaud, D. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*, 2019.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psych.*, 66(5):688–701, 1974.

Rubin, D. B. Which ifs have causal answers. *J. Am. Stat. Assoc.*, 81(396):961–962, 1986.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Schölkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.

Schroeder, A. R., Dehghan, M., Newman, T. B., Bentley, J. P., and Park, K. T. Association of opioid prescriptions from dental clinicians for US adolescents and young adults with subsequent opioid use and abuse. *JAMA Intern Med.*, in press.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. L., and Rauschecker, H. F. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994. doi: 10.1200/jco.1994.12.10.2086.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Song, Z., Henao, R., Carlson, D., and Carin, L. Learning sigmoid belief networks via monte carlo expectation maximization. In *Artificial Intelligence and Statistics*, pp. 1347–1355, 2016.

Soyka, M. Alcohol use disorders in opioid maintenance therapy: Prevalence, clinical correlates and treatment. *Eur. Addict. Res.*, 21(2):78–87, January 2015.

Springer, S. A., Korthuis, P. T., and Del Rio, C. Integrating treatment at the intersection of opioid use disorder and infectious disease epidemics in medical settings. *Ann Intern Med*, 169(5):335–336, 2018.

Stepanova, M. and Thomas, L. Survival analysis methods for personal loan data. *Operations Research*, 50(2): 277–289, 2002.

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Merz, C. N. B., Blum, C. B., Eckel, R. H., Goldberg, A. C., Gordon, D., Levy, D., Lloyd-Jones, D. M., et al. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B): 2889–2934, 2014.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Uno, H., Cai, T., Tian, L., and Wei, L.-J. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Ustun, B. and Rudin, C. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1125–1134, 2017.

Ustun, B. and Rudin, C. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

van Houwelingen, H. and Putter, H. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

Van Houwelingen, H. C. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

Vinzamuri, B. and Reddy, C. K. Cox regression with correlation based regularization for electronic health records. In *2013 IEEE 13th International Conference on Data Mining*, pp. 757–766. IEEE, 2013.

Vinzamuri, B., Li, Y., and Reddy, C. K. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 241–250. ACM, 2014.

Vittinghoff, E., McCulloch, C. E., Woo, C., and Cummings, S. R. Estimating long-term effects of treatment from placebo-controlled trials with an extension period, using virtual twins. *Statistics in Medicine*, 29(10): 1127–1136, 2010.

Volkow, N. D. and McLellan, A. T. Opioid abuse in chronic pain — misconceptions and mitigation strategies. *N. Engl. J. Med.*, 374(13):1253–1263, 2016.

Vyas, D. A., Eisenstein, L. G., and Jones, D. S. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wang, J., Li, C., Han, S., Sarkar, S., and Zhou, X. Predictive maintenance based on event-log analysis: A case study. *IBM Journal of Research and Development*, 61(1):11–121, 2017.

Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., and Naumann, T. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, 2020.

Wang, T. and Rudin, C. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing*, 2022.

Wei, G. C. and Tanner, M. A. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., et al. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248, 2018.

Wu, H., Tan, S., Li, W., Garrard, M., Obeng, A., Dimmery, D., Singh, S., Wang, H., Jiang, D., and Bakshy, E. Interpretable personalized experimentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4173–4183, 2022.

Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., and Azen, S. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

Xiao, C., Choi, E., and Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*.

Xiu, Z., Tao, C., and Henao, R. Variational learning of individual survival distributions. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 10–18, 2020.

Xu, Y., Ignatiadis, N., Sverdrup, E., Fleming, S., Wager, S., and Shah, N. Treatment heterogeneity with survival outcomes. *arXiv preprint arXiv:2207.07758*, 2022.

Yadlowsky, S., Hayward, R. A., Sussman, J. B., McClelland, R. L., Min, Y.-I., and Basu, S. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of internal medicine*, 169(1):20–29, 2018.

Yadlowsky, S., Basu, S., and Tian, L. A calibration metric for risk scores with survival data. In *Machine Learning for Healthcare Conference*, pp. 424–450, 2019.

Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A., Payton, M., and Tchounwou, P. B. Health and racial disparity in breast cancer. In *Breast Cancer Metastasis and Drug Resistance*, pp. 31–49. Springer, 2019.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24, 2011.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, 68 (1):49–67, 2006.

Zhang, J., Iyengar, V. S., Wei, D., Vinzamuri, B., Bastani, H., Macalalad, A. R., Fischer, A. E., Yuen-Reed, G., Mojsilović, A., and Varshney, K. R. Exploring the causal relationships between initial opioid prescriptions and outcomes. In *AMIA Workshop Data Min. Med. Informat.*, Washington, DC, USA, November 2017.

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2015.

Zhu, X., Yao, J., and Huang, J. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 544–547. IEEE, 2016.