Journal of Machine Learning Research 25 (2024) 1-85 Submitted 7/23; Revised 5/24; Published 8/24

Fairness in Survival Analysis with Distributionally Robust Optimization

Shu Hu* HU968@PURDUE.EDU
Department of Computer and Information Technology
Purdue University
Indianapolis, IN, 46202, USA

George H. Chen*† GEORGECHEN@CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Editor: David Sontag

Abstract

We propose a general approach for encouraging fairness in survival analysis models that is based on minimizing a worst-case error across *all* subpopulations that are "large enough" (occurring with at least a user-specified probability threshold). This approach can be used to convert a wide variety of existing survival analysis models into ones that simultaneously encourage fairness, *without* requiring the user to specify which attributes or features to treat as sensitive in the training loss function. From a technical standpoint, our approach applies recent methodological developments of *distributionally robust optimization* (DRO) to survival analysis. The complication is that existing DRO theory uses a training loss function that decomposes across contributions of individual data points, i.e., any term that shows up in the loss function depends only on a single training point. This decomposition does not hold for commonly used survival loss functions, including for the standard Cox proportional hazards model, its deep neural network variants, and many other recently developed survival analysis models that use loss functions involving ranking or similarity score calculations. We address this technical hurdle using a sample splitting strategy. We demonstrate our sample splitting DRO approach by using it to create fair versions of a diverse set of existing survival analysis models including the classical Cox model (and its deep neural network variant DeepSurv), the discrete-time model DeepHit, and the neural ODE model SODEN. We also establish a finite-sample theoretical guarantee to show what our sample splitting DRO loss converges to. Specifically for the Cox model, we further derive an exact DRO approach that does not use sample splitting. For all the survival models that we convert into DRO variants, we show that the DRO variants often score better on recently established fairness metrics (without incurring a significant drop in accuracy) compared to existing survival analysis fairness regularization techniques, including ones which directly use sensitive demographic information in their training loss functions.

Our code is available at: `https://github.com/discovershu/DRO_survival`.

Keywords: survival analysis, fairness, distributionally robust optimization

∗. equal contribution
†. corresponding author

1. Introduction

Survival analysis aims to model time durations before a critical event happens. Examples of such critical events include a patient dying, a convicted criminal reoffending, or a customer cancelling a subscription service. Predicting such time durations accurately could help plan patient treatments, make bail decisions, or target subscription pricing promotions. If a survival analysis model is to be used in high-stakes decision making, fairness could be an important design criterion. For example, in the case of making bail decisions with the help of predicted time durations until a criminal reoffends, we may want a survival analysis model that produces these predictions to be similarly accurate across different races.

One of the major recent advances in encouraging fairness for machine learning models is to minimize a worst-case error over *all* subpopulations that are "large enough" (e.g., Hashimoto et al. 2018; Duchi and Namkoong 2021; Li et al. 2021; Duchi et al. 2022; Hu et al. 2022a). In particular, a modeler specifies a minimum probability threshold α. The goal then is to ensure that all subpopulations that occur with probability at least α have low average error, whereas we make no promises for subpopulations that occur with probability less than α. The modeler need not provide a list of subpopulations to account for. This problem can be tractably solved in practice and is called *distributionally robust optimization* (DRO).

We emphasize that curating a list of all subpopulations to account for can be challenging for various reasons. For example, one major challenge is *intersectionality*: subpopulations that a machine learning model yields the worst accuracy scores for can be defined by complex intersections of sensitive attributes (such as age, race, and gender simultaneously taking on specific values) (Buolamwini and Gebru, 2018). Some of these attributes might require discretization (e.g., dividing age into bins), for which choosing the "best" discretization strategy might not be straightforward. Moreover, if there is a large number of features and we suspect that the sensitive attributes (encoded by specific features) could possibly be correlated with other features (not flagged as sensitive), then there is a question of whether these other features should also be accounted for in a listing of what the sensitive attributes are. DRO provides a theoretically sound alternative to having to specify which attributes to treat as sensitive in a training loss function.

Our main contribution in this paper is to show how to apply DRO to survival analysis. Specifically, we propose a general strategy for converting a wide variety of survival analysis models into ones that simultaneously encourage fairness. Our strategy supports all survival analysis models we are aware of that minimize a loss function (details on the general form of survival analysis models that our approach supports are in Section 3).

The key technical challenge is that existing DRO theory assumes that the overall training loss is the sum of individual loss terms, where each such term only depends on a single data point. This assumption fails to hold for commonly used survival analysis loss functions—including that of the standard Cox proportional hazards model (Cox, 1972)—that involve pairwise comparisons from ranking or similarity score evaluations (e.g., Steck et al. 2007; Lee et al. 2018; Chen 2020; Wu et al. 2021). In particular, there are loss terms that arise that incorporate information from multiple data points at once. We propose a sample splitting approach to address this technical challenge, and we establish a finite-sample theoretical guarantee on what our sample splitting DRO loss converges to. We point out that there are also parametric survival analysis models with loss functions that directly adhere to existing

DRO theory (e.g., parametric accelerated failure time models (Klein and Moeschberger, 2003, Chapter 12) or, as a more exotic example, the recently proposed neural ordinary differential equation (ODE) model called SODEN (Tang et al., 2022b)); such models can trivially be modified to use DRO without the sample splitting approach that we propose.

We specifically show how to derive DRO variants of the standard Cox model (Cox, 1972) (and its deep neural network variant DeepSurv (Faraggi and Simon, 1995; Katzman et al., 2018)), the discrete-time DeepHit model (Lee et al., 2018), and the neural ODE model called SODEN (Tang et al., 2022b). Again, we emphasize that our strategy for converting an existing survival analysis model to its DRO variant is fairly general and is not limited to only the few models that we showcase as illustrative examples.

We further derive an exact DRO approach specific to the Cox model that does not require sample splitting. In particular, by introducing additional parameters to optimize over for the Cox model's standard negative partial likelihood loss, it is possible to convert this loss function into one that decouples across training points. This derivation is specific to the Cox model though and does not easily generalize to other survival models.

On three standard datasets that have been previously used for research on fair survival analysis, we show that our DRO modification often outperforms various baseline fairness regularization techniques in terms of existing fairness metrics that focus on user-specified sensitive attributes. Most of these baselines require the user to specify which attributes to treat as sensitive attributes within the added regularization term. As with other fairness methods recently developed for survival analysis (e.g., Keya et al. (2021); Rahman and Purushotham (2022)), our approach also results in a drop in accuracy (compared to using a loss that does not encourage fairness). Note that our paper does not aim to find which survival model is the most accurate or the most fair. In fact, per survival model, there is in general a tradeoff between accuracy and fairness that can be tuned by the modeler. We show how to visualize this tradeoff using a plot inspired by an ROC curve.

Related work on fair survival analysis Despite many recent advances in survival analysis methodology (see, for instance, the survey by Wang et al. (2019)), very few of these advances study fairness (Keya et al., 2021; Zhang and Weiss, 2022; Sonabend et al., 2022; Rahman and Purushotham, 2022). We provide an overview of these existing papers, and we discuss how they differ from our work.

Keya et al. (2021) adapted existing fairness definitions to the survival analysis setting and showed how to encourage different notions of fairness by adding fairness regularization terms. Specifically, Keya et al. (2021) came up with individual (Dwork et al., 2012), group (Dwork et al., 2012), and intersectional (Foulds et al., 2020) fairness definitions specialized to Cox models. Keya et al. define individual fairness in terms of model predictions being similar for similar individuals, and group fairness in terms of different user-specified groups having similar average predicted outcomes. Intersectional fairness further considers subgroups defined by intersections of protected groups (e.g., individuals of a specific race and simultaneously a specific gender). However, a major limitation of the notions of fairness defined by Keya et al. is that they focus on predicted model outputs and do not actually use any of the ground truth label information. For example, if one uses age as a sensitive attribute and suppose we discretize age into two groups, then the notion of group fairness by Keya et al. would ask for the predicted outcomes of the two age groups to be similar, which

for healthcare problems often does not make sense (since age is often highly predictive of different health outcomes). Instead, in such a scenario, a more desirable notion of fairness is that the model's *accuracy* for the different age groups be similar.

To account for model accuracy, Zhang and Weiss (2022) introduced a fairness metric called *concordance imparity* that computes a quantity similar to the standard survival analysis accuracy metric of *concordance index* (Harrell et al., 1982) for different groups and then looks at the worst-case difference between any two groups' accuracy scores. Meanwhile, Rahman and Purushotham (2022) directly modified the fairness definitions of Keya et al. (2021) to account for ground truth label information, and also generalized these definitions to survival models beyond Cox models.

Separately, Sonabend et al. (2022) empirically explored how well existing survival analysis accuracy and calibration metrics measure bias by synthetically modifying datasets (e.g., undersampling disadvantaged groups). However, they do not propose any new fairness metric or survival model that encourages fairness.

The papers mentioned above that propose new methods for learning fair survival models all either require user-specified demographic information to treat as sensitive (possibly as a list of subpopulations or groups to account for) or are simply adding a regularization term that encourages smoothness in the model outputs (the individual fairness regularization by Keya et al. (2021) and Rahman and Purushotham (2022) are directly related to encouraging Lipschitz continuity; for details, see Appendix F). In contrast, our proposed DRO approach does not require the user to indicate which attributes to treat as sensitive in the training loss function, and is not simply encouraging the model output to be Lipschitz continuous.

Bibliographical note This paper significantly extends our previous conference paper (Hu and Chen, 2022) in methodological development and in experiments. For methodological development, whereas our conference paper only considered Cox models, we show in this journal paper version how to convert a much wider class of survival analysis models into their DRO variants that encourage fairness. In fact, this wider class of models consists of all survival models we are aware of that are learned by minimizing an overall loss function. Furthermore, this journal paper extension includes theoretical analysis of our sample splitting DRO approach and also an exact DRO approach for the Cox model without sample splitting; neither of these contributions were in our conference paper. For experiments, we demonstrate our conversion strategy on not only Cox models but also on DeepHit and SODEN models. Our experiments are overall more extensive, and the SEER dataset we now use is much larger ($\sim$28k data points in this version vs $\sim$4k in the conference paper). Lastly, we also add a new visualization for seeing the tradeoff between accuracy and fairness across multiple models within a single plot.

Outline The rest of the paper is organized as follows. We provide background on survival analysis, existing research on fairness in survival analysis, and DRO in Section 2. We then present our strategy for converting a wide family of existing survival analysis models into their corresponding DRO variants that encourage fairness in Section 3; notably, this section introduces a sample splitting DRO approach and formally establishes its rate of convergence. Specifically for the Cox model, we present an exact DRO Cox model without sample splitting in Section 4. We conduct experiments to compare DRO variants of Cox, DeepHit, and SODEN models to their original non-DRO variants as well as to variants of

these models that encourage fairness using non-DRO baseline regularization strategies. We conclude the paper in Section 6.

2. Background

We begin by reviewing the basic survival analysis problem setup in Section 2.1 and then provide three examples of survival analysis models (Cox, DeepHit, and SODEN) in Section 2.2. We then review DRO in Section 2.3. Throughout the paper, we frequently use the notation $[\ell] \triangleq \{1, 2, \ldots, \ell\}$ for any positive integer ℓ.

2.1 Survival Analysis Setup

Survival analysis aims to model the amount of time that will elapse before a critical event of interest happens. We assume that we have training data $\{(X_i, Y_i, \Delta_i)\}_{i=1}^n$, where training data point $i \in [n]$ has raw input $X_i \in \mathcal{X}$, observed duration $Y_i \geq 0$, and event indicator $\Delta_i \in \{0, 1\}$. If $\Delta_i = 1$ (i.e., the critical event of interest happened for the i-th data point), then Y_i is the time until the event happens. Otherwise, if $\Delta_i = 0$, then Y_i is the time until censoring for the i-th point, i.e., the true time until event is unknown but we know that it is at least Y_i. The raw input space $\mathcal{X}$ could be any input space supported by standard neural network software (e.g., tabular data, images, time series).

Each training data point (X_i, Y_i, Δ_i) is assumed to be generated as follows:

1. Sample raw input X_i from a raw input distribution $\mathbb{P}_X$.
2. Sample nonnegative time duration T_i (this is the true time until the critical event happens) from a conditional distribution $\mathbb{P}_{T|X=X_i}$.
3. Sample nonnegative time duration C_i (this is the true time until the data point is censored) from a conditional distribution $\mathbb{P}_{C|X=X_i}$.
4. If $T_i \leq C_i$ (the critical event happens before censoring), then set $Y_i = T_i$ and $\Delta_i = 1$. Otherwise, set $Y_i = C_i$ and $\Delta_i = 0$.

Distributions $\mathbb{P}_X$, $\mathbb{P}_{T|X}$, and $\mathbb{P}_{C|X}$ are shared across data points and are unknown. We assume that the random variables T_i and C_i are independent given raw input X_i (since the training data are i.i.d., this means that conditioned on the raw input, the censoring times are random and independent of each other, and they are also independent of the true survival times). We denote the CDF of distribution $\mathbb{P}_{T|X=x}$ as $F(\cdot|x)$.

Prediction A standard prediction task is to estimate the probability that a data point with raw input $x \in \mathcal{X}$ survives beyond time t. Formally, this is defined as the *conditional survival function*

$$S(t|x) \triangleq \mathbb{P}(T > t|X = x) = 1 - F(t|x) \qquad \text{for } t \geq 0. \tag{1}$$

Importantly, for raw input x, we are predicting an entire probability distribution (since $S(\cdot|x)$ encodes the same information as the CDF $F(\cdot|x)$).

Some survival analysis models, such as the Cox proportional hazards model (Cox, 1972), estimate a transformed version of $S(\cdot|x)$ called the *hazard function*, given by

$$h(t|x) \triangleq -\frac{\partial}{\partial t} \log S(t|x) \qquad \text{for } t \geq 0. \tag{2}$$

From negating both sides of this equation, integrating over time, and exponentiating, we get $S(t|x) = \exp(-\int_0^t h(u|x)\mathrm{d}u)$. Thus, if we have an estimate of $h(\cdot|x)$, then we can readily estimate the conditional survival function $S(\cdot|x)$.

2.2 Examples of Survival Analysis Models

We now review three examples of survival analysis models (Cox, DeepHit, and SODEN) that can be modified to encourage fairness using DRO. In reviewing these models, we focus on aspects most relevant to our exposition later for how to convert these models into their DRO variants. For all three examples, we denote the neural network to be learned as $f(\cdot;\theta)$, where θ denotes the parameters of the neural network. The domain and range of f depends on the specific survival model we look at. Meanwhile, the architecture of f is up to the modeler to specify, where standard strategies could be used (e.g., if the raw inputs are tabular data, then a multilayer perceptron could be used; if the raw inputs are images, a convolutional neural network could be used; etc).

2.2.1 Classical and Deep Cox Models

The classical Cox model assumes that the hazard function has the factorization

$$h(t|x) = h_0(t)\exp(f(x;\theta)), \tag{3}$$

where h_0 is called the baseline hazard function (h_0 maps a nonnegative time $t \geq 0$ to a nonnegative number), and neural network $f(\cdot;\theta)$ maps a raw input from $\mathcal{X}$ to a single real number (i.e., $f(\cdot;\theta)$ has domain $\mathcal{X}$ and range $\mathbb{R}$). In particular, $f(x;\theta)$ models the so-called *log partial hazard function* and could be thought of as assigning a real-valued "risk score" to raw input $x \in \mathcal{X}$: when $f(x;\theta)$ is higher, then x has a higher risk of the critical event happening, so that the survival time of x will tend to be lower.

The original Cox model (Cox, 1972) defines f to be a dot product: $f(x;\theta) = \theta^T x$, where θ and x are in the same Euclidean vector space. More recently, researchers replaced f with a neural network (Faraggi and Simon, 1995; Katzman et al., 2018), resulting in a method called DeepSurv (which could be viewed as a generalization of the original Cox model in that the classical definition $f(x;\theta) = \theta^T x$ is a simple neural network consisting of a linear layer with no bias and no nonlinear activation). In either case, the standard approach for learning a Cox model is to first learn the neural network parameters θ by minimizing the negative log partial likelihood:

$$L^{\mathrm{Cox}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_i^{\mathrm{Cox}}(\theta), \tag{4}$$

where the i-th data point's loss is

$$L_i^{\mathrm{Cox}}(\theta) \triangleq -\Delta_i\left[f(X_i;\theta) - \log \sum_{j\in[n] \text{ s.t. } Y_j \geq Y_i} \exp(f(X_j;\theta))\right]. \tag{5}$$

If the i-th data point is censored (i.e., $\Delta_i = 0$), then $L_i^{\mathrm{Cox}}(\theta) = 0$. Thus, the overall loss $L^{\mathrm{Cox}}(\theta)$ could be viewed as weighting the *uncensored* training points equally. After learning

θ, we then estimate h_0; as this step is not essential to our exposition, we explain it in Appendix A.1, along with details on constructing the final estimate of $S(\cdot|x)$.

We remark that the factorization in equation (3) is referred to as the *proportional hazards assumption*: regardless of what the input x is, the hazard function $h(\cdot|x)$ is always proportional to the baseline hazard function h_0. A consequence of this assumption is that the resulting conditional survival function $S(\cdot|x)$ is heavily constrained in terms of its shape. In particular, regardless of what x is, $S(t|x)$ must be a power of the function $S_0(t) \triangleq \exp(-\int_0^t h_0(u)\mathrm{d}u)$ (for details, see Appendix A.2). The next two survival analysis models that we describe do not have this assumption and can more flexibly estimate $S(\cdot|x)$.

2.2.2 DeepHit

A wide class of survival analysis models directly estimate (some transformed version of) the conditional survival function $S(\cdot|x)$ along a discretized time grid, without requiring the proportional hazards assumption. The time grid itself is up to the modeler to choose and can depend on the observed time Y_i and event indicator Δ_i variables in the training data. For example, we could use a uniformly-spaced time grid between the minimum and maximum observed times (for some user-specified number of discretized time steps), or we could have the time grid consist of all unique times in the training data in which the critical event happened (in fact, this how the classical Kaplan-Meier estimator (Kaplan and Meier, 1958) discretizes time). Some other time grids are discussed by Kvamme and Borgan (2021).

An example of a model that uses a discretized time grid is DeepHit (Lee et al., 2018). Note that DeepHit supports the so-called *competing risks* setting where there are multiple critical events of interest. For simplicity, we review DeepHit where we only present the case where there is a single critical event of interest, which reduces the problem setup to the same one we specified in Section 2.1.

Let $t_1 < t_2 < \cdots < t_m$ denote the m discretized time points based on some user-specified grid. We assume that these are the only time points in which the critical event or censoring could happen (if a critical event or censoring happens at some other time point, we quantize it to one of these m time points). Then DeepHit parameterizes the following conditional probability mass function using a neural network:

$$\mathbb{P}(T = t_j | X = x) = f_j(x; \theta) \qquad \text{for } j \in [m], \tag{6}$$

where neural network $f(x;\theta) = \big(f_1(x;\theta), f_2(x;\theta), \ldots, f_m(x;\theta)\big) \in [0,1]^m$ has parameters θ and maps a raw input $x \in \mathcal{X}$ to a probability distribution over the m time steps. In other words, the domain of $f(\cdot;\theta)$ is $\mathcal{X}$ and the range of $f(\cdot;\theta)$ is the probability simplex $\big\{z \in \mathbb{R}^m : z_j \geq 0 \text{ for all } j \in [m] \text{ and } \sum_{j=1}^m z_j = 1\big\}$. For example, when working with tabular data, f could be a multilayer perceptron, where the last linear layer outputs m numbers and has softmax activation.

Because of the parameterization in equation (6), we can write the conditional survival function $S(t|x)$ at any discrete time point t_j in terms of the neural network $f(\cdot;\theta)$:

$$S_j(x;\theta) \triangleq S(t_j|x) = \mathbb{P}(T > t_j | X = x) = \sum_{\ell=j+1}^{m} f_\ell(x;\theta) \quad \text{for } j \in [m].$$

To learn θ, DeepHit uses the sum of two loss terms, corresponding to a negative log likelihood term and, separately, a ranking loss term. In what follows, we use the notation $\kappa(Y_i) \in [m]$

to denote the time step index corresponding to the i-th training point's observed time Y_i (i.e., Y_i gets quantized to integer time step $\kappa(Y_i)$). For example, one way to define the function $\kappa : [t_1, \infty) \to [m]$ is as follows:[1]

$$\kappa(t) \triangleq \begin{cases} \ell & \text{if there exists time index } \ell \in [m] \text{ s.t. } t_\ell = t, \\ \max\{\ell \in [m] : t_\ell < t\} & \text{otherwise.} \end{cases} \tag{7}$$

Then the overall DeepHit loss is

$$\begin{aligned} L^{\text{DeepHit}}(\theta) \triangleq \beta \cdot \overbrace{\frac{1}{n}\sum_{i=1}^{n}\big[-\Delta_i \log(f_{\kappa(Y_i)}(X_i;\theta)) - (1-\Delta_i)\log(S_{\kappa(Y_i)}(X_i;\theta))\big]}^{\text{negative log likelihood loss term}} \\ + (1-\beta)\cdot \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\Delta_i \sum_{j\in[n] \text{ s.t. } \kappa(Y_j)>\kappa(Y_i)} \exp\left(\frac{S_{\kappa(Y_i)}(X_i;\theta) - S_{\kappa(Y_i)}(X_j;\theta)}{\sigma}\right)}_{\text{ranking loss term}}, \end{aligned} \tag{8}$$

where $\beta \in [0,1]$ and $\sigma > 0$ are hyperparameters. Note that this formulation of the overall loss follows the implementation of DeepHit by Kvamme et al. (2019) in the now-standard `pycox` software package and is slightly different from the original formulation by Lee et al. (2018) (the only difference is in the weights used to combine the two main loss terms).

For how we convert DeepHit into its DRO variant later, it will be helpful to rewrite the DeepHit loss in terms of individual losses:

$$L^{\text{DeepHit}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} L_i^{\text{DeepHit}}(\theta), \tag{9}$$

where the i-th individual loss is

$$\begin{aligned} L_i^{\text{DeepHit}}(\theta) = \beta \cdot \big[-\Delta_i \log(f_{\kappa(Y_i)}(X_i;\theta)) - (1-\Delta_i)\log(S_{\kappa(Y_i)}(X_i;\theta))\big] \\ + (1-\beta)\cdot \frac{1}{n}\cdot\Delta_i \sum_{j\in[n] \text{ s.t. } \kappa(Y_j)>\kappa(Y_i)} \exp\left(\frac{S_{\kappa(Y_i)}(X_i;\theta) - S_{\kappa(Y_i)}(X_j;\theta)}{\sigma}\right). \end{aligned} \tag{10}$$

2.2.3 SODEN

Recently, a number of researchers have considered a differential-equation approach to setting up a survival analysis model that can avoid the proportional hazards assumption while also not requiring the modeler to explicitly specify a discrete time grid (Groha et al., 2020; Moon et al., 2022; Tang et al., 2022a,b). We review one such model called SODEN (Survival model

1. Note that it is possible to instead define the domain of κ to be $[0, \infty)$, where we either require $t_1 \triangleq 0$ or alternatively we define a special time point $t_0 \triangleq 0$ (and have $\kappa(t) = 0$ when $t < t_1$, where t_1 is assumed to be positive). In the latter case where we introduce time point t_0, the range of κ would of course be $\{0, 1, \ldots, m\}$ instead of $[m] = \{1, 2, \ldots, m\}$.

through Ordinary Differential Equation Networks), proposed by Tang et al. (2022b). Note that we review a special case that is easier to describe and that corresponds to our survival analysis problem setup in Section 2.1, where survival times are all nonnegative.

In what follows, we denote $H(t|x) \triangleq -\log S(t|x)$. From how we defined the hazard function $h(t|x)$ in equation (2), we have $h(t|x) = \frac{\partial}{\partial t} H(t|x)$, so $H(t|x) = \int_0^t h(u|x)\mathrm{d}u$; this integral expression reveals why $H(t|x)$ is commonly called the *cumulative hazard function*.

SODEN uses a neural network $f(\cdot;\theta)$ to parameterize the hazard function as the solution to an ordinary differential equation (ODE):

$$\begin{cases} \frac{\partial}{\partial t} H(t|x) = h(t|x) = f\big((t, H(t|x), x); \theta\big) & \text{for } t > 0, \\ H(0|x) = 0 \qquad \text{(initial condition at time 0)}, \end{cases} \tag{11}$$

where the neural network $f(\cdot;\theta)$ has domain $[0,\infty) \times [0,\infty) \times \mathcal{X}$ and range $\mathbb{R}$. Specifically $f(\cdot;\theta)$ takes as input time $t \geq 0$, a cumulative hazard value $H(t|x)$ (which is nonnegative), and a raw input $x \in \mathcal{X}$, and $f(\cdot;\theta)$ outputs a single real number that is $h(t|x)$. For example, $f(\cdot;\theta)$ could concatenate all its inputs to form a single vector of numbers that is then treated as the input to a multilayer perceptron, where the final linear layer outputs a single number and has softplus activation (to ensure that the output is always positive). The initial condition follows from the fact that $H(0|x) = \int_0^0 h(u|x)\mathrm{d}u = 0$.

Learning neural networks in terms of ODEs (as in equation (11)) is possible thanks to the landmark paper by Chen et al. (2018). Importantly, using any user-specified ODE solver, given any raw input $x \in \mathcal{X}$ and neural network parameters θ, we can numerically solve the ODE in equation (11) (going from time 0 to any user-specified time $t > 0$) to obtain an estimate for $H(t|x)$; we denote this estimate as $H_{\text{ODE-solve}}(t|x;\theta)$. In particular, a major result of Chen et al. (2018) is that the loss function we use to train the neural network can contain terms involving $h(t|x) = f((t, H(t|x), x);\theta)$ and $H(t|x)$, where we replace $H(t|x)$ with $H_{\text{ODE-solve}}(t|x;\theta)$. Backpropagation is possible with the help of any ODE solver.

To train the SODEN model, Tang et al. (2022b) use the overall loss function

$$L^{\text{SODEN}}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} L_i^{\text{SODEN}}(\theta), \tag{12}$$

where the i-th individual loss is

$$L_i^{\text{SODEN}}(\theta) = -\Delta_i \log f\big((Y_i, H_{\text{ODE-solve}}(Y_i|X_i;\theta), X_i);\theta\big) + H_{\text{ODE-solve}}(Y_i|X_i;\theta). \tag{13}$$

Note that the overall loss (12) is just a negative log likelihood expression, so that minimizing this loss corresponds to solving a maximum likelihood problem.

2.3 Distributionally Robust Optimization (DRO)

DRO uses a worst-case average error over "large enough" subpopulations. Note that there are now a number of different versions of DRO (e.g., Hashimoto et al. 2018; Sagawa et al. 2020; Duchi and Namkoong 2021; Duchi et al. 2022). We specifically use the one by Hashimoto et al. (2018). Even though existing literature on DRO does not consider survival analysis to the best of our knowledge, we intentionally review DRO here using survival analysis notation that we have introduced in Section 2.1. In fact, existing DRO theory actually

works with many existing survival analysis loss functions already, without modification. In particular, survival analysis models for which each data point's individual loss does not depend on any other data points could trivially use existing DRO machinery. Examples of such survival analysis models include DeepHit when $\beta = 1$ (see equation (10)), SODEN (see equation (13)), as well as exponential, Weibull, log-logistic, log-normal, and generalized Gamma accelerated failure time models (Klein and Moeschberger, 2003, Chapter 12).

Problem setup Let $\mathbb{P}$ denote the joint distribution over each data point (X_i, Y_i, Δ_i). This joint distribution corresponds to the generative procedure described in Section 2.1. We assume that there are K groups that comprise $\mathbb{P}$. In particular, $\mathbb{P}$ is a mixture of K distributions $\mathbb{P} \triangleq \sum_{k=1}^{K} \pi_k \mathbb{P}_k$, where the k-th group occurs with probability $\pi_k \in (0, 1)$ and has associated distribution $\mathbb{P}_k$. Moreover, $\sum_{k=1}^{K} \pi_k = 1$. We assume that we do not know $\{(\pi_k, \mathbb{P}_k)\}_{k=1}^{K}$, nor do we know K. This setting, for instance, handles the case where we do not exhaustively know all subpopulations to consider. The smallest minority group corresponds to whichever group has the smallest π_k value. A simple special case would be when $K = 2$, where data are drawn from either a minority group or a majority group.

We would like to minimize the risk

$$R_{\max}(\theta) \triangleq \max_{k=1,\ldots,K} \mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}_k}[L_{\text{indiv}}(\theta; X, Y, \Delta)],$$

where L_{indiv} is a loss function that depends only on the parameters θ (for a survival analysis model that we aim to learn) and on a single data point (X, Y, Δ). However, minimizing $R_{\max}(\theta)$ is not possible as we do not know any of the latent groups. Nevertheless, it turns out that there is an optimization problem that we can tractably solve that minimizes an empirical version of an upper bound on $R_{\max}(\theta)$. We explain what the upper bound is in Section 2.3.1 and how to empirically minimize the upper bound in Section 2.3.2.

2.3.1 Upper Bound on the Risk $R_{\max}(\theta)$ Using DRO

For a set of distributions $\mathcal{B}_r(\mathbb{P})$ to be defined shortly, we consider minimizing the following alternative risk instead:

$$R_{\text{DRO}}(\theta; r) \triangleq \sup_{\mathbb{Q}\in\mathcal{B}_r(\mathbb{P})} \mathbb{E}_{(X,Y,\Delta)\sim\mathbb{Q}}[L_{\text{indiv}}(\theta; X, Y, \Delta)]. \tag{14}$$

This is the worst-case expected loss when we sample from any distribution in $\mathcal{B}_r(\mathbb{P})$.

The definition for $\mathcal{B}_r(\mathbb{P})$ is somewhat technical; we first give its precise definition and then state how to choose r so that $R_{\text{DRO}}(\theta; r)$ is an upper bound on $R_{\max}(\theta)$. Importantly, we will be able to efficiently minimize an empirical version of $R_{\text{DRO}}(\theta; r)$.

Definition 1 *The set $\mathcal{B}_r(\mathbb{P})$ consists of all distributions $\mathbb{Q}$ that have the same (or smaller) support as $\mathbb{P}$ and have χ^2-divergence at most r from distribution $\mathbb{P}$. Formally,*

$$\mathcal{B}_r(\mathbb{P}) \triangleq \{\textit{distribution } \mathbb{Q} \textit{ such that } \mathbb{Q} \ll \mathbb{P} \textit{ and } D_{\chi^2}(\mathbb{Q}\|\mathbb{P}) \le r\},$$

where, using standard measure theory notation, "$\mathbb{Q} \ll \mathbb{P}$" means that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, and $D_{\chi^2}(\mathbb{Q}\|\mathbb{P}) \triangleq \int (\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} - 1)^2 \mathrm{d}\mathbb{P}$.

Working with $\mathcal{B}_r(\mathbb{P})$ turns out to be straightforward so long as we have a lower bound on the smallest group's probability (i.e., a lower bound on $\min_{k=1,\ldots,K} \pi_k$).

Proposition 2 *(Directly follows from Proposition 2 of Hashimoto et al. (2018)) Suppose that we have a lower bound $\alpha > 0$ on the K latent groups' probabilities of occurring (i.e., $\alpha \leq \min_{k=1,\ldots,K} \pi_k$). Then $R_{DRO}(\theta; r_{\max}) \geq R_{\max}(\theta)$, where $r_{\max} \triangleq (\frac{1}{\alpha} - 1)^2$.*

In other words, if we have a guess for $\alpha \in (0, \min_{k=1,\ldots,K} \pi_k]$, then it suffices to choose r for $\mathcal{B}_r(\mathbb{P})$ to be $r_{\max} = (\frac{1}{\alpha} - 1)^2$. Furthermore, the risk $R_{\text{DRO}}(\theta; r_{\max})$ is an upper bound on $R_{\max}(\theta)$. In practice, $\alpha \in (0, 1)$ is a user-specified hyperparameter since we do not know $\pi_1, \ldots, \pi_K$ nor K. Choosing α to be smaller means that we want to ensure that groups with smaller probabilities of occurring also have low expected loss. For example, setting $\alpha = 0.1$ means that the "rarest" group that we want to ensure low expected loss for occurs with probability at least 0.1.

To provide some more detail, if there is some underlying true K unknown subpopulation distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K$ (where the value of K itself is also unknown), it is important to keep in mind that often times it suffices to consider there to simply be two subpopulations: a minority subpopulation which we could without loss of generality take to be $\mathbb{P}_1$ (since we can reorder the subpopulations so that the minority subpopulation of interest is the first one) and everyone else (the combination of $\mathbb{P}_2, \ldots, \mathbb{P}_K$, i.e., $\sum_{k=2}^{K} \pi_k \mathbb{P}_k$); note that this is a mixture of two distributions now. Then we would be tuning α with the hope that if $\mathbb{P}_1$ occurs with probability π_1 that is at least α, then we ensure low expected loss for $\mathbb{P}_1$. However, if $\alpha > \pi_1$, then we would no longer ensure that $\mathbb{P}_1$ has low expected loss. However, if the combination of $\mathbb{P}_1$ and any of the other mixture components, say, $\mathbb{P}_2$ have probability $\pi_1 + \pi_2 \geq \alpha$, then we would be ensuring that this larger subpopulation of $\pi_1 \mathbb{P}_1 + \pi_2 \mathbb{P}_2$ has low expected loss. In this manner, as we increase α, we are ensuring low expected loss across a larger subpopulation, where this subpopulation could be the combination of multiple of the $\mathbb{P}_k$ distributions. More precisely, for any choice of α, if $\mathcal{K}$ is any subset of $[K]$ such that $\sum_{k \in \mathcal{K}} \pi_k \geq \alpha$, then we would be ensuring low expected loss for the subpopulation $\sum_{k \in \mathcal{K}} \pi_k \mathbb{P}_k$.

2.3.2 Empirical DRO Risk

The next issue is how to minimize the risk $R_{\text{DRO}}(\theta; r_{\max})$, which at a first glance might appear daunting due to the supremum over all distributions in $\mathcal{B}_{r_{\max}}(\mathbb{P})$. However, a fundamental theoretical result from DRO literature is that $R_{\text{DRO}}(\theta; r_{\max})$ can be written in a form that is amenable to computation.

Proposition 3 *(Lemma 1 in Duchi and Namkoong (2021)) Suppose $\widehat{\ell}(\theta; X, Y, \Delta)$ is upper semi-continuous with respect to θ. Let $[\cdot]_+$ denote the ReLU function (i.e., $[a]_+ \triangleq \max\{a, 0\}$ for any $a \in \mathbb{R}$), and $C_\alpha \triangleq \sqrt{2(\frac{1}{\alpha} - 1)^2 + 1}$. Then*

$$R_{DRO}(\theta; r_{\max}) = \inf_{\eta \in \mathbb{R}} \left\{ C_\alpha \sqrt{\mathbb{E}_{(X,Y,\Delta) \sim \mathbb{P}} \left[[L_{indiv}(\theta; X, Y, \Delta) - \eta]_+^2 \right]} + \eta \right\}, \tag{15}$$

where, as a reminder, $r_{\max} = (\frac{1}{\alpha} - 1)^2$.

Algorithm 1: DRO

Input: A training dataset $\{(X_i, Y_i, \Delta_i)\}_{i=1}^n$, minimum subpopulation probability hyperparameter α, learning rate ξ, max_iterations

Output: Survival model parameters $\widehat{\theta}$

Obtain initial survival model parameters $\widehat{\theta}_0$ (e.g., using default PyTorch parameter initialization).
for $\ell = 0$ *to* max_iterations **do**
  for $i = 1$ *to* n **do**
    Set $u_i \leftarrow L_{\text{indiv}}(\widehat{\theta}_\ell; X_i, Y_i, \Delta_i)$.
  end
  Set $\widehat{\eta} \leftarrow \arg\min_{\eta \in \mathbb{R}} \left\{ \left(\sqrt{2\left(\frac{1}{\alpha} - 1\right)^2 + 1} \right) \sqrt{\frac{1}{n} \sum_{i=1}^n [u_i - \eta]_+^2} + \eta \right\}$, where this minimization is solved using binary search. (This step directly corresponds to minimizing $L_{\text{DRO}}(\widehat{\theta}_\ell, \eta)$ as given in equation (16).)
  Set $\widehat{\theta}_{\ell+1} \leftarrow \widehat{\theta}_\ell - \xi \cdot \nabla_\theta L_{\text{DRO}}(\widehat{\theta}_\ell, \widehat{\eta})$.
end
return $\widehat{\theta} \leftarrow \widehat{\theta}_{\text{max\_iterations}+1}$

The right-hand side of equation (15) could be interpreted as follows. Suppose that we have achieved the optimal value η^*. Then the loss from a data point will be ignored if it is less than η^* (due to the ReLU function). Thus, only the data points with losses above η^* are considered for learning the survival model.

Note that as we vary the model parameters θ, the different data points' losses change. Thus, as a function of θ, the DRO risk $R_{\text{DRO}}(\theta; r_{\max})$ dynamically adjusts which data points to focus on, always prioritizing the points with the highest loss values (again, we only consider the points with a loss greater than the optimal value of η).

We can readily minimize an empirical version of $R_{\text{DRO}}(\theta; r_{\max})$. Specifically, we replace the expectation on the right-hand side of equation (15) with an empirical average to arrive at the following optimization problem:

$$\min_{\theta \in \Theta, \eta \in \mathbb{R}} \Bigg(\underbrace{C_\alpha \sqrt{\frac{1}{n} \sum_{i=1}^n [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+^2} + \eta}_{\triangleq L_{\text{DRO}}(\theta, \eta)} \Bigg), \tag{16}$$

where Θ denotes the feasible set of the model parameters.

Numerical optimization The optimization problem in equation (16) can be solved with an iterative gradient descent approach (Hu et al., 2020, 2021, 2022b). Specifically, we first initialize the model parameters θ. Then, following Hashimoto et al. (2018), we alternate between two steps:

- We fix θ and update η by finding the value of η that minimizes $L_{\text{DRO}}(\theta, \eta)$. To do this, we use binary search to find the global optimum of η since $L_{\text{DRO}}(\theta, \eta)$ is a convex function with respect to η.
- We fix η and update θ by minimizing $L_{\text{DRO}}(\theta, \eta)$ (e.g., using gradient descent).

We stop iterating after user-specified stopping criteria are reached (e.g., maximum number of iterations reached, early stopping due to no improvement in a validation metric after a pre-specified number of epochs). The pseudocode can be found in Algorithm 1.

3. Converting Existing Survival Analysis Models into DRO Variants

Throughout this section, we assume that the training points $\{(X_i, Y_i, \Delta_i)\}_{i=1}^n$ are generated by the procedure stated in Section 2.1. We describe the general class of survival models that we can convert into DRO variants in Section 3.1. For some models (such as SODEN), the existing DRO approach stated in Section 2.3 directly works without modification. For other survival models (such as Cox models), existing DRO theory does not work as advertised and we propose a sample splitting DRO approach in Section 3.2 to obtain an approximate loss to minimize that does comply with DRO theory. We establish theoretical guarantees for this sample splitting DRO approach in Section 3.3.

3.1 Class of Survival Models Convertible Into DRO Variants

Our technique for converting a survival model into its DRO variant works with any survival model that minimizes a loss of the form

$$L_{\text{average}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_i(\theta; \mathcal{A}_i), \tag{17}$$

where the i-th loss term L_i depends on training point $i \in [n]$ as well as possibly other training points $\mathcal{A}_i \subseteq [n] \setminus \{i\}$. We refer to $\mathcal{A}_i$ as the *adjacency set* for the i-th training point, where $\mathcal{A}_i$ can be empty. The basic idea is that the i-th training point's loss term could potentially depend on training points aside from the i-th training point.

Importantly, in what follows, we assume that we have access to a function $\mathcal{A}^*$ that tell us for any data point what its adjacency set is, and we also have access to a function $L^*(\cdot, \cdot; \theta)$ (with parameter variable θ) that tells us for any data point what its individual loss term is. Note that θ contains all the underlying survival model's parameters and is thus shared across data points. The functions $\mathcal{A}^*$ and $L^*(\cdot, \cdot; \theta)$ (to be defined shortly) can be evaluated even for points that are not in the training data. We first formally describe these functions and then we state what they are for the survival models we presented in Section 2.2.

Adjacency function Let $\mathcal{Z} \triangleq \mathcal{X} \times [0, \infty) \times \{0, 1\}$ denote the set of possible data points (for instance, each training point (X_i, Y_i, Δ_i) belongs to $\mathcal{Z}$). We assume that given any $(x, y, \delta) \in \mathcal{Z}$ and any set $\mathcal{C}$ of data points from $\mathcal{Z}$, there is a function $\mathcal{A}^*$ that tells us which points in $\mathcal{C}$ are adjacent to (x, y, δ); namely, $\mathcal{A}^*((x, y, \delta), \mathcal{C})$ denotes the subset of points in $\mathcal{C}$ that are adjacent to (x, y, δ). This means that for the i-th training point (X_i, Y_i, Δ_i), the training points adjacent to the i-th point (and excluding the i-th point itself) is given by

$$\mathcal{Z}_i \triangleq \mathcal{A}^*\big((X_i, Y_i, \Delta_i), \{(X_j, Y_j, \Delta_j) \text{ for } j \in [n] \setminus \{i\}\}\big).$$

Then the adjacency set $\mathcal{A}_i$ is precisely defined as the set of training data indices corresponding to the data points in $\mathcal{Z}_i$:

$$\mathcal{A}_i \triangleq \{j \in [n] \text{ such that } (X_j, Y_j, \Delta_j) \in \mathcal{Z}_i\}. \tag{18}$$

Individual loss function Next, individual loss functions are determined using a function $L^*(\cdot, \cdot; \theta)$, where parameter variable θ. Similar to the function $\mathcal{A}^*$, $L^*(\cdot, \cdot; \theta)$ takes two inputs: a single data point from $\mathcal{Z}$ and a (possibly empty) set of data points from $\mathcal{Z}$. Specifically,

for any $(x, y, \delta) \in \mathcal{Z}$ and any set $\mathcal{C}$ of data points from $\mathcal{Z}$, we use $L^*((x, y, \delta), \mathcal{C}; \theta)$ to denote the individual loss for data point (x, y, δ). We then define

$$L_i(\theta; \mathcal{I}) \triangleq L^*\big((X_i, Y_i, \Delta_i), \{(X_j, Y_j, \Delta_j) \text{ for } j \in \mathcal{I}\}; \theta\big) \qquad \text{for any } \mathcal{I} \subseteq [n]. \tag{19}$$

In particular, the i-th data point's loss in equation (17) is taken to be $L_i(\theta; \mathcal{A}_i)$.

We now give explicit examples for the functions $\mathcal{A}^*$ and $L^*(\cdot, \cdot; \theta)$.

Example 1 (Cox models) *For any $(x, y, \delta) \in \mathcal{Z}$ and for any (possibly empty) set $\mathcal{C}$ of data points from $\mathcal{Z}$, define the adjacency function*

$$\mathcal{A}^*\big((x, y, \delta), \mathcal{C}\big) \triangleq \begin{cases} \varnothing & \text{if } \delta = 0, \\ \big\{(x', y', \delta') \in \mathcal{C} \text{ such that } y' \geq y\big\} & \text{otherwise,} \end{cases}$$

and the individual loss function

$$L^*((x, y, \delta), \mathcal{C}; \theta) \triangleq -\delta \bigg[f(x; \theta) - \log\bigg(\exp(f(x; \theta)) + \sum_{(x', y', \delta') \in \mathcal{C}} \exp(f(x'; \theta)) \bigg) \bigg].$$

One can verify that plugging in these choices for $\mathcal{A}^*$ and $L^*(\cdot, \cdot; \theta)$ into equations (18) and (19) to obtain $\mathcal{A}_i$ and $L_i(\theta; \mathcal{A}_i)$, and subsequently plugging $\mathcal{A}_i$ and $L_i(\theta; \mathcal{A}_i)$ into equation (17), we recover the Cox loss from equation (4).

Example 2 (DeepHit) *Recall that DeepHit discretizes time so as to use the user-specified grid $t_1 < t_2 < \cdots < t_m$, and $\kappa : [t_1, \infty) \to [m]$ maps from a continuous time to one of the discrete time indices as given in equation (7). For any $(x, y, \delta) \in \mathcal{Z}$ and for any (possibly empty) set $\mathcal{C}$ of data points from $\mathcal{Z}$, define the adjacency function*

$$\mathcal{A}^*\big((x, y, \delta), \mathcal{C}\big) \triangleq \begin{cases} \varnothing & \text{if } \delta = 0, \\ \big\{(x', y', \delta') \in \mathcal{C} \text{ such that } \kappa(y') \geq \kappa(y)\big\} & \text{otherwise,} \end{cases}$$

and the individual loss function

$$\begin{aligned} L^*((x, y, \delta), \mathcal{C}; \theta) \triangleq\ & \beta \cdot \big[-\delta \log(f_{\kappa(y)}(x; \theta)) - (1 - \delta) \log(S_{\kappa(y)}(x; \theta)) \big] \\ & + (1 - \beta) \cdot \frac{1}{n} \cdot \delta \sum_{(x', y', \delta') \in \mathcal{C}} \exp\Big(\frac{S_{\kappa(y)}(x; \theta) - S_{\kappa(y)}(x'; \theta)}{\sigma} \Big). \end{aligned} \tag{20}$$

Plugging in these choices for $\mathcal{A}^*$ and $L^*(\cdot, \cdot; \theta)$ into equations (18) and (19) to obtain $\mathcal{A}_i$ and $L_i(\theta; \mathcal{A}_i)$, and subsequently plugging $\mathcal{A}_i$ and $L_i(\theta; \mathcal{A}_i)$ into equation (17), we recover the DeepHit loss from equation (9).

Specifically when $\beta = 1$, note that the second term (i.e., the ranking loss) in equation (20) becomes 0, so that in this special case, the adjacency function $\mathcal{A}^*$ can just always be set to output the empty set. Conceptually, the ranking loss is the only reason that the DeepHit loss has terms that couple different data points.

If $\mathcal{A}_i = \varnothing$ for all $i \in [n]$, then we can directly use the existing DRO optimization (16); the overall loss decouples across the different data points so we do not run into issues where multiple data points get "coupled".

Example 3 (SODEN) *For the SODEN model, the overall loss function* (12) *actually has no coupling across training points, so it suffices to define the adjacency function $\mathcal{A}^*$ to always output the empty set. Meanwhile, for any $(x, y, \delta) \in \mathcal{Z}$ and any (possibly empty) set of data points from $\mathcal{Z}$, we define the individual loss function to be*

$$L^*((x, y, \delta), \mathcal{C}; \theta) \triangleq -\delta \log f\big((y, H_{\text{ODE-solve}}(y|x; \theta), x); \theta\big) + H_{\text{ODE-solve}}(y|x; \theta).$$

With these definitions of $\mathcal{A}^*$ and $L^*(\cdot, \cdot; \theta)$, one can show that equation (17) *becomes the SODEN loss from equation* (13)*.*

As our examples above illustrate, the loss function in equation (17) can vary quite a bit across different survival models. For Cox models, the loss function corresponds to a negative partial log likelihood (it is called "partial" since it excludes the baseline hazard function; we discuss this in more detail in Section 4). For the DeepHit model, the loss function corresponds to the sum of a negative log likelihood term and a ranking loss term (e.g., if hyperparameter β is set equal to 0, then we would only be using the ranking loss term). For the SODEN model, the loss function is just a negative log likelihood. In particular, the loss function is not required to be a negative log likelihood, such as in the case of using only the DeepHit ranking loss (without the negative log likelihood term).

3.2 Applying DRO When Adjacency Sets Can be Nonempty

We now discuss how to use DRO when $\mathcal{A}_i$ is not guaranteed to be empty.

3.2.1 Heuristic Approach

To convert a survival analysis model that minimizes the loss (17) into one that uses DRO, a heuristic approach that does not comply with existing DRO theory would be to solve the DRO optimization problem (16), ignoring the fact that the individual loss terms are not guaranteed to depend only on a single data point each. To be clear, existing DRO theory effectively requires that the sets $\mathcal{A}_i$ are all empty. As a preview of our experimental results, we mention that this heuristic approach actually works well in practice but we lack any justification as to why it should be expected to work well.

3.2.2 Sample Splitting Approach

We now propose a sample splitting approach that creates an approximate loss function that complies with existing DRO theory. We divide the training data into two sets $\mathcal{D}_1 \subset [n]$ and $\mathcal{D}_2 \triangleq [n] \setminus \mathcal{D}_1$ of sizes $n_1 \triangleq |\mathcal{D}_1|$ and $n_2 \triangleq |\mathcal{D}_2| = n - n_1$. The basic idea is that we treat the data points in $\mathcal{D}_2$ as fixed, and then define a DRO loss only over data points in $\mathcal{D}_1$. For each $i \in \mathcal{D}_1$, we replace its original individual loss $L_i(\theta; \mathcal{A}_i)$ with an approximate version $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$ that only depends on the i-th point along with points in $\mathcal{D}_2$. Specifically, we minimize the new DRO loss function

$$L_{\text{DRO}}^{\text{split}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) \triangleq C_\alpha \sqrt{\frac{1}{|\mathcal{D}_1|} \sum_{i \in \mathcal{D}_1} [L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2) - \eta]_+^2} + \eta. \tag{21}$$

The key observation is that conditioned on the points in $\mathcal{D}_2$, the loss terms $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$ appear i.i.d. across $i \in \mathcal{D}_1$ and the i-th loss only depends on the i-th data point (and possibly points in $\mathcal{D}_2$ which are treated as fixed). Hence, the original DRO theory applies. More formally, we can state the following.

Proposition 4 *Suppose that we condition on indices $\mathcal{D}_2$ and the data $\{(X_i, Y_i, \Delta_i) : i \in \mathcal{D}_2\}$. Then the individual losses $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$ for $i \in \mathcal{D}_1$ appear i.i.d. Consequently, we can directly apply Propositions 2 and 3, where*

$$L_{indiv}(\theta; X, Y, \Delta) \triangleq L^*\big((X, Y, \Delta), \underbrace{\mathcal{A}^*\big((X, Y, \Delta), \{(X_j, Y_j, \Delta_j) : j \in \mathcal{D}_2\}\big)}_{\substack{\text{points in } \mathcal{D}_2 \\ \text{adjacent to } (X,Y,\Delta)}}; \theta\big). \tag{22}$$

Clearly this sample splitting strategy is introducing an approximation since we replace $L_i(\theta; \mathcal{A}_i)$ with $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$. However, it is unclear how to quantify the approximation error of the resulting individual loss in equation (22). The technical challenge is that to measure this approximation error, we need to state what the target individual loss function is that we are measuring the error from. However, the problem is that DRO theory, to the best of our knowledge, does not work with an individual loss function that has coupling across points. In short, it is unclear what the "correct" L_{indiv} function (that is compliant with DRO theory) should even be for the general class of survival models that we consider.

One could view Proposition 4 as positing an individual loss function that can be used with DRO theory. In particular, suppose that we treat $n_2 = |\mathcal{D}_2|$ as fixed, and we sample $\{(X'_j, Y'_j, \Delta'_j)\}_{j=1}^{n_2}$ i.i.d. from $\mathbb{P}$ (the same distribution that the training data are sampled from), where these freshly sampled points are also independent of the training data. Then equation (22) could be viewed as an empirical estimate of the individual risk

$$R_{\text{indiv}}(\theta; X, Y, \Delta) \triangleq \mathbb{E}_{\{(X'_j, Y'_j, \Delta'_j)\}_{j=1}^{n_2}}\big[L^*\big((X, Y, \Delta), \mathcal{A}^*((X, Y, \Delta), \{(X'_j, Y'_j, \Delta'_j)\}_{j=1}^{n_2}); \theta\big)\big]. \tag{23}$$

We refer to R_{indiv} as a risk rather than a loss since it cannot be computed exactly in practice due to the expectation. This risk says that for any individual data point, we measure its error in comparison to n_2 randomly sampled reference data points.

We point out that our sample splitting strategy is somewhat inspired by the "case control" strategy by Kvamme et al. (2019), where instead of using the original Cox loss, they approximate each individual data point's loss (which could depend on many other data points) to only depend on a *single* other randomly sampled reference data point. Kvamme et al. found that by optimizing this modified loss, the resulting model's prediction accuracy is often about as good as using the original Cox loss.

Returning to the earlier question of quantifying the "approximation error" of replacing $L_i(\theta; \mathcal{A}_i)$ with $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$, we reiterate that we do not know of a clear way to do this. If we state that our goal is to minimize the individual risk $R_{\text{indiv}}(\theta; X, Y, \Delta)$, then clearly we would use the empirical version of this individual risk as given by $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$, and in particular, it would not make sense to use $L_i(\theta; \mathcal{A}_i)$, which we suspect does not correspond to any individual loss or risk that works with DRO theory when $\mathcal{A}_i$ is nonempty.

Cross-fitting Although minimizing $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ is compliant with DRO theory, it uses data "less effectively" since at most n_1 data points (rather than n) are used to compute the empirical average inside the square root in equation (21) (as compared to the empirical average inside the square root of $L_{\text{DRO}}(\theta, \eta)$ in equation (16)); as reminder, some individual loss terms might actually be zero (in the case of the Cox model, individual loss terms are zero for censored data). Moreover, in the new split loss $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$, each individual loss within the empirical average is computed using only a subset of each individual's original adjacency set (for each $i \in \mathcal{D}_1$, we approximate individual loss $L_i(\theta; \mathcal{A}_i)$ by $L_i(\theta; \mathcal{A}_i \cap \mathcal{D}_2)$).

To "more effectively" use data, we use the basic idea from cross-fitting (e.g., Schick 1986; Chernozhukov et al. 2018). Whereas the loss $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ treats $\mathcal{D}_2$ as fixed and computes an average over $\mathcal{D}_1$, we also do the opposite: we treat $\mathcal{D}_1$ as fixed and compute an average over $\mathcal{D}_2$, which would corresponds precisely to using the loss $L^{\text{split}}_{\text{DRO}}(\theta, \eta', \mathcal{D}_2 \mid \mathcal{D}_1)$; note that we use a different variable η' than the variable η used in $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$. Overall, we minimize the loss

$$L^{\text{split}}_{\text{DRO}}(\theta, \eta, \eta') \triangleq \frac{1}{2} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) + \frac{1}{2} L^{\text{split}}_{\text{DRO}}(\theta, \eta', \mathcal{D}_2 \mid \mathcal{D}_1)$$

via coordinate descent, alternating between the following steps:

- Treating η' and θ as fixed, we update η by finding the value of η that minimizes $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \eta')$. This amounts to solving $\min_{\eta \in \mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ using binary search (since $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ is convex w.r.t. η).
- Treating η and θ as fixed, we update η' by finding the value of η' that minimizes $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \eta')$. This amounts to solving $\min_{\eta' \in \mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta, \eta', \mathcal{D}_2 \mid \mathcal{D}_1)$ using binary search.
- Treating η and η' as fixed, we update θ by minimizing $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \eta')$ (e.g., using gradient descent).

We provide the pseudocode in Algorithm 2.

Note that it is possible to do cross-fitting where we partition the training data into more than 2 sets $\mathcal{D}_1$ and $\mathcal{D}_2$, similar to how for K-fold cross-validation, one could use more than 2 folds. We explain how to do this in Appendix B. However, for simplicity, we only use 2-fold cross-fitting in our experiments later.

3.3 Theoretical Guarantees for the Sample Splitting Approach

We now derive a theoretical guarantee for our sample splitting approach. We begin by stating assumptions on the survival data generating process:

- **A1 (compact raw input space).** We assume that the raw input space $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, and we denote its ε-covering number in Euclidean distance by $\mathbb{N}(\varepsilon, \mathcal{X})$.
- **A2 (discrete time).** We assume that the survival and censoring times are discrete along a grid $t_1 < t_2 < \cdots < t_m$ (with $m \geq 2$), and that all of these time points are used in the sense that there exists a positive constant $\zeta > 0$ such that

  $$\mathbb{P}(Y = t_\ell) \geq \zeta \qquad \text{for all time indices } \ell \in [m].$$

  In other words, the probability of an observed time being equal to t_ℓ is never 0.

Algorithm 2: DRO (SPLIT)

Input: A training dataset $\{(X_i, Y_i, \Delta_i)\}_{i=1}^n$, minimum subpopulation probability hyperparameter α, subset size n_1, learning rate ξ, max_iterations

Output: Survival model parameters $\widehat{\theta}$

Obtain initial survival model parameters $\widehat{\theta}_0$ (e.g., using default PyTorch parameter initialization).
Set $\mathcal{D}_1 \leftarrow \{1, 2, \ldots, n_1\}$ and $\mathcal{D}_2 \leftarrow \{n_1 + 1, \ldots, n\}$.
for $\ell = 0$ *to* max_iterations **do**
  for $i \in \mathcal{D}_1$ **do**
    Set $u_i \leftarrow L_i(\widehat{\theta}_\ell; \mathcal{A}_i \cap \mathcal{D}_2)$.
  end
  Set $\widehat{\eta} \leftarrow \arg\min_{\eta \in \mathbb{R}} \left\{ \left(\sqrt{2\left(\frac{1}{\alpha} - 1\right)^2 + 1} \right) \sqrt{\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} [u_i - \eta]_+^2} + \eta \right\}$, where this minimization is solved using binary search.
  for $i \in \mathcal{D}_2$ **do**
    Set $v_i \leftarrow L_i(\widehat{\theta}_\ell; \mathcal{A}_i \cap D_1)$.
  end
  Set $\widehat{\eta}' \leftarrow \arg\min_{\eta' \in \mathbb{R}} \left\{ \left(\sqrt{2\left(\frac{1}{\alpha} - 1\right)^2 + 1} \right) \sqrt{\frac{1}{n_2} \sum_{i \in \mathcal{D}_2} [v_i - \eta']_+^2} + \eta' \right\}$, where this minimization is solved using binary search.
  Set $\widehat{\theta}_{\ell+1} \leftarrow \widehat{\theta}_\ell - \frac{\xi}{2} \left(\nabla_\theta L_{\text{DRO}}^{\text{split}}(\widehat{\theta}_\ell, \widehat{\eta}, \mathcal{D}_1 \mid \mathcal{D}_2) + \nabla_\theta L_{\text{DRO}}^{\text{split}}(\widehat{\theta}_\ell, \widehat{\eta}', \mathcal{D}_2 \mid \mathcal{D}_1) \right)$.
end
return $\widehat{\theta} \leftarrow \widehat{\theta}_{\text{max\_iterations}+1}$

Next, we state assumptions on the adjacency function $\mathcal{A}^*$ and the loss function L^* (note that special instances of Cox and DeepHit models satisfy these conditions, as we explain later):

- **A3 (adjacency function).** We assume that the adjacency function is as stated for the DeepHit model (in fact, under ssumption A2, the adjacency function for the Cox model would be equivalent but for simplicity, we use the version stated for the DeepHit model that explicitly has time discretized).

- **A4 (loss function).** We assume that the individual loss function L^* is of the form

$$L^*((x, y, \delta), \mathcal{C}; \theta) = \phi_{\text{indiv}}((x, y, \delta); \theta) + \delta \cdot \phi_{\text{transform}}\left(\sum_{(x', y', \delta') \in \mathcal{C}} \phi_{\text{couple}}((x, y, \delta), (x', y', \delta'); \theta) \right)$$

  for some functions $\phi_{\text{indiv}}(\cdot; \theta) : \mathcal{Z} \to \mathbb{R}$, $\phi_{\text{couple}}(\cdot, \cdot; \theta) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, and $\phi_{\text{transform}} : \mathbb{R} \to \mathbb{R}$. These functions satisfy the following conditions:

  (a) There exist constants $M_{\text{indiv}} \in [0, \infty)$ and $M_{\text{couple-min}}, M_{\text{couple-max}} \in (0, \infty)$ such that

$$\phi_{\text{indiv}}((x, y, \delta); \theta) \in [0, M_{\text{indiv}}] \qquad \text{for all } (x, y, \delta) \in \mathcal{Z} \text{ and } \theta \in \Theta,$$

  and

$$\phi_{\text{couple}}((x, y, \delta), (x', y', \delta'); \theta) \in [M_{\text{couple-min}}, M_{\text{couple-max}}] \quad \text{for all } (x, y, \delta), (x', y', \delta') \in \mathcal{Z} \text{ and } \theta \in \Theta.$$

Importantly, when a coupling term appears, it is nontrivial in the sense that it is not just equal to 0, whereas we allow for the possibility that $M_{\text{indiv}} = 0$.

(b) The function $\phi_{\text{transform}}$ is either (a) the identity function $\phi_{\text{transform}}(s) = s$, or (b) the function $\phi_{\text{transform}}(s) = \log(1+s)$.

(c) For any fixed $y \in \{t_1, t_2, \ldots, t_m\}$, $\delta \in \{0,1\}$, set of data points $\mathcal{C}$ from $\mathcal{Z}$, and parameter setting $\theta \in \Theta$, the map $x \mapsto L^*((x,y,\delta),\mathcal{C};\theta)$ is $\mathcal{L}$-Lipschitz with respect to Euclidean norm, i.e., for all $x, x' \in \mathcal{X}$,

$$|L^*((x,y,\delta),\mathcal{C};\theta) - L^*((x',y,\delta),\mathcal{C};\theta)| \leq \mathcal{L}\|x - x'\|_2.$$

We define

$$R^{\text{split}}_{\text{DRO}}(\theta,\eta) \triangleq C_\alpha \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]} + \eta,$$

where, as a reminder, $C_\alpha = \sqrt{2(\frac{1}{\alpha}-1)^2 + 1}$, and R_{indiv} is defined in a manner that depends on taking the expectation with respect to a fresh sample $\{(X_i', Y_i', \Delta_i')\}_{i=1}^{n_2}$ with n_2 treated as a constant. Put another way, $R^{\text{split}}_{\text{DRO}}(\theta,\eta)$ is simply the population-level version of $L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)$. In particular, $R^{\text{split}}_{\text{DRO}}(\theta,\eta)$ does not depend on the training data.

We are now ready to state our main theoretical result.

Theorem 5 *Fix $n \in \mathbb{N}$ even and randomly split the training data into $\mathcal{D}_1$ and $\mathcal{D}_2$ of sizes $n_1 = n_2 = n/2$. Let $\omega > 0$. Suppose that Assumptions A1–A4 hold. If $\phi_{transform}(s) = s$, then define*

$$M \triangleq M_{indiv} + \frac{M_{couple\text{-}max}}{2} n,$$
$$M' \triangleq (M_{couple\text{-}max} - M_{couple\text{-}min})\sqrt{\frac{\omega n}{2\zeta}}.$$

If instead $\phi_{transform}(s) = \log(1+s)$, then define

$$M \triangleq M_{indiv} + \log\Big(1 + \frac{M_{couple\text{-}max}}{2} n\Big),$$
$$M' \triangleq \frac{4(M_{couple\text{-}max} - M_{couple\text{-}min})}{\zeta M_{couple\text{-}min}}\sqrt{\frac{\omega}{n}}.$$

Then with probability at least

$$1 - 2\left[\frac{M}{(C_\alpha - 1)\big[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}M + (2\mathcal{L}+1)M'\big]} + \mathbb{N}(M', \mathcal{X})\right] e^{-\omega} - m e^{-\frac{n\zeta}{16}} \tag{24}$$

over randomness in the training data, we have

$$\Big|\inf_{\eta\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{split}_{DRO}(\theta,\eta)\Big|$$
$$\leq 10 C_\alpha^2 \left[\sqrt{\frac{2}{n}}\max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\}\big(\sqrt{2\omega}+1\big)M + (2\mathcal{L}+1)M'\right]. \tag{25}$$

We defer the proof of this theorem to Appendix C, where we state a slightly more general result in which n need not be even, and n_1 need not equal n_2 (we present the theorem in the manner above to prevent the notation from getting more unwieldy). To help provide intuition for Theorem 5, we illustrate its use for special cases of Cox and DeepHit models, where we see that as $n \to \infty$, the probability in equation (24) goes to 1 and the right-hand side of bound (25) goes to 0.

Corollary 6 *(Special case of a Cox model) Suppose that the raw input space $\mathcal{X} \subseteq \mathbb{R}^d$ and the parameter vector space $\Theta \in \mathbb{R}^d$ are both set to be the unit ball in $\mathbb{R}^d$, where $d \geq 5$. Consider the standard Cox model where $f(x;\theta) = \theta^\top x$, and time is discrete along the finite grid $t_1 < t_2 < \cdots < t_m$ that satisfies Assumption A2. This setup implies that Assumptions A1, A3, and A4 are also satisfied. Specifically for Assumption A4, we have*

$$\begin{aligned}
\phi_{indiv}((x,y,\delta);\theta) &= \underbrace{0}_{M_{indiv}}, \\
\phi_{couple}((x,y,\delta),(x',y',\delta');\theta) &= e^{\theta^\top(x-x')} \in [\ \underbrace{e^{-2}}_{M_{couple\text{-}min}}\ ,\ \underbrace{e^{2}}_{M_{couple\text{-}max}}\ ], \\
\phi_{transform}(s) &= \log(1+s).
\end{aligned}$$

Furthermore, a valid Lipschitz constant in Assumption A4(c) in this case is $\mathcal{L} = 1$.

Now assume that the number of training data is sufficiently large, namely

$$\begin{aligned}
n \geq \max\Bigg\{ &\Big(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\Big)^2\Big(\frac{d+1}{2}\Big)e^{\sqrt{2\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)-1}}, \\
&2e^{-\frac{6(e^4-1)}{\zeta\max\{2,\frac{C_\alpha}{C_\alpha-1}\}}-2}, \quad e^{\sqrt{2(\log\frac{d+1}{2}-1)}+\log\frac{d+1}{2}}, \quad e^{\frac{2}{d+1}}\Bigg\}.
\end{aligned}$$

Define $\Upsilon \triangleq 2\big[\frac{1}{2\max\left\{2(C_\alpha-1),C_\alpha\right\}} + \big(\frac{3\zeta e^{-2}}{4(e^2-e^{-2})}\big)^d\big]$, which is constant with respect to n. Then with probability at least

$$1 - \frac{\Upsilon}{\sqrt{n}} - me^{-\frac{n\zeta}{16}} \tag{26}$$

over randomness in the training data, we have

$$\begin{aligned}
&\left|\inf_{\eta\in\mathbb{R}} L_{DRO}^{split}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R_{DRO}^{split}(\theta,\eta)\right| \\
&\quad\leq 10C_\alpha^2\Bigg[\sqrt{\frac{2}{n}}\max\Big\{2,\frac{C_\alpha}{C_\alpha-1}\Big\}\big(\sqrt{(d+1)\log n}+1\big)\log\Big(1+\frac{e^2}{2}n\Big) \\
&\qquad\qquad + \frac{12(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{(d+1)\log n}{2n}}\Bigg] \\
&\quad= \widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{n}}\Big),
\end{aligned} \tag{27}$$

where $\widetilde{\mathcal{O}}$ is big O notation ignoring log factors.

We provide the proof in Appendix D. The key idea is that we set $\omega = \frac{d+1}{2}\log n$ for Theorem 5 and we further impose constraints on n and d that enable us to simplify the probability bound in equation (24).

Corollary 7 *(Special case of a DeepHit model) Just as in Corollary 6, we assume $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Theta \in \mathbb{R}^d$ are both set to be the unit ball in $\mathbb{R}^d$,where $d \geq 5$. We consider a DeepHit model defined over a discrete time grid with $m \geq 3$ time points, where the conditional probabilities of each time index satisfy the bound*

$$f_j(x;\theta) = \mathbb{P}(T = t_j \mid X = x) \geq \varrho \qquad \text{for all } x \in \mathcal{X}, j \in [m], \theta \in \Theta.$$

Furthermore, we assume that the very last time step t_m is special in that it is used to mean "any time strictly after all the observed training times" (so that $t_m > Y_j$ for all $j \in [n]$); of course this last time step also obeys the bound above of $f_m(x;\theta) \geq \varrho$. Moreover, we assume that $f_j(\cdot;\theta)$ is 1-Lipschitz:

$$|f_j(x;\theta) - f_j(x';\theta)| \leq \|x - x'\|_2 \qquad \text{for all } x, x' \in \mathcal{X}, j \in [m], \theta \in \Theta.$$

We further suppose that Assumption A2 holds, that the number of training data is sufficiently large

$$n \geq \max\Bigg\{(1-\beta)^2\big(e^{(1-\varrho)/\sigma} - e^{(\varrho-1)/\sigma}\big)^2\Big(\frac{d+1}{4\zeta}\Big)e^{\sqrt{2\log\left((1-\beta)^2(e^{(1-\varrho)/\sigma}-e^{(\varrho-1)/\sigma})^2\left(\frac{d+1}{4\zeta}\right)\right)}}, \\ e^{\sqrt{2(\log\frac{d+1}{2}-1)}+\log\frac{d+1}{2}}, \quad e^{\frac{2}{d+1}}\Bigg\},$$

and that

$$\log\Big((1-\beta)^2(e^{(1-\varrho)/\sigma} - e^{(\varrho-1)/\sigma})^2\Big(\frac{d+1}{4\zeta}\Big)\Big) > 1.$$

Then this setup as stated also implies that Assumptions A1, A3, and A4 are satisfied. Specifically for Assumption A4, we have

$$\begin{aligned}
\phi_{indiv}((x,y,\delta);\theta) &= \beta \cdot \big[-\delta\log(f_{\kappa(y)}(x;\theta)) - (1-\delta)\log(S_{\kappa(y)}(x;\theta))\big], \\
\phi_{couple}((x,y,\delta),(x',y',\delta'),\mathcal{C};\theta) &= (1-\beta)\cdot\frac{1}{n}\cdot\exp\Big(\frac{S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta)}{\sigma}\Big), \\
\phi_{transform}(s) &= s,
\end{aligned}$$

where one can verify that

$$\begin{aligned}
M_{indiv} &= \beta\log\frac{1}{\varrho}, \\
M_{couple\text{-}min} &= \Big(\frac{1-\beta}{n}\Big)e^{\frac{\varrho-1}{\sigma}}, \\
M_{couple\text{-}max} &= \Big(\frac{1-\beta}{n}\Big)e^{\frac{1-\varrho}{\sigma}},
\end{aligned}$$

and that $x \mapsto L^*((x,y,\delta),\mathcal{C};\theta)$ has Lipschitz constant $\mathcal{L} = \frac{2\beta(m-1)}{\varrho}$.

Define the constant

$$\Psi \triangleq \frac{2}{(C_\alpha-1)\Big[2\max\{2,\frac{C_\alpha}{C_\alpha-1}\}+(2\mathcal{L}+1)\Big(\frac{\left((1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}\right)}{\left(2\beta\log\frac{1}{\varrho}+(1-\beta)e^{(1-\varrho)/\sigma}\right)}\sqrt{\frac{2}{\zeta}}\Big)\Big]} + 2\bigg(\frac{3\sqrt{2\zeta}}{(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}}\bigg)^d.$$

Then with probability at least

$$1-\frac{\Psi}{\sqrt{n}}-me^{-\frac{n\zeta}{16}},$$

we have

$$\begin{aligned}&\left|\inf_{\eta\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta,\mathcal{D}_1\mid\mathcal{D}_2)-\inf_{\eta\in\mathbb{R}} R^{split}_{DRO}(\theta,\eta)\right|\\ &\le 10C_\alpha^2\bigg[\sqrt{\frac{2}{n}}\max\Big\{2,\frac{C_\alpha}{C_\alpha-1}\Big\}\big(\sqrt{(d+1)\log n}+1\big)\Big(\beta\log\frac{1}{\varrho}+\frac{(1-\beta)e^{(1-\varrho)/\sigma}}{2}\Big)\\ &\qquad+\Big(\frac{4\beta(m-1)}{\varrho}+1\Big)\sqrt{\frac{(d+1)\log n}{\zeta n}}\frac{(1-\beta)(e^{(1-\varrho)/\sigma}-e^{(\varrho-1)/\sigma})}{2}\bigg]\\ &=\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

The proof is in Appendix E and uses similar ideas as that of Corollary 6. The constants change since M and M' are different, but we again set $\omega=\frac{d+1}{2}\log n$ in the statement of Theorem 5.

Corollary 8 *(Cross-fitting) We assume the same setting as Theorem 5. For a fixed $\theta\in\Theta$, the cross-fitting approach solves*

$$\inf_{\eta,\eta'\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta,\eta')=\frac{1}{2}\inf_{\eta\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta,\mathcal{D}_1\mid\mathcal{D}_2)+\frac{1}{2}\inf_{\eta'\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta',\mathcal{D}_2\mid\mathcal{D}_1).$$

With probability at least

$$1-4\bigg[\frac{M}{(C_\alpha-1)\big[2\sqrt{\frac{\omega}{n}}\max\{2,\frac{C_\alpha}{C_\alpha-1}\}M+(2\mathcal{L}+1)M'\big]}+\mathbb{N}(M',\mathcal{X})\bigg]e^{-\omega}-2me^{-\frac{n\zeta}{16}},$$

we have

$$\begin{aligned}&\left|\inf_{\eta,\eta'\in\mathbb{R}} L^{split}_{DRO}(\theta,\eta,\eta')-\inf_{\eta\in\mathbb{R}} R^{split}_{DRO}(\theta,\eta)\right|\\ &\le 10C_\alpha^2\bigg[\sqrt{\frac{2}{n}}\max\Big\{2,\frac{C_\alpha}{C_\alpha-1}\Big\}\big(\sqrt{2\omega}+1\big)M+(2\mathcal{L}+1)M'\bigg].\end{aligned}$$

Proof The proof is straightforward and amounts to applying Theorem 5 for each of the two folds separately and then union bounding over the bad events of the two folds. This

union bound just multiplies the bad event's probability in bound (24) by 2. Then since this bad event does not happen for either fold, we have

$$\max\left\{\left|\inf_{\eta\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right|, \left|\inf_{\eta\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_2 \mid \mathcal{D}_1) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right|\right\}$$
$$\leq 10C_\alpha^2\left[\sqrt{\frac{2}{n}}\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\}\left(\sqrt{2\omega}+1\right)M + (2\mathcal{L}+1)M'\right].$$

Then

$$\begin{aligned}
&\left|\inf_{\eta,\eta'\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\eta') - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right| \\
&\quad= \left|\frac{1}{2}\inf_{\eta\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) + \frac{1}{2}\inf_{\eta'\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta',\mathcal{D}_1 \mid \mathcal{D}_2)\right. \\
&\qquad\left. - \frac{1}{2}\inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta) - \frac{1}{2}\inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right| \\
&\quad= \frac{1}{2}\left|\inf_{\eta\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right. \\
&\qquad\left. + \inf_{\eta'\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta',\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right| \\
&\quad\leq \frac{1}{2}\left|\inf_{\eta\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right| \\
&\qquad\frac{1}{2}\left|\inf_{\eta'\in\mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta,\eta',\mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta,\eta)\right| \\
&\quad\leq 10C_\alpha^2\left[\sqrt{\frac{2}{n}}\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\}\left(\sqrt{2\omega}+1\right)M + (2\mathcal{L}+1)M'\right].
\end{aligned}$$

■

4. An Exact DRO Cox Approach

We now derive an exact DRO approach for Cox models. The rough idea is that we reparameterize the Cox model in such a way that the resulting loss function decouples across training data points, removing the coupling issue. Our derivation here is specific to the Cox model and, as far as we are aware, does not easily generalize to other survival models with nonempty adjacency sets.

To obtain an exact approach for using the Cox model with DRO that does not require sample splitting, we turn to a standard derivation of the Cox partial likelihood loss. Specifically, Breslow (1972) showed that the Cox log partial likelihood could be derived by assuming that the baseline hazard function h_0 is piecewise constant. First, denote the unique times in which the critical event happened in the training data as $t_1 < t_2 < \cdots < t_m$ (so that there are m unique times in which the event happened), with the convention that $t_0 \triangleq 0$

(note that we are reusing notation used for the discrete time grid for DeepHit; however, the difference is that for the Cox model, the time grid is typically set based on the unique critical event times whereas for DeepHit, the time grid is user-specified and need not be the unique critical event times). Then we parameterize the baseline hazard function as

$$h_0(t;\psi) \triangleq \begin{cases} e^{\psi_\ell} & \text{if } t_{\ell-1} < t \leq t_\ell \quad \text{for } \ell \in [m], \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where $\psi_1, \psi_2, \ldots, \psi_m \in \mathbb{R}$ are parameters to be learned, and $\psi \triangleq (\psi_1, \ldots, \psi_m)$.[2]

Next, let $\kappa(Y_i, \Delta_i) \in [m]$ denote the discrete time index that Y_i corresponds to in a manner that depends also on Δ_i: if $\Delta_i = 1$, then $\kappa(Y_i, \Delta_i)$ is set equal to the index ℓ such that $t_\ell = Y_i$, and if $\Delta_i = 0$ (so that Y_i is a censoring time), then we set $\kappa(Y_i, \Delta_i)$ to be the largest time index corresponding to when a critical event happened strictly before Y_i (i.e., we use the largest index in $\{\ell \in \{0, 1, \ldots, m\} : t_\ell < Y_i\}$). Then the full negative Cox log likelihood can be written as

$$L^{\text{Cox-full}}(\theta, \psi) \triangleq \frac{1}{n} \sum_{i=1}^{n} L_i^{\text{Cox-full}}(\theta, \psi), \tag{29}$$

where

$$L_i^{\text{Cox-full}}(\theta, \psi) \triangleq -\Delta_i [f(X_i;\theta) + \psi_{\kappa(Y_i,\Delta_i)}] + e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell}. \tag{30}$$

Then a standard result is as follows.

Proposition 9 (slight variant of Breslow 1972) *Suppose that the baseline hazard function is piecewise constant as stated in equation* (28)*. Suppose that we preprocess the data so that for each training point $i \in [n]$ that is censored (i.e., $\Delta_i = 0$), we set $Y_i \triangleq t_{\kappa(Y_i,0)}$ (we do not modify the observed times for the uncensored training points). Then the partial Cox loss L^{Cox} (from equation* (4)*) is related to the full Cox loss $L^{Cox\text{-}full}$ (from equation* (29)*) by*

$$L^{Cox}(\theta) = \min_{\psi \in \mathbb{R}^m} L^{Cox\text{-}full}(\theta, \psi) + \textit{constant w.r.t. } \theta.$$

Hence, $\arg\min_{\theta\in\Theta} L^{Cox}(\theta) = \arg\min_{\theta\in\Theta}\{\min_{\psi\in\mathbb{R}^m} L^{Cox\text{-}full}(\theta, \psi)\}$*, where Θ is the feasible set of model parameters.*

While the proof is standard (Breslow, 1972), to keep the paper relatively self-contained, we provide it in Appendix A.3, where we also provide a little bit of background on how the expression for individual loss $L_i^{\text{Cox-full}}(\theta, \psi)$ (from equation (30)) is derived. We separately point out that, as far as we are aware, the full Cox loss in equation (29) is typically not used in practice and is instead mainly used for theoretically justifying the standard Cox loss (equation (4)) that actually is extremely commonly used in practice.

2. Note that in equation (28), the exponential function can be replaced with any differentiable, strictly increasing, positive activation function (e.g., instead of e^{ψ_ℓ}, we could use the softplus function $\log(1 + \exp(\psi_\ell))$). For ease of exposition, we stick to using the exponential function.

An immediate consequence of Proposition 9 is that we could apply DRO to the loss $L^{\text{Cox-full}}(\theta, \psi)$ (using the individual losses given by $L_i^{\text{Cox-full}}(\theta, \psi)$ in equation (30)), which does not involve coupling across training points. The high-level idea is that whereas $L^{\text{Cox}}(\theta)$ had the coupling issue, by introducing an additional parameter variable ψ, we remove the dependence between the training points' contributions.

Numerical optimization For completeness, we now state how to use DRO with the full Cox loss. We first define

$$L_{\text{DRO-Cox-exact}}(\theta, \psi, \eta) \triangleq C_\alpha \sqrt{\frac{1}{n} \sum_{i=1}^{n} [L_i^{\text{Cox-full}}(\theta, \psi) - \eta]_+^2} + \eta. \tag{31}$$

Then we alternate between the following two steps until convergence:

- Treating θ and ψ as fixed, we update η by finding the value of η that minimizes $L_{\text{DRO-Cox-exact}}(\theta, \psi, \eta)$. As before, this step is done using binary search to find the global minimum since $L_{\text{DRO-Cox-exact}(\theta,\psi,\eta)}$ is convex with respect to η.
- Treating η as fixed, we update (θ, ψ) by minimizing $L_{\text{DRO-Cox-exact}}(\theta, \psi, \eta)$ (e.g., using gradient descent).

This procedure corresponds to using Algorithm 1, where L_{indiv} is set to be $L_i^{\text{Cox-full}}(\theta, \psi)$ (equation (30)), and the survival model parameter variable θ is replaced by (θ, ψ). Note that we intentionally specified the parameter variable ψ so that it remains unconstrained so that it could be optimized along with θ using standard gradient descent variants.

5. Experiments

To see how well our general proposed DRO conversion strategy works in practice (the heuristic approach without guarantees and, separately, our sample splitting DRO approach), we now conduct extensive experiments to evaluate the accuracy and fairness of DRO variants of different survival models compared to the original versions of these models, as well as to versions of these models modified to encourage fairness using existing fairness regularizers. Specifically for the Cox model, we also show how well our exact Cox DRO approach works in practice.

We describe the datasets we use in Section 5.1, the experimental setup in Section 5.2, the evaluation metrics in Section 5.3, and the models evaluated in Section 5.4. We then present our experimental results in Section 5.5. Lastly, we show how to compare across multiple models using a plot inspired by ROC curves in Section 5.6.

5.1 Datasets

We use three standard, publicly available survival analysis datasets:

- The **FLC** dataset (Dispenzieri et al., 2012) is from a study on the relationship between serum free light chain (FLC) and mortality of Olmsted County residents aged 50 or higher. We treat discretized age (age$\leq$65 and age$>$65) and gender (women and men) as sensitive attributes.
- The **SUPPORT** dataset (Knaus et al., 1995) is from a study at Vanderbilt University on understanding prognoses, preferences, outcomes, and risks of treatment by

Table 1: Basic dataset characteristics.

	FLC	SUPPORT	SEER
# samples	7,874	9,105	28,018
# features	6 (9*)	14 (19*)	11
Censoring rate	0.725	0.319	0.654
Sensitive attributes	age, gender	age, race, gender	age, race

* indicates the number before preprocessing (preprocessing removes some features)

analyzing survival times of severely ill hospitalized patients. We treat discretized age (age$\leq$65 and age>65), race (white and non-white), and gender (women and men) as sensitive attributes.

- The **SEER** dataset is on breast cancer patients from the Sureillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. We collected this dataset using the data extraction software from the official SEER program of the National Cancer Institute. We used 11 covariates that also appear in an existing snapshot of the SEER dataset (Teng, 2019) that only contained 4024 data points. We also treat discretized age (age$\leq$65 and age>65) and race (white and non-white) as sensitive attributes.

These datasets have appeared in existing fair survival analysis research (e.g., Keya et al. 2021; Rahman and Purushotham 2022; Zhang and Weiss 2022) although not always with all three of these appearing within the same paper. Basic characteristics of these datasets are reported in Table 1.

5.2 Experimental Setup

For all models, we first use a random 80%/20% train/test split to hold out a test set that will be the same across experimental repeats for all datasets. Then we repeat the following basic experiment 10 times: (1) We hold out 20% of the training data to treat as a validation set, which is used to tune hyperparameters. (2) We then compute evaluation metrics across the same test set. We describe the evaluation metrics and how hyperparameter tuning works shortly. When we report our experimental results, we provide the mean and standard deviation of each metric across the 10 experimental repeats. More hyperparameter settings can be found in Appendix G.

5.3 Evaluation Metrics

For accuracy metrics, we use Time-dependent concordance index (C^{td}, higher is better) (Antolini et al., 2005) and Integrated IPCW Brier Score (IBS, lower is better) (Graf et al., 1999). For fairness metrics, we use the concordance imparity (CI) fairness metric by (Zhang and Weiss, 2022), Censoring-based individual fairness (F_{CI}) (Rahman and Purushotham, 2022), and Censoring-based group fairness (F_{CG}) (Rahman and Purushotham, 2022). For these fairness metrics, lower is better. Definitions of these fairness metrics are in Appendix F.

Note that the fairness metrics CI and F_{CG} require us to specify groups. For the FLC dataset, we use (discretized) age and, separately, gender (i.e., we first run experiments using

only age in evaluating CI and F_{CG}; we then re-run experiments using gender instead of age). For the SUPPORT dataset, we separately use gender, age, and race. For the SEER dataset, we separately use race and age.

5.4 Models Evaluated

Working off our running examples from Section 2.2, we consider Cox models (classical and deep), DeepHit, and SODEN. For each of these, we compare the original model with its DRO variants using our conversion strategy (the heuristic approach and also the sample splitting approach stated in Section 3.2; for Cox models, we also compare with the exact DRO Cox approach, and for SODEN, there is no need to do sample splitting and the heuristic approach is actually exact). We also try versions of the original models modified to encourage fairness using existing fairness regularizers.

Note that when we use our sample splitting DRO approach, for simplicity, we randomly split the training data so that $\mathcal{D}_1$ and $\mathcal{D}_2$ are (approximately) the same size and, moreover, we stratify the sampling so that $\mathcal{D}_1$ and $\mathcal{D}_2$ have (approximately) the same censoring rates. We include some supplemental experiments that consider deviations of both of these in Appendix H that focus specifically on Cox models.

Cox models We separately experiment on the classical **linear** setting (the log partial hazard function is $f(x;\theta) = \theta^T x$) or the "deep" **nonlinear** setting in which f is a multilayer perceptron (MLP). In the linear case, we denote the heuristic DRO variant as DRO-COX and the sample splitting DRO variant as DRO-COX (SPLIT). For the nonlinear case, we add the prefix "Deep" to these names for clarity.

In terms of baselines, we use the unregularized linear Cox model (Cox, 1972) (denoted as "Cox" in tables later), whereas the unregularized nonlinear Cox model (Katzman et al., 2018) is denoted as "DeepSurv". As baselines, we use regularized versions of either the standard Cox or DeepSurv models, using different fairness regularization terms. When we use individual, group, or intersectional regularization terms by Keya et al. (2021) (we discuss these in Appendix F), we add the suffix "$_I$(Keya et al.)", "$_G$(Keya et al.)", or "$_\cap$(Keya et al.)" respectively to a model name; for example, "DeepSurv$_G$(Keya et al.)" corresponds to DeepSurv with group fairness regularization by Keya et al. (2021). When we use the individual or group fairness regularization terms that account for censoring information (Rahman and Purushotham, 2022), we instead use the suffix "$_I$(R&P)" or "$_G$(R&P)".[3] Note that group fairness regularization (suffixes "$_G$(Keya et al.)" and "$_G$(R&P)") uses the same groups that test set CI and F_{CG} fairness metrics use. As additional baselines, we also use the pseudo value-based approaches proposed by Rahman and Purushotham (2022), namely their Fair DeepPseudo and Fair PseudoNAM methods (abbreviated as "FIDP" and "FIPNAM" respectively; note that these abbreviations are the same as the ones used by Rahman and Purushotham (2022) and, moreover, following Rahman and Purushotham's paper and implementation, FIDP and FIPNAM specifically use individual fairness regularization).

3. Rahman and Purushotham (2022) did not propose an intersectional fairness regularizer and technically did not try regularized versions of Cox models using their fairness definitions. However, it is straightforward to adapt their individual and group fairness definitions as regularization terms for a Cox model, especially as their work is directly modifying definitions by Keya et al. (2021).

In terms of hyperparameter tuning, we use the strategy by Keya et al. (2021): the final hyperparameter setting used per dataset and per method is determined based on a preset rule in practice that allows up to a 5% degradation in the validation set C^{td} from the classical Cox model (for the linear setting) or DeepSurv (for the nonlinear setting) while minimizing the validation set CI fairness metric or F_{CG} fairness metric (for details, see Appendix H).

DeepHit and SODEN For DeepHit (Tang et al., 2022b), we denote its heuristic DRO variant as DRO-DEEPHIT and its sample splitting DRO variant as DRO-DEEPHIT (SPLIT). For SODEN (Tang et al., 2022b), there is only one DRO variant to consider which we denote as DRO-SODEN.

In terms of baselines, we consider the original DeepHit and SODEN models that do not account for fairness. We further adapt the group-based fairness regularization that accounts for censoring from Rahman and Purushotham (2022) to each of DeepHit and SODEN separately as additional baselines (DEEPHIT$_G$(R&P) and SODEN$_G$(R&P)).

The hyperparameter setting used per dataset and per method is also determined based on a preset rule in practice that allows up to a 5% degradation in the validation set C^{td} from the original model (that does not encourage fairness) while minimizing the validation set CI fairness metric or F_{CG} fairness metric. Hyperparameter grids for all methods are in Appendix G, where we also provide information on the compute environment that we used.

5.5 Experimental Results

Cox models We compare DRO-COX and DRO-COX (SPLIT) against various baselines using a similar experimental setup as Keya et al. (2021). Specifically, we report the test set evaluation metrics for FLC (using age to evaluate CI and F_{CG}) in Table 2, SUPPORT (gender) in Table 3, and SEER (race) in Table 4. Experimental results using other sensitive attributes for the datasets have similar trends and are in Appendix H. From these tables, we have the following observations:

- Among linear methods, the heuristic DRO-COX method consistently outperforms baselines in terms of the CI fairness metric (and often on the other fairness metrics too) while still achieving reasonably high accuracy scores. A similar trend holds among nonlinear methods for the heuristic deep DRO-COX variant.
- The performance difference (in terms of both accuracy and fairness) between the heuristic DRO-COX and sample-splitting-based DRO-COX (SPLIT) is not clear cut; sometimes one performs better than the other and vice versa. This holds for their linear variants as well as, separately, their nonlinear (deep) variants.
- As expected, the unregularized Cox and DeepSurv models often have (among) the highest accuracy scores but tend to have poor performance on fairness metrics.
- The baselines that are regularized variants of Cox and DeepSurv typically do not simultaneously achieve low scores across all fairness metrics. Even though some of these can work well with some of the metrics by Keya et al. (2021), they clearly do not work as well as our DRO-COX variants when it comes to the CI fairness metric that actually accounts for accuracy.

Effect of α. To show how α trades off between fairness and accuracy, we show results for DRO-COX in the linear setting across all datasets (using age for evaluating F_G and CI) in Figure 1, where we use c-index as the accuracy metric. It is clear that accuracy tends

Table 2: Cox model test set accuracy and fairness metrics on the FLC (age) dataset. We report mean and standard deviation (in parentheses) across 10 experimental repeats (each repeat holds out a different 20% of the training data as a validation set for hyperparameter tuning; the test set is the same across experimental repeats). Higher is better for metrics with "↑", while lower is better for metrics with "↓". The best results are shown in bold for linear and, separately, nonlinear models. When one of our methods outperforms all baselines (in linear and, separately, nonlinear models), we highlight the corresponding cell in green. Evaluation metrics are reported to 4 decimal places unless the number is exactly equal to 0 (in which case we just state 0 without using a decimal point) or smaller than 10^{-4} (in which case we report the number in scientific notation). Note that achieving F_{CI} or F_{CG} scores that are exactly 0 is due to the manner in which these fairness metrics are defined.[5]

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$\mathrm{C}^{td}\uparrow$	IBS↓	CI(%)↓	$\mathrm{F}_{CI}\downarrow$	$\mathrm{F}_{CG}\downarrow$	$\mathrm{C}^{td}\uparrow$	IBS↓	CI(%)↓	$\mathrm{F}_{CI}\downarrow$	$\mathrm{F}_{CG}\downarrow$
Linear	Cox	**0.8032** (**0.0002**)	0.1739 (0.0004)	0.5350 (0.0413)	0.0249 (0.0002)	0.0044 (2.8919e-05)	**0.8032** (**0.0002**)	0.1739 (0.0004)	0.5350 (0.0413)	0.0249 (0.0002)	0.0044 (2.8919e-05)
Linear	Cox_I(Keya et al.)	0.7937 (0.0068)	0.1414 (0.0073)	0.5400 (0.3270)	0.0129 (0.0028)	0.0021 (0.0005)	0.7923 (0.0074)	0.1334 (0.0034)	0.4010 (0.2631)	0.0068 (0.0006)	0.0010 (0.0001)
Linear	Cox_I(R&P)	0.8029 (0.0005)	0.1735 (0.0023)	0.4660 (0.1551)	0.0247 (0.0009)	0.0043 (0.0002)	0.8020 (0.0007)	0.1700 (0.0034)	0.2530 (0.2658)	0.0233 (0.0014)	0.0040 (0.0003)
Linear	Cox_G(Keya et al.)	0.7974 (0.0117)	0.1492 (0.0077)	0.3410 (0.3011)	0.0123 (0.0043)	0.0024 (0.0007)	0.7862 (0.0133)	0.1413 (0.0035)	0.5360 (0.3888)	0.0079 (0.0029)	0.0016 (0.0004)
Linear	Cox_G(R&P)	0.8029 (0.0005)	0.1735 (0.0023)	0.4660 (0.1551)	0.0247 (0.0009)	0.0043 (0.0002)	0.8015 (0.0003)	0.1673 (0.0004)	**0.0390** (**0.0243**)	0.0222 (0.0002)	0.0038 (3.3934e-05)
Linear	$\mathrm{Cox}_\cap$(Keya et al.)	0.7870 (0.0029)	0.1400 (0.0005)	1.0790 (0.1098)	0.0073 (0.0002)	0.0016 (0.0001)	0.7875 (0.0021)	0.1402 (0.0004)	1.1190 (0.1073)	0.0073 (0.0002)	0.0016 (0.0001)
Linear	DRO-COX	0.7959 (0.0036)	0.1408 (0.0050)	**0.0510** (**0.0401**)	0.0078 (0.0051)	0.0012 (0.0008)	0.7958 (0.0049)	**0.1330** (**0.0002**)	0.1620 (0.1132)	**0** (**0**)	**0** (**0**)
Linear	DRO-COX (SPLIT)	0.7964 (0.0045))	**0.1389** (**0.0008**)	0.2350 (0.1277)	**0** (**0**)	**0** (**0**)	0.7964 (0.0045))	**0.1389** (**0.0008**)	0.2350 (0.1277)	**0** (**0**)	**0** (**0**)
Linear	EXACT DRO-COX	0.7821 (0.0142)	0.3916 (0.0487)	0.9838 (0.4567)	0.0094 (0.0016)	0.0019 (0.0003)	0.7821 (0.0142)	0.3916 (0.0487)	0.9838 (0.4567)	0.0094 (0.0016)	0.0019 (0.0003)
Nonlinear	DeepSurv	0.8070 (0.0014)	0.1767 (0.0018)	0.2940 (0.2147)	0.0259 (0.0004)	0.0050 (0.0003)	0.8070 (0.0014)	0.1767 (0.0018)	0.2940 (0.2147)	0.0259 (0.0004)	0.0050 (0.0003)
Nonlinear	$\mathrm{DeepSurv}_I$(Keya et al.)	0.7884 (0.0070)	0.1441 (0.0130)	0.3700 (0.2523)	0.0127 (0.0080)	0.0025 (0.0017)	0.7994 (0.0069)	0.1672 (0.0051)	0.6310 (0.5316)	0.0245 (0.0014)	0.0050 (0.0005)
Nonlinear	$\mathrm{DeepSurv}_I$(R&P)	0.8070 (0.0033)	0.1736 (0.0086)	0.2300 (0.1471)	0.0246 (0.0040)	0.0047 (0.0008)	0.8086 (0.0015)	0.1766 (0.0024)	0.1560 (0.0956)	0.0258 (0.0011)	0.0050 (0.0002)
Nonlinear	$\mathrm{DeepSurv}_G$(Keya et al.)	0.7990 (0.0120)	0.4190 (0.2487)	0.2490 (0.1646)	0.0071 (0.0069)	0.0015 (0.0013)	0.8061 (0.0020)	0.4713 (0.2142)	0.2700 (0.2260)	0.0070 (0.0081)	0.0014 (0.0016)
Nonlinear	$\mathrm{DeepSurv}_G$(R&P)	0.8069 (0.0033)	0.1735 (0.0086)	0.2580 (0.1661)	0.0245 (0.0040)	0.0047 (0.0008)	**0.8086** (**0.0015**)	0.1766 (0.0024)	**0.1560** (**0.0956**)	0.0258 (0.0011)	0.0050 (0.0002)
Nonlinear	$\mathrm{DeepSurv}_\cap$(Keya et al.)	0.7751 (0.0018)	0.1357 (0.0002)	0.4300 (0.1091)	0.0037 (0.0001)	0.0008 (1.3494e-05)	0.7751 (0.0018)	0.1357 (0.0002)	0.4300 (0.1091)	0.0037 (0.0001)	0.0008 (1.3494e-05)
Nonlinear	FIDP	**0.8077** (**0.0022**)	**0.1228** (**0.0019**)	0.2530 (0.0974)	0.0239 (0.0018)	0.0048 (0.0004)	0.8077 (0.0022)	**0.1228** (**0.0019**)	0.2530 (0.0974)	0.0239 (0.0018)	0.0048 (0.0004)
Nonlinear	FIPNAM	0.7829 (0.0037)	0.1810 (0.0050)	0.3660 (0.0508)	0.0251 (0.0006)	0.0052 (0.0004)	0.7829 (0.0037)	0.1810 (0.0050)	0.3660 (0.0508)	0.0251 (0.0006)	0.0052 (0.0004)
Nonlinear	Deep DRO-COX	0.8068 (0.0024)	0.1595 (0.0135)	**0.0730** (**0.0822**)	0.0189 (0.0056)	0.0036 (0.0013)	0.7781 (0.0091)	0.1331 (0.0002)	2.4300 (0.3462)	0.0001 (3.1257e-05)	9.9660e-06 (3.5999e-06)
Nonlinear	Deep DRO-COX (SPLIT)	0.7784 (0.0092)	0.1647 (0.0037)	2.3210 (0.3590)	**0** (**0**)	**0** (**0**)	0.7784 (0.0092)	0.1647 (0.0037)	2.3210 (0.3590)	**0** (**0**)	**0** (**0**)
Nonlinear	Deep EXACT DRO-COX	0.8048 (0.0011)	0.1363 (0.0016)	0.5050 (0.2489)	0.0197 (0.0005)	0.0038 (0.0001)	0.8048 (0.0011)	0.1363 (0.0016)	0.5050 (0.2489)	0.0197 (0.0005)	0.0038 (0.0001)

to increase when α increases from 0.1 to 0.3 on FLC and SEER, and from 0.3 to 0.5 on SUPPORT. However, the increase in α results in worse scores across fairness metrics.

Additional experiments. Across all methods, instead of minimizing the validation set CI fairness metric during hyperparameter tuning (tolerating a small degradation in the validation set C^{td}), we also tried instead minimizing the validation set F_{CG} metric and found similar results (see the rightmost columns under the heading "F_{CG}-based Tuning" in Tables 2, 3, and 4).

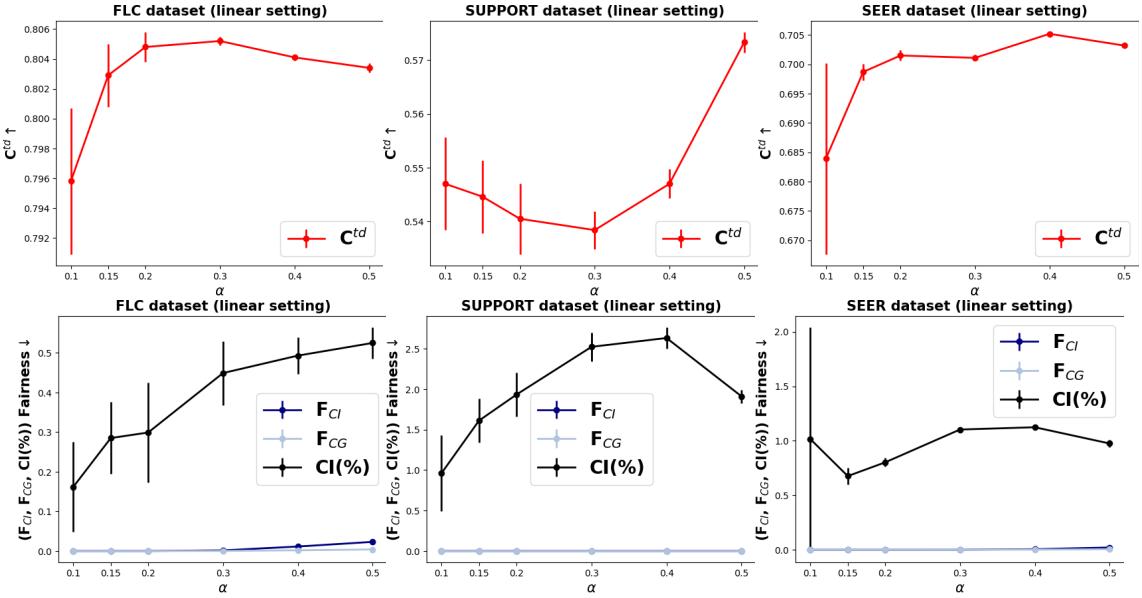
5. Censoring-based individual and group fairness metrics (F_{CI} and F_{CG} respectively) by Rahman and Purushotham (2022)—which we formally define in Appendix F—depend on a user-specified scale constant $\gamma > 0$ that must be specified in advance (of running experiments). A higher value of γ makes it easier

Table 3: Cox model test set scores on the SUPPORT (gender) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	0.6025 (0.0005)	0.2304 (0.0015)	1.4300 (0.0654)	0.0054 (0.0002)	0.0028 (0.0001)	**0.6025 (0.0005)**	0.2304 (0.0015)	1.4300 (0.0654)	0.0054 (0.0002)	0.0028 (0.0001)
	Cox_I(Keya et al.)	0.5881 (0.0114)	**0.2157 (0.0060)**	0.9650 (0.6126)	0.0004 (0.0004)	0.0002 (0.0002)	0.5829 (0.0099)	**0.2147 (0.0063)**	1.1330 (0.6846)	**0 (0)**	**0 (0)**
	Cox_I(R&P)	0.6018 (0.0016)	0.2309 (0.0011)	1.4390 (0.1077)	0.0056 (0.0002)	0.0029 (0.0001)	0.6022 (0.0005)	0.2307 (0.0012)	1.4060 (0.0932)	0.0055 (0.0002)	0.0028 (0.0001)
	Cox_G(Keya et al.)	**0.6030 (0.0007)**	0.2297 (0.0018)	1.4190 (0.0632)	0.0051 (0.0003)	0.0026 (0.0001)	0.6024 (0.0006)	0.2284 (0.0009)	1.4360 (0.0674)	0.0047 (0.0001)	0.0025 (4.2335e-05)
	Cox_G(R&P)	0.6018 (0.0016)	0.2309 (0.0011)	1.4390 (0.1077)	0.0056 (0.0002)	0.0029 (0.0001)	0.6024 (0.0007)	0.2307 (0.0012)	1.4010 (0.0931)	0.0055 (0.0002)	0.0028 (0.0001)
	$Cox_\cap$(Keya et al.)	0.5715 (0.0062)	0.2275 (0.0016)	1.1270 (0.2457)	0.0028 (0.0003)	0.0015 (0.0002)	0.5631 (0.0070)	0.2264 (0.0017)	0.8650 (0.2958)	0.0024 (0.0003)	0.0012 (0.0002)
	DRO-COX	0.5734 (0.0019)	0.2210 (0.0010)	0.4350 (0.0674)	0.0002 (2.3882e-05)	0.0001 (1.3621e-05)	0.5641 (0.0105)	0.2211 (0.0010)	**0.3840 (0.1830)**	0.0001 (0.0001)	0.0001 (4.8271e-05)
	DRO-COX (SPLIT)	0.5701 (0.0056)	0.4569 (0.1314)	**0.3860 (0.1163)**	**1.1922e-07 (2.6445e-07)**	**9.6779e-08 (2.1315e-07)**	0.5701 (0.0056)	0.4570 (0.1314)	0.3860 (0.1163)	1.1922e-07 (2.6445e-07)	9.6779e-08 (2.1315e-07)
	EXACT DRO-COX	0.5884 (0.0063)	0.3122 (0.0068)	0.8580 (0.2434)	8.1822e-06 (8.1542e-06)	5.2437e-06 (5.0535e-06)	0.5884 (0.0063)	0.3122 (0.0068)	0.8580 (0.2434)	8.1822e-06 (8.1542e-06)	5.2437e-06 (5.0535e-06)
Nonlinear	DeepSurv	0.6108 (0.0029)	0.2417 (0.0016)	1.6220 (0.3303)	0.0090 (0.0002)	0.0046 (0.0001)	0.6108 (0.0029)	0.2417 (0.0016)	1.6220 (0.3303)	0.0090 (0.0002)	0.0046 (0.0001)
	$DeepSurv_I$(Keya et al.)	0.5984 (0.0124)	0.2376 (0.0182)	1.3280 (0.7670)	0.0061 (0.0036)	0.0031 (0.0019)	0.6031 (0.0059)	0.2459 (0.0102)	**1.1590 (0.8626)**	0.0090 (0.0007)	0.0046 (0.0004)
	$DeepSurv_I$(R&P)	0.6100 (0.0070)	0.2383 (0.0075)	1.6100 (0.3374)	0.0080 (0.0023)	0.0041 (0.0012)	**0.6115 (0.0051)**	0.2444 (0.0036)	1.5410 (0.4066)	0.0097 (0.0009)	0.0050 (0.0004)
	$DeepSurv_G$(Keya et al.)	0.5982 (0.0109)	0.2436 (0.0121)	1.6540 (0.3892)	0.0090 (0.0036)	0.0046 (0.0019)	0.6034 (0.0037)	0.2499 (0.0024)	1.2390 (0.4314)	0.0111 (0.0003)	0.0057 (0.0001)
	$DeepSurv_G$(R&P)	**0.6105 (0.0055)**	0.2408 (0.0067)	1.5410 (0.3661)	0.0087 (0.0019)	0.0045 (0.0010)	0.6115 (0.0051)	0.2444 (0.0036)	1.5410 (0.4066)	0.0097 (0.0009)	0.0050 (0.0004)
	$DeepSurv_\cap$(Keya et al.)	0.6015 (0.0069)	0.2378 (0.0053)	1.4110 (0.2129)	0.0066 (0.0017)	0.0034 (0.0009)	0.5912 (0.0012)	0.2309 (0.0011)	1.5390 (0.1303)	0.0043 (0.0002)	0.0023 (0.0001)
	FIDP	0.5811 (0.0090)	0.2356 (0.0023)	1.2670 (0.4179)	0.0059 (0.0005)	0.0029 (0.0003)	0.5811 (0.0090)	0.2356 (0.0023)	1.2670 (0.4179)	0.0059 (0.0005)	0.0029 (0.0003)
	FIPNAM	0.5760 (0.0039)	0.2330 (0.0005)	1.0360 (0.0448)	0.0021 (0.0001)	0.0009 (0.0001)	0.5760 (0.0039)	0.2330 (0.0005)	1.0360 (0.0448)	0.0021 (0.0001)	0.0009 (0.0001)
	Deep DRO-COX	0.5829 (0.0067)	**0.2240 (0.0010)**	**1.2600 (0.4412)**	0.0019 (0.0006)	0.0010 (0.0003)	0.5754 (0.0120)	**0.2227 (0.0011)**	1.5550 (0.4622)	0.0010 (0.0005)	0.0005 (0.0003)
	Deep DRO-COX (SPLIT)	0.5772 (0.0093)	0.6387 (0.0007)	1.5530 (0.4682)	**0 (0)**	**0 (0)**	0.5772 (0.0093)	0.6387 (0.0007)	1.5530 (0.4682)	**0 (0)**	**0 (0)**
	Deep EXACT DRO-COX	0.5811 (0.0065)	0.2621 (0.0098)	2.0490 (0.4989)	0.0062 (0.0020)	0.0033 (0.0010)	0.5811 (0.0065)	0.2621 (0.0098)	2.0490 (0.4989)	0.0062 (0.0020)	0.0033 (0.0010)

Table 4: Cox model test set scores on the SEER (race) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	0.7025 (0.0003)	0.2128 (0.0009)	2.5200 (0.0431)	0.0256 (0.0006)	0.0204 (0.0005)	**0.7025 (0.0003)**	0.2128 (0.0009)	2.5200 (0.0431)	0.0256 (0.0006)	0.0204 (0.0005)
Linear	Cox_I(Keya et al.)	0.6894 (0.0046)	**0.1837 (0.0027)**	1.9750 (0.6480)	0.0005 (0.0001)	0.0004 (0.0001)	0.6894 (0.0046)	**0.1837 (0.0027)**	1.9750 (0.6480)	0.0005 (0.0001)	0.0004 (0.0001)
Linear	Cox_I(R&P)	0.7032 (0.0025)	0.2103 (0.0031)	2.4590 (0.0886)	0.0235 (0.0022)	0.0186 (0.0017)	0.7035 (0.0025)	0.2097 (0.0031)	2.4520 (0.0862)	0.0231 (0.0021)	0.0183 (0.0016)
Linear	Cox_G(Keya et al.)	0.6952 (0.0146)	0.2073 (0.0049)	2.8690 (0.5267)	0.0216 (0.0041)	0.0175 (0.0035)	0.6952 (0.0146)	0.2073 (0.0049)	2.8690 (0.5267)	0.0216 (0.0041)	0.0175 (0.0035)
Linear	Cox_G(R&P)	**0.7037 (0.0025)**	0.2089 (0.0020)	2.4790 (0.0611)	0.0226 (0.0017)	0.0179 (0.0014)	**0.7037 (0.0025)**	0.2089 (0.0020)	2.4790 (0.0611)	0.0226 (0.0017)	0.0179 (0.0014)
Linear	$\text{Cox}_\cap$(Keya et al.)	0.6494 (0.0016)	0.1963 (0.0012)	2.1290 (0.2573)	0.0107 (0.0010)	0.0087 (0.0008)	0.6494 (0.0016)	0.1963 (0.0012)	2.1290 (0.2573)	0.0107 (0.0010)	0.0087 (0.0008)
Linear	DRO-COX	0.6927 (0.0069)	0.1868 (0.0004)	2.3090 (0.5215)	**0 (0)**	**0 (0)**	0.6927 (0.0069)	0.1868 (0.0004)	2.3090 (0.5215)	**0 (0)**	**0 (0)**
Linear	DRO-COX (SPLIT)	0.6872 (0.0047)	0.1869 (0.0004)	2.8280 (0.7434)	**0 (0)**	**0 (0)**	0.6872 (0.0047)	0.1869 (0.0004)	2.8280 (0.7434)	**0 (0)**	**0 (0)**
Linear	EXACT DRO-COX	0.6833 (0.0060)	0.2422 (0.0044)	**1.3020 (0.3474)**	0.0056 (0.0005)	0.0045 (0.0004)	0.6833 (0.0060)	0.2422 (0.0044)	**1.3020 (0.3474)**	0.0056 (0.0005)	0.0045 (0.0004)
Nonlinear	DeepSurv	**0.7095 (0.0014)**	0.2200 (0.0012)	2.5990 (0.1189)	0.0309 (0.0006)	0.0249 (0.0004)	**0.7095 (0.0014)**	0.2200 (0.0012)	2.5990 (0.1189)	0.0309 (0.0006)	0.0249 (0.0004)
Nonlinear	DeepSurv_I(Keya et al.)	0.6982 (0.0045)	0.2127 (0.0032)	1.5740 (0.6970)	0.0291 (0.0014)	0.0235 (0.0012)	0.6982 (0.0045)	0.2127 (0.0032)	1.5740 (0.6970)	0.0291 (0.0014)	0.0235 (0.0012)
Nonlinear	DeepSurv_I(R&P)	0.7064 (0.0021)	0.2168 (0.0012)	2.5120 (0.1847)	0.0288 (0.0006)	0.0233 (0.0004)	0.7064 (0.0021)	0.2168 (0.0012)	2.5120 (0.1847)	0.0288 (0.0006)	0.0233 (0.0004)
Nonlinear	DeepSurv_G(Keya et al.)	0.7034 (0.0016)	0.2154 (0.0007)	2.5920 (0.1468)	0.0278 (0.0010)	0.0229 (0.0008)	0.7034 (0.0016)	0.2154 (0.0007)	2.5920 (0.1468)	0.0278 (0.0010)	0.0229 (0.0008)
Nonlinear	DeepSurv_G(R&P)	0.7062 (0.0017)	0.2169 (0.0010)	2.5010 (0.1626)	0.0289 (0.0005)	0.0234 (0.0004)	0.7062 (0.0017)	0.2169 (0.0010)	2.5010 (0.1626)	0.0289 (0.0005)	0.0234 (0.0004)
Nonlinear	$\text{DeepSurv}_\cap$(Keya et al.)	0.6537 (0.0054)	0.1998 (0.0008)	**1.0480 (0.4252)**	0.0136 (0.0012)	0.0111 (0.0010)	0.6537 (0.0054)	0.1998 (0.0008)	**1.0480 (0.4252)**	0.0136 (0.0012)	0.0111 (0.0010)
Nonlinear	FIDP	0.7086 (0.0030)	**0.1824 (0.0033)**	2.3290 (0.2906)	0.0168 (0.0055)	0.0120 (0.0040)	0.7086 (0.0030)	0.1824 (0.0033)	2.3290 (0.2906)	0.0168 (0.0055)	0.0120 (0.0040)
Nonlinear	FIPNAM	0.7022 (0.0118)	0.2226 (0.0019)	2.3480 (0.2087)	0.0181 (0.0020)	0.0129 (0.0016)	0.7022 (0.0118)	0.2226 (0.0019)	2.3480 (0.2087)	0.0181 (0.0020)	0.0129 (0.0016)
	Deep DRO-COX	0.6830 (0.0050)	0.1869 (0.0004)	2.5810 (0.5244)	**5.3651e-06 (6.3580e-06)**	**5.3233e-06 (6.2580e-06)**	0.6830 (0.0050)	0.1869 (0.0004)	2.5810 (0.5244)	**5.3651e-06 (6.3580e-06)**	**5.3233e-06 (6.2580e-06)**
	Deep DRO-COX (SPLIT)	0.6829 (0.0049)	0.1881 (0.0012)	2.4880 (0.5154)	6.3123e-06 (7.2058e-06)	6.2466e-06 (7.0785e-06)	0.6829 (0.0049)	0.1881 (0.0012)	2.4880 (0.5154)	6.3123e-06 (7.2058e-06)	6.2466e-06 (7.0785e-06)
	Deep EXACT DRO-COX	0.7057 (0.0014)	**0.1597 (0.0003)**	2.5030 (0.2540)	0.0277 (0.0004)	0.0225 (0.0003)	0.7057 (0.0014)	**0.1597 (0.0003)**	2.5030 (0.2540)	0.0277 (0.0004)	0.0225 (0.0003)



(a) FLC (age) (b) SUPPORT (age) (c) SEER (age)

Figure 1: Effect of α on test set accuracy (c-index; higher is better) and fairness metrics (F_{CI}, F_{CG}, and CI; lower is better for all fairness metrics) of DRO-COX on four datasets.

We further conduct a number of supplemental experiments that we summarize the findings for now. First, we show that our DRO-COX (SPLIT) procedure is somewhat robust to the choice of $n_1 = |\mathcal{D}_1|$ and $n_2 = |\mathcal{D}_2|$. Second, we show what happens if $\mathcal{D}_1$ and $\mathcal{D}_2$ have different censoring rates, where the main finding is that DRO-COX (SPLIT) can still work well when there is a large imbalance in censoring rates between $\mathcal{D}_1$ and $\mathcal{D}_2$. Third, if DRO-COX (SPLIT) did not use both losses $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ and $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_2 \mid \mathcal{D}_1)$ (i.e., if it only used one of these), then it performs worse. For details on these additional experiments including a formal definition of a quantity that controls the amount of imbalance in censoring rates between $\mathcal{D}_1$ and $\mathcal{D}_2$, see Appendix H.

DeepHit We now compare DRO-DEEPHIT and DRO-DEEPHIT (SPLIT) to the original DeepHit method (Lee et al., 2018) and the regularized variant DeepHit$_G$(R&P). We report the test performance on all three datasets in Table 5. According to the results in Table 5, we have the following observations:

- Our DRO variants can achieve better CI performance than the original DeepHit method on most of datasets with different sensitive attributes when using a CI-based hyperparameter tuning strategy. It is also clear that DRO-DEEPHIT and DRO-DEEPHIT (SPLIT) can achieve lower values on F_{CI} and F_{CG} on all datasets when using an F_{CG}-based hyperparameter

for a survival model to achieve an F_{CI} or F_{CG} score that is *exactly* and not just approximately equal to 0. We set $\gamma = 0.01$ for all datasets and it turns out that in this case, for the Cox model, it is possible for our DRO variants to achieve F_{CI} or F_{CG} scores that are exactly 0. We point out that we have found that if we decrease γ, then we no longer get exactly 0 for F_{CI} or F_{CG}. Ultimately, this issue of F_{CI} or F_{CG} being exactly 0 is due to how they are defined by Rahman and Purushotham (2022), and if one did not want these scores to be exactly 0, one would have to tune on γ in a manner that could depend on the dataset. For more details, see Appendix F.

Table 5: DeepHit test set scores on the FLC, SUPPORT, SEER datasets when hyperparameter tuning is based on CI and F_{CG}.

Datasets	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
FLC (age)	DeepHit	0.7937 (0.0080)	0.1560 (0.0204)	1.1950 (0.7885)	0.0108 (0.0012)	0.0022 (0.0003)	**0.7937 (0.0080)**	0.1560 (0.0204)	1.1950 (0.7885)	0.0108 (0.0012)	0.0022 (0.0003)
	DeepHit$_G$(R&P)	0.7825 (0.0237)	**0.1449 (0.0201)**	1.2340 (0.6046)	0.0097 (0.0023)	0.0021 (0.0003)	0.7446 (0.0051)	**0.1326 (0.0027)**	2.1110 (0.4770)	0.0064 (0.0001)	0.0018 (0.0001)
	DRO-DEEPHIT	**0.7956 (0.0051)**	0.1971 (0.0543)	1.0430 (0.4835)	0.0084 (0.0028)	0.0017 (0.0006)	0.7821 (0.0101)	0.2754 (0.0076)	1.0180 (0.5330)	**0.0026 (0.0005)**	**0.0006 (0.0001)**
	DRO-DEEPHIT (SPLIT)	0.7748 (0.0189)	0.2264 (0.0623)	**0.9950 (0.4556)**	**0.0067 (0.0036)**	**0.0015 (0.0007)**	0.7622 (0.0122)	0.2734 (0.0092)	**0.9270 (0.5411)**	0.0027 (0.0009)	0.0007 (0.0002)
FLC (gender)	DeepHit	0.7937 (0.0080)	0.1560 (0.0204)	0.4990 (0.3792)	0.0108 (0.0012)	0.0055 (0.0006)	**0.7937 (0.0080)**	**0.1560 (0.0204)**	**0.4990 (0.3792)**	0.0108 (0.0012)	0.0055 (0.0006)
	DeepHit$_G$(R&P)	0.7840 (0.0245)	**0.1489 (0.0212)**	0.5170 (0.3982)	0.0099 (0.0021)	0.0050 (0.0011)	**0.7937 (0.0080)**	**0.1560 (0.0204)**	**0.4990 (0.3792)**	0.0108 (0.0012)	0.0055 (0.0006)
	DRO-DEEPHIT	**0.7956 (0.0051)**	0.1971 (0.0543)	**0.4320 (0.4786)**	0.0084 (0.0028)	0.0043 (0.0014)	0.7821 (0.0101)	0.2754 (0.0076)	1.3700 (0.6702)	**0.0026 (0.0005)**	0.0013 (0.0002)
	DRO-DEEPHIT (SPLIT)	0.7748 (0.0189)	0.2264 (0.0623)	1.3100 (0.9915)	**0.0067 (0.0036)**	**0.0034 (0.0018)**	0.7622 (0.0122)	0.2734 (0.0092)	1.9350 (0.7234)	0.0027 (0.0009)	**0.0014 (0.0004)**
SUPPORT (age)	DeepHit	**0.6029 (0.0071)**	0.2151 (0.0067)	3.5910 (0.3987)	0.0055 (0.0008)	0.0026 (0.0004)	**0.6029 (0.0071)**	0.2151 (0.0067)	3.5910 (0.3987)	0.0055 (0.0008)	0.0026 (0.0004)
	DeepHit$_G$(R&P)	0.5775 (0.0050)	**0.2123 (0.0009)**	**1.1940 (0.8221)**	0.0046 (0.0006)	0.0023 (0.0003)	0.5766 (0.0033)	**0.2126 (0.0007)**	**1.0230 (0.4416)**	0.0044 (0.0002)	0.0022 (0.0001)
	DRO-DEEPHIT	0.5932 (0.0159)	0.2447 (0.0147)	2.9160 (0.8347)	**0.0014 (0.0009)**	**0.0007 (0.0004)**	0.5899 (0.0154)	0.2493 (0.0159)	3.3740 (0.6078)	**0.0007 (0.0002)**	**0.0003 (0.0001)**
	DRO-DEEPHIT (SPLIT)	0.5753 (0.0236)	0.2225 (0.0112)	2.7280 (0.9570)	0.0044 (0.0013)	0.0021 (0.0006)	0.5792 (0.0234)	0.2392 (0.0268)	3.5270 (0.7331)	0.0037 (0.0019)	0.0018 (0.0009)
SUPPORT (gender)	DeepHit	**0.6029 (0.0071)**	0.2151 (0.0067)	0.5880 (0.2895)	0.0055 (0.0008)	0.0028 (0.0004)	**0.6029 (0.0071)**	0.2151 (0.0067)	**0.5880 (0.2895)**	0.0055 (0.0008)	0.0028 (0.0004)
	DeepHit$_G$(R&P)	0.5767 (0.0034)	**0.2126 (0.0008)**	0.6960 (0.3183)	0.0044 (0.0002)	0.0022 (0.0001)	0.5773 (0.0039)	**0.2125 (0.0007)**	0.7600 (0.2994)	0.0043 (0.0002)	0.0022 (0.0001)
	DRO-DEEPHIT	0.5932 (0.0159)	0.2447 (0.0147)	1.1980 (0.6834)	**0.0014 (0.0009)**	**0.0007 (0.0005)**	0.5899 (0.0154)	0.2493 (0.0159)	1.4460 (0.4235)	**0.0007 (0.0002)**	**0.0004 (0.0001)**
	DRO-DEEPHIT (SPLIT)	0.5753 (0.0236)	0.2225 (0.0112)	**0.5160 (0.3942)**	0.0044 (0.0013)	0.0022 (0.0006)	0.5792 (0.0234)	0.2392 (0.0268)	0.7550 (0.5022)	0.0037 (0.0019)	0.0019 (0.0010)
SUPPORT (race)	DeepHit	**0.6029 (0.0071)**	0.2151 (0.0067)	1.2250 (0.4454)	0.0055 (0.0008)	0.0033 (0.0005)	**0.6029 (0.0071)**	0.2151 (0.0067)	1.2250 (0.4454)	0.0055 (0.0008)	0.0033 (0.0005)
	DeepHit$_G$(R&P)	0.5767 (0.0031)	**0.2126 (0.0008)**	**0.7290 (0.4122)**	0.0044 (0.0002)	0.0026 (0.0001)	0.5813 (0.0108)	**0.2144 (0.0041)**	**0.7400 (0.4211)**	0.0043 (0.0003)	0.0026 (0.0002)
	DRO-DEEPHIT	0.5932 (0.0159)	0.2447 (0.0147)	1.0630 (0.5174)	**0.0014 (0.0009)**	**0.0009 (0.0005)**	0.5899 (0.0154)	0.2493 (0.0159)	1.4220 (0.4302)	**0.0007 (0.0002)**	**0.0004 (0.0001)**
	DRO-DEEPHIT (SPLIT)	0.5753 (0.0236)	0.2225 (0.0112)	1.1930 (0.4449)	0.0044 (0.0013)	0.0027 (0.0008)	0.5792 (0.0234)	0.2392 (0.0268)	1.5640 (0.6744)	0.0037 (0.0019)	0.0022 (0.0012)
SEER (age)	DeepHit	**0.7156 (0.0047)**	**0.1715 (0.0038)**	1.4450 (0.2901)	0.0122 (0.0011)	0.0038 (0.0003)	**0.7156 (0.0047)**	**0.1715 (0.0038)**	1.4450 (0.2901)	0.0122 (0.0011)	0.0038 (0.0003)
	DeepHit$_G$(R&P)	0.7122 (0.0086)	0.1743 (0.0064)	1.4160 (0.2443)	0.0105 (0.0029)	0.0034 (0.0007)	0.6987 (0.0025)	0.1801 (0.0021)	2.0960 (0.4633)	0.0046 (0.0002)	0.0019 (0.0001)
	DRO-DEEPHIT	0.7112 (0.0084)	0.2794 (0.0871)	**0.7990 (0.3281)**	**0.0061 (0.0041)**	**0.0020 (0.0013)**	0.6951 (0.0051)	0.4122 (0.0304)	1.3800 (0.5574)	**0.0002 (0.0002)**	**0.0001 (0.0001)**
	DRO-DEEPHIT (SPLIT)	0.6969 (0.0211)	0.2073 (0.0464)	1.0240 (0.3449)	0.0107 (0.0016)	0.0038 (0.0004)	0.6963 (0.0224)	0.2063 (0.0419)	**1.2630 (0.7467)**	0.0098 (0.0025)	0.0034 (0.0005)
SEER (race)	DeepHit	**0.7156 (0.0047)**	**0.1715 (0.0038)**	3.2820 (0.5958)	0.0122 (0.0011)	0.0099 (0.0009)	**0.7156 (0.0047)**	**0.1715 (0.0038)**	3.2820 (0.5958)	0.0122 (0.0011)	0.0099 (0.0009)
	DeepHit$_G$(R&P)	0.7132 (0.0073)	0.1728 (0.0045)	3.1330 (0.8321)	0.0113 (0.0028)	0.0091 (0.0023)	0.6987 (0.0048)	0.1806 (0.0028)	**1.6760 (0.6385)**	0.0045 (0.0023)	0.0034 (0.0019)
	DRO-DEEPHIT	0.7112 (0.0084)	0.2794 (0.0871)	3.0120 (0.5652)	**0.0061 (0.0041)**	**0.0049 (0.0033)**	0.6951 (0.0051)	0.4122 (0.0304)	3.2520 (1.7820)	**0.0002 (0.0002)**	**0.0002 (0.0002)**
	DRO-DEEPHIT (SPLIT)	0.6969 (0.0211)	0.2073 (0.0464)	**2.7700 (0.5636)**	0.0107 (0.0016)	0.0085 (0.0014)	0.6963 (0.0224)	0.2063 (0.0419)	3.0070 (0.8355)	0.0098 (0.0025)	0.0078 (0.0021)

tuning strategy. These results indicate that our DRO methods can encourage fairness for DeepHit and can obtain better fairness scores than DeepHit$_G$(R&P).

- We find that our DRO variants outperform DeepHit on F_{CI} and F_{CG} metrics when using CI-based hyperparameter tuning. However, we find that our DRO variants cannot always achieve the best scores on the CI fairness metric when using F_{CG}-based hyperparameter tuning. We conclude that the CI metric may reflect fairness in the F_{CG} fairness metric but the reverse may not be true.
- It is hard to distinguish which method is better between DRO-DEEPHIT and DRO-DEEPHIT (SPLIT). For both methods, as expected, they have slightly lower performance than the

Table 6: SODEN test set scores on the FLC, SUPPORT, SEER datasets when hyperparameter tuning is based on CI and F_{CG}.

Datasets	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
FLC (age)	SODEN	0.7785 (0.0175)	0.1482 (0.0138)	1.1790 (0.6098)	0.0004 (0.0009)	0.0001 (0.0002)	0.7785 (0.0175)	0.1482 (0.0138)	**1.1790 (0.6098)**	0.0004 (0.0009)	0.0001 (0.0002)
	SODEN$_G$(R&P)	0.7832 (0.0138)	0.1454 (0.0134)	0.8248 (0.6491)	0.0006 (0.0007)	0.0002 (0.0002)	**0.7807 (0.0140)**	**0.1454 (0.0134)**	1.7324 (1.2715)	0.0001 (0.0001)	1.2814e-05 (2.2052e-05)
	DRO-SODEN	**0.7857 (0.0124)**	**0.1434 (0.0141)**	**1.0401 (0.7724)**	**1.6903e-05 (3.9865e-05)**	**6.2555e-06 (1.4369e-05)**	0.7787 (0.0134)	0.1619 (0.0247)	1.4140 (0.7545)	**2.4385e-06 (5.8775e-06)**	**8.9596e-07 (2.1956e-06)**
FLC (gender)	SODEN	0.7785 (0.0175)	0.1482 (0.0138)	1.3822 (0.6028)	**0.0004 (0.0009)**	**0.0002 (0.0005)**	0.7785 (0.0175)	0.1482 (0.0138)	1.3822 (0.6028)	0.0004 (0.0009)	0.0002 (0.0005)
	SODEN$_G$(R&P)	0.7824 (0.0126)	0.1496 (0.0117)	0.8252 (0.3665)	0.0005 (0.0006)	0.0003 (0.0003)	**0.7832 (0.0126)**	**0.1452 (0.0131)**	**1.0564 (0.4984)**	0.0001 (0.0001)	2.5966e-05 (4.4394e-05)
	DRO-SODEN	**0.7857 (0.0100)**	**0.1350 (0.0069)**	**0.7115 (0.3545)**	0.0008 (0.0023)	0.0004 (0.0011)	0.7811 (0.0131)	0.1592 (0.0258)	1.5226 (0.7068)	**2.4385e-06 (5.8775e-06)**	**1.5323e-06 (3.5413e-06)**
SUPPORT (age)	SODEN	**0.6276 (0.0101)**	**0.1933 (0.0013)**	2.6275 (0.2490)	0.0081 (0.0007)	0.0035 (0.0003)	**0.6276 (0.0101)**	**0.1933 (0.0013)**	2.6275 (0.2490)	0.0081 (0.0007)	0.0035 (0.0003)
	SODEN$_G$(R&P)	0.6162 (0.0118)	0.2073 (0.0134)	1.9914 (0.4342)	0.0059 (0.0018)	0.0025 (0.0006)	0.6070 (0.0080)	0.2147 (0.0099)	**1.6135 (0.2891)**	0.0043 (0.0008)	0.0020 (0.0004)
	DRO-SODEN	0.6080 (0.0161)	0.2002 (0.0095)	**1.9901 (0.4576)**	**0.0045 (0.0022)**	**0.0021 (0.0012)**	0.5996 (0.0128)	0.2041 (0.0082)	1.9980 (0.4681)	**0.0031 (0.0011)**	**0.0019 (0.0011)**
SUPPORT (gender)	SODEN	**0.6276 (0.0101)**	**0.1933 (0.0013)**	1.7548 (0.1958)	0.0081 (0.0007)	0.0041 (0.0003)	**0.6276 (0.0101)**	**0.1933 (0.0013)**	1.7548 (0.1958)	0.0081 (0.0007)	0.0041 (0.0003)
	SODEN$_G$(R&P)	0.6263 (0.0077)	0.1960 (0.0057)	1.6308 (0.3285)	0.0074 (0.0012)	0.0038 (0.0006)	0.6083 (0.0070)	0.2147 (0.0099)	**1.2239 (0.1634)**	0.0043 (0.0008)	0.0023 (0.0004)
	DRO-SODEN	0.6177 (0.0118)	0.1943 (0.0023)	**1.6282 (0.2094)**	**0.0065 (0.0016)**	**0.0033 (0.0008)**	0.5996 (0.0128)	0.2041 (0.0082)	1.5995 (0.1943)	**0.0031 (0.0011)**	**0.0016 (0.0006)**
SUPPORT (race)	SODEN	**0.6276 (0.0101)**	**0.1933 (0.0013)**	1.6910 (0.2182)	0.0081 (0.0007)	0.0048 (0.0004)	**0.6276 (0.0101)**	**0.1933 (0.0013)**	**1.6910 (0.2182)**	0.0081 (0.0007)	0.0048 (0.0004)
	SODEN$_G$(R&P)	0.6137 (0.0085)	0.2045 (0.0115)	1.6304 (0.1344)	0.0058 (0.0016)	0.0036 (0.0009)	0.6089 (0.0073)	0.2164 (0.0101)	1.7705 (0.4111)	0.0043 (0.0008)	0.0027 (0.0005)
	DRO-SODEN	0.6113 (0.0143)	0.1993 (0.0093)	**1.3418 (0.4286)**	**0.0052 (0.0025)**	**0.0032 (0.0015)**	0.5996 (0.0128)	0.2041 (0.0082)	1.1979 (0.4273)	**0.0031 (0.0011)**	**0.0019 (0.0006)**
SEER (age)	SODEN	**0.7132 (0.0017)**	**0.1550 (0.0009)**	**0.8531 (0.0940)**	0.0280 (0.0014)	0.0075 (0.0006)	**0.7132 (0.0017)**	**0.1550 (0.0009)**	**0.8531 (0.0940)**	0.0280 (0.0014)	0.0075 (0.0006)
	SODEN$_G$(R&P)	0.7131 (0.0017)	0.1556 (0.0011)	0.8541 (0.1562)	0.0277 (0.0011)	0.0075 (0.0006)	0.7122 (0.0009)	0.1561 (0.0012)	0.9110 (0.0948)	0.0276 (0.0013)	0.0072 (0.0004)
	DRO-SODEN	0.7026 (0.0116)	0.1757 (0.0293)	1.1275 (0.3644)	**0.0227 (0.0054)**	**0.0071 (0.0010)**	0.6980 (0.0108)	0.2008 (0.0367)	1.4280 (0.5242)	**0.0161 (0.0095)**	**0.0049 (0.0021)**
SEER (race)	SODEN	**0.7132 (0.0017)**	**0.1550 (0.0009)**	2.4948 (0.1341)	0.0280 (0.0014)	0.0227 (0.0011)	**0.7132 (0.0017)**	**0.1550 (0.0009)**	2.4948 (0.1341)	0.0280 (0.0014)	0.0227 (0.0011)
	SODEN$_G$(R&P)	0.7124 (0.0016)	0.1558 (0.0011)	2.4390 (0.1775)	0.0273 (0.0013)	0.0220 (0.0011)	0.7123 (0.0016)	0.1561 (0.0013)	2.4747 (0.1869)	0.0271 (0.0013)	0.0218 (0.0011)
	DRO-SODEN	0.6913 (0.0109)	0.2079 (0.0373)	**1.6398 (0.4948)**	**0.0167 (0.0057)**	**0.0132 (0.0047)**	0.6893 (0.0055)	0.2191 (0.0269)	**1.7676 (0.4458)**	**0.0126 (0.0059)**	**0.0099 (0.0047)**

DeepHit method on accuracy metrics. However, DRO-DEEPHIT method has the best C^{td} performance on the FLC dataset in Table 5.

SODEN We conduct experiments to compare the accuracy and fairness of SODEN and SODEN$_G$(R&P) to DRO-SODEN. Our experimental results are reported in Table 6 (CI-based hyperparameter tuning and F_{CG}-based hyperparameter tuning). From these results, we have the following observations:

- When tuning hyperparameters based on CI, it is clear that DRO-SODEN outperforms the other methods on the CI fairness metric for FLC and SUPPORT datasets. Meanwhile, F_{CI} and F_{CG} are also reduced by using DRO-SODEN while accuracy scores become a little lower than those of SODEN. However, we find DRO-SODEN can achieve a slightly higher C^{td} scores on the FLC dataset.
- When tuning hyperparameters based on F_{CG}, we find DRO-SODEN also can achieve better performance on F_{CG} than the corresponding values that are from the CI-based tuning since we tune hyperparameters based on this metric. In addition, DRO-SODEN can obtain the best F_{CG} and F_{CI} scores compared to the baselines.

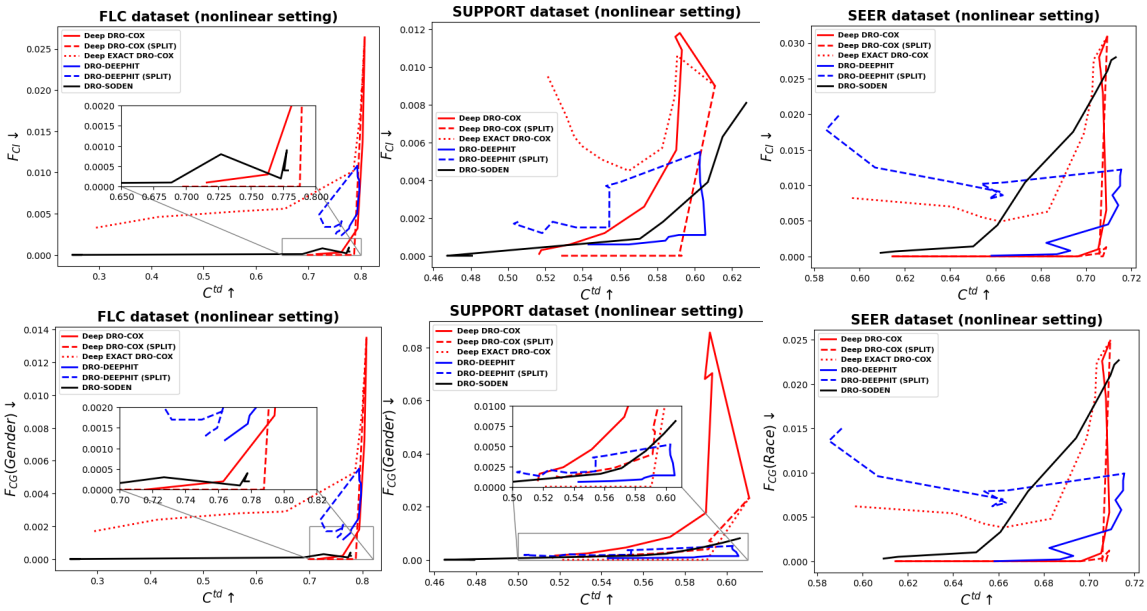


Figure 2: Comparison of all proposed fairness methods in terms of F_{CI} (first row) and F_{CG} (second row) with C^{td} on FLC, SUPPORT, and SEER datasets. Each line is drawn based on the various values of α (from 0.1 to 1.0). In each subfigure, the closer the curve is to the lower right corner, the better the performance.

5.6 Accuracy Fairness Tradeoff Comparison Across DRO Variants of Different Survival Models

We can compare the tradeoff between accuracy and tradeoff across different DRO variants. Specifically, for the DEEP DRO-COX, DEEP DRO-COX (SPLIT), DRO-DEEPHIT, DRO-DEEPHIT (SPLIT), and DRO-SODEN models, we test them under the nonlinear setting on FLC, SUPPORT, and SEER datasets. We evaluate the F_{CI}, F_{CG}, and C^{td} scores of all methods using different values of α from 0.1 to 1.0 and then plot accuracy vs fairness curves for each dataset, as shown in Figure 2. Note that in each plot, being closer to the lower right is considered better, corresponding to a model having an α value that achieves as low of a fairness metric score (either F_{CI} or F_{CG}) as possible (which is considered more fair) and as high of a C^{td} score as possible. From the figure, we find that the DRO-DEEPHIT method outperforms the other methods.

6. Discussion

We have shown a general strategy for converting a wide class of survival models into DRO variants that encourage fairness. The key idea is to write the overall loss in terms of individual losses, which in turn could be used in a DRO framework. When there is coupling so that an individual loss technically is not "individual" as it depends on multiple data points, we introduced a sample splitting approach that is compliant with DRO theory. We also showed that the heuristic approach that ignores this coupling problem and naively runs an

existing DRO algorithm (that assumes that there is no coupling) works in practice about as well as the sample splitting version. When a survival model used does not have this coupling issue (such as SODEN or the Cox model when using the full Cox loss as in Section 4), then existing DRO machinery directly works; there is no need to use any sample splitting. Specifically for the Cox model, we derive an exact DRO approach that does not use any sample splitting, where the trick is to lift the problem to a higher dimensional (in terms of the number of survival model parameters) space where the coupling issue vanishes.

We now discuss some extensions of our work as well as open questions.

Competing risks In various time-to-event prediction problems, we aim to predict the time until the earliest of multiple competing events happen along with which such event happens. For instance, consider hospitalized patients who are in a coma. We may be interested in predicting their time until awakening. However, it could be that they die before awakening. Thus, the two critical events that compete as to which happens first is awakening vs death. Meanwhile, by the time we stop collecting training data, some patients could still be in a coma (so that their outcome is censored). Such a setting is referred to as a *competing risks* problem (see, for instance, Chapter 8 of the textbook by Kalbfleisch and Prentice (2002)).

Our DRO conversion framework can easily accommodate the competing risks setting. To illustrate this, consider the DeepHit model (Lee et al., 2018) that was originally designed to handle competing risks (and that we actually simplified in our exposition in Section 2.2.2 and Example 2 to be for the standard survival analysis setting). Suppose that there are $\delta_{\max} \in \mathbb{N}$ competing events, which we simply label as the events $1, 2, \ldots, \delta_{\max}$. Now each training point still is represented by the triple (X_i, Y_i, Δ_i) but Y_i is the time until the earliest critical event happens (or the censoring time if censoring happened prior to any critical event happening), and $\Delta_i \in \{0, 1, \ldots, \delta_{\max}\}$ indicates which critical event happened first (with the special value of 0 meaning that censoring happened first).

Then in general, DeepHit aims to estimate the so-called *cumulative incidence function* (CIF) (Gray, 1988; Fine and Gray, 1999) that is specific to each event $\delta \in [\delta_{\max}]$:

$$F_\delta(t|x) \triangleq \mathbb{P}(Y \leq t, \Delta = \delta \mid X = x).$$

To estimate the CIF, DeepHit uses a user-specified discrete time grid $t_1 < t_2 < \cdots < t_m$. Letting random variable T be the time until the earliest event happens for a data point with feature vector X, and letting random variable Δ indicate which of the critical events is the earliest to happen (also for feature vector X), we define

$$\mathbb{P}(T = t_j, \Delta = \delta \mid X = x) \triangleq f_{\delta,j}(x;\theta) \quad \text{for } \delta \in [\delta_{\max}] \text{ and } j \in [m], \tag{32}$$

where neural network

$$\begin{aligned} f(x;\theta) = \big(&f_{1,1}(x;\theta), f_{1,2}(x;\theta), \ldots, f_{1,m}(x;\theta), \\ &f_{2,1}(x;\theta), f_{2,2}(x;\theta), \ldots, f_{2,m}(x;\theta), \\ &\ldots, \\ &f_{\delta_{\max},1}(x;\theta), f_{\delta_{\max},2}(x;\theta), \ldots, f_{\delta_{\max},m}(x;\theta)\big) \in [0,1]^{\delta_{\max} \cdot m} \end{aligned}$$

has parameters θ and maps a raw input $x \in \mathcal{X}$ to a probability distribution over $\delta_{\max} \cdot m$ entries. Note that when there is only one critical event of interest (so that $\delta_{\max} = 1$), then equation (32) reduces to equation (6)).

In particular, DeepHit's estimate of the CIF is given by

$$\widehat{F}_\delta(t_\ell|x) \triangleq \sum_{j=1}^{\ell} f_{\delta,j}(x;\theta) \qquad \text{for } \delta \in [\delta_{\max}] \text{ and } \ell \in [m].$$

Moreover, the loss function of DeepHit in this case is

$$L^{\text{DeepHit-general}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} L_i^{\text{DeepHit-general}}(\theta),$$

where the i-th individual loss is

$$\begin{aligned}
&L_i^{\text{DeepHit-general}}(\theta) \\
&= \beta \cdot \left[-\mathbf{1}\{\Delta_i \neq 0\} \log(f_{\Delta_i,\kappa(Y_i)}(X_i;\theta)) - \mathbf{1}\{\Delta_i = 0\} \log\left(1 - \sum_{\delta=1}^{\delta_{\max}} \sum_{\ell=1}^{\kappa(Y_i)} f_{\delta,\ell}(X_i;\theta)\right)\right] \\
&\quad + (1-\beta) \cdot \frac{1}{n} \cdot \mathbf{1}\{\Delta_i \neq 0\} \sum_{\substack{j\in[n] \text{ s.t.} \\ \kappa(Y_j)>\kappa(Y_i)}} \exp\left(\frac{\sum_{\ell=1}^{\kappa(Y_i)}[f_{\Delta_i,\ell}(X_j;\theta) - f_{\Delta_i,\ell}(X_i;\theta)]}{\sigma}\right).
\end{aligned}$$

Note that the model hyperparameters are the same as what we had presented earlier in Section 2.2.2, with the only minor difference being that in practice, for the ranking loss, it could be helpful to weight the contributions of different competing events differently (i.e., for the i-th point to have a ranking loss contribution, it needs to have $\Delta_i \neq 0$, in which case we multiply by a scalar weight hyperparameter specific to the event type $\Delta_i \in [\delta_{\max}]$).

In particular, because we can write the loss function as the average of individual loss terms, we can convert this model into a DRO variant using either the heuristic or sample splitting approaches we presented in Section 3.2.

Tuning subpopulation probability threshold α In using DRO, tuning the subpopulation probability threshold α can significantly impact the results. In our experiments, we tuned α using one of two different fairness metrics (CI or F_{CG}) on a validation set. While DRO (whether heuristic or using sample splitting) itself does not require the user to specify which features to treat as sensitive in the training loss, we are effectively using some information about which features to treat as sensitive as it shows up in computing the validation set fairness metric. We do this primarily because this is how other researchers have tuned hyperparameters for fair survival models. An open question thus arises of whether we could tune α in some other way in practice that either does not use a validation set or which does not require using a validation set fairness metric that knows which features to treat as sensitive. Fundamentally this is about coming up with other fairness evaluation metrics that can be used for the validation set.

Choosing "optimal" splits For simplicity, in how we presented our split DRO approach, we used 2-fold cross-fitting, where a key step is randomly splitting the training data into the two sets $\mathcal{D}_1$ and $\mathcal{D}_2$ (as a reminder, we explain how to use more than 2 folds in Appendix B). Furthermore, our main theoretical guarantee for split DRO (Theorem 5) assumes that $|\mathcal{D}_1| =$

$|\mathcal{D}_2|$. We defer identifying an "optimal" split to future work. Some interesting problems that arise include coming up with some notion of optimality of a split, and figuring out the number of folds that would be optimal. It could even be that instead of randomly splitting the training data, there could be some better non-random optimization-based approach for data splitting.

Impact of DRO on evaluation metrics that are not about fairness Lastly, we point out that it would be interesting to empirically study how converting a survival model into its DRO variant impacts other metrics aside from the accuracy or fairness metrics we considered, such as calibration metrics (Haider et al., 2020; Goldstein et al., 2020). Ultimately, we suspect that DRO variants of survival models potentially have interesting properties that make them useful beyond encouraging fairness.

Acknowledgments

Shu Hu is supported by the U.S. National Science Foundation (NSF) CRII award IIS-2434967, the National Artificial Intelligence Research Resource (NAIRR) Pilot, and TACC Lonestar6. George H. Chen is supported by NSF CAREER award #2047981. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the NSF or the NAIRR Pilot.

Appendix A. More Details on Cox Models

A.1 Estimating the Baseline Hazard and Conditional Survival Function

After learning the log partial hazard function $f(\cdot;\theta)$ (or, equivalently, learning the parameters θ), a standard approach to estimating the baseline hazard function h_0 is to use the so-called Breslow method (Breslow, 1972). In what follows, we use $\widehat{\theta}$ to denote the learned estimate of θ.

The Breslow method estimates a discretized version of h_0. Specifically, let $t_1 < t_2 < \cdots < t_m$ denote the unique times when critical event happened in the training data. Let d_j denote the number of critical events that occurred at time t_j for $j \in [m]$. Then we compute the following estimate of h_0 at the j-th time step:

$$\widehat{h}_{0,j} \triangleq \frac{d_j}{\sum_{i\in[n] \text{ s.t. } Y_i \geq t_j} \exp(f(x_i;\widehat{\theta}))} \qquad \text{for } j \in [m].$$

After estimating the baseline hazard function, estimating the survival function is straightforward. Recall that $S(t|x) = \exp\big(-\int_0^t h(u|x)\mathrm{d}u\big)$. Then combining this equation with the proportional hazards assumption (i.e., the factorization in equation (3)), we get

$$S(t|x) = \exp\Big(-\int_0^t h_0(u)\exp(f(x;\theta))\mathrm{d}u\Big) = \exp\Big(\Big[\underbrace{-\int_0^t h_0(u)\mathrm{d}u}_{\text{abbreviate as } H_0(t)}\Big]\exp(f(x;\theta))\Big). \quad \text{(A.1)}$$

We can estimate $H_0(t)$ via a summation in place of an integration:

$$\widehat{H}_0(t) \triangleq \sum_{j\in[m] \text{ s.t. } t_j \leq t} \widehat{h}_{0,j} \qquad \text{for } t \geq 0.$$

Thus, by plugging in $\widehat{H}_0$ in place of H_0 and $\widehat{\theta}$ in place of θ in equation (A.1), we obtain the conditional survival function estimate $\widehat{S}(t|x) \triangleq \exp(-\widehat{H}_0(t)\exp(f(x;\widehat{\theta})))$.

A.2 The Proportional Hazards Assumption and the Shape of the Conditional Survival Function

The proportional hazards assumption constrains the shape of the conditional survival function. Recall that for any two real numbers $a, b \in \mathbb{R}$, we have $\exp(a \cdot b) = (\exp(a))^b$. Then equation (A.1) (which was derived using the proportional hazard assumption) is equal to

$$S(t|x) = \exp\big(H_0(t)\exp(f(x;\theta))\big) = [\underbrace{\exp(H_0(t))}_{\triangleq S_0(t)}]^{\exp(f(x;\theta))}.$$

In other words, under the proportional hazards assumption, the conditional survival function $S(\cdot|x)$ must necessarily be a power of the so-called baseline survival function $S_0(\cdot)$.

A.3 Details on the Full Cox Loss

Throughout this section, we assume that we have done the preprocessing stated for Proposition 9, namely that for $i \in [n]$ such that $\Delta_i = 0$, we have set $Y_i = t_{\kappa(Y_i,0)}$.

Deriving the full Cox loss function For a survival model specified in terms of a hazard function, the full likelihood (see, e.g., Section 3.2 of Kalbfleisch and Prentice (2002)) is

$$\text{likelihood} \triangleq \prod_{i=1}^{n} \left\{ [h(Y_i|X_i)]^{\Delta_i} \exp\left(- \int_0^{Y_i} h(u|X_i) du \right) \right\}. \tag{A.2}$$

For the Cox model, recall that the hazard function is given by

$$h(t|x) = h_0(t) e^{f(x;\theta)}.$$

Under the assumption that h_0 is piecewise constant, as given in equation (28), we have

$$h(t|x) = \begin{cases} e^{\psi_\ell + f(x;\theta)} & \text{if } t \in (t_{\ell-1}, t_\ell] \text{ for } \ell \in [m], \\ 0 & \text{otherwise.} \end{cases}$$

Plugging this expression for $h(t|x)$ into the full likelihood (equation (A.2)), we get

$$\text{likelihood} = \prod_{i=1}^{n} \left\{ [e^{\psi_\ell + f(X_i;\theta)}]^{\Delta_i} \exp\left(- e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \right) \right\}, \tag{A.3}$$

where we have crucially used the preprocessing of the censoring data's observed times in evaluating the integral. (If we did not do the preprocessing, we could still come up with an expression for the integral but the math gets messy.)

Taking the negative log of both sides of equation (A.3), we get:

$$-\text{log likelihood} = -\sum_{i=1}^{n} \left\{ \Delta_i [\psi_{\kappa(Y_i,\Delta_i)} + f(X_i;\theta)] - e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \right\}$$

Multiplying both sides by $\frac{1}{n}$, we get the full loss (i.e., equation (29)), which we reproduce here for convenience:

$$L^{\text{Cox-full}}(\theta, \psi) = \frac{1}{n} \sum_{i=1}^{n} \left\{ - \Delta_i [\psi_{\kappa(Y_i,\Delta_i)} + f(X_i;\theta)] + e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \right\}.$$

Proof of Proposition 9 First, we do some re-indexing (to introduce summation over the unique times in which critical events happen):

$$\begin{aligned}
&L^{\text{Cox-full}}(\theta, \psi) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ - \Delta_i \psi_{\kappa(Y_i,\Delta_i)} - \Delta_i f(X_i;\theta) + e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \right\} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \psi_{\kappa(Y_i,\Delta_i)} - \frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{i=1}^{n} e^{f(X_i;\theta)} \sum_{\ell=1}^{\kappa(Y_i,\Delta_i)} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \\
&= -\frac{1}{n} \sum_{\ell=1}^{m} \underbrace{\sum_{j=1}^{n} \Delta_j \mathbf{1}\{Y_j = t_\ell\}}_{\triangleq d_\ell} \psi_\ell - \frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{\ell=1}^{m} (t_\ell - t_{\ell-1}) e^{\psi_\ell} \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)}.
\end{aligned} \tag{A.4}$$

Taking the derivative of $L^{\text{Cox-full}}(\theta, \psi)$ with respect to ψ_ℓ for $\ell \in [m]$, we get

$$\frac{\partial L^{\text{Cox-full}}(\theta, \psi)}{\partial \psi_\ell} = -\frac{d_\ell}{n} + \left[\frac{1}{n}(t_\ell - t_{\ell-1}) \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq \ell\} e^{f(X_j;\theta)}\right] e^{\psi_\ell}.$$

By setting this derivative to 0, we get that the optimal value of ψ_ℓ is

$$\widehat{\psi}_\ell = \log \frac{d_\ell}{(t_\ell - t_{\ell-1}) \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)}}.$$

One can verify that indeed $[\frac{\partial^2 L^{\text{Cox-full}}(\theta,\psi)}{\partial \psi_\ell^2}]_{\psi_\ell = \widehat{\psi}_\ell} > 0$ so that this optimal value corresponds to a minimum.

Finally, we plug in $\widehat{\psi} \triangleq (\widehat{\psi}_1, \ldots, \widehat{\psi}_m)$ in place of $\psi = (\psi_1, \ldots, \psi_m)$ in $L^{\text{Cox-full}}(\theta, \psi)$ (using equation (A.4)):

$$\begin{aligned}
&L^{\text{Cox-full}}(\theta, \widehat{\psi}) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) - \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \widehat{\psi}_\ell + \frac{1}{n} \sum_{\ell=1}^{m} (t_\ell - t_{\ell-1}) e^{\widehat{\psi}_\ell} \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) - \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \log \frac{d_\ell}{(t_\ell - t_{\ell-1}) \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)}} + \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) - \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \left[\log \frac{d_\ell}{t_\ell - t_{\ell-1}} - \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)}\right] + \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{\ell=1}^{m} d_\ell \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)} + \underbrace{\text{constant}}_{\text{w.r.t. } \theta} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{\ell=1}^{m} \left[\sum_{i=1}^{n} \Delta_i \mathbf{1}\{Y_i = t_\ell\}\right] \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)} + \text{constant} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{i=1}^{n} \Delta_i \sum_{\ell=1}^{m} \mathbf{1}\{Y_i = t_\ell\} \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_\ell\} e^{f(X_j;\theta)} + \text{constant} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \Delta_i f(X_i;\theta) + \frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq Y_i\} e^{f(X_j;\theta)} + \text{constant} \\
&= \frac{1}{n} \sum_{i=1}^{n} -\Delta_i \left[f(X_i;\theta) - \log \sum_{j=1}^{n} \mathbf{1}\{Y_j \geq Y_i\} e^{f(X_j;\theta)}\right] + \text{constant} \\
&= L^{\text{Cox}}(\theta) + \text{constant}.
\end{aligned}$$

■

Appendix B. Cross-Fitting With More Than Two Folds

For example, for some pre-specified number of folds K_{folds}, we could randomly partition the training data into K_{folds} roughly equal-size sets $\mathcal{D}_1, \ldots, \mathcal{D}_{K_{\text{folds}}}$. Then for each $k \in [K_{\text{folds}}]$,

we could either set

$$L_{\mathrm{DRO}}^{\mathrm{split}}(\theta, \eta^{(1)}, \ldots, \eta^{(K_{\mathrm{folds}})}) \triangleq \sum_{k=1}^{K_{\mathrm{folds}}} L_{\mathrm{DRO}}^{\mathrm{split}}\big(\theta, \eta^{(k)}, \mathcal{D}_k \mid ([n] \setminus \mathcal{D}_k)\big),$$

or

$$L_{\mathrm{DRO}}^{\mathrm{split}}(\theta, \eta^{(1)}, \ldots, \eta^{(K_{\mathrm{folds}})}) \triangleq \sum_{k=1}^{K_{\mathrm{folds}}} L_{\mathrm{DRO}}^{\mathrm{split}}\big(\theta, \eta^{(k)}, ([n] \setminus \mathcal{D}_k) \mid \mathcal{D}_k\big).$$

Appendix C. Proof of Theorem 5

We prove the following.

Proposition 10 *(Slightly more general version of Theorem 5) Let $n \geq 2$ and randomly split the training data into $\mathcal{D}_1$ and $\mathcal{D}_2$ of sizes $n_1 \geq 1$ and $n_2 = n - n_1$. Let $\omega > 0$. Suppose that Assumptions A1–A6 hold. If $\phi_{transform}(s) = s$, then define*

$$M \triangleq M_{indiv} + n_2 M_{couple\text{-}max},$$
$$M' \triangleq (M_{couple\text{-}max} - M_{couple\text{-}min})\sqrt{\frac{\omega n_2}{\zeta}}.$$

If instead $\phi_{transform}(s) = \log(1+s)$, then define

$$M \triangleq M_{indiv} + \log(1 + n_2 M_{couple\text{-}max}),$$
$$M' \triangleq \frac{(M_{couple\text{-}max} - M_{couple\text{-}min})}{\zeta M_{couple\text{-}min}}\sqrt{\frac{8\omega}{n_2}}.$$

Then with probability at least

$$1 - 2\left[\frac{M}{(C_\alpha - 1)\big[\sqrt{\frac{2\omega}{n_1}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}M + (2\mathcal{L}+1)M'\big]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n_2\zeta}{8}}$$

over randomness in the training data, we have

$$\left|\inf_{\eta\in\mathbb{R}} L_{DRO}^{split}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta\in\mathbb{R}} R_{DRO}^{split}(\theta, \eta)\right|$$
$$\leq 10C_\alpha^2\left[\frac{1}{\sqrt{n_1}}\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\}\big(\sqrt{2\omega}+1\big)M + (2\mathcal{L}+1)M'\right].$$

Theorem 5 corresponds to Proposition 10, where we assume n to be even and we set $n_1 = n_2 = n/2$.

We define

$$L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta, \eta) \triangleq C_\alpha\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\mathrm{indiv}}(\theta; X, Y, \Delta) - \eta]_+^2\big]} + \eta.$$

The main goal in the proof is to bound

$$\begin{aligned}
&|L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - R_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta)| \\
&\quad = |L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)] \\
&\qquad + \mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)] - L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) \\
&\qquad + L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - R_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta)| \\
&\quad \le \underbrace{|L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - \mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)]|}_{\triangleq \spadesuit} \\
&\qquad + \underbrace{|\mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)] - L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2)|}_{\triangleq \heartsuit} \\
&\qquad + \underbrace{|L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) - R_{\mathrm{DRO}}^{\mathrm{split}}(\theta,\eta)|}_{\triangleq \clubsuit},
\end{aligned}$$

where we have used the triangle inequality. The bulk of the proof is in upper-bounding ♠, ♡, and ♣. Prior to bounding these, we collect two important lemmas. The first establishes that $L_{\mathrm{indiv}}(\theta;x,y,\delta)$ and $R_{\mathrm{indiv}}(\theta;x,y,\Delta)$ are bounded. Note that we defer all proofs of lemmas to subsections immediately following this main proof outline.

Lemma 11 *Under Assumption A4, for all $(x,y,\delta) \in \mathcal{Z}$, if $\phi_{transform}$ is the identity function, then*

$$L_{indiv}(\theta;x,y,\delta), R_{indiv}(\theta;x,y,\delta) \in [0, M_{indiv} + n_2 M_{couple\text{-}max}].$$

Otherwise if $\phi_{transform}(s) = \log(1+s)$, then

$$L_{indiv}(\theta;x,y,\delta), R_{indiv}(\theta;x,y,\delta) \in [0, M_{indiv} + \log(1 + n_2 M_{couple\text{-}max})].$$

In fact, M is defined in Proposition 10 precisely based on the upper bounds in Lemma 11.

The next lemma says that even though we are optimizing over $\eta \in \mathbb{R}$, we actually only need to consider η within a closed interval that depends on M.

Lemma 12 *(Slight variant of Lemma 9 of Duchi and Namkoong (2021)) We have*

$$\inf_{\eta\in\mathbb{R}} L_{DRO}^{split}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2) = \inf_{\eta\in[-\frac{1}{C_\alpha-1}M, M]} L_{DRO}^{split}(\theta,\eta,\mathcal{D}_1 \mid \mathcal{D}_2),$$

and similarly

$$\inf_{\eta\in\mathbb{R}} R_{DRO}^{split}(\theta,\eta) = \inf_{\eta\in[-\frac{1}{C_\alpha-1}M, M]} R_{DRO}^{split}(\theta,\eta).$$

Now we present the bounds on ♠, ♡, and ♣ in three successive lemmas.

Lemma 13 *(Bound on ♠; appears in the proof of Theorem 2 of Duchi and Namkoong (2021)) Let $\omega > 0$. We have*

$$\mathbb{P}\Bigg(\underbrace{\spadesuit \ge C_\alpha \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M\sqrt{\frac{2\omega}{n_1}}}_{\mathcal{E}_{bad\ spade}(\eta)}\Bigg) \le 2e^{-\omega}.$$

Note that for this probabilistic bound, the event $\mathcal{E}_{bad\ spade}(\eta)$, as the notation suggests, depends on η as $\spadesuit$ depends on η (later we union bound $\mathcal{E}_{bad\ spade}(\eta)$ over a finite choice of options of η).

The statement of this lemma depends on Lemma 11 (since the constant M shows up), and the proof crucially uses the fact that from Lemma 12, we know that it suffices to only consider $\eta \in [-\frac{1}{C_\alpha - 1}M, M]$.

Lemma 14 *(Bound on $\heartsuit$; appears in the proof of Theorem 2 of Duchi and Namkoong (2021)) We have*

$$\heartsuit \leq C_\alpha \sqrt{\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\} M} \cdot \frac{1}{\sqrt{n_1}}.$$

Once again, the statement of this lemma depends on Lemma 11 (due to the constant M showing up), and the proof again uses the fact that $\eta \in [-\frac{1}{C_\alpha - 1}M, M]$.

Lemma 15 *(Bound on $\clubsuit$) Under Assumptions A1, A2, and A4, we have*

$$\begin{aligned}\clubsuit &\leq C_\alpha \sup_{(x,y,\delta)\in\mathcal{Z}} |L_{indiv}(\theta; x, y, \delta) - R_{indiv}(\theta; x, y, \delta)| \\ &= C_\alpha \underbrace{\max_{(x,y,\delta)\in\mathcal{Z}} |L_{indiv}(\theta; x, y, \delta) - R_{indiv}(\theta; x, y, \delta)|}_{\triangleq \diamondsuit}.\end{aligned}$$

Importantly, the supremum is attained by a specific point in the set $\mathcal{Z} \triangleq \mathcal{X} \times \{t_1, t_2, \ldots, t_m\} \times \{0, 1\}$.

To help upper-bound $\diamondsuit$, we make use of the following lemma.

Lemma 16 *(Enough data points in $\mathcal{D}_2$ for every time index) Define the bad event*

$$\begin{aligned}&\mathcal{E}_{bad\ time} \\ &\triangleq \bigcup_{\ell=1}^{m} \left\{ \text{the number of points in } \{(X_i, Y_i, \Delta_i)\}_{i\in\mathcal{D}_2} \text{ with observed time equal to } t_\ell \text{ is} \leq \frac{n_2\zeta}{2} \right\} \\ &= \bigcup_{\ell=1}^{m} \left\{ \sum_{i\in\mathcal{D}_2} \mathbf{1}\{Y_i = t_\ell\} \leq \frac{n_2\zeta}{2} \right\}.\end{aligned}$$

Under Assumption A2,

$$\mathbb{P}(\mathcal{E}_{bad\ time}) \leq m e^{-\frac{n_2\zeta}{8}}.$$

The reason that Lemma 16 is helpful is that it ensures that the adjacency sets we get are large enough. Specifically, note that by Assumption A3, we use

$$\mathcal{A}^*((x, y, \delta), \mathcal{C}) = \begin{cases} \varnothing & \text{if } \delta = 0, \\ \{(x', y', \delta') \in \mathcal{C} : \kappa(y') \geq \kappa(y) & \text{otherwise.} \end{cases}$$

In particular, for any time index t_ℓ for $\ell \in [m]$, the set

$$\begin{aligned}\mathcal{A}^*((x, t_\ell, 1), \{(X_i, Y_i, \Delta_i)\}_{i\in\mathcal{D}_2}) &= \{(X_i, Y_i, \Delta_i) : i \in \mathcal{D}_2 \text{ and } \kappa(Y_i) \geq \ell\} \\ &= \{(X_i, Y_i, \Delta_i) : i \in \mathcal{D}_2 \text{ and } Y_i \geq t_\ell\},\end{aligned}$$

has cardinality

$$\sum_{\widetilde{\ell}=\ell}^{m} \sum_{i\in\mathcal{D}_2} \mathbf{1}\{Y_i = t_{\widetilde{\ell}}\} \geq \sum_{i\in\mathcal{D}_2} \mathbf{1}\{Y_i = t_\ell\} > \frac{n_2\zeta}{2},$$

where the strict inequality occurs when $\mathcal{E}_{\text{bad time}}$does not happen.

Before we bound $\diamondsuit$ from Lemma 15, we collect one more lemma.

Lemma 17 *Let $\omega > 0$. Let $(x, y, \delta) \in \mathcal{Z}$. Under Assumptions A1—A4, when $\delta = 0$, trivially*

$$|L_{indiv}(\theta; x, y, \delta) - R_{indiv}(\theta; x, y, \delta)| = 0.$$

Otherwise, suppose that the bad event $\mathcal{E}_{bad\ time}$ in Lemma 16 does not happen:

- *If $\phi_{transform}(s) = s$, then*

$$\mathbb{P}\left(|L_{indiv}(\theta; x, y, \delta) - R_{indiv}(\theta; x, y, \delta)| \geq (M_{couple\text{-}max} - M_{couple\text{-}min})\sqrt{\frac{\omega n_2}{\zeta}}\right) \leq 2e^{-\omega}.$$

- *If $\phi_{transform}(s) = \log(1 + s)$, then*

$$\mathbb{P}\left(|L_{indiv}(\theta; x, y, \delta) - R_{indiv}(\theta; x, y, \delta)| \geq \frac{(M_{couple\text{-}max} - M_{couple\text{-}min})}{\zeta M_{couple\text{-}min}}\sqrt{\frac{8\omega}{n_2}}\right) \leq 2e^{-\omega}.$$

Note that M' from Proposition 10 is precisely defined based on the bounds in Lemma 17. Moreover, we now define the bad event based on Lemma 17:

$$\mathcal{E}_{\text{bad couples}}(x, y, \delta) \triangleq \left\{|L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| \geq M'\right\}.$$

Note that this bad event depends on (x, y, δ). By compactness of $\mathcal{X}$ (Assumption A1) and the time grid being finite (Assumption A2), let $\mathcal{Q}$ be an M'-cover of minimal size for $\mathcal{X}$ in Euclidean norm (so that $|\mathcal{Q}| = \mathbb{N}(M', \mathcal{X})$); denote the elements of $\mathcal{Q}$ by $q_1, q_2, \ldots, q_{\mathbb{N}(M',\mathcal{X})}$, and for $x \in \mathcal{X}$, let $j(x) \in [\mathbb{N}(\varepsilon, \mathcal{X})]$ be the index of the closest point (in Euclidean distance) from $\mathcal{Q}$ to x (breaking ties arbitrarily). Then we shall union bound over $\mathcal{E}_{\text{bad couples}}(x, y, \delta)$ for all $x \in \mathcal{Q}$, $y \in \{t_1, \ldots, t_m\}$, and $\delta \in \{0, 1\}$. Importantly, by ensuring that $\mathcal{E}_{\text{bad couples}}(x, y, \delta)$ holds at all these coordinates means that

$$\begin{aligned}\diamondsuit &= \max_{(x,y,\delta)} |L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| \\ &= \max_{(x,y,\delta)} \big\{|L_{\text{indiv}}(\theta; x, y, \delta) - L_{\text{indiv}}(\theta; q_{j(x)}, y, \delta)| \\ &\qquad + |L_{\text{indiv}}(\theta; q_{j(x)}, y, \delta) - R_{\text{indiv}}(\theta; q_{j(x)}, y, \delta)| \\ &\qquad + |R_{\text{indiv}}(\theta; q_{j(x)}, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)|\big\} \\ &\leq \max_{(x,y,\delta)} \{\mathcal{L}M' + M' + \mathcal{L}M'\} \\ &= (2\mathcal{L} + 1)M',\end{aligned}$$

where the inequality uses the coordinate-based Lipschitz continuity of L^* (and thus also L_{indiv} and R_{indiv}) to obtain the two different $\mathcal{L}M'$ terms, whereas the M' term comes from the bound from Lemma 17.

At this point, when the bad events of Lemma 13 (this bad event depends on η), Lemma 16, and Lemma 17 (this bad event depends on (x, y, δ)) do not happen,

$$\begin{aligned}
&|L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta)| \\
&\quad \le \spadesuit + \heartsuit + \clubsuit \\
&\quad \le C_\alpha \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M \sqrt{\frac{2\omega}{n_1}} + C_\alpha \sqrt{\max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M \cdot \frac{1}{\sqrt{n_1}}} + C_\alpha(2\mathcal{L} + 1)M' \\
&\quad \le \underbrace{C_\alpha \left[\frac{1}{\sqrt{n_1}} \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} \big(\sqrt{2\omega} + 1\big) M + (2\mathcal{L} + 1)M' \right]}_{\triangleq \varepsilon_{\omega, n_1, n_2}}.
\end{aligned}$$

We now apply yet another covering argument. For positive integers $i \le \frac{C_\alpha}{C_\alpha - 1} \frac{M}{\varepsilon_{\omega,n_1,n_2}}$, we define

$$\eta_i \triangleq -\frac{1}{C_\alpha - 1} M + i \varepsilon_{\omega, n_1, n_2}.$$

By construction, $\{\eta_1, \eta_2 \dots\}$ forms an $\varepsilon_{\omega,n_1,n_2}$-cover of $[-\frac{1}{C_\alpha - 1} M, M]$, meaning that for any $\eta \in [-\frac{1}{C_\alpha - 1} M, M]$, there exists an integer $i(\eta) \in [1, \frac{C_\alpha}{C_\alpha - 1} \frac{M}{\varepsilon_{\omega,n_1,n_2}}]$ such that $|\eta - \eta_{i(\eta)}| \le \varepsilon_{\omega,n_1,n_2}$. The size of this $\varepsilon_{\omega,n_1,n_2}$-cover is bounded above by $\frac{C_\alpha}{C_\alpha - 1} \frac{M}{\varepsilon_{\omega,n_1,n_2}}$. We have

$$\begin{aligned}
&\sup_{\eta \in [-\frac{1}{C_\alpha - 1} M, M]} |L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta)| \\
&\le \sup_{\eta \in [-\frac{1}{C_\alpha - 1} M, M]} \big| L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - L^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}, \mathcal{D}_1 \mid \mathcal{D}_2) \\
&\qquad\qquad + L^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}) + R^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}) - R^{\text{split}}_{\text{DRO}}(\theta, \eta) \big| \\
&\le \sup_{\eta \in [-\frac{1}{C_\alpha - 1} M, M]} \Big\{ \big| L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - L^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}, \mathcal{D}_1 \mid \mathcal{D}_2) \big| \\
&\qquad\qquad + \big| L^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}) \big| + \big| R^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}) - R^{\text{split}}_{\text{DRO}}(\theta, \eta) \big| \Big\} \\
&\le \sup_{i \in [1, \frac{C_\alpha}{C_\alpha - 1} \frac{M}{\varepsilon_{\omega,n_1,n_2}}]} \big\{ (1 + C_\alpha) \varepsilon_{\omega,n_1,n_2} + \big| L^{\text{split}}_{\text{DRO}}(\theta, \eta_i, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta_{i(\eta)}) \big| + (1 + C_\alpha) \varepsilon_{\omega,n_1,n_2} \big\},
\end{aligned}$$

where the first and third terms inside the supremum objective have been bounded in the last step using the fact that $\eta \mapsto L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ and $\eta \mapsto R^{\text{split}}_{\text{DRO}}(\theta, \eta)$ are each $(1 + C_\alpha)$-Lipschitz. Meanwhile, the second term in the objective is upper-bounded by $\varepsilon_{\omega,n_1,n_2}$ when none of the bad events happen, where for the bad event of Lemma 13 (the probabilistic bound for $\spadesuit$) we now have to union bound over it not happening across all the points

$\eta_1, \eta_2, \ldots$ in the $\varepsilon_{\omega,n_1,n_2}$-cover. We thus conclude that with probability at least

$$
\begin{aligned}
&1 - \underbrace{\frac{C_\alpha}{C_\alpha - 1}\frac{M}{\varepsilon_{\omega,n_1,n_2}}}_{\substack{\text{upper bound on size of} \\ \varepsilon_{\omega,n_1,n_2}\text{-cover}}} \cdot \underbrace{2e^{-\omega}}_{\text{from Lemma 13}} - \underbrace{me^{-\frac{n_2\zeta}{8}}}_{\text{from Lemma 16}} - \mathbb{N}(M', \mathcal{X}) \underbrace{2e^{-\omega}}_{\text{from Lemma 17}} \\
&\geq 1 - 2\Big[\frac{C_\alpha}{C_\alpha - 1}\frac{M}{\varepsilon_{\omega,n_1,n_2}} + \mathbb{N}(M', \mathcal{X})\Big]e^{-\omega} - me^{-\frac{n_2\zeta}{8}} \\
&= 1 - 2\left[\frac{C_\alpha}{C_\alpha - 1} \cdot \frac{M}{C_\alpha\big[\frac{1}{\sqrt{n_1}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}\big(\sqrt{2\omega} + 1\big)M + (2\mathcal{L} + 1)M'\big]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n_2\zeta}{8}} \\
&= 1 - 2\left[\frac{M}{(C_\alpha - 1)\big[\frac{1}{\sqrt{n_1}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}\big(\sqrt{2\omega} + 1\big)M + (2\mathcal{L} + 1)M'\big]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n_2\zeta}{8}} \\
&\geq 1 - 2\left[\frac{M}{(C_\alpha - 1)\big[\sqrt{\frac{2\omega}{n_1}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}M + (2\mathcal{L} + 1)M'\big]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n_2\zeta}{8}},
\end{aligned}
$$

we have

$$
\begin{aligned}
&\sup_{\eta \in [-\frac{1}{C_\alpha - 1}M, M]} |L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta)| \\
&\quad \leq 2(1 + C_\alpha)\varepsilon_{\omega,n_1,n_2} + \varepsilon_{\omega,n_1,n_2} \\
&\quad = (3 + 2C_\alpha)\varepsilon_{\omega,n_1,n_2} \\
&\quad = (3 + 2C_\alpha)C_\alpha\left[\frac{1}{\sqrt{n_1}}\max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\}\big(\sqrt{2\omega} + 1\big)M + (2\mathcal{L} + 1)M'\right] \\
&\quad \leq 10C_\alpha^2\left[\frac{1}{\sqrt{n_1}}\max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\}\big(\sqrt{2\omega} + 1\big)M + (2\mathcal{L} + 1)M'\right].
\end{aligned}
$$

At this point, using Lemma 12, we have

$$
\begin{aligned}
&\left|\inf_{\eta \in \mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta \in \mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta, \eta)\right| \\
&\quad = \left|\inf_{\eta \in [-\frac{1}{C_\alpha - 1}M, M]} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \inf_{\eta \in [-\frac{1}{C_\alpha - 1}M, M]} R^{\text{split}}_{\text{DRO}}(\theta, \eta)\right| \\
&\quad \leq \sup_{\eta \in [-\frac{1}{C_\alpha - 1}M, M]} |L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - R^{\text{split}}_{\text{DRO}}(\theta, \eta)| \\
&\quad \leq 10C_\alpha^2\left[\frac{1}{\sqrt{n_1}}\max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\}\big(\sqrt{2\omega} + 1\big)M + (2\mathcal{L} + 1)M'\right],
\end{aligned}
$$

where the first inequality is a standard result that holds when optimizing over bounded functions. This finishes the proof of Proposition 10. ■

C.1 Proof of Lemma 11

Under Assumption A4, when $\phi_{\text{transform}}$ is the identity function,

$$\begin{aligned} L_{\text{indiv}}(\theta; x, y, \delta) &= L^*\big((x, y, \delta), \mathcal{A}^*\big((x, y, \delta), \{(X_j, Y_j, \Delta_j) : j \in \mathcal{D}_2\}\big); \theta\big) \\ &\le M_{\text{indiv}} + |\mathcal{D}_2| M_{\text{couple-max}} \\ &= M_{\text{indiv}} + n_2 M_{\text{couple-max}}. \end{aligned}$$

Assumption A4 trivially also implies that $L_{\text{indiv}}(\theta; x, y, \delta) \ge 0$. Hence, $L_{\text{indiv}}(\theta; x, y, \delta) \in [0, M_{\text{indiv}} + n_2 M_{\text{couple-max}}]$. By a similar argument, $R_{\text{indiv}}(\theta; x, y, \delta) \in [0, M_{\text{indiv}} + n_2 M_{\text{couple-max}}]$.

If instead $\phi_{\text{transform}}(s) = \log(1 + s)$, then

$$\begin{aligned} L_{\text{indiv}}(\theta; x, y, \delta) &= L^*\big((x, y, \delta), \mathcal{A}^*\big((x, y, \delta), \{(X_j, Y_j, \Delta_j) : j \in \mathcal{D}_2\}\big); \theta\big) \\ &\le M_{\text{indiv}} + \log(1 + |\mathcal{D}_2| M_{\text{couple-max}}) \\ &= M_{\text{indiv}} + \log(1 + n_2 M_{\text{couple-max}}). \end{aligned}$$

Again, we have $L_{\text{indiv}}(\theta; x, y, \delta) \ge 0$, so $L_{\text{indiv}}(\theta; x, y, \delta) \in [0, M_{\text{indiv}} + \log(1 + n_2 M_{\text{couple-max}})]$. Similarly, $R_{\text{indiv}}(\theta; x, y, \delta) \in [0, M_{\text{indiv}} + \log(1 + n_2 M_{\text{couple-max}})]$. ■

C.2 Proof of Lemma 12

Using Proposition 4, we have

$$L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) = C_\alpha \sqrt{\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+^2} + \eta.$$

Since $L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) \in [0, M]$ (from Lemma 11), this means that when $\eta \ge M$, we have $[L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+ = 0$ for all $i \in \mathcal{D}_1$ in which case $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) = \eta$.

Meanwhile,

$$\begin{aligned} L^{\text{split}}_{\text{DRO}}\Big(\theta, -\frac{1}{C_\alpha - 1} M, \mathcal{D}_1 \mid \mathcal{D}_2\Big) &= C_\alpha \sqrt{\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) + \frac{1}{C_\alpha - 1} M]_+^2} - \frac{1}{C_\alpha - 1} M \\ &\ge C_\alpha \sqrt{\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} [0 + \frac{1}{C_\alpha - 1} M]_+^2} - \frac{1}{C_\alpha - 1} M \\ &= \frac{C_\alpha}{C_\alpha - 1} M - \frac{1}{C_\alpha - 1} M \\ &= M. \end{aligned}$$

Since $\eta \mapsto L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ is convex, and $L^{\text{split}}_{\text{DRO}}(\theta, -\frac{1}{C_\alpha - 1} M, \mathcal{D}_1 \mid \mathcal{D}_2) = M$ and $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) = \eta$ for all $\eta \ge M$, then it must be that

$$\inf_{\eta \in \mathbb{R}} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) = \inf_{\eta \in [-\frac{1}{C_\alpha - 1} M, M]} L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2).$$

Using the same reasoning,

$$\inf_{\eta \in \mathbb{R}} R^{\text{split}}_{\text{DRO}}(\theta, \eta) = \inf_{\eta \in [-\frac{1}{C_\alpha - 1} M, M]} R^{\text{split}}_{\text{DRO}}(\theta, \eta).$$ ■

C.3 Proof of Lemma 13

We define

$$\Xi_i \triangleq [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+ \quad \text{for } i \in \mathcal{D}_1.$$

As a consequence of Lemma 12, it suffices to only consider $\eta \in [-\frac{1}{C_\alpha - 1} M, M]$. Hence,

$$\begin{aligned}
\Xi_i &= [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+ \\
&\le |L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta| \\
&\le |L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i)| + |\eta| \\
&\le M + |\eta| \\
&\le M + \max\Big\{\frac{1}{C_\alpha - 1} M, M\Big\} \\
&= \max\Big\{\frac{1}{C_\alpha - 1} M + M, 2M\Big\} \\
&= \max\Big\{\frac{C_\alpha}{C_\alpha - 1} M, 2M\Big\}.
\end{aligned}$$

Meanwhile, trivially $\Xi_i \ge 0$, so

$$\Xi_i \in \Big[0, \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M\Big]. \tag{C.1}$$

Next, by Lemma 7 of Duchi and Namkoong (2021), $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ is $\frac{C_\alpha}{\sqrt{n_1}}$ Lipschitz with respect to the vector $(\Xi_i)_{i \in \mathcal{D}_1}$ in Euclidean norm. Then by Lemma 6 of Duchi and Namkoong (2021), for any $\widetilde{\omega} > 0$,

$$\begin{aligned}
&\mathbb{P}\big(|L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \mathbb{E}[L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)]| \ge \widetilde{\omega}\big) \\
&\quad \le 2 \exp\bigg(- \frac{\widetilde{\omega}^2 n_1}{2 C_\alpha^2 \big(\max\{2, \frac{C_\alpha}{C_\alpha - 1}\} M\big)^2} \bigg).
\end{aligned}$$

We do a change of variables. Let $\omega > 0$. Plugging in

$$\widetilde{\omega} = C_\alpha \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M \sqrt{\frac{2\omega}{n_1}},$$

we get that

$$\begin{aligned}
&\mathbb{P}\Bigg(\overbrace{|L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \mathbb{E}[L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)]|}^{\spadesuit} \ge C_\alpha \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M \sqrt{\frac{2\omega}{n_1}} \Bigg) \\
&\quad \le 2e^{-\omega}.
\end{aligned}$$

■

C.4 Proof of Lemma 14

Recall from bound (C.1) that for $i \in \mathcal{D}_1$, the variable $\Xi_i = [L_{\text{indiv}}(\theta; X_i, Y_i, \Delta_i) - \eta]_+$ satisfies

$$\Xi_i \in \Big[0, \max\Big\{2, \frac{C_\alpha}{C_\alpha - 1}\Big\} M\Big].$$

Thus, we trivially have

$$\mathbb{E}[|\Xi_i|^4] \leq \left(\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\} M \right)^2 \mathbb{E}[|\Xi_i|^2],$$

which means that applying Lemma 8 of Duchi and Namkoong (2021), we get

$$\mathbb{E}\left[\sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}|\Xi_i|^2}\right] \geq \sqrt{\mathbb{E}[|\Xi_i|^2]} - \sqrt{\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\} M} \cdot \frac{1}{\sqrt{n_1}}.$$

This means that

$$\begin{aligned}
&L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) - \mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)] \\
&\quad = C_\alpha \left(\sqrt{\mathbb{E}[|\Xi_i|^2]} - \mathbb{E}\left[\sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}|\Xi_i|^2}\right] \right) \\
&\quad \leq C_\alpha \sqrt{\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\} M} \cdot \frac{1}{\sqrt{n_1}}.
\end{aligned}$$

Separately, by Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)] - L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) \\
&\quad = C_\alpha \left(\mathbb{E}\left[\sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}|\Xi_i|^2}\right] - \sqrt{\mathbb{E}[|\Xi_i|^2]} \right) \\
&\quad \leq C_\alpha \left(\sqrt{\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}|\Xi_i|^2\right]} - \sqrt{\mathbb{E}[|\Xi_i|^2]} \right) \\
&\quad = C_\alpha \left(\sqrt{\mathbb{E}[|\Xi_i|^2]} - \sqrt{\mathbb{E}[|\Xi_i|^2]} \right) \\
&\quad = 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\heartsuit &= \left| \mathbb{E}[L_{\mathrm{DRO}}^{\mathrm{split}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)] - L_{\mathrm{DRO}}^{\mathrm{split},*}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2) \right| \\
&\leq C_\alpha \sqrt{\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\} M} \cdot \frac{1}{\sqrt{n_1}}.
\end{aligned}$$

■

C.5 Proof of Lemma 15

First off, under Assumptions A1, A2, and A4, we have

$$\begin{aligned}
&\sup_{(x,y,\delta)\in\mathcal{Z}} |L_{\mathrm{indiv}}(\theta; x, y, \delta) - R_{\mathrm{indiv}}(\theta; x, y, \delta)| \\
&\quad = \max_{(x,y,\delta)\in\mathcal{Z}} |L_{\mathrm{indiv}}(\theta; x, y, \delta) - R_{\mathrm{indiv}}(\theta; x, y, \delta)| \\
&\quad = \max_{y\in\{t_1,\ldots,t_m\}} \max_{\delta\in\{0,1\}} \max_{x\in\mathcal{X}} |L_{\mathrm{indiv}}(\theta; x, y, \delta) - R_{\mathrm{indiv}}(\theta; x, y, \delta)|.
\end{aligned}$$

The reason the supremum is attained (so that it is equal to the max) is because not only do we discretize time to a finite grid (Assumption A2) so that maximizing over the m values of y and the 2 values of δ does not present any issues in the supremum being attained, we further assume that $\mathcal{X}$ is compact (Assumption A1) and $L^*((x,y,\delta),\mathcal{C};\theta)$ is continuous in the coordinate x with respect to Euclidean norm (Assumption A4(c)), which ensures that the supremum over $x \in \mathcal{X}$ is equal to the max over $x \in \mathcal{X}$.

Next, using the fact that for any $a, b, c, d \in \mathbb{R}$, we have $\max\{a+b, c+d\} \leq \max\{a,c\} + \max\{b,d\}$ (which implies that $\max\{a+b,0\} \leq \max\{a,0\} + \max\{b,0\} = [a]_+ + [b]_+$),

$$\begin{aligned}
&\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big] \\
&\quad = \mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta) + R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big] \\
&\quad \leq \mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[\big([L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+ + [R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+\big)^2\big].
\end{aligned}$$

Taking the square root of both sides, we get

$$\begin{aligned}
&\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]} \\
&\quad \leq \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[\big([L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+ + [R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+\big)^2\big]}.
\end{aligned} \tag{C.2}$$

Applying Minkowski's inequality,

$$\begin{aligned}
&\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[\big([L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+ + [R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+\big)^2\big]} \\
&\quad \leq \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+^2\big]} \\
&\qquad + \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]}.
\end{aligned} \tag{C.3}$$

Next, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+^2\big]} \\
&\quad \leq \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[(L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta))^2\big]} \\
&\quad \leq \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[\max_{(x,y,\delta)\in\mathcal{Z}}(L_{\text{indiv}}(\theta;x,y,\delta) - R_{\text{indiv}}(\theta;x,y,\delta))^2\big]} \\
&\quad = \sqrt{\max_{(x,y,\delta)\in\mathcal{Z}}(L_{\text{indiv}}(\theta;x,y,\delta) - R_{\text{indiv}}(\theta;x,y,\delta))^2} \\
&\quad = \max_{(x,y,\delta)\in\mathcal{Z}} |L_{\text{indiv}}(\theta;x,y,\delta) - R_{\text{indiv}}(\theta;x,y,\delta)|.
\end{aligned} \tag{C.4}$$

Combining inequalities (C.2), (C.3), and (C.4), we obtain

$$\begin{aligned}
&\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]} \\
&\quad \leq \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta) - R_{\text{indiv}}(\theta;X,Y,\Delta)]_+^2\big]} \\
&\qquad + \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]} \\
&\quad \leq \max_{(x,y,\delta)\in\mathcal{Z}} |L_{\text{indiv}}(\theta;x,y,\delta) - R_{\text{indiv}}(\theta;x,y,\delta)| + \sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta) - \eta]_+^2\big]},
\end{aligned}$$

i.e.,

$$\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}-\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}\\ \leq \max_{(x,y,\delta)\in\mathcal{Z}}|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|. \tag{C.5}$$

Repeating the same proof ideas but with the roles of L_{indiv} and R_{indiv} swapped, we would instead obtain the bound

$$\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}-\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}\\ \leq \max_{(x,y,\delta)\in\mathcal{Z}}|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|. \tag{C.6}$$

Thus, inequalities (C.5) and (C.6) together imply that

$$\left|\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}-\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}\right|\\ \leq \max_{(x,y,\delta)\in\mathcal{Z}}|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|.$$

Therefore,

$$\begin{aligned}&|L_{\text{DRO}}^{\text{split},*}(\theta,\eta)-R_{\text{DRO}}^{\text{split}}(\theta,\eta)|\\ &\quad=C_\alpha\left|\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[L_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}-\sqrt{\mathbb{E}_{(X,Y,\Delta)\sim\mathbb{P}}\big[[R_{\text{indiv}}(\theta;X,Y,\Delta)-\eta]_+^2\big]}\right|\\ &\quad\leq C_\alpha\max_{(x,y,\delta)\in\mathcal{Z}}|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|.\end{aligned}$$ ■

C.6 Proof of Lemma 16

For each time index $\ell\in[m]$, by the multiplicative Chernoff bound and Assumption A2,

$$\mathbb{P}\bigg(\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}\leq\frac{1}{2}n_2\mathbb{P}(Y=t_\ell)\bigg)\leq e^{-\frac{\mathbb{E}[\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}]}{8}}=e^{-\frac{n_2\mathbb{P}(Y=t_\ell)}{8}}\leq e^{-\frac{n_2\zeta}{8}}.$$

Note that $\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}\leq\frac{1}{2}n_2\zeta$ implies that $\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}\leq\frac{1}{2}n_2\mathbb{P}(Y=t_\ell)$ since $\mathbb{P}(Y=t_\ell)\geq\zeta$ by Assumption A2. This means that

$$\mathbb{P}\bigg(\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}\leq\frac{1}{2}n_2\zeta\bigg)\leq\mathbb{P}\bigg(\sum_{i\in\mathcal{D}_2}\mathbf{1}\{Y_i=t_\ell\}\leq\frac{1}{2}n_2\mathbb{P}(Y=t_\ell)\bigg)\leq e^{-\frac{n_2\zeta}{8}}.$$

Union-bounding over all m time indices yields the claim. ■

C.7 Proof of Lemma 17

Let $(x,y,\delta)\in\mathcal{Z}$. If $\delta=0$, then we obtain the trivial equality

$$|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|=0,$$

since by Assumption A4, $L_{\text{indiv}}(\theta; x, y, 0)$ has no coupling terms, in which case it is exactly equal to $R_{\text{indiv}}(\theta; x, y, 0)$.

For the remainder of this lemma's proof, we assume that $\delta \neq 0$. In this case, the adjacency set of (x, y, δ) could be nonempty. First, we introduce the shorthand notation where $\mathcal{N}_{\mathcal{D}_2} \subseteq [n]$ denotes the indices of training data in $\mathcal{D}_2$ that are considered adjacent to data point (x, y, δ), and $\mathcal{N}_{\text{fresh}} \in [n_2]$ is analogously defined but for the fresh sample of n_2 data points (used in the definition of R_{indiv}). Formally,

$$\mathcal{N}_{\mathcal{D}_2} \triangleq \Big\{ i \in \mathcal{D}_2 : (X_i, Y_i, \Delta_i) \in \mathcal{A}^*\big((x, y, \delta), \{(X_j, Y_j, \Delta_j) : j \in \mathcal{D}_2\}\big) \Big\},$$
$$\mathcal{N}_{\text{fresh}} \triangleq \Big\{ i \in [n_2] : (X'_i, Y'_i, \Delta'_i) \in \mathcal{A}^*\big((x, y, \delta), \{(X'_j, Y'_j, \Delta'_j) : j \in [n_2]\}\big) \Big\}.$$

When the event $\mathcal{E}_{\text{bad time}}$ in Lemma 16 does not happen, we are guaranteed that $|\mathcal{N}_{\mathcal{D}_2}| > \frac{n_2\zeta}{2} > 0$ (by the definition of DeepHit's adjacency function, when $\delta = 1$, $\mathcal{N}_{\mathcal{D}_2}$ would at least contain all points in $\mathcal{D}_2$ with the same time index as y, for which there are more than $\frac{n_2\zeta}{2}$ such data points).

Then when $\phi_{\text{transform}}$ is the identity function,

$$\begin{aligned}
&|L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| \\
&\quad = \Bigg| \phi_{\text{indiv}}((x, y, \delta); \theta) + \sum_{j \in \mathcal{N}^*_{\mathcal{D}_2}} \phi_{\text{couple}}((x, y, \delta), (X_j, Y_j, \Delta_j); \theta) \\
&\qquad - \mathbb{E}_{\{(X'_i, Y'_i, \Delta'_i)\}_{i=1}^{n_2}} \Bigg[\phi_{\text{indiv}}((x, y, \delta); \theta) + \sum_{j \in \mathcal{N}^*_{\text{fresh}}} \phi_{\text{couple}}((x, y, \delta), (X'_j, Y'_j, \Delta'_j); \theta) \Bigg] \Bigg| \\
&\quad = \Bigg| \sum_{j \in \mathcal{N}^*_{\mathcal{D}_2}} \phi_{\text{couple}}((x, y, \delta), (X_j, Y_j, \Delta_j); \theta) \\
&\qquad - \mathbb{E}_{\{(X'_i, Y'_i, \Delta'_i)\}_{i=1}^{n_2}} \Bigg[\sum_{j \in \mathcal{N}^*_{\text{fresh}}} \phi_{\text{couple}}((x, y, \delta), (X'_j, Y'_j, \Delta'_j); \theta) \Bigg] \Bigg|.
\end{aligned}$$

The key observation is that by construction, $\sum_{j \in \mathcal{N}^*_{\mathcal{D}_2}} \phi_{\text{couple}}((x, y, \delta), (X_j, Y_j, \Delta_j); \theta)$ has the same distribution as $\sum_{j \in \mathcal{N}^*_{\text{fresh}}} \phi_{\text{couple}}((x, y, \delta), (X'_j, Y'_j, \Delta'_j); \theta)$, so

$$\begin{aligned}
&\mathbb{E}_{\{(X_i, Y_i, \Delta_i)\}_{i \in \mathcal{D}_2}} \Bigg[\sum_{j \in \mathcal{N}^*_{\mathcal{D}_2}} \phi_{\text{couple}}((x, y, \delta), (X_j, Y_j, \Delta_j); \theta) \Bigg] \\
&\qquad = \mathbb{E}_{\{(X'_i, Y'_i, \Delta'_i)\}_{i=1}^{n_2}} \Bigg[\sum_{j \in \mathcal{N}^*_{\text{fresh}}} \phi_{\text{couple}}((x, y, \delta), (X'_j, Y'_j, \Delta'_j); \theta) \Bigg].
\end{aligned}$$

Hence, denoting

$$\Phi \triangleq \sum_{j \in \mathcal{N}^*_{\mathcal{D}_2}} \phi_{\text{couple}}((x, y, \delta), (X_j, Y_j, \Delta_j); \theta), \tag{C.7}$$

we have

$$|L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| = |\Phi - \mathbb{E}[\Phi]|.$$

When event $\mathcal{E}_{\text{bad time}}$ does not happen, we know that $|\mathcal{N}_{\mathcal{D}_2}| \geq \lceil \frac{n_2\zeta}{2} \rceil$. This means that Φ is a nonempty sum of i.i.d. nonnegative random variables each bounded within $[M_{\text{couple-min}}, M_{\text{couple-max}}]$. Then by Hoeffding's inequality,

for any $\widetilde{\omega} > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(|\Phi - \mathbb{E}[\Phi]| \geq \widetilde{\omega}|\mathcal{N}_{\mathcal{D}_2}| \;\Big|\; |\mathcal{N}_{\mathcal{D}_2}| \geq \Big\lceil \frac{n_2\zeta}{2} \Big\rceil\Big) \\
&= \frac{\sum_{\ell=\lceil \frac{n_2\zeta}{2} \rceil}^{n_2} \mathbb{P}\big(|\Phi - \mathbb{E}[\Phi]| \leq \widetilde{\omega}|\mathcal{N}_{\mathcal{D}_2}| \;\big|\; |\mathcal{N}_{\mathcal{D}_2}| = \ell\big)\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| = \ell)}{\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| \geq \lceil \frac{n_2\zeta}{2} \rceil)} \\
&\leq \frac{\sum_{\ell=\lceil \frac{n_2\zeta}{2} \rceil}^{n_2} 2\exp\Big(-\frac{2(\widetilde{\omega}\ell)^2}{\ell(M_{\text{couple-max}} - M_{\text{couple-min}})^2}\Big)\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| = \ell)}{\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| \geq \lceil \frac{n_2\zeta}{2} \rceil)} \\
&= \frac{\sum_{\ell=\lceil \frac{n_2\zeta}{2} \rceil}^{n_2} 2\exp\Big(-\frac{2\widetilde{\omega}^2\ell}{(M_{\text{couple-max}} - M_{\text{couple-min}})^2}\Big)\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| = \ell)}{\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| \geq \lceil \frac{n_2\zeta}{2} \rceil)} \\
&\leq \frac{\sum_{\ell=\lceil \frac{n_2\zeta}{2} \rceil}^{n_2} 2\exp\Big(-\frac{2\widetilde{\omega}^2(\frac{n_2\zeta}{2})}{(M_{\text{couple-max}} - M_{\text{couple-min}})^2}\Big)\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| = \ell)}{\mathbb{P}(|\mathcal{N}_{\mathcal{D}_2}| \geq \lceil \frac{n_2\zeta}{2} \rceil)} \\
&= 2\exp\Big(-\frac{\widetilde{\omega}^2 n_2\zeta}{(M_{\text{couple-max}} - M_{\text{couple-min}})^2}\Big).
\end{aligned}$$

Now we do a change of variables. Let $\omega > 0$, and set

$$\widetilde{\omega} = (M_{\text{couple-max}} - M_{\text{couple-min}})\sqrt{\frac{\omega}{\zeta n_2}}.$$

Then we have

$$\begin{aligned}
&\mathbb{P}\Big(|\Phi - \mathbb{E}[\Phi]| \geq (M_{\text{couple-max}} - M_{\text{couple-min}})\sqrt{\frac{\omega}{\zeta n_2}}|\mathcal{N}_{\mathcal{D}_2}| \;\Big|\; |\mathcal{N}_{\mathcal{D}_2}| \geq \Big\lceil \frac{n_2\zeta}{2} \Big\rceil\Big) \\
&\quad \leq 2e^{-\omega}.
\end{aligned}$$

In summary, when $\mathcal{E}_{\text{bad time}}$ does not happen, with probability at least $1 - 2e^{-\omega}$, we have

$$\begin{aligned}
|L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| &= |\Phi - \mathbb{E}[\Phi]| \\
&\leq (M_{\text{couple-max}} - M_{\text{couple-min}})\sqrt{\frac{\omega}{\zeta n_2}}|\mathcal{N}_{\mathcal{D}_2}| \\
&\leq (M_{\text{couple-max}} - M_{\text{couple-min}})\sqrt{\frac{\omega}{\zeta n_2}} \cdot n_2 \\
&= (M_{\text{couple-max}} - M_{\text{couple-min}})\sqrt{\frac{\omega n_2}{\zeta}}.
\end{aligned}$$

Now let's consider when instead $\phi_{\text{transform}}(s) = \log(1+s)$, and as a reminder we assume $\delta \neq 0$. Then

$$
\begin{aligned}
&|L_{\text{indiv}}(\theta; x, y, \delta) - R_{\text{indiv}}(\theta; x, y, \delta)| \\
&= \Bigg| \phi_{\text{indiv}}((x,y,\delta);\theta) + \log\left(1 + \sum_{j \in \mathcal{N}_{\mathcal{D}_2}} \phi_{\text{couple}}((x,y,\delta),(X_j,Y_j,\Delta_j);\theta)\right) \\
&\quad - \mathbb{E}_{\{(X_i',Y_i',\Delta_i')\}_{i=1}^{n_2}}\left[\phi_{\text{indiv}}((x,y,\delta);\theta) + \delta \log\left(1 + \sum_{j \in \mathcal{N}_{\text{fresh}}} \phi_{\text{couple}}((x,y,\delta),(X_j',Y_j',\Delta_j');\theta)\right)\right]\Bigg| \\
&= \Bigg| \log\left(1 + \sum_{j \in \mathcal{N}_{\mathcal{D}_2}} \phi_{\text{couple}}((x,y,\delta),(X_j,Y_j,\Delta_j);\theta)\right) \\
&\quad - \mathbb{E}_{\{(X_i',Y_i',\Delta_i')\}_{i=1}^{n_2}}\left[\log\left(1 + \sum_{j \in \mathcal{N}_{\text{fresh}}} \phi_{\text{couple}}((x,y,\delta),(X_j',Y_j',\Delta_j');\theta)\right)\right]\Bigg|.
\end{aligned}
$$

By a similar argument as we used for proving the case where $\phi_{\text{transform}}$ is the identity function, the key observation is that

$$
\begin{aligned}
&\mathbb{E}_{\{(X_i,Y_i,\Delta_i)\}_{i\in\mathcal{D}_2}}\left[\log\left(1 + \sum_{j \in \mathcal{N}_{\mathcal{D}_2}} \phi_{\text{couple}}((x,y,\delta),(X_j,Y_j,\Delta_j);\theta)\right)\right] \\
&\quad = \mathbb{E}_{\{(X_i',Y_i',\Delta_i')\}_{i=1}^{n_2}}\left[\log\left(1 + \sum_{j \in \mathcal{N}_{\text{fresh}}} \phi_{\text{couple}}((x,y,\delta),(X_j',Y_j',\Delta_j');\theta)\right)\right].
\end{aligned}
$$

We now define $\Gamma_j \triangleq \phi_{\text{couple}}((x,y,\delta),(X_j,Y_j,\Delta_j);\theta)$ for each $j \in \mathcal{N}_{\mathcal{D}_2}$. Again, when event $\mathcal{E}_{\text{bad time}}$ does not happen, we are guaranteed that $|\mathcal{N}_{\mathcal{D}_2}| \geq \frac{n_2\zeta}{2}$, i.e., $\mathcal{N}_{\mathcal{D}_2}$ is nonempty. Note that the map $(\Gamma_j)_{j\in\mathcal{N}_{\mathcal{D}_2}} \mapsto \log(1 + \sum_{j\in\mathcal{N}_{\mathcal{D}_2}} \Gamma_j)$ is concave. We now show that this map is Lipschitz continuous with respect to the Euclidean norm by showing what a valid Lipschitz constant is for the map. Note that for $i \in \mathcal{N}_{\mathcal{D}_2}$,

$$
\frac{\partial \log(1 + \sum_{j\in\mathcal{N}_{\mathcal{D}_2}} \Gamma_j)}{\partial \Gamma_i} = \frac{1}{1 + \sum_{j\in\mathcal{N}_{\mathcal{D}_2}} \Gamma_j}.
$$

Then

$$\begin{aligned}\Big\|\nabla \log\Big(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j\Big)\Big\|_2 &= \sqrt{\sum_{i\in\mathcal{N}_{\mathcal{D}_2}}\Big(\frac{\partial \log(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j)}{\partial\Gamma_i}\Big)^2}\\ &= \sqrt{\frac{|\mathcal{N}_{\mathcal{D}_2}|}{(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j)^2}}\\ &\le \sqrt{\frac{|\mathcal{N}_{\mathcal{D}_2}|}{(\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j)^2}}\\ &\le \sqrt{\frac{|\mathcal{N}_{\mathcal{D}_2}|}{(\frac{n_2\zeta}{2}M_{\text{couple-min}})^2}}\\ &\le \sqrt{\frac{n_2}{(\frac{n_2\zeta}{2}M_{\text{couple-min}})^2}}\\ &= \frac{2}{\zeta M_{\text{couple-min}}}\cdot\frac{1}{\sqrt{n_2}}.\end{aligned}$$

In other words, the map $(\Gamma_j)_{j\in\mathcal{N}_{\mathcal{D}_2}}\mapsto \log(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j)$ has Lipschitz constant $\frac{2}{\zeta M_{\text{couple-min}}}\cdot\frac{1}{\sqrt{n_2}}$ when event $\mathcal{E}_{\text{bad time}}$ does not happen. Then applying Lemma 6 of Duchi and Namkoong (2021), for any $\widetilde{\omega}>0$,

$$\begin{aligned}&\mathbb{P}\Big(\Big|\log\Big(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j\Big)-\mathbb{E}\Big[\log\Big(1+\sum_{j\in\mathcal{N}_{\mathcal{D}_2}}\Gamma_j\Big)\Big]\Big|\ge\widetilde{\omega}\,\Big|\,|\mathcal{N}_{\mathcal{D}_2}|\ge\frac{n_2\zeta}{2}\Big)\\ &\le 2\exp\Big(-\frac{\widetilde{\omega}^2}{2(\frac{2}{\zeta M_{\text{couple-min}}}\cdot\frac{1}{\sqrt{n_2}})^2(M_{\text{couple-max}}-M_{\text{couple-min}})^2}\Big)\\ &= 2\exp\Big(-\frac{\widetilde{\omega}^2\zeta^2M_{\text{couple-min}}^2n_2}{8(M_{\text{couple-max}}-M_{\text{couple-min}})^2}\Big).\end{aligned}$$

Let $\omega>0$ and set $\widetilde{\omega}=\frac{(M_{\text{couple-max}}-M_{\text{couple-min}})}{\zeta M_{\text{couple-min}}}\sqrt{\frac{8\omega}{n_2}}$. Then

$$\begin{aligned}&\mathbb{P}\Bigg(\overbrace{\Big|\log\Big(1+\sum_{j\in\mathcal{N}^*_{\mathcal{D}_2}}\Gamma_j\Big)-\mathbb{E}\Big[\log\Big(1+\sum_{j\in\mathcal{N}^*_{\mathcal{D}_2}}\Gamma_j\Big)\Big]\Big|}^{=|L_{\text{indiv}}(\theta;x,y,\delta)-R_{\text{indiv}}(\theta;x,y,\delta)|}\ge\frac{(M_{\text{couple-max}}-M_{\text{couple-min}})}{\zeta M_{\text{couple-min}}}\sqrt{\frac{8\omega}{n_2}}\Bigg)\\ &\le 2e^{-\omega}.\end{aligned}$$

■

Appendix D. Proof of Corollary 6

The proof of this corollary consists of two main parts. First, we check that the Assumptions A1–A4 needed by Theorem 5 hold. Then we apply Theorem 5, where we impose constraints on n and d so that we can simplify the probability bound in equation (24).

Verifying Assumptions A1–A4 Assumption A1 clearly holds since $\mathcal{X}$ is the unit ball in $\mathbb{R}^d$, which is compact. There is no need to check Assumption A2 in that we are directly assuming it. Similarly, Assumption A3 also trivially holds (for discrete time, the adjacency function for the Cox model is the same as for DeepHit).

We proceed to verify Assumption A4. For the Cox model where $f(x;\theta) = \theta^\top x$, we have

$$\begin{aligned}
L^*((x,y,\delta),\mathcal{C};\theta) &= -\delta\left[\theta^\top x - \log\left(\exp(\theta^\top x) + \sum_{(x',y',\delta')\in\mathcal{C}} \exp(\theta^\top x')\right)\right] \\
&= -\delta\left[\log\exp(\theta^\top x) - \log\left(\exp(\theta^\top x) + \sum_{(x',y',\delta')\in\mathcal{C}} \exp(\theta^\top x')\right)\right] \\
&= -\delta\log\left(\frac{\exp(\theta^\top x)}{\exp(\theta^\top x) + \sum_{(x',y',\delta')\in\mathcal{C}}\exp(\theta^\top x')}\right) \\
&= \delta\log\left(\frac{\exp(\theta^\top x) + \sum_{(x',y',\delta')\in\mathcal{C}}\exp(\theta^\top x')}{\exp(\theta^\top x)}\right) \\
&= \delta\log\Big(1 + \sum_{(x',y',\delta')\in\mathcal{C}} \exp\big(\theta^\top(x'-x)\big)\Big),
\end{aligned}$$

which corresponds to Assumption A4 where $\phi_{\text{transform}}(s) = \log(1+s)$, ϕ_{indiv} always outputs 0 (so $M_{\text{indiv}} = 0$), and

$$\phi_{\text{couple}}((x,y,\delta),(x',y',\delta');\theta) = \exp\big(\theta^\top(x'-x)\big).$$

In this case, since $\mathcal{X}$ and Θ are constrained to be within the unit ball, by the Cauchy-Schwarz inequality,

$$|\theta^\top(x'-x)| \le \underbrace{\|\theta\|_2}_{\le 1} \underbrace{\|x'-x\|_2}_{\substack{\le 2 \text{ since a unit ball} \\ \text{has diameter 2}}} \le 2.$$

In particular,

$$\theta^\top(x'-x) \in [-2,2],$$

so the largest $\phi_{\text{couple}}((x,y,\delta),(x',y',\delta');\theta)$ can be is

$$\exp\big(\theta^\top(x'-x)\big) \le \exp(2) \triangleq M_{\text{couple-max}},$$

whereas the smallest is

$$\exp\big(\theta^\top(x'-x)\big) \ge \exp(-2) \triangleq M_{\text{couple-min}}.$$

Meanwhile, to check that $L^*((x,y,\delta),\mathcal{C};\theta)$ satisfies Lipschitz continuity, first note that when $\delta = 0$, the $L^*((x,y,\delta),\mathcal{C};\theta) = 0$, so there is nothing to show. When $\delta \neq 0$, we have

$$L^*((x,y,\delta),\mathcal{C};\theta) = \log\Big(1 + \sum_{(x',y',\delta')\in\mathcal{C}} \exp\big(\theta^\top(x'-x)\big)\Big).$$

Taking the gradient with respect to x, we get

$$\frac{\partial L^*((x,y,\delta),\mathcal{C};\theta)}{\partial x} = -\frac{\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}{1+\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}\theta.$$

Then

$$\begin{aligned}\left\|\frac{\partial L^*((x,y,\delta),\mathcal{C};\theta)}{\partial x}\right\|_2 &= \left\| -\frac{\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}{1+\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}\theta\right\|_2 \\ &= \underbrace{\frac{\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}{1+\sum_{(x',y',\delta')\in\mathcal{C}} \exp\left(\theta^\top(x'-x)\right)}}_{\leq 1 \text{ (this is a probability from a softmax calculation)}} \underbrace{\|\theta\|_2}_{\leq 1} \\ &\leq 1.\end{aligned}$$

Thus, $L^*((x,y,\delta),\mathcal{C};\theta)$ is 1-Lipschitz (i.e., the constant $\mathcal{L}$ in Assumption A4(c) is 1). At this point we have verified that Assumption A4 holds.

Applying Theorem 5 We now apply Theorem 5. In this case, we have

$$M \triangleq \log\left(1+\frac{e^2}{2}n\right) \qquad \text{and} \qquad M' \triangleq \frac{4(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\omega}{n}}. \tag{D.1}$$

By a standard result (see, for instance, Corollary 4.2.13 of Vershynin (2018)), for all $\varepsilon \in (0,1]$,

$$\mathbb{N}(\varepsilon,\mathcal{X}) \leq \left(\frac{3}{\varepsilon}\right)^d. \tag{D.2}$$

Since the Theorem 5's probability bound (24) depends on $\mathbb{N}(M',\mathcal{X})$, we first verify that $M' \leq 1$ so that inequality (D.2) holds. To do this, we make use of the lower branch W_{-1} of the Lambert W function and the standard result that

$$-1-\sqrt{2s}-s < W_{-1}(-e^{-s-1}) \quad \text{for} \quad s>0. \tag{D.3}$$

By assumption,

$$\begin{aligned} n &\geq \left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)e^{\sqrt{2\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)-1}} \\ &= e^{\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)+\sqrt{2\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)-1}}. \end{aligned}$$

Inequality (D.3) (with $s = \log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)-1$) implies that

$$e^{\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)+\sqrt{2\log\left(\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)\right)-1}} > e^{-W_{-1}\left(-\frac{1}{\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)}\right)},$$

so that

$$n \geq e^{-W_{-1}\left(-\frac{1}{\left(\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\right)^2\left(\frac{d+1}{2}\right)}\right)}.$$

This implies that

$$M' = \frac{4(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\omega}{n}} = \frac{4(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\frac{d+1}{2}\log n}{n}} \leq 1$$

as desired. Then using inequality (D.2), the probability bound in equation (24) satisfies the bound

$$\begin{aligned}
&1 - 2\left[\frac{M}{(C_\alpha - 1)\left[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}M + (2\mathcal{L}+1)M'\right]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n\zeta}{16}} \\
&= 1 - 2\left[\frac{\log\left(1+\frac{e^2}{2}n\right)}{(C_\alpha - 1)\left[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}\log\left(1+\frac{e^2}{2}n\right) + \frac{12(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\omega}{n}}\right]} + \mathbb{N}(M', \mathcal{X})\right]e^{-\omega} - me^{-\frac{n\zeta}{16}} \\
&\geq 1 - 2\left[\frac{\log\left(1+\frac{e^2}{2}n\right)}{(C_\alpha - 1)\left[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}\log\left(1+\frac{e^2}{2}n\right) + \frac{12(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\omega}{n}}\right]} + \left(\frac{3}{\frac{4(e^2-e^{-2})}{\zeta e^{-2}}\sqrt{\frac{\omega}{n}}}\right)^d\right]e^{-\omega} \\
&\quad - me^{-\frac{n\zeta}{16}} \\
&= 1 - 2\left[\frac{\log\left(1+\frac{e^2}{2}n\right)}{2(C_\alpha - 1)\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}\log\left(1+\frac{e^2}{2}n\right) + \frac{12(C_\alpha-1)(e^2-e^{-2})}{\zeta e^{-2}}}\sqrt{\frac{n}{\omega}} + \left(\frac{3\zeta e^{-2}}{4(e^2-e^{-2})}\right)^d\left(\frac{n}{\omega}\right)^{d/2}\right]e^{-\omega} \\
&\quad - me^{-\frac{n\zeta}{16}} \\
&= 1 - 2\left[\frac{\log\left(1+\frac{e^2}{2}n\right)}{\Upsilon_1\log\left(1+\frac{e^2}{2}n\right) + \Upsilon_2}\left(\frac{n}{\omega}\right)^{1/2} + \Upsilon_3\left(\frac{n}{\omega}\right)^{d/2}\right]e^{-\omega} - me^{-\frac{n\zeta}{16}},
\end{aligned}$$

where

$$\begin{aligned}
\Upsilon_1 &\triangleq 2(C_\alpha - 1)\max\left\{2, \frac{C_\alpha}{C_\alpha - 1}\right\}, \\
\Upsilon_2 &\triangleq \frac{12(C_\alpha - 1)(e^2 - e^{-2})}{\zeta e^{-2}}, \\
\Upsilon_3 &\triangleq \left(\frac{3\zeta e^{-2}}{4(e^2 - e^{-2})}\right)^d.
\end{aligned}$$

Recall that we have the assumption

$$n \geq 2e^{-\frac{6(e^4-1)}{\zeta\max\{2, \frac{C_\alpha}{C_\alpha-1}\}}-2} = \frac{2e^{-\Upsilon_2/\Upsilon_1}}{e^2}.$$

This implies that

$$n > \frac{2(e^{-\Upsilon_2/\Upsilon_1} - 1)}{e^2},$$

which further implies that

$$\frac{\log\left(1+\frac{e^2}{2}n\right)}{\Upsilon_1\log\left(1+\frac{e^2}{2}n\right)+\Upsilon_2}\leq\frac{1}{\Upsilon_1}.$$

Consequently,

$$\begin{aligned}&1-2\left[\frac{\log\left(1+\frac{e^2}{2}n\right)}{\Upsilon_1\log\left(1+\frac{e^2}{2}n\right)+\Upsilon_2}\left(\frac{n}{\omega}\right)^{1/2}+\Upsilon_3\left(\frac{n}{\omega}\right)^{d/2}\right]e^{-\omega}-me^{-\frac{n\zeta}{16}}\\&\geq 1-2\left[\frac{1}{\Upsilon_1}\left(\frac{n}{\omega}\right)^{1/2}+\Upsilon_3\left(\frac{n}{\omega}\right)^{d/2}\right]e^{-\omega}-me^{-\frac{n\zeta}{16}}.\end{aligned}$$

Next, we use the fact that we set $\omega=\frac{d+1}{2}\log n$. Let W_{-1} to be the lower branch of the Lambert W function. The statement of the corollary also assumes that

$$n\geq e^{\sqrt{2(\log\frac{d+1}{2}-1)}+\log\frac{d+1}{2}}.$$

A standard bound on W_{-1} is that for any $s>0$, we have $-1-\sqrt{2s}-s<W_{-1}(-e^{-s-1})$. Plugging in $s=\log\frac{d+1}{2}-1$ (which is guaranteed to be positive since we assume that $d\geq 5>2e-1$), we get that

$$n\geq e^{\sqrt{2(\log\frac{d+1}{2}-1)}+\log\frac{d+1}{2}}\geq e^{-W_{-1}(-\frac{2}{d+1})}.$$

This in turn implies that $n\geq\omega=\frac{d+1}{2}\log n$. Then since $n\geq\omega$, we have

$$\begin{aligned}&1-2\left[\frac{1}{\Upsilon_1}\left(\frac{n}{\omega}\right)^{1/2}+\Upsilon_3\left(\frac{n}{\omega}\right)^{d/2}\right]e^{-\omega}-me^{-\frac{n\zeta}{16}}\\&\geq 1-2\Big(\frac{1}{\Upsilon_1}+\Upsilon_3\Big)\left(\frac{n}{\omega}\right)^{d/2}e^{-\omega}-me^{-\frac{n\zeta}{16}}.\end{aligned}$$

Lastly, because we assume that $n\geq e^{\frac{2}{d+1}}$ and $d\geq 5>0$, then these two conditions imply that

$$\frac{d}{2}\log n-\frac{d}{2}\log\omega-\omega\leq-\frac{1}{2}\log n,$$

which means that

$$\underbrace{\left(\frac{n}{\omega}\right)^{d/2}e^{-\omega}}_{=n^{d/2}\omega^{-d/2}e^{-\omega}}\leq n^{-1/2}.$$

Thus,

$$\begin{aligned}&1-2\Big(\frac{1}{\Upsilon_1}+\Upsilon_3\Big)\left(\frac{n}{\omega}\right)^{d/2}e^{-\omega}-me^{-\frac{n\zeta}{16}}\\&\geq 1-2\Big(\frac{1}{\Upsilon_1}+\Upsilon_3\Big)\frac{1}{\sqrt{n}}-me^{-\frac{n\zeta}{16}}.\end{aligned}$$

This results in the simplified probability bound in equation (26).

Finally, we plug in M and M' from equation (D.1) as well as $\omega=\frac{d+1}{2}\log n$ into bound (25) to arrive at (27), which completes the proof of the corollary. ■

Appendix E. Proof of Corollary 7

Verifying Assumptions A1–A4 Since $\mathcal{X}$ is the unit ball in $\mathbb{R}^2$, Assumption A1 is satisfied. There is no need to check Assumptions A2 or A3 (we directly assume A2, and A3 says that we are using the DeepHit adjacency function, which is the case since we are analyzing DeepHit). As for Assumption A4, we now describe how the bounds on M_{indiv}, $M_{\text{couple-min}}$, and $M_{\text{couple-max}}$ are obtained.

First, let's look at

$$\phi_{\text{indiv}}((x,y,\delta);\theta) = \beta \cdot \big[-\delta \log(f_{\kappa(y)}(x;\theta)) - (1-\delta)\log(S_{\kappa(y)}(x;\theta))\big].$$

Note that this function is nonnegative since log probabilities are negative and are at most 0, i.e., $\log(f_{\kappa(y)}(x;\theta)) \le 0$ and $\log(S_{\kappa(y)}(x;\theta)) \le 0$. Since $f_{\kappa(y)}(x;\theta) \ge \varrho$, this means that

$$-\beta \cdot \delta \cdot \log(f_{\kappa(y)}(x;\theta)) \le -\beta \log(f_{\kappa(y)}(x;\theta)) \le -\beta \log \varrho = \beta \log \frac{1}{\varrho}.$$

Meanwhile,

$$-\beta(1-\delta)\log(S_{\kappa(y)}(x;\theta)) \le -\beta \log(S_{\kappa(y)}(x;\theta)) \le -\beta \log \varrho = \beta \log \frac{1}{\varrho},$$

where the last inequality holds because $S_j(x;\theta)$ monotonically decreases as we go to later time indices; the smallest it gets is $S_{m-1}(x;\theta) = f_m(x;\theta) \ge \varrho$ (where we have used the assumption that within the training data, no observed time corresponds to index m). Thus, we can take $M_{\text{indiv}} = \beta \log \frac{1}{\varrho}$.

Next, we look at

$$\phi_{\text{couple}}((x,y,\delta),(x',y',\delta'),\mathcal{C};\theta) = (1-\beta)\cdot\frac{1}{n}\cdot\exp\Big(\frac{S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta)}{\sigma}\Big).$$

Here, the main observation is that $S_j(x;\theta) \in [\varrho, 1]$ for $j \in [m-1]$. Hence, $S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta) \in [\varrho - 1, 1-\varrho]$, from which we conclude that

$$\phi_{\text{couple}}((x,y,\delta),(x',y',\delta'),\mathcal{C};\theta) \in [\underbrace{(1-\beta)\cdot\frac{1}{n}\cdot e^{(\varrho-1)/\sigma}}_{M_{\text{couple-min}}}, \underbrace{(1-\beta)\cdot\frac{1}{n}\cdot e^{(1-\varrho)/\sigma}}_{M_{\text{couple-max}}}].$$

Now we check the Lipschitz constant. When $\delta = 0$, then

$$\begin{aligned} L^*((x,y,\delta),\mathcal{C};\theta) &= \phi_{\text{indiv}}((x,y,\delta);\theta) + \delta\phi_{\text{transform}}\Big(\sum_{(x',y',\delta')\in\mathcal{C}} \phi_{\text{couple}}((x,y,\delta),(x',y',\delta');\theta)\Big) \\ &= \phi_{\text{indiv}}((x,y,\delta);\theta) \\ &= -\beta\log(f_{\kappa(y)}(x;\theta)). \end{aligned}$$

Note that $s \mapsto \log s$ defined on the interval $[\varrho,\infty)$ has Lipschitz constant $\frac{1}{\varrho}$. We are composing $s \mapsto \log s$ with $f_{\kappa(y)}(x;\theta)$, which we assumed is 1-Lipschitz, so $\log(f_{\kappa(y)}(x;\theta))$ is $\frac{1}{\varrho}$-Lipschitz. Finally by multiplying by $-\beta$, we have that $L^*((x,y,\delta),\mathcal{C};\theta)$ is $\frac{\beta}{\varrho}$-Lipschitz.

Next, we consider when $\delta = 1$. In this case,

$$L^*((x,y,\delta),\mathcal{C};\theta) = \phi_{\text{indiv}}((x,y,\delta);\theta) + \delta\phi_{\text{transform}}\Big(\sum_{(x',y',\delta')\in\mathcal{C}} \phi_{\text{couple}}((x,y,\delta),(x',y',\delta');\theta)\Big)$$
$$= -\beta\log(S_{\kappa(y)}(x;\theta)) + (1-\beta)\cdot\frac{1}{n}\cdot\exp\Big(\frac{S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta)}{\sigma}\Big).$$

First off, we show that $-\beta\log(S_{\kappa(y)}(x;\theta))$ is $\frac{\beta(m-1)}{\varrho}$-Lipschitz. Note that $S_j(x;\theta) = \sum_{\ell=j+1}^{m} f_\ell(x;\theta)$ is the sum of at most $(m-1)$ functions that are each 1-Lipschitz, so it is $(m-1)$-Lipschitz. As stated earlier, $S_{\kappa(y)}(x;\theta) \geq \varrho$, and $s \mapsto \log(s)$ defined over $[\varrho,\infty)$ is $\frac{1}{\varrho}$-Lipschitz. Thus, $x \mapsto \log(S_{\kappa(y)}(x;\theta))$ is $\frac{m-1}{\varrho}$-Lipschitz. Finally, $x \mapsto -\beta\log(S_{\kappa(y)}(x;\theta))$ is $\frac{\beta(m-1)}{\varrho}$-Lipschitz.

Now we consider the term $(1-\beta)\cdot\frac{1}{n}\cdot\exp\Big(\frac{S_{\kappa(y)}(x;\theta)-S_{\kappa(y)}(x';\theta)}{\sigma}\Big)$. Note that $S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta)$ is the difference of two $(m-1)$-Lipschitz functions, so it is $2(m-1)$-Lipschitz. Next, $\frac{S_{\kappa(y)}(x;\theta)-S_{\kappa(y)}(x';\theta)}{\sigma}$ is $\frac{2(m-1)}{\sigma}$-Lipschitz. Note that

$$\frac{S_{\kappa(y)}(x;\theta) - S_{\kappa(y)}(x';\theta)}{\sigma} \in \Big[\frac{\varrho-1}{\sigma}, \frac{1-\varrho}{\sigma}\Big].$$

Observe that the map $s \mapsto \exp(s)$ defined over the interval $[\frac{\varrho-1}{\sigma}, \frac{1-\varrho}{\sigma}]$ is Lipschitz with Lipschitz constant

$$\frac{e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}}}{\frac{1-\varrho}{\sigma} - \frac{\varrho-1}{\sigma}} = \frac{\sigma(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})}{2(1-\varrho)}.$$

Then $x \mapsto \exp\big(\frac{S_{\kappa(y)}(x;\theta)-S_{\kappa(y)}(x';\theta)}{\sigma}\big)$ has Lipschitz constant

$$\frac{\sigma(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})}{2(1-\varrho)} \cdot \frac{2(m-1)}{\sigma} = \frac{(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})(m-1)}{(1-\varrho)}.$$

Finally, $x \mapsto (1-\beta)\cdot\frac{1}{n}\cdot\exp\big(\frac{S_{\kappa(y)}(x;\theta)-S_{\kappa(y)}(x';\theta)}{\sigma}\big)$ has Lipschitz constant

$$\frac{(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})(m-1)}{n(1-\varrho)}.$$

We conclude that $x \mapsto L^*((x,y,\delta),\mathcal{C};\theta)$ has Lipschitz constant

$$\frac{\beta(m-1)}{\varrho} + \frac{(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})(m-1)}{n(1-\varrho)}.$$

Under the assumption that

$$n \geq \frac{\varrho(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})}{(1-\varrho)\beta},$$

we have

$$\begin{aligned}
&\frac{\beta(m-1)}{\varrho} + \frac{(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})(m-1)}{n(1-\varrho)} \\
&\quad\leq \frac{\beta(m-1)}{\varrho} + \frac{(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})(m-1)}{\frac{\varrho(1-\beta)(e^{\frac{1-\varrho}{\sigma}} - e^{\frac{\varrho-1}{\sigma}})}{(1-\varrho)\beta}(1-\varrho)} \\
&\quad= \frac{2\beta(m-1)}{\varrho} \\
&\quad\triangleq \mathcal{L}.
\end{aligned}$$

Note that for simplicity, we have used a somewhat loose bound on the Lipschitz constant. This finishes our verification of Assumptions A1–A4.

Applying Theorem 5 We begin by noting that in this setup,

$$\begin{aligned}
M &= M_{\text{indiv}} + \frac{M_{\text{couple-max}}}{2} n \\
&= \beta \log \frac{1}{\varrho} + \frac{(1-\beta) \cdot \frac{1}{n} \cdot e^{(1-\varrho)/\sigma}}{2} n \\
&= \beta \log \frac{1}{\varrho} + \frac{(1-\beta) e^{(1-\varrho)/\sigma}}{2},
\end{aligned}$$

and

$$\begin{aligned}
M' &= (M_{\text{couple-max}} - M_{\text{couple-min}}) \sqrt{\frac{\omega n}{2\zeta}} \\
&= \Big((1-\beta) \cdot \frac{1}{n} \cdot e^{(1-\varrho)/\sigma} - (1-\beta) \cdot \frac{1}{n} \cdot e^{(\varrho-1)/\sigma} \Big) \sqrt{\frac{\omega n}{2\zeta}} \\
&= \sqrt{\frac{\omega}{n}} \Big(\frac{(1-\beta) e^{(1-\varrho)/\sigma} - (1-\beta) e^{(\varrho-1)/\sigma}}{\sqrt{2\zeta}} \Big).
\end{aligned}$$

Just as in the proof of Corollary 6, we begin by showing that $M' \leq 1$, which ensures that

$$\mathbb{N}(M', \mathcal{X}) \leq \Big(\frac{3}{M'} \Big)^d.$$

We have

$$\begin{aligned}
M' &= \sqrt{\frac{\omega}{n}} \Big(\frac{(1-\beta) e^{(1-\varrho)/\sigma} - (1-\beta) e^{(\varrho-1)/\sigma}}{\sqrt{2\zeta}} \Big) \\
&= \sqrt{\frac{\frac{d+1}{2} \log n}{n}} \Big(\frac{(1-\beta) e^{(1-\varrho)/\sigma} - (1-\beta) e^{(\varrho-1)/\sigma}}{\sqrt{2\zeta}} \Big) \\
&= \sqrt{\frac{\log n}{n}} \left(\frac{(1-\beta) e^{(1-\varrho)/\sigma} - (1-\beta) e^{(\varrho-1)/\sigma}}{2} \sqrt{\frac{d+1}{\zeta}} \right).
\end{aligned}$$

Under the assumptions that

$$n \geq [(1-\beta)e^{(1-\varrho)/\sigma} - (1-\beta)e^{(\varrho-1)/\sigma}]^2\big(\frac{d+1}{4\zeta}\big)e^{\sqrt{2\log\left([(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}]^2\left(\frac{d+1}{4\zeta}\right)\right)}}$$
$$= e^{\log\left([(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}]^2\left(\frac{d+1}{4\zeta}\right)\right)+\sqrt{2\log\left([(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}]^2\left(\frac{d+1}{4\zeta}\right)\right)}},$$

and that

$$\log[(1-\beta)e^{(1-\varrho)/\sigma} - (1-\beta)e^{(\varrho-1)/\sigma}]^2\Big(\frac{d+1}{4\zeta}\Big) > 1,$$

then by inequality (D.3) (with $s = \log[(1-\beta)e^{(1-\varrho)/\sigma} - (1-\beta)e^{(\varrho-1)/\sigma}]^2\Big(\frac{d+1}{4\zeta}\Big) - 1$), we have

$$n \geq e^{-W_{-1}\left(-\frac{1}{[(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}]^2\left(\frac{d+1}{4\zeta}\right)}\right)},$$

which implies that $M' \leq 1$.

Now that we have shown that $M' \leq 1$ so that $\mathbb{N}(M', \mathcal{X}) \leq (\frac{3}{M'})^d$, the probability in equation (24) satisfies the bound

$$1 - 2\bigg[\frac{M}{(C_\alpha - 1)\big[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha - 1}\}M + (2\mathcal{L}+1)M'\big]} + \mathbb{N}(M', \mathcal{X})\bigg]e^{-\omega} - me^{-\frac{n\zeta}{16}}$$
$$= 1 - 2\bigg[\frac{\beta\log\frac{1}{\varrho} + \frac{(1-\beta)e^{(1-\varrho)/\sigma}}{2}}{(C_\alpha - 1)\big[2\sqrt{\frac{\omega}{n}}\max\{2, \frac{C_\alpha}{C_\alpha-1}\}(\beta\log\frac{1}{\varrho} + \frac{(1-\beta)e^{(1-\varrho)/\sigma}}{2}) + (2\mathcal{L}+1)\sqrt{\frac{\omega}{n}}\Big(\frac{(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}}{\sqrt{2\zeta}}\Big)\big]}$$
$$+ \mathbb{N}(M', \mathcal{X})\bigg]e^{-\omega} - me^{-\frac{n\zeta}{16}}$$
$$= 1 - 2\bigg[\frac{1}{(C_\alpha - 1)\Big[2\max\{2, \frac{C_\alpha}{C_\alpha-1}\} + (2\mathcal{L}+1)\Big(\frac{\left((1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}\right)}{\left(2\beta\log\frac{1}{\varrho}+(1-\beta)e^{(1-\varrho)/\sigma}\right)}\sqrt{\frac{2}{\zeta}}\Big)\Big]}\sqrt{\frac{n}{\omega}} + \mathbb{N}(M', \mathcal{X})\bigg]e^{-\omega}$$
$$- me^{-\frac{n\zeta}{16}}$$
$$\leq 1 - 2\bigg[\frac{1}{(C_\alpha - 1)\Big[2\max\{2, \frac{C_\alpha}{C_\alpha-1}\} + (2\mathcal{L}+1)\Big(\frac{\left((1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}\right)}{\left(2\beta\log\frac{1}{\varrho}+(1-\beta)e^{(1-\varrho)/\sigma}\right)}\sqrt{\frac{2}{\zeta}}\Big)\Big]}\sqrt{\frac{n}{\omega}}$$
$$+ \left(\frac{3}{\sqrt{\frac{\omega}{n}}\Big(\frac{(1-\beta)e^{(1-\varrho)/\sigma}-(1-\beta)e^{(\varrho-1)/\sigma}}{\sqrt{2\zeta}}\Big)}\right)^d\bigg]e^{-\omega} - me^{-\frac{n\zeta}{16}}$$
$$= 1 - 2\bigg[\Psi_1\sqrt{\frac{n}{\omega}} + \Psi_2\Big(\frac{n}{\omega}\Big)^{d/2}\bigg]e^{-\omega} - me^{-\frac{n\zeta}{16}},$$

where

$$\Psi_1 \triangleq \frac{1}{(C_\alpha - 1)\Big[2\max\{2, \frac{C_\alpha}{C_\alpha - 1}\} + (2\mathcal{L}+1)\Big(\frac{\Big((1-\beta)e^{(1-\varrho)/\sigma} - (1-\beta)e^{(\varrho-1)/\sigma}\Big)}{\Big(2\beta\log\frac{1}{\varrho} + (1-\beta)e^{(1-\varrho)/\sigma}\Big)}\sqrt{\frac{2}{\zeta}}\Big)\Big]},$$

$$\Psi_2 \triangleq \left(\frac{3\sqrt{2\zeta}}{(1-\beta)e^{(1-\varrho)/\sigma} - (1-\beta)e^{(\varrho-1)/\sigma}}\right)^d.$$

Using the same reasoning as in the proof of Corollary 6, when $n \geq e^{\sqrt{2(\log\frac{d+1}{2}-1)}+\log\frac{d+1}{2}}$ (which we assume in the corollary statement), we are guaranteed that $n \geq \omega$. Hence, we have

$$\begin{aligned}
&1 - 2\left[\Psi_1\sqrt{\frac{n}{\omega}} + \Psi_2\Big(\frac{n}{\omega}\Big)^{d/2}\right]e^{-\omega} - me^{-\frac{n\zeta}{16}} \\
&\geq 1 - 2(\Psi_1 + \Psi_2)\Big(\frac{n}{\omega}\Big)^{d/2}e^{-\omega} - me^{-\frac{n\zeta}{16}}.
\end{aligned}$$

Moreover, when $n \geq e^{\frac{2}{d+1}}$ and $d \geq 5 > 0$ (we assume both of these), using the same reasoning as the proof of Corollary 6,

$$\Big(\frac{n}{\omega}\Big)^{d/2}e^{-\omega} \leq \frac{1}{\sqrt{n}}.$$

Hence,

$$\begin{aligned}
&1 - 2(\Psi_1 + \Psi_2)\Big(\frac{n}{\omega}\Big)^{d/2}e^{-\omega} - me^{-\frac{n\zeta}{16}} \\
&\geq 1 - 2(\Psi_1 + \Psi_2)\frac{1}{\sqrt{n}} - me^{-\frac{n\zeta}{16}}.
\end{aligned}$$

In the statement of the corollary, $\Psi \triangleq 2(\Psi_1 + \Psi_2)$. As for the loss bound (25), we simply plug in the values of M and M' specific to the DeepHit setup here, and we also plug in $\omega = \frac{d+1}{2}\log n$. This finishes the proof. ■

Appendix F. Fairness Metrics

In this paper, we use the individual, group, and intersectional fairness metrics defined by Keya et al. (2021), the concordance imparity (CI) metric by Zhang and Weiss (2022), and also censoring-based individual and censoring-based group fairness metrics by Rahman and Purushotham (2022). For all of these fairness metrics, lower is considered better, where the minimum possible value is 0. We point out that the fairness metrics by Keya et al. (2021) and Rahman and Purushotham (2022) can readily be treated as regularizers (i.e., they could be included as additional loss terms during model training). Moreover, the individual fairness metric by Keya et al. (2021) and the censoring-based individual and censoring-based group fairness metrics by Rahman and Purushotham (2022) crucially depend on a scaling constant $\gamma > 0$ that must be set by the user in advance: if γ is set to be higher, then it becomes easier for a survival model to achieve a score of exactly (and not just approximately) 0 for these particular fairness metrics.

Note that in Section 5 of the main paper, we use the fairness metrics by Keya et al. as regularizers in baseline methods and not as evaluation metrics. However, we include additional experimental results that use the individual and group fairness metrics by Keya et al. as evaluation metrics in Appendix H (specifically, see Tables H.5–H.13).

We begin by explaining the fairness metrics proposed by Keya et al. (2021) as these were the earliest fairness metrics we are aware of that were developed for survival analysis. Note that Keya et al. focused on Cox proportional hazards models. For such models, we can take the predicted outcome for a feature vector x to be the so-called *partial hazard* $\widetilde{h}(x) \triangleq \exp(f(x;\theta))$; this is the same as the hazard function given in equation (3) except where we exclude the baseline hazard factor $h_0(t)$. Note that once we exclude $h_0(t)$, then $\widetilde{h}$ no longer depends on time t. We state the fairness metrics in terms of a collection of N_{test} test patients with data $(X_1^{\text{test}}, Y_1^{\text{test}}, \Delta_1^{\text{test}}), \ldots, (X_{N_{\text{test}}}^{\text{test}}, Y_{N_{\text{test}}}^{\text{test}}, \Delta_{N_{\text{test}}}^{\text{test}})$. Note that the fairness metrics by Keya et al. (2021) only use the test feature vectors $X_1^{\text{test}}, \ldots, X_{N_{\text{test}}}^{\text{test}}$ and ignores the test patients' observed times and event indicators. Also, at the end of this section, we point out that the individual and group fairness metrics by Keya et al. (2021) are sensitive to the scale of the log partial hazard $f(\cdot;\theta)$.

Individual fairness Roughly, Keya et al. (2021) consider a model to be fair across individuals (patients) if similar individuals have similar predicted outcomes. To operationalize this notion of fairness in the context of Cox models, Keya et al. define the individual fairness metric

$$\mathrm{F}_I \triangleq \sum_{i=1}^{N_{\text{test}}} \sum_{j=i+1}^{N_{\text{test}}} \big[|\widetilde{h}(X_i^{\text{test}}) - \widetilde{h}(X_j^{\text{test}})| - \gamma \|X_i^{\text{test}} - X_j^{\text{test}}\|\big]_+,$$

where γ is a predefined scale factor (0.01 in our experiments). As a reminder, $[\,\cdot\,]_+$ is the ReLU function (so that $[a]_+ = \max\{0, a\}$ for any $a \in \mathbb{R}$). Importantly, we point out that by setting γ to be larger, then more terms in the summation become 0 (since within the ReLU expression, we are subtracting a larger quantity, making it more likely that after applying ReLU, we get 0). If γ is set to be too large, then it is possible that all terms in the summation become 0 (i.e., the fairness metric becomes exactly and not just approximately equal to 0).

Note that this individual fairness metric is actually just penalizing $\widetilde{h}$ for not being Lipschitz continuous (as empirically evaluated over the test data). Specifically, $\widetilde{h}$ is defined to be γ-Lipschitz continuous if

$$|\widetilde{h}(x) - \widetilde{h}(x')| \leq \gamma \|x - x'\| \quad \text{for all } x, x' \in \mathcal{X}.$$

Meanwhile, when F_I is equal to 0, then it means that

$$|\widetilde{h}(X_i^{\text{test}}) - \widetilde{h}(X_j^{\text{test}})| \leq \gamma \|X_i^{\text{test}} - X_j^{\text{test}}\| \quad \text{for all } i, j \in \{1, \ldots, N_{\text{test}}\}.$$

As a technical remark, in the definition of F_I and also γ-Lipschitz continuity, the metric used to measure distances between feature vectors does not have to be Euclidean. For example, we can replace $\|X_i^{\text{test}} - X_j^{\text{test}}\|$ with $\rho(X_i^{\text{test}}, X_j^{\text{test}})$, where $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a user-specified metric.

The individual fairness metric by Keya et al. (2021) can be modified to support survival models that do not assume the proportional hazards assumption (such as DeepHit and

SODEN) in a straightforward manner: we simply replace the hazard function $\widetilde{h}(x)$ by the estimated survival function $\widehat{S}(t|x)$ to obtain the following time-dependent fairness metric:

$$\mathrm{F}_I(t) \triangleq \sum_{i=1}^{N_{\text{test}}} \sum_{j=i+1}^{N_{\text{test}}} \left[|\widehat{S}(t|X_i^{\text{test}}) - \widehat{S}(t|X_j^{\text{test}})| - \gamma \|X_i^{\text{test}} - X_j^{\text{test}}\| \right]_+.$$

Group fairness Keya et al. (2021) consider a model is fair across a user-specified set of groups if these different groups have similar predicted outcomes. Keya et al. define the group fairness metric F_G to look at the maximum deviation of a group's average predicted outcome to the overall population's average predicted outcome. Specifically, let $\mathcal{G}$ be the user-specified set of groups to consider (for example, there could be two groups: everyone with age at most 65 years, and everyone older than 65 years), where each group $g \in \mathcal{G}$ is a subset of the test set indices $\{1, \ldots, N_{\text{test}}\}$ (so that using this notation, group g has size $|g|$); the different groups should form a partition of the test set (so that the groups are disjoint and their union is the entire test set). Then

$$\mathrm{F}_G \triangleq \max_{g \in \mathcal{G}} \left| \underbrace{\frac{1}{|g|} \sum_{i \in g} \widetilde{h}(X_i^{\text{test}})}_{\substack{\text{average predicted} \\ \text{outcome of group } g}} - \underbrace{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \widetilde{h}(X_i^{\text{test}})}_{\substack{\text{average predicted} \\ \text{outcome of population}}} \right|.$$

Once again, for survival models that do not assume a proportional hazards assumption (such as DeepHit and SODEN), we can instead replace $\widetilde{h}(x)$ with $\widetilde{S}(t|x)$ to obtain the following time-dependent group fairness metric:

$$\mathrm{F}_G(t) \triangleq \max_{g \in \mathcal{G}} \left| \underbrace{\frac{1}{|g|} \sum_{i \in g} \widehat{S}(t|X_i^{\text{test}})}_{\substack{\text{average predicted} \\ \text{outcome of group } g}} - \underbrace{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \widehat{S}(t|X_i^{\text{test}})}_{\substack{\text{average predicted} \\ \text{outcome of population}}} \right|.$$

Intersectional fairness Keya et al. (2021) consider a notion of intersectional fairness that accounts for multiple sensitive attributes. For example, in the FLC dataset, we have 2 different sensitive attributes, age and gender. For each of these sensitive attributes, we can partition the test set into groups. Specifically, let $\mathcal{G}_1$ be a partition of the test set into different age groups (for example, two groups: at most 65 years old and over 65 years old), and let $\mathcal{G}_2$ be a partition of the test set into different gender groups (for example, two groups: female and male). Then intersectional fairness looks at every intersection of age/gender groups (continuing from the previous examples, we would have four intersectional subgroups: at most 65 years old and female, at most 65 years and male, over 65 years old and female, over 65 years old and male).

The notation here is a bit more involved. The set of all intersectional subgroups of $\mathcal{G}_1$ and $\mathcal{G}_2$ is given by the Cartesian product $\mathcal{G}_1 \times \mathcal{G}_2$. Note that $s \in \mathcal{G}_1 \times \mathcal{G}_2$ means that $s = (s_1, s_2)$, where $s_1 \in \mathcal{G}_1$ and $s_2 \in \mathcal{G}_2$. More generally, if there are J sensitive attributes, corresponding to groupings $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_J$, then the set of all intersectional subgroups would be $\mathcal{S} \triangleq \mathcal{G}_1 \times \mathcal{G}_2 \times \cdots \mathcal{G}_J$. Now $s \in \mathcal{S}$ is a list consisting of J different subsets of test patients

(i.e., $s = (s_1, s_2, \ldots, s_J)$, where $s_1 \in \mathcal{G}_1$, $\ldots$, $s_J \in \mathcal{G}_J$). The intersection of these J subsets (i.e., $\cap_{j=1}^{J} s_j \subset \{1, \ldots, N_{\text{test}}\}$) is precisely the set of test patients that intersectional subgroup s corresponds to. Then the average predicted outcome for intersectional subgroup s is

$$\widetilde{\mathbf{h}}(s) \triangleq \frac{1}{|\cap_{j=1}^{J} s_j|} \sum_{i \in \cap_{j=1}^{J} s_j} \widetilde{h}(X_i^{\text{test}}).$$

Then the intersection fairness metric $\mathrm{F}_\cap$ by Keya et al. (2021) is the worst-case log ratio of expected predicted outcomes between two intersectional subgroups:

$$\mathrm{F}_\cap \triangleq \max_{s,s' \in \mathcal{S}} \Big| \log \frac{\widetilde{\mathbf{h}}(s)}{\widetilde{\mathbf{h}}(s')} \Big|.$$

Concordance imparity We now describe an alternative metric for group fairness called concordance imparity (CI) that asks that a survival analysis model achieves similar prediction accuracy for different groups. For ease of exposition, we only state the CI metric by Zhang and Weiss (2022) in terms of a single sensitive attribute that has already been discretized (e.g., the attribute is already discrete or we have a pre-specified discretization rule); this special case is sufficient for our experiments. We denote the set of possible discretized values of this sensitive attribute as $\mathcal{A}$. For example, $\mathcal{A}$ could correspond to age and we could have $\mathcal{A} = \{\text{“age} \leq 65\text{”}, \text{“age} > 65\text{”}\}$, i.e., $\mathcal{A}$ consists of the different groups to consider. We refer the reader to the Zhang and Weiss's original paper for their more general definition of CI that can handle a continuous sensitive attribute via an automatic discretization strategy that they propose.

Assuming that the sensitive attribute has already been discretized into the set $\mathcal{A}$, the CI metric looks at a variant of the standard survival analysis accuracy metric of concordance index (Harrell et al., 1982) that Zhang and Weiss call the *concordance fraction* (CF), which is specific to each sensitive attribute value $a \in \mathcal{A}$. The CI metric is then defined to be the worst-case difference between the CF scores of any two $a, a' \in \mathcal{A}$ where $a \neq a'$. The pseudocode can be found in Algorithm 3; note that to keep the notation from getting clunky, we drop the superscript “test” from the test feature vectors, observed times, and event indicators in the pseudocode but we still use N_{test} to denote the number of test patients. Also, in the pseudocode, we let $A_i \in \mathcal{A}$ denote the sensitive attribute value for the i-th test patient, where we assume that A_i can directly be computed based on the i-th test patient's feature vector. For example, when age (which is not discretized) is one of the features and $\mathcal{A}$ consists of the two age groups previously stated (≤ 65 or > 65), then since we know the discretization rule used, we can readily determine which age group in $\mathcal{A}$ that any test patient is in.

Importantly, to calculate the CI metric, a way to calculate a risk score is required to compute the CF scores. How to define a risk score differs across models. For Cox models, we can take the risk score to be the log partial hazard function $f(\cdot;\theta)$. For DeepHit and SODEN models, we take the risk score to be the estimated survival probability $\widehat{S}(t|x)$ and therefore we need to replace $f(\cdot;\theta)$ with $\widehat{S}(t|x)$ before using Algorithm 3. Since different values of time t can have different estimated $\widehat{S}(t|x)$ values, we would obtain different value of the CI fairness metric for different t. We test three different values of t (the 25^{th}, 50^{th}, and 75^{th} percentile of the observed times in the test data) and use the average value for the final CI score.

Censoring-based individual fairness Individual fairness F_I does not consider censoring information that is available. Rahman and Purushotham (2022) defined a censoring-based individual fairness metric as follows:

$$\mathrm{F}_{CI}(t) \triangleq \frac{1}{|N_c| \times |N_{uc}|} \sum_{\substack{i \in N_c, j \in N_{uc} \\ \text{s.t. } Y_j \geq Y_i}} \left[|\widehat{S}(t|X_i^{\text{test}}) - \widehat{S}(t|X_j^{\text{test}})| - \gamma \|X_i^{\text{test}} - X_j^{\text{test}}\| \right]_+,$$

where N_c and N_{uc} are the index sets of censored and uncensored data, and $\widehat{S}(t|X)$ is the estimated survival probability at time t for patient X. Similar to in F_I, the scale factor γ is a predefined (0.01 in our experiments). This fairness metric ensures that a censored patient and an uncensored patient who have similar features should also have similar predictions whenever the observed time from the uncensored patient is larger than that of the censored patient. Similar to the CI fairness metric and following the experimental settings of Rahman and Purushotham (2022), we test three different t values (25^{th}, 50^{th}, 75^{th} percentile of the observed times in the test data) and use their average value to calculate the final F_{CI} score.

As a warning, just as with the individual fairness metric F_I by Keya et al. (2021) that we described earlier, if γ is set higher, then it becomes easier for the $\mathrm{F}_{CI}(t)$ metric to become exactly and not just approximately equal to 0.

Censoring-based group fairness Rahman and Purushotham (2022) also modified the F_G metric by Keya et al. (2021) to account for censoring information. Reusing notation from the definitions of F_G and $\mathrm{F}_{CI}(t)$, we now define the censoring-based group fairness metric

$$\mathrm{F}_{CG}(t) \triangleq \frac{1}{|N_c| \times |N_{uc}|} \sum_{g \in \mathcal{G}} \sum_{\substack{i \in N_{c,g}, j \in N_{uc,g} \\ \text{s.t. } Y_j \geq Y_i}} \left[|\widehat{S}(t|X_i^{\text{test}}) - \widehat{S}(t|X_j^{\text{test}})| - \gamma \|X_i^{\text{test}} - X_j^{\text{test}}\| \right]_+,$$

where $N_{c,g}$ and $N_{uc,g}$ are the index sets of censored and uncensored in group g, and $\widehat{S}(t|X)$ is the estimated survival probability at time t for patient X. Similar to the setting in censoring-based individual fairness, we use three different t to test the value of $\mathrm{F}_{CG}(t)$ and use their average for the final reported F_{CG} score. Once again, if γ is set too large, then it becomes easier for $\mathrm{F}_{CG}(t)$ to be exactly 0.

Scale Issues with F_I and F_G

We point out that the F_I and F_G fairness metrics by Keya et al. (2021) are sensitive to the scale of the log partial hazard function $f(\cdot;\theta)$, and thus also the scale of the partial hazard $\widetilde{h}(x) = \exp(f(x;\theta))$ if they are calculated by using $\widetilde{h}(x)$. For instance, consider a standard linear Cox model with $f(x;\theta) = \theta^T x$, where the parameters θ have already been learned. Then one way to make the model appear fairer according to the F_I and F_G metrics is to just scale all values in θ by any positive constant smaller than 1; doing so, the standard accuracy metric of concordance index (Harrell et al., 1982) would actually remain unchanged for the model as it only depends on the ranking of the different individuals' (log) partial hazard values. However, an accuracy score that considers each individual's survival function estimate (e.g., integrated IPCW Brier Score (Graf et al., 1999)) would be affected.

Algorithm 3: Concordance Imparity (CI) with a discrete sensitive attribute

Input: Test dataset $\{(X_i, Y_i, \Delta_i)\}_{i=1}^{N_{\text{test}}}$, risk score $f(\cdot;\theta)$ (from an already trained model), set of sensitive attribute values $\mathcal{A}$ (so that each $a \in \mathcal{A}$ corresponds to a different group), $A_1, \ldots, A_{N_{\text{test}}} \in \mathcal{A}$ says which sensitive attribute value each test patient has

Output: CI score

for $a \in \mathcal{A}$ **do**
  Initialize the numerator count $\mathbf{N}(a) \leftarrow 0$ and denominator count $\mathbf{D}(a) \leftarrow 0$.
end
for $i = 1, \ldots, N_{test}$ **do**
  for $j = 1, \ldots, N_{test}$ *s.t.* $j \neq i$ **do**
    if *(*$Y_i < Y_j$ *and* $\Delta_i == 0$*) or (*$Y_j < Y_i$ *and* $\Delta_j == 0$*) or (*$Y_i == Y_j$ *and* $\Delta_i == 0$ *and* $\Delta_j == 0$*)* **then**
      continue
    else
      Set $\mathbf{D}(A_i) \leftarrow \mathbf{D}(A_i) + 1$.
    end
    if $Y_i < Y_j$ **then**
      if $f(X_i;\theta) > f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 1$.
      else if $f(X_i;\theta) == f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 0.5$.
      end
    else if $Y_i > Y_j$ **then**
      if $f(X_i;\theta) < f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 1$.
      else if $f(X_i;\theta) == f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 0.5$.
      end
    else if $Y_i == Y_j$ **then**
      if $\Delta_i == 1$ *and* $\Delta_j == 1$ **then**
        if $f(X_i;\theta) == f(X_j;\theta)$ **then**
          Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 1$.
        else
          Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 0.5$.
        end
      else if $\Delta_i == 0$ *and* $\Delta_j == 1$ *and* $f(X_i;\theta) < f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 1$.
      else if $\Delta_i == 1$ *and* $\Delta_j == 0$ *and* $f(X_i;\theta) > f(X_j;\theta)$ **then**
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 1$.
      else
        Set $\mathbf{N}(A_i) \leftarrow \mathbf{N}(A_i) + 0.5$.
      end
    end
  end
end
for $a \in \mathcal{A}$ **do**
  Set the concordance fraction of a: $\mathbf{CF}(a) \leftarrow \frac{\mathbf{N}(a)}{\mathbf{D}(a)}$.
end
return CI $\leftarrow \max_{a,a' \in \mathcal{A} \text{ s.t. } a \neq a'} |\mathbf{CF}(a) - \mathbf{CF}(a')|$

Appendix G. Hyperparameter Tuning and Compute Environment Details

Hyperparameters *Cox models*: for nonlinear Cox models, we always use a two-layer MLP with ReLU as the activation function and 24 as the number of hidden units. All models (linear and nonlinear) are trained using Adam (Kingma and Ba, 2014) in PyTorch 1.7.1 in a batch setting for 500 iterations (except in the case of the exact DRO Cox model on the FLC dataset, where we use 5000 iterations as it took more iterations for the model to converge), only using a CPU and no GPU.

DeepHit models: we use three-layer MLP with ReLU activation, batch normalization, and dropout (in 0.1). The number of hidden units is 32. The original DeepHit and DRO-DEEPHIT models are trained using Adam in PyTorch 1.7.1 in a mini-batch 256 setting for 500 epochs. However, the DRO-DEEPHIT (SPLIT) model is trained using a batch setting for 500 iterations.

SODEN models: for the FLC dataset, we use an MLP with 4 layers and 16 hidden units. For SUPPORT and SEER datasets, we use an MLP with 2 layers and 26 hidden units. In addition, RMSprop (Tieleman et al., 2012) in 128 batch size with a maximum 100 epochs is used to train all models.

We tune on the following hyperparameter grid:

- To find the optimal learning rate for each Cox model, we conducted a sweep over values of 0.01, 0.001, and 0.0001. Specifically for the exact DRO Cox model, we used a fixed learning rate of 0.1. For the FIDP, FIPNAM, and DeepHit models, we used a fixed learning rate of 0.01. In the case of SODEN models, the learning rates applied were 0.01, 0.002, and 0.002 for the FLC, SUPPORT, and SEER datasets, respectively.
- λ (only used for baselines; a hyperparameter that controls the tradeoff between the original Cox loss and fairness regularization term): 1, 0.7, 0.4. We set $\lambda = 0.1$ for FIDP and FIPNAM.
- α: 0.1, 0.15, 0.2, 0.3, 0.4, 0.5 for DRO-COX/DRO-COX (SPLIT)/EXACT DRO-COX variants; 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 for DRO-DEEPHIT/DRO-DEEPHIT (SPLIT) variants and DRO-SODEN.

In addition, for DRO-COX (SPLIT) and DRO-DEEPHIT (SPLIT), we choose $n_1 = n_2 = n/2$ (rounding as needed when n is odd, so that n_1 might not equal n_2).

Compute environment All models are implemented with Python 3.8.3, and they are trained and tested on identical compute instances, each with an Intel Core i9-10900K CPU (3.70GHz with 64 GB RAM) and a Quadro RTX 4000 GPU.

Appendix H. Additional Experiments

Using other sensitive attributes in evaluating CI and F_{CG} in Cox models In Section 5 of the main paper, for the Cox model, we only showed test set performance metrics (C^{td} and IBS accuracy metrics, and CI, F_{CI}, and F_{CG} fairness metrics) for FLC, SUPPORT, and SEER using age, gender, race, and race respectively (in Tables 2, 3, and 4). We now provide results using gender for FLC (Table H.1), age and separately race for SUPPORT (Tables H.2 and H.3), and age for SEER (Table H.4). Our main findings still hold for these additional results.

We point out that for DeepHit and SODEN models, in Section 5, we had already shown results for FLC, SUPPORT, and SEER where per dataset, we consider different sensitive attributes (see Tables 5 and 6).

Using individual and group fairness evaluation metrics by Keya et al. (2021) Whereas in the main paper, we focused on evaluating test data using CI, F_{CI}, and F_{CG} fairness metrics, we now also show results where we use the F_I and F_G fairness metrics by Keya et al. (2021) instead. See Tables H.5–H.13. Our main findings still hold using these fairness metrics by Keya et al.

Effect of changing n_1 (or n_2) for DRO-COX (SPLIT) In the above experiments, we set $n_1 = n_2 = n/2$ (rounding as needed). To evaluate the sensitivity of this setting, we test the model performance using DRO-COX (SPLIT) under the linear and nonlinear settings, where we set $n_2 = 0.1n, 0.2n, 0.3n, 0.4n, 0.5n$ (corresponding to $n_1 = 0.9n, 0.8n, 0.7n, 0.6n, 0.5n$). We report the test set performance metrics for the FLC dataset (using age for evaluation) in Table H.14. From the table, we find that per metric, different settings for n_1 and n_2 lead to results that, while slightly different, are not dramatically different, i.e., the performance of DRO-COX (SPLIT) does not appear very sensitive w.r.t. the choice of n_1 and n_2.

Effect of changing imbalance in censoring rates across training data splits for DRO-COX (SPLIT) For our split DRO strategy, to see what happens when the two subsets of the training data $\mathcal{D}_1$ and $\mathcal{D}_2$ have different censoring rates, we conduct the following experiment. We first randomly divide the training data into 50/50 pieces $\mathcal{D}_1/\mathcal{D}_2$ where we stratify on the censoring rate so that $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same censoring rate. Then, we introduce an censoring rate imbalance by trading, for instance, a randomly chosen censored point from $\mathcal{D}_2$ with a randomly chosen uncensored point from $\mathcal{D}_1$. We could of course trade multiple points.

To formalize a notion of how much imbalance we are introducing, we define a censoring rate *imbalance ratio* as follows. First, note that using the above strategy of trading points between $\mathcal{D}_1$ and $\mathcal{D}_2$ that we stated, the maximum number of points we could possibly trade is given by the *minimum* of the number of uncensored points in $\mathcal{D}_1$ and the number of censored points in $\mathcal{D}_2$. Let's call this maximum number of points we could trade as $n_{\text{max trade}}$. Then we define the imbalance ratio to be a percentage of $n_{\text{max trade}}$ points that we trade. Thus, an imbalance ratio of 80% means that we trade $0.8n_{\text{max trade}}$ randomly chosen censored points from $\mathcal{D}_2$ with $0.8n_{\text{max trade}}$ randomly chosen uncensored points from $\mathcal{D}_1$.

We repeat the same experiment that resulted in Table 2 specifically for DRO-COX (SPLIT) (i.e., for simplicity, we only consider the FLC dataset treating age as sensitive), where the only difference now is that we re-train DRO-COX (SPLIT) using imbalance ratios of 0%, 20%, 40%, 60%, 80%, and 100% (per imbalance ratio, we re-run experiments 10 times). The resulting test set accuracy and fairness metrics are reported in Table H.15.

The most important takeaway from Table H.15 is that our split DRO approach still can work well even with high censoring rate imbalance ratios. For instance, in the linear setting, accounting for the standard deviations that have been reported in the table, at an imbalance ratio of 100%, the resulting accuracy and fairness metrics are actually within noise of using an imbalance ratio of 0%. In the nonlinear setting, at an imbalance ratio of 80%, the model achieves a better mean CI fairness score compared to an imbalance ratio of 0% while

achieving the highest mean C^{td} score (although the mean IBS score increases). Meanwhile, still in the nonlinear setting, at an imbalance of 100%, the model achieves the lowest IBS and CI fairness scores (within the nonlinear setting). To recap, these findings suggest that our split DRO approach can still work well even at high censoring rate imbalance ratios.

As for whether we should favor low or high censoring rate imbalance ratios, Table H.15 suggests that in practice, we should just tune on this imbalance ratio since an intermediate imbalance ratio could achieve the best tradeoff of accuracy and fairness scores. For simplicity though, in the main paper, we do not tune on the censoring rate imbalance ratio and stick to just using an imbalance ratio of 0%. We defer a more thorough investigation of the impact of the imbalance ratio to future work.

The effect of using two losses for DRO-COX (SPLIT) rather than only one Recall that DRO-COX (SPLIT) minimizes the sum of two losses $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ and $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_2 \mid \mathcal{D}_1)$. Towards the end of Section 3.2, we said that an approach that only minimizes one of these losses would not use the data as effectively compared to minimizing the sum of these losses. We conducted an experiment to verify this claim, where we refer to the version of DRO-COX (SPLIT) that only minimizes $L^{\text{split}}_{\text{DRO}}(\theta, \eta, \mathcal{D}_1 \mid \mathcal{D}_2)$ as DRO-COX (SPLIT, ONE SIDE). Specifically, we compare DRO-COX (SPLIT, ONE SIDE) and DRO-COX (SPLIT) under linear and nonlinear settings on the FLC dataset using age for evaluation. We report the resulting test set performance metrics in Table H.16. From the table, we find that DRO-COX (SPLIT) outperforms DRO-COX (SPLIT, ONE SIDE) on most metrics. This experimental finding supports our hypothesis that DRO-COX (SPLIT, ONE SIDE) uses data less effectively.

References

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Norman Breslow. Discussion of the paper by David R Cox (1972), cited below. *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

George H Chen. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, pages 537–565. PMLR, 2020.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Table H.1: Cox model test set scores on the FLC (gender) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	**0.8032**	0.1739	0.8610	0.0249	0.0128	**0.8032**	0.1739	0.8610	0.0249	0.0128
		(0.0002)	(0.0004)	(0.0197)	(0.0002)	(0.0001)	**(0.0002)**	(0.0004)	(0.0197)	(0.0002)	(0.0001)
	Cox_I(Keya et al.)	0.7932	0.1368	1.6750	0.0096	0.0051	0.7859	0.1334	1.9650	0.0067	0.0035
		(0.0083)	(0.0052)	(0.7969)	(0.0031)	(0.0016)	(0.0220)	(0.0034)	(1.3510)	(0.0006)	(0.0004)
	Cox_I(R&P)	0.8024	0.1714	0.7850	0.0239	0.0123	0.8016	0.1679	0.7350	0.0224	0.0115
		(0.0007)	(0.0033)	(0.0648)	(0.0014)	(0.0007)	(0.0006)	(0.0020)	(0.0329)	(0.0008)	(0.0004)
	Cox_G(Keya et al.)	0.8011	0.1619	0.7020	0.0195	0.0100	0.8003	0.1567	0.6350	0.0172	0.0089
		(0.0015)	(0.0077)	(0.1081)	(0.0035)	(0.0018)	(0.0004)	(0.0004)	(0.0081)	(0.0001)	(0.0001)
	Cox_G(R&P)	0.8022	0.1707	0.7730	0.0236	0.0121	0.8015	0.1673	0.7210	0.0222	0.0114
		(0.0006)	(0.0033)	(0.0628)	(0.0015)	(0.0007)	(0.0003)	(0.0004)	(0.0137)	(0.0002)	(0.0001)
	$\text{Cox}_\cap$(Keya et al.)	0.7868	0.1400	0.4830	0.0073	0.0039	0.7868	0.1400	**0.4830**	0.0073	0.0039
		(0.0018)	(0.0005)	(0.1020)	(0.0001)	(0.0001)	(0.0018)	(0.0005)	**(0.1020)**	(0.0001)	(0.0001)
	DRO-COX	0.7605	**0.1350**	**0.3040**	0.0018	0.0010	0.7958	**0.1330**	1.0780	**0**	**0**
		(0.0096)	**(0.0003)**	**(0.1569)**	(0.0006)	(0.0003)	(0.0049)	**(0.0002)**	(0.0739))	**(0)**	**(0)**
	DRO-COX (SPLIT)	0.7964	0.1389	1.0120	**0**	**0**	0.7964	0.1389	1.0120	**0**	**0**
		(0.0045)	(0.0008)	(0.1369)	**(0)**	**(0)**	(0.0045)	(0.0008)	(0.1369)	**(0)**	**(0)**
	EXACT DRO-COX	0.7821	0.3916	1.3025	0.0094	0.0049	0.7821	0.3916	1.3025	0.0094	0.0049
		(0.0142)	(0.0487)	(0.3796)	(0.0016)	(0.0008)	(0.0142)	(0.0487)	(0.3796)	(0.0016)	(0.0008)
Nonlinear	DeepSurv	0.8070	0.1767	1.0760	0.0259	0.0133	0.8070	0.1767	1.0760	0.0259	0.0133
		(0.0014)	(0.0018)	(0.1702)	(0.0004)	(0.0002)	(0.0014)	(0.0018)	(0.1702)	(0.0004)	(0.0002)
	DeepSurv_I(Keya et al.)	0.7916	0.1548	1.4610	0.0176	0.0091	0.7994	0.1673	1.4660	0.0245	0.0126
		(0.0121)	(0.0176)	(0.7342)	(0.0088)	(0.0045)	(0.0069)	(0.0051)	(0.8459)	(0.0014)	(0.0008)
	DeepSurv_I(R&P)	0.8066	0.1736	1.0520	0.0245	0.0126	0.8086	0.1766	1.1210	0.0258	0.0132
		(0.0033)	(0.0087)	(0.1533)	(0.0039)	(0.0020)	(0.0015)	(0.0024)	(0.0964)	(0.0011)	(0.0005)
	DeepSurv_G(Keya et al.)	0.7964	0.1576	0.9420	0.0161	0.0083	0.8017	0.1655	1.0310	0.0201	0.0103
		(0.0117)	(0.0196)	(0.2229)	(0.0097)	(0.0050)	(0.0114)	(0.0182)	(0.2034)	(0.0091)	(0.0046)
	DeepSurv_G(R&P)	0.8054	0.1704	1.0420	0.0231	0.0119	**0.8086**	0.1766	1.1210	0.0258	0.0132
		(0.0039)	(0.0113)	(0.1463)	(0.0051)	(0.0026)	**(0.0015)**	(0.0024)	(0.0964)	(0.0011)	(0.0005)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.7804	0.1399	0.8440	0.0062	0.0033	0.7751	0.1357	**0.7400**	0.0037	0.0020
		(0.0119)	(0.0086)	(0.2581)	(0.0052)	(0.0026)	(0.0018)	(0.0002)	**(0.0671)**	(0.0001)	(4.1949e-05)
	FIDP	**0.8077**	**0.1228**	1.2500	0.0239	0.0118	0.8077	**0.1228**	1.2500	0.0239	0.0118
		(0.0022)	**(0.0019)**	(0.1186)	(0.0018)	(0.0009)	(0.0022)	**(0.0019)**	(0.1186)	(0.0018)	(0.0009)
	FIPNAM	0.7829	0.1810	0.9750	0.0251	0.0127	0.7829	0.1810	0.9750	0.0251	0.0127
		(0.0037)	(0.0050)	(0.0246)	(0.0006)	(0.0004)	(0.0037)	(0.0050)	(0.0246)	(0.0006)	(0.0004)
	Deep DRO-COX	0.7699	0.1336	**0.4870**	0.0010	0.0006	0.7781	0.1331	0.9050	0.0001	0.0001
		(0.0147)	(0.0004)	**(0.2540)**	(0.0008)	(0.0004)	(0.0091)	(0.0002)	(0.2372)	(3.1257e-05)	(2.4246e-05)
	Deep DRO-COX (SPLIT)	0.7784	0.1647	1.0500	**0**	**0**	0.7784	0.1647	1.0500	**0**	**0**
		(0.0092)	(0.0037)	(0.3409)	**(0)**	**(0)**	(0.0092)	(0.0037)	(0.3409)	**(0)**	**(0)**
	Deep EXACT DRO-COX	0.8048	0.1363	0.9660	0.0197	0.0102	0.8048	0.1363	0.9660	0.0197	0.0102
		(0.0011)	(0.0016)	(0.1395)	(0.0005)	(0.0002)	(0.0011)	(0.0016)	(0.1395)	(0.0005)	(0.0002)

Table H.2: Cox model test set scores on the SUPPORT (age) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	**0.6025** (**0.0005**)	0.2304 (0.0015)	2.2240 (0.1078)	0.0054 (0.0002)	0.0023 (0.0001)	**0.6025** (**0.0005**)	0.2304 (0.0015)	2.2240 (0.1078)	0.0054 (0.0002)	0.0023 (0.0001)
	Cox_I(Keya et al.)	0.5820 (0.0116)	**0.2153** (**0.0076**)	1.3120 (0.7623)	0.0001 (0.0003)	3.6626e-05 (0.0001)	0.5829 (0.0099)	**0.2147** (**0.0063**)	1.3600 (0.8532)	**0** (**0**)	**0** (**0**)
	Cox_I(R&P)	0.6019 (0.0008)	0.2310 (0.0014)	2.1220 (0.2937)	0.0056 (0.0003)	0.0024 (0.0001)	0.6017 (0.0012)	0.2308 (0.0013)	2.2370 (0.2882)	0.0055 (0.0003)	0.0023 (0.0001)
	Cox_G(Keya et al.)	0.5875 (0.0013)	0.2315 (0.0014)	2.2030 (0.0986)	0.0045 (0.0002)	0.0023 (0.0001)	0.5862 (0.0009)	0.2292 (0.0010)	2.2070 (0.0958)	0.0038 (0.0001)	0.0020 (0.0001)
	Cox_G(R&P)	0.6019 (0.0008)	0.2310 (0.0014)	2.1220 (0.2937)	0.0056 (0.0003)	0.0024 (0.0001)	0.6017 (0.0012)	0.2308 (0.0013)	2.2370 (0.2882)	0.0055 (0.0003)	0.0023 (0.0001)
	$\text{Cox}_\cap$(Keya et al.)	0.5664 (0.0061)	0.2273 (0.0016)	2.8030 (0.2551)	0.0027 (0.0003)	0.0014 (0.0002)	0.5631 (0.0070)	0.2264 (0.0017)	2.8350 (0.2498)	0.0024 (0.0003)	0.0013 (0.0002)
	DRO-COX	0.5722 (0.0031)	0.2210 (0.0010)	1.8310 (0.2546)	0.0001 (0.0001)	0.0001 (2.5317e-05)	0.5641 (0.0105)	0.2211 (0.0010)	1.8490 (0.6025)	0.0001 (0.0001)	4.7704e-05 (3.9221e-05)
	DRO-COX (SPLIT)	0.5701 (0.0056)	0.4569 (0.1314)	1.7240 (0.3998)	**1.1922e-07** (**2.6445e-07**)	**3.2988e-08** (**8.8067e-08**)	0.5701 (0.0056)	0.4570 (0.1314)	1.7210 (0.3977)	1.1922e-07 (2.6445e-07)	3.2988e-08 (8.8067e-08)
	EXACT DRO-COX	0.5884 (0.0063)	0.3122 (0.0068)	**0.8540** (**0.3189**)	8.1822e-06 (8.1542e-06)	5.1031e-06 (4.9357e-06)	0.5884 (0.0063)	0.3122 (0.0068)	**0.8540** (**0.3189**)	8.1822e-06 (8.1542e-06)	5.1031e-06 (4.9357e-06)
Nonlinear	DeepSurv	0.6108 (0.0029)	0.2417 (0.0016)	2.1170 (0.2107)	0.0090 (0.0002)	0.0041 (0.0001)	**0.6108** (**0.0029**)	0.2417 (0.0016)	2.1170 (0.2107)	0.0090 (0.0002)	0.0041 (0.0001)
	DeepSurv_I(Keya et al.)	0.5950 (0.0116)	0.2316 (0.0188)	1.6330 (0.5036)	0.0048 (0.0041)	0.0021 (0.0018)	0.6031 (0.0059)	0.2459 (0.0102)	1.8950 (0.6473)	0.0090 (0.0007)	0.0040 (0.0003)
	DeepSurv_I(R&P)	0.6034 (0.0089)	0.2334 (0.0078)	2.0940 (0.4228)	0.0063 (0.0027)	0.0028 (0.0013)	**0.6115** (**0.0051**)	0.2444 (0.0036)	1.9370 (0.5165)	0.0097 (0.0009)	0.0044 (0.0003)
	DeepSurv_G(Keya et al.)	0.5869 (0.0122)	0.2372 (0.0131)	1.6760 (0.4326)	0.0062 (0.0045)	0.0031 (0.0021)	0.5966 (0.0048)	0.2543 (0.0032)	1.9710 (0.4498)	0.0117 (0.0006)	0.0057 (0.0003)
	DeepSurv_G(R&P)	0.6039 (0.0094)	0.2329 (0.0074)	2.0890 (0.4199)	0.0061 (0.0025)	0.0027 (0.0012)	0.6115 (0.0051)	0.2444 (0.0036)	1.9370 (0.5165)	0.0097 (0.0009)	0.0044 (0.0003)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.5979 (0.0063)	0.2345 (0.0036)	2.4300 (0.2338)	0.0055 (0.0012)	0.0028 (0.0006)	0.5912 (0.0012)	0.2309 (0.0011)	2.4750 (0.1695)	0.0043 (0.0002)	0.0022 (0.0001)
	FIDP	0.5811 (0.0090)	0.2356 (0.0023)	1.4920 (0.3806)	0.0059 (0.0005)	0.0027 (0.0003)	0.5811 (0.0090)	0.2356 (0.0023)	1.4920 (0.3806)	0.0059 (0.0005)	0.0027 (0.0003)
	FIPNAM	0.5760 (0.0039)	0.2330 (0.0005)	2.1960 (0.1062)	0.0021 (0.0001)	0.0008 (0.0001)	0.5760 (0.0039)	0.2330 (0.0005)	2.1960 (0.1062)	0.0021 (0.0001)	0.0008 (0.0001)
	Deep DRO-COX	0.5833 (0.0088)	**0.2231** (**0.0015**)	**0.7590** (**0.3395**)	0.0012 (0.0004)	0.0006 (0.0002)	0.5754 (0.0120)	**0.2227** (**0.0011**)	**0.8240** (**0.3554**)	0.0010 (0.0005)	0.0005 (0.0003)
	Deep DRO-COX (SPLIT)	0.5772 (0.0093)	0.6387 (0.0007)	0.8660 (0.3260)	**0** (**0**)	**0** (**0**)	0.5772 (0.0093)	0.6387 (0.0007)	0.8660 (0.3260)	**0** (**0**)	**0** (**0**)
	Deep EXACT DRO-COX	0.5811 (0.0065)	0.2621 (0.0098)	1.7720 (0.7390)	0.0062 (0.0020)	0.0031 (0.0009)	0.5811 (0.0065)	0.2621 (0.0098)	1.7720 (0.7390)	0.0062 (0.0020)	0.0031 (0.0009)

Table H.3: Cox model test set scores on the SUPPORT (race) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	**0.6025**	0.2304	1.4160	0.0054	0.0034	**0.6025**	0.2304	1.4160	0.0054	0.0034
		(0.0005)	(0.0015)	(0.0696)	(0.0002)	(0.0001)	**(0.0005)**	(0.0015)	(0.0696)	(0.0002)	(0.0001)
	Cox_I(Keya et al.)	0.5905	**0.2161**	1.1230	0.0005	0.0003	0.5829	**0.2147**	1.1820	**0**	**0**
		(0.0086)	**(0.0054)**	(0.6621)	(0.0005)	(0.0004)	(0.0099)	**(0.0063)**	(0.5238)	**(0)**	**(0)**
	Cox_I(R&P)	0.6023	0.2309	1.3590	0.0055	0.0035	0.6024	0.2307	1.3800	0.0055	0.0035
		(0.0009)	(0.0012)	(0.1882)	(0.0003)	(0.0002)	(0.0007)	(0.0012)	(0.1880)	(0.0002)	(0.0001)
	Cox_G(Keya et al.)	0.6013	0.2282	1.3610	0.0047	0.0030	0.6011	0.2279	1.3610	0.0046	0.0029
		(0.0008)	(0.0017)	(0.0647)	(0.0003)	(0.0002)	(0.0006)	(0.0009)	(0.0650)	(0.0001)	(0.0001)
	Cox_G(R&P)	0.6023	0.2309	1.3590	0.0055	0.0035	0.6024	0.2307	1.3800	0.0055	0.0035
		(0.0009)	(0.0012)	(0.1882)	(0.0003)	(0.0002)	(0.0007)	(0.0012)	(0.1880)	(0.0002)	(0.0001)
	$Cox_\cap$(Keya et al.)	0.5681	0.2271	1.4020	0.0027	0.0017	0.5631	0.2264	1.3670	0.0024	0.0015
		(0.0079)	(0.0018)	(0.1743)	(0.0004)	(0.0003)	(0.0070)	(0.0017)	(0.1406)	(0.0003)	(0.0002)
	DRO-COX	0.5735	0.2210	**0.4640**	0.0002	0.0001	0.5641	0.2211	0.6660	0.0001	0.0001
		(0.0018)	(0.0010)	**(0.0790)**	(2.4047e-05)	(1.6044e-05)	(0.0105)	(0.0010)	(0.3208)	(0.0001)	(0.0001)
	DRO-COX (SPLIT)	0.5701	0.4569	0.6450	**1.1922e-07**	**1.0222e-07**	0.5701	0.4570	0.6440	**0**	**0**
		(0.0056)	(0.1314)	(0.3222)	**(2.6445e-07)**	**(2.2769e-07)**	(0.0056)	(0.1314)	(0.3228)	**(0)**	**(0)**
	EXACT DRO-COX	0.5884	0.3122	0.6010	8.1822e-06	6.3444e-06	0.5884	0.3122	**0.6010**	8.1822e-06	6.3444e-06
		(0.0063)	(0.0068)	(0.2146)	(8.1542e-06)	(6.1862e-06)	(0.0063)	(0.0068)	**(0.2146)**	(8.1542e-06)	(6.1862e-06)
Nonlinear	DeepSurv	**0.6108**	0.2417	1.7440	0.0090	0.0056	0.6108	0.2417	1.7440	0.0090	0.0056
		(0.0029)	(0.0016)	(0.2649)	(0.0002)	(0.0001)	(0.0029)	(0.0016)	(0.2649)	(0.0002)	(0.0001)
	$DeepSurv_I$(Keya et al.)	0.5927	0.2316	1.0380	0.0044	0.0029	0.6031	0.2459	1.2450	0.0090	0.0058
		(0.0082)	(0.0166)	(0.5996)	(0.0039)	(0.0025)	(0.0059)	(0.0102)	(0.7264)	(0.0007)	(0.0004)
	$DeepSurv_I$(R&P)	0.6087	0.2379	1.4840	0.0076	0.0047	0.6115	0.2444	1.4410	0.0097	0.0060
		(0.0083)	(0.0081)	(0.3203)	(0.0024)	(0.0014)	(0.0051)	(0.0036)	(0.4472)	(0.0009)	(0.0005)
	$DeepSurv_G$(Keya et al.)	0.5941	0.2369	1.2780	0.0068	0.0043	0.6056	0.2485	1.4490	0.0107	0.0066
		(0.0145)	(0.0117)	(0.3894)	(0.0041)	(0.0025)	(0.0044)	(0.0023)	(0.4958)	(0.0005)	(0.0003)
	$DeepSurv_G$(R&P)	0.6103	0.2393	1.4420	0.0081	0.0050	**0.6115**	0.2444	1.4410	0.0097	0.0060
		(0.0075)	(0.0075)	(0.3373)	(0.0021)	(0.0012)	**(0.0051)**	(0.0036)	(0.4472)	(0.0009)	(0.0005)
	$DeepSurv_\cap$(Keya et al.)	0.5992	0.2357	1.4230	0.0059	0.0037	0.5912	0.2309	1.1590	0.0043	0.0028
		(0.0072)	(0.0042)	(0.4286)	(0.0015)	(0.0009)	(0.0012)	(0.0011)	(0.1338)	(0.0002)	(0.0001)
	FIDP	0.5811	0.2356	0.9400	0.0059	0.0033	0.5811	0.2356	0.9400	0.0059	0.0033
		(0.0090)	(0.0023)	(0.3875)	(0.0005)	(0.0004)	(0.0090)	(0.0023)	(0.3875)	(0.0005)	(0.0004)
	FIPNAM	0.5760	0.2330	1.0380	0.0021	0.0011	0.5760	0.2330	1.0380	0.0021	0.0011
		(0.0039)	(0.0005)	(0.0519)	(0.0001)	(0.0001)	(0.0039)	(0.0005)	(0.0519)	(0.0001)	(0.0001)
	Deep DRO-COX	0.5798	**0.2234**	0.7900	0.0015	0.0009	0.5754	**0.2227**	0.7140	0.0010	0.0006
		(0.0101)	**(0.0017)**	(0.4283)	(0.0007)	(0.0004)	(0.0120)	**(0.0011)**	(0.4094)	(0.0005)	(0.0003)
	Deep DRO-COX (SPLIT)	0.5772	0.6387	**0.7100**	**0**	**0**	0.5772	0.6387	**0.7100**	**0**	**0**
		(0.0093)	(0.0007)	**(0.4386)**	**(0)**	**(0)**	(0.0093)	(0.0007)	**(0.4386)**	**(0)**	**(0)**
	Deep EXACT DRO-COX	0.5811	0.2621	0.7600	0.0062	0.0038	0.5811	0.2621	0.7600	0.0062	0.0038
		(0.0065)	(0.0098)	(0.4098)	(0.0020)	(0.0013)	(0.0065)	(0.0098)	(0.4098)	(0.0020)	(0.0013)

Table H.4: Cox model test set scores on the SEER (age) dataset, in the same format as Table 2.

	Methods	CI-based Tuning					F_{CG}-based Tuning				
		Accuracy Metrics		Fairness Metrics			Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$	$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	Cox	0.7025 (0.0003)	0.2128 (0.0009)	0.9420 (0.0271)	0.0256 (0.0006)	0.0063 (0.0001)	0.7025 (0.0003)	0.2128 (0.0009)	0.9420 (0.0271)	0.0256 (0.0006)	0.0063 (0.0001)
Linear	Cox_I(Keya et al.)	0.6911 (0.0049)	0.1910 (0.0041)	0.6840 (0.4589)	0.0114 (0.0044)	0.0026 (0.0011)	0.6877 (0.0065)	**0.1838 (0.0027)**	0.7570 (0.4637)	0.0005 (0.0002)	3.2960e-06 (4.6576e-06)
Linear	Cox_I(R&P)	0.7032 (0.0011)	0.2128 (0.0035)	1.0080 (0.0456)	0.0254 (0.0024)	0.0062 (0.0004)	0.7037 (0.0025)	0.2090 (0.0025)	0.9970 (0.0986)	0.0226 (0.0017)	0.0057 (0.0001)
Linear	Cox_G(Keya et al.)	0.6517 (0.0023)	0.1986 (0.0005)	3.2470 (0.0794)	0.0129 (0.0003)	0.0068 (0.0001)	0.6517 (0.0023)	0.1986 (0.0005)	3.2470 (0.0794)	0.0129 (0.0003)	0.0068 (0.0001)
Linear	Cox_G(R&P)	**0.7040 (0.0010)**	0.2107 (0.0039)	1.0220 (0.0564)	0.0238 (0.0026)	0.0061 (0.0005)	**0.7040 (0.0024)**	0.2083 (0.0015)	1.0090 (0.1030)	0.0222 (0.0012)	0.0057 (0.0001)
Linear	$\text{Cox}_\cap$(Keya et al.)	0.6494 (0.0016)	0.1963 (0.0012)	2.2630 (0.1127)	0.0107 (0.0010)	0.0056 (0.0005)	0.6494 (0.0016)	0.1963 (0.0012)	2.2630 (0.1127)	0.0107 (0.0010)	0.0056 (0.0005)
Linear	DRO-COX	0.6927 (0.0069)	**0.1868 (0.0004)**	0.6340 (0.2865)	**0 (0)**	**0 (0)**	0.6927 (0.0069)	0.1868 (0.0004)	0.6340 (0.2865)	**0 (0)**	**0 (0)**
Linear	DRO-COX (SPLIT)	0.6872 (0.0047)	0.1869 (0.0004)	**0.5010 (0.3020)**	**0 (0)**	**0 (0)**	0.6872 (0.0047)	0.1869 (0.0004)	**0.5010 (0.3020)**	**0 (0)**	**0 (0)**
Linear	EXACT DRO-COX	0.6833 (0.0060)	0.2422 (0.0044)	1.6980 (0.2181)	0.0056 (0.0005)	0.0026 (0.0001)	0.6833 (0.0060)	0.2422 (0.0044)	1.6980 (0.2181)	0.0056 (0.0005)	0.0026 (0.0001)
Nonlinear	DeepSurv	**0.7095 (0.0014)**	0.2200 (0.0012)	0.9800 (0.1702)	0.0309 (0.0006)	0.0094 (0.0003)	**0.7095 (0.0014)**	0.2200 (0.0012)	0.9800 (0.1702)	0.0309 (0.0006)	0.0094 (0.0003)
Nonlinear	DeepSurv_I(Keya et al.)	0.6985 (0.0041)	0.2123 (0.0035)	0.7700 (0.3175)	0.0289 (0.0014)	0.0082 (0.0003)	0.6982 (0.0045)	0.2127 (0.0032)	0.7640 (0.3397)	0.0291 (0.0014)	0.0082 (0.0004)
Nonlinear	DeepSurv_I(R&P)	0.7062 (0.0017)	0.2169 (0.0010)	0.7220 (0.1638)	0.0289 (0.0005)	0.0078 (0.0004)	0.7059 (0.0012)	0.2169 (0.0011)	**0.6970 (0.1434)**	0.0289 (0.0005)	0.0077 (0.0002)
Nonlinear	DeepSurv_G(Keya et al.)	0.7076 (0.0015)	0.2397 (0.0822)	0.9810 (0.1653)	0.0234 (0.0060)	0.0077 (0.0020)	0.7076 (0.0015)	0.2397 (0.0822)	0.9810 (0.1653)	0.0234 (0.0060)	0.0077 (0.0020)
Nonlinear	DeepSurv_G(R&P)	0.7062 (0.0017)	0.2169 (0.0010)	**0.7220 (0.1638)**	0.0289 (0.0005)	0.0078 (0.0004)	0.7062 (0.0017)	0.2169 (0.0010)	0.7220 (0.1638)	0.0289 (0.0005)	0.0078 (0.0004)
Nonlinear	$\text{DeepSurv}_\cap$(Keya et al.)	0.6537 (0.0054)	0.1998 (0.0008)	2.0120 (0.1339)	0.0136 (0.0012)	0.0075 (0.0006)	0.6537 (0.0054)	0.1998 (0.0008)	2.0120 (0.1339)	0.0136 (0.0012)	0.0075 (0.0006)
Nonlinear	FIDP	0.7086 (0.0030)	0.1824 (0.0033)	1.1460 (0.2295)	0.0168 (0.0055)	0.0044 (0.0018)	0.7086 (0.0030)	0.1824 (0.0033)	1.1460 (0.2295)	0.0168 (0.0055)	0.0044 (0.0018)
Nonlinear	FIPNAM	0.7022 (0.0118)	0.2226 (0.0019)	0.8610 (0.1067)	0.0181 (0.0020)	0.0047 (0.0012)	0.7022 (0.0118)	0.2226 (0.0019)	0.8610 (0.1067)	0.0181 (0.0020)	0.0047 (0.0012)
Nonlinear	Deep DRO-COX	0.6830 (0.0050)	0.1869 (0.0004)	0.7250 (0.3413)	**5.3651e-06 (6.3580e-06)**	**3.6414e-06 (2.0671e-06)**	0.6830 (0.0050)	0.1869 (0.0004)	0.7250 (0.3413)	**5.3651e-06 (6.3580e-06)**	**3.6414e-06 (2.0671e-06)**
Nonlinear	Deep DRO-COX (SPLIT)	0.6829 (0.0049)	0.1881 (0.0012)	0.7700 (0.3233)	6.3123e-06 (7.2058e-06)	4.1674e-06 (2.1908e-06)	0.6829 (0.0049)	0.1881 (0.0012)	0.7700 (0.3233)	6.3123e-06 (7.2058e-06)	4.1674e-06 (2.1908e-06)
Nonlinear	Deep EXACT DRO-COX	0.7057 (0.0014)	**0.1597 (0.0003)**	0.9670 (0.1247)	0.0277 (0.0004)	0.0076 (0.0003)	0.7057 (0.0014)	**0.1597 (0.0003)**	0.9670 (0.1247)	0.0277 (0.0004)	0.0076 (0.0003)

Table H.5: Cox model test set individual and group fairness metrics on the FLC (age) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0964 (0.0006)	0.1912 (0.0012)	0.0964 (0.0006)	0.1912 (0.0012)
	Cox_I(Keya et al.)	0.0496 (0.0121)	0.1079 (0.0213)	0.0256 (0.0018)	0.0658 (0.0030)
	Cox_I(R&P)	0.0958 (0.0030)	0.1899 (0.0064)	0.0912 (0.0043)	0.1803 (0.0092)
	Cox_G(Keya et al.)	0.0526 (0.0167)	0.1033 (0.0321)	0.0353 (0.0109)	0.0716 (0.0236)
	Cox_G(R&P)	0.1033 (0.0030)	0.1899 (0.0064)	0.0951 (0.0006)	0.1726 (0.0010)
	$\text{Cox}_\cap$(Keya et al.)	0.0325 (0.0006)	0.0652 (0.0025)	0.0326 (0.0006)	0.0661 (0.0023)
	DRO-COX	0.0317 (0.0207)	0.0695 (0.0441)	**0 (0)**	0.0021 (0.0001)
	DRO-COX (SPLIT)	**0 (0)**	**0.0017 (0.0001)**	**0 (0)**	**0.0017 (0.0001)**
	EXACT DRO-COX	0.0094 (0.0016)	0.0019 (0.0003)	0.0094 (0.0016)	0.0019 (0.0003)
Nonlinear	DeepSurv	0.1001 (0.0018)	0.1922 (0.0010)	0.1001 (0.0018)	0.1922 (0.0010)
	DeepSurv_I(Keya et al.)	0.0486 (0.0288)	0.1020 (0.0505)	0.0931 (0.0024)	0.1758 (0.0106)
	DeepSurv_I(R&P)	0.0955 (0.0141)	0.1838 (0.0248)	0.1002 (0.0032)	0.1902 (0.0060)
	DeepSurv_G(Keya et al.)	0.0289 (0.0274)	0.0572 (0.0522)	0.0278 (0.0318)	0.0540 (0.0602)
	DeepSurv_G(R&P)	0.1028 (0.0141)	0.1839 (0.0249)	0.1076 (0.0031)	0.1902 (0.0060)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0158 (0.0003)	0.0421 (0.0007)	0.0158 (0.0003)	0.0421 (0.0007)
	FIDP	0.0899 (0.0064)	0.1630 (0.0108)	0.0899 (0.0064)	0.1630 (0.0108)
	FIPNAM	0.1071 (0.0031)	0.1802 (0.0026)	0.1071 (0.0031)	0.1802 (0.0026)
	Deep DRO-COX	0.0754 (0.0215)	0.1468 (0.0353)	0.0002 (0.0001)	0.0097 (0.0017)
	Deep DRO-COX (SPLIT)	**7.1710e-11 (2.1182e-10)**	**0.0015 (0.0003)**	**7.1710e-11 (2.1182e-10)**	**0.0015 (0.0003)**
	Deep EXACT DRO-COX	0.0197 (0.0005)	0.0038 (0.0001)	0.0197 (0.0005)	0.0038 (0.0001)

Table H.6: Cox model test set individual and group fairness metrics on the FLC (gender) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0964 (0.0006)	0.0167 (0.0015)	0.0964 (0.0006)	0.0167 (0.0015)
	Cox_I(Keya et al.)	0.0367 (0.0118)	0.0048 (0.0031)	0.0253 (0.0019)	0.0065 (0.0058)
	Cox_I(R&P)	0.0931 (0.0045)	0.0150 (0.0041)	0.0885 (0.0024)	0.0122 (0.0026)
	Cox_G(Keya et al.)	0.0784 (0.0119)	0.0127 (0.0047)	0.0708 (0.0005)	0.0097 (0.0011)
	Cox_G(R&P)	0.0997 (0.0045)	0.0141 (0.0034)	0.0951 (0.0006)	0.0117 (0.0018)
	$\text{Cox}_\cap$(Keya et al.)	0.0325 (0.0006)	0.0022 (0.0012)	0.0325 (0.0006)	0.0022 (0.0012)
	DRO-COX	0.0091 (0.0022)	0.0058 (0.0010)	**0 (0)**	0.0010 (0.0002)
	DRO-COX (SPLIT)	**0 (0)**	**0.0009 (0.0001)**	**0 (0)**	**0.0009 (0.0001)**
	EXACT DRO-COX	0.0094 (0.0016)	0.0049 (0.0008)	0.0094 (0.0016)	0.0049 (0.0008)
Nonlinear	DeepSurv	0.1001 (0.0018)	0.0186 (0.0014)	0.1001 (0.0018)	0.0186 (0.0014)
	DeepSurv_I(Keya et al.)	0.0676 (0.0329)	0.0093 (0.0068)	0.0932 (0.0025)	0.0166 (0.0063)
	DeepSurv_I(R&P)	0.0953 (0.0138)	0.0184 (0.0037)	0.1002 (0.0032)	0.0202 (0.0044)
	DeepSurv_G(Keya et al.)	0.0636 (0.0370)	0.0126 (0.0074)	0.0785 (0.0346)	0.0157 (0.0071)
	DeepSurv_G(R&P)	0.0977 (0.0182)	0.0180 (0.0035)	0.1076 (0.0031)	0.0202 (0.0044)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0265 (0.0219)	0.0025 (0.0029)	0.0158 (0.0003)	0.0015 (0.0005)
	FIDP	0.0899 (0.0064)	0.0125 (0.0057)	0.0899 (0.0064)	0.0125 (0.0057)
	FIPNAM	0.1071 (0.0031)	0.0154 (0.0017)	0.1071 (0.0031)	0.0154 (0.0017)
	Deep DRO-COX	0.0042 (0.0030)	0.0043 (0.0018)	0.0002 (0.0001)	0.0025 (0.0007)
	Deep DRO-COX (SPLIT)	**7.1710e-11 (2.1182e-10)**	**0.0008 (0.0001)**	**7.1710e-11 (2.1182e-10)**	**0.0008 (0.0001)**
	Deep EXACT DRO-COX	0.0197 (0.0005)	0.0102 (0.0002)	0.0197 (0.0005)	0.0102 (0.0002)

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, and L Joseph Melton III. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, pages 517–523. Elsevier, 2012.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 2022.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Table H.7: Cox model test set individual and group fairness metrics on the SUPPORT (age) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0461 (0.0020)	0.0498 (0.0021)	0.0461 (0.0020)	0.0498 (0.0021)
	Cox_I(Keya et al.)	0.0010 (0.0031)	0.0041 (0.0067)	**0 (0)**	0.0006 (0.0004)
	Cox_I(R&P)	0.0479 (0.0020)	0.0513 (0.0042)	0.0473 (0.0016)	0.0517 (0.0039)
	Cox_G(Keya et al.)	0.0409 (0.0019)	0.0031 (0.0013)	0.0350 (0.0009)	0.0031 (0.0008)
	Cox_G(R&P)	0.0639 (0.0021)	0.0513 (0.0042)	0.0632 (0.0016)	0.0517 (0.0039)
	$\text{Cox}_\cap$(Keya et al.)	0.0242 (0.0031)	0.0035 (0.0015)	0.0214 (0.0034)	0.0034 (0.0014)
	DRO-COX	0.0013 (0.0005)	0.0068 (0.0022)	0.0009 (0.0007)	0.0052 (0.0032)
	DRO-COX (SPLIT)	**1.8200e-06 (4.4671e-06)**	0.0020 (0.0012)	1.8199e-06 (4.4672e-06)	0.0020 (0.0012)
	EXACT DRO-COX	8.1822e-06 (8.1542e-06)	**5.1031e-06 (4.9357e-06)**	8.1822e-06 (8.1542e-06)	**5.1031e-06 (4.9357e-06)**
Nonlinear	DeepSurv	0.0674 (0.0012)	0.0548 (0.0033)	0.0674 (0.0012)	0.0548 (0.0033)
	DeepSurv_I(Keya et al.)	0.0417 (0.0348)	0.0407 (0.0266)	0.0759 (0.0050)	0.0631 (0.0125)
	DeepSurv_I(R&P)	0.0487 (0.0180)	0.0485 (0.0098)	0.0726 (0.0060)	0.0608 (0.0104)
	DeepSurv_G(Keya et al.)	0.0476 (0.0320)	0.0204 (0.0237)	0.0856 (0.0037)	0.0487 (0.0036)
	DeepSurv_G(R&P)	0.0636 (0.0178)	0.0491 (0.0099)	0.0892 (0.0061)	0.0608 (0.0104)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0470 (0.0078)	0.0048 (0.0034)	0.0390 (0.0015)	0.0024 (0.0010)
	FIDP	0.0456 (0.0043)	0.0197 (0.0087)	0.0456 (0.0043)	0.0197 (0.0087)
	FIPNAM	0.0190 (0.0011)	0.0190 (0.0015)	0.0190 (0.0011)	0.0190 (0.0015)
	Deep DRO-COX	0.0097 (0.0030)	0.0097 (0.0015)	0.0076 (0.0040)	0.0083 (0.0030)
	Deep DRO-COX (SPLIT)	**0 (0)**	**0.0006 (5.9147e-06)**	**0 (0)**	**0.0006 (5.9147e-06)**
	Deep EXACT DRO-COX	0.0062 (0.0020)	0.0031 (0.0009)	0.0062 (0.0020)	0.0031 (0.0009)

Table H.8: Cox model test set individual and group fairness metrics on the SUPPORT (gender) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0461 (0.0020)	0.0142 (0.0017)	0.0461 (0.0020)	0.0142 (0.0017)
	Cox_I(Keya et al.)	0.0042 (0.0042)	0.0026 (0.0020)	**0 (0)**	0.0003 (0.0001)
	Cox_I(R&P)	0.0475 (0.0010)	0.0145 (0.0055)	0.0470 (0.0014)	0.0137 (0.0034)
	Cox_G(Keya et al.)	0.0437 (0.0027)	0.0010 (0.0006)	0.0407 (0.0005)	0.0009 (0.0006)
	Cox_G(R&P)	0.0635 (0.0011)	0.0145 (0.0055)	0.0629 (0.0014)	0.0133 (0.0031)
	$\text{Cox}_\cap$(Keya et al.)	0.0256 (0.0034)	0.0038 (0.0012)	0.0214 (0.0034)	0.0031 (0.0008)
	DRO-COX	0.0015 (0.0002)	0.0008 (0.0005)	0.0009 (0.0007)	0.0007 (0.0006)
	DRO-COX (SPLIT)	**1.8200e-06 (4.4671e-06)**	0.0005 (0.0002)	1.8200e-06 (4.4671e-06)	0.0005 (0.0002)
	EXACT DRO-COX	8.1822e-06 (8.1542e-06)	**5.2437e-06 (5.0535e-06)**	8.1822e-06 (8.1542e-06)	**5.2437e-06 (5.0535e-06)**
Nonlinear	DeepSurv	0.0674 (0.0012)	0.0121 (0.0026)	0.0674 (0.0012)	0.0121 (0.0026)
	DeepSurv_I(Keya et al.)	0.0531 (0.0312)	0.0125 (0.0079)	0.0759 (0.0050)	0.0156 (0.0128)
	DeepSurv_I(R&P)	0.0604 (0.0151)	0.0123 (0.0055)	0.0726 (0.0060)	0.0106 (0.0070)
	DeepSurv_G(Keya et al.)	0.0666 (0.0256)	0.0024 (0.0021)	0.0816 (0.0016)	0.0018 (0.0019)
	DeepSurv_G(R&P)	0.0819 (0.0130)	0.0133 (0.0067)	0.0892 (0.0061)	0.0106 (0.0070)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0538 (0.0107)	0.0016 (0.0017)	0.0390 (0.0015)	0.0014 (0.0006)
	FIDP	0.0456 (0.0043)	0.0053 (0.0025)	0.0456 (0.0043)	0.0053 (0.0025)
	FIPNAM	0.0190 (0.0011)	0.0031 (0.0013)	0.0190 (0.0011)	0.0031 (0.0013)
	Deep DRO-COX	0.0148 (0.0044)	0.0018 (0.0005)	0.0076 (0.0040)	0.0018 (0.0009)
	Deep DRO-COX (SPLIT)	**0 (0)**	**0.0006 (3.0118e-06)**	**0 (0)**	**0.0006 (3.0118e-06)**
	Deep EXACT DRO-COX	0.0062 (0.0020)	0.0033 (0.0010)	0.0062 (0.0020)	0.0033 (0.0010)

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995.

Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-cal: Explicit calibration for survival analysis. *Advances in neural information processing systems*, 33:18296–18307, 2020.

Table H.9: Cox model test set individual and group fairness metrics on the SUPPORT (race) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0461 (0.0020)	0.0115 (0.0042)	0.0461 (0.0020)	0.0115 (0.0042)
	Cox_I(Keya et al.)	0.0050 (0.0052)	0.0053 (0.0072)	**0 (0)**	0.0008 (0.0005)
	Cox_I(R&P)	0.0475 (0.0016)	0.0118 (0.0061)	0.0470 (0.0014)	0.0116 (0.0047)
	Cox_G(Keya et al.)	0.0404 (0.0024)	0.0011 (0.0006)	0.0396 (0.0006)	0.0010 (0.0006)
	Cox_G(R&P)	0.0634 (0.0017)	0.0118 (0.0061)	0.0629 (0.0014)	0.0116 (0.0047)
	$\text{Cox}_\cap$(Keya et al.)	0.0242 (0.0038)	0.0045 (0.0016)	0.0214 (0.0034)	0.0042 (0.0015)
	DRO-COX	0.0015 (0.0002)	0.0033 (0.0009)	0.0009 (0.0007)	0.0025 (0.0015)
	DRO-COX (SPLIT)	**1.8200e-06 (4.4671e-06)**	0.0017 (0.0008)	1.8199e-06 (4.4672e-06)	0.0017 (0.0008)
	EXACT DRO-COX	8.1822e-06 (8.1542e-06)	**6.3444e-06 (6.1862e-06)**	8.1822e-06 (8.1542e-06)	**6.3444e-06 (6.1862e-06)**
Nonlinear	DeepSurv	0.0674 (0.0012)	0.0126 (0.0053)	0.0674 (0.0012)	0.0126 (0.0053)
	DeepSurv_I(Keya et al.)	0.0126 (0.0053)	0.0230 (0.0208)	0.0759 (0.0050)	0.0440 (0.0160)
	DeepSurv_I(R&P)	0.0586 (0.0169)	0.0157 (0.0119)	0.0726 (0.0060)	0.0164 (0.0147)
	DeepSurv_G(Keya et al.)	0.0515 (0.0288)	0.0069 (0.0087)	0.0789 (0.0027)	0.0179 (0.0062)
	DeepSurv_G(R&P)	0.0784 (0.0154)	0.0174 (0.0137)	0.0892 (0.0061)	0.0164 (0.0147)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0495 (0.0103)	0.0095 (0.0040)	0.0390 (0.0015)	0.0110 (0.0011)
	FIDP	0.0456 (0.0043)	0.0154 (0.0083)	0.0456 (0.0043)	0.0154 (0.0083)
	FIPNAM	0.0190 (0.0011)	0.0055 (0.0028)	0.0190 (0.0011)	0.0055 (0.0028)
	Deep DRO-COX	0.0115 (0.0054)	0.0018 (0.0010)	0.0076 (0.0040)	0.0014 (0.0009)
	Deep DRO-COX (SPLIT)	**0 (0)**	**0.0012 (1.6940e-06)**	**0 (0)**	**0.0012 (1.6940e-06)**
	Deep EXACT DRO-COX	0.0062 (0.0020)	0.0038 (0.0013)	0.0062 (0.0020)	0.0038 (0.0013)

Table H.10: Cox model test set individual and group fairness metrics on the SEER (age) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0642 (0.0014)	0.1676 (0.0048)	0.0642 (0.0014)	0.1676 (0.0048)
	Cox_I(Keya et al.)	0.0289 (0.0110)	0.0923 (0.0268)	0.0012 (0.0004)	0.0229 (0.0032)
	Cox_I(R&P)	0.0640 (0.0055)	0.1670 (0.0156)	0.0574 (0.0037)	0.1495 (0.0139)
	Cox_G(Keya et al.)	0.0367 (0.0007)	0.0026 (0.0012)	0.0367 (0.0007)	0.0026 (0.0012)
	Cox_G(R&P)	0.0727 (0.0062)	0.1557 (0.0164)	0.0686 (0.0023)	0.1462 (0.0102)
	$\text{Cox}_\cap$(Keya et al.)	0.0298 (0.0029)	0.0046 (0.0012)	0.0298 (0.0029)	0.0046 (0.0012)
	DRO-COX	**0 (0)**	0.0025 (0.0013)	**0 (0)**	0.0025 (0.0013)
	DRO-COX (SPLIT)	**0 (0)**	**0.0018 (0.0002)**	**0 (0)**	**0.0018 (0.0002)**
	EXACT DRO-COX	0.0056 (0.0005)	0.0026 (0.0001)	0.0056 (0.0005)	0.0026 (0.0001)
Nonlinear	DeepSurv	0.0799 (0.0015)	0.1793 (0.0021)	0.0799 (0.0015)	0.1793 (0.0021)
	DeepSurv_I(Keya et al.)	0.0753 (0.0018)	0.1798 (0.0063)	0.0759 (0.0018)	0.1811 (0.0065)
	DeepSurv_I(R&P)	0.0737 (0.0014)	0.1796 (0.0056)	0.0736 (0.0014)	0.1807 (0.0036)
	DeepSurv_G(Keya et al.)	0.0607 (0.0154)	0.1289 (0.0328)	0.0607 (0.0154)	0.1289 (0.0328)
	DeepSurv_G(R&P)	0.0861 (0.0014)	0.1796 (0.0056)	0.0861 (0.0014)	0.1796 (0.0056)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0386 (0.0033)	**0.0058 (0.0041)**	0.0386 (0.0033)	**0.0058 (0.0041)**
	FIDP	0.0433 (0.0138)	0.0976 (0.0233)	0.0433 (0.0138)	0.0976 (0.0233)
	FIPNAM	0.0455 (0.0055)	0.1022 (0.0038)	0.0455 (0.0055)	0.1022 (0.0038)
	Deep DRO-COX	**1.4579e-05 (1.6833e-05)**	0.0061 (0.0009)	**1.4579e-05 (1.6833e-05)**	0.0061 (0.0009)
	Deep DRO-COX (SPLIT)	1.7208e-05 (1.8956e-05)	0.0063 (0.0010)	1.7208e-05 (1.8956e-05)	0.0063 (0.0010)
	Deep EXACT DRO-COX	0.0277 (0.0004)	0.0076 (0.0003)	0.0277 (0.0004)	0.0076 (0.0003)

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

Robert J Gray. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, pages 1141–1154, 1988.

Stefan Groha, Sebastian M Schmon, and Alexander Gusev. A general framework for survival analysis and multi-state modelling. *arXiv preprint arXiv:2006.04893*, 2020.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.*, 21(85):1–63, 2020.

Table H.11: Cox model test set individual and group fairness metrics on the SEER (race) dataset, in the same format as Table 2.

	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
Linear	Cox	0.0642 (0.0014)	0.0597 (0.0027)	0.0642 (0.0014)	0.0597 (0.0027)
	Cox_I(Keya et al.)	0.0011 (0.0002)	0.0072 (0.0015)	0.0011 (0.0002)	0.0072 (0.0015)
	Cox_I(R&P)	0.0595 (0.0050)	0.0612 (0.0135)	0.0586 (0.0048)	0.0634 (0.0110)
	Cox_G(Keya et al.)	0.0541 (0.0098)	0.0096 (0.0168)	0.0541 (0.0098)	0.0096 (0.0168)
	Cox_G(R&P)	0.0695 (0.0035)	0.0571 (0.0087)	0.0696 (0.0038)	0.0610 (0.0086)
	$\text{Cox}_\cap$(Keya et al.)	0.0298 (0.0029)	0.0024 (0.0010)	0.0298 (0.0029)	0.0024 (0.0010)
	DRO-COX	**0 (0)**	0.0008 (0.0006)	**0 (0)**	0.0008 (0.0006)
	DRO-COX (SPLIT)	**0 (0)**	**0.0003 (0.0001)**	**0 (0)**	**0.0003 (0.0001)**
	EXACT DRO-COX	0.0056 (0.0005)	0.0045 (0.0004)	0.0056 (0.0005)	0.0045 (0.0004)
Nonlinear	DeepSurv	0.0799 (0.0015)	0.0731 (0.0043)	0.0799 (0.0015)	0.0731 (0.0043)
	DeepSurv_I(Keya et al.)	0.0759 (0.0018)	0.0681 (0.0056)	0.0759 (0.0018)	0.0681 (0.0056)
	DeepSurv_I(R&P)	0.0736 (0.0014)	0.0681 (0.0101)	0.0736 (0.0014)	0.0681 (0.0101)
	DeepSurv_G(Keya et al.)	0.0697 (0.0028)	0.0111 (0.0039)	0.0697 (0.0028)	0.0111 (0.0039)
	DeepSurv_G(R&P)	0.0861 (0.0014)	0.0676 (0.0115)	0.0861 (0.0014)	0.0676 (0.0115)
	$\text{DeepSurv}_\cap$(Keya et al.)	0.0386 (0.0033)	**0.0023 (0.0013)**	0.0386 (0.0033)	**0.0023 (0.0013)**
	FIDP	0.0433 (0.0138)	0.0417 (0.0130)	0.0433 (0.0138)	0.0417 (0.0130)
	FIPNAM	0.0455 (0.0055)	0.0446 (0.0057)	0.0455 (0.0055)	0.0446 (0.0057)
	Deep DRO-COX	**1.4579e-05 (1.6833e-05)**	0.0035 (0.0006)	**1.4579e-05 (1.6833e-05)**	0.0035 (0.0006)
	Deep DRO-COX (SPLIT)	1.7208e-05 (1.8956e-05)	0.0035 (0.0005)	1.7208e-05 (1.8956e-05)	0.0035 (0.0005)
	Deep EXACT DRO-COX	0.0277 (0.0004)	0.0225 (0.0003)	0.0277 (0.0004)	0.0225 (0.0003)

Table H.12: DeepHit test set individual and group fairness on the FLC, SUPPORT, SEER datasets when hyperparameter tuning is based on CI and F_{CG}.

Datasets	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
FLC (age)	DeepHit	0.0330 (0.0049)	0.0721 (0.0078)	0.0330 (0.0049)	0.0721 (0.0078)
	DeepHit$_G$(R&P)	0.0289 (0.0089)	0.0615 (0.0193)	0.0173 (0.0006)	0.0331 (0.0012)
	DRO-DEEPHIT	0.0233 (0.0111)	0.0561 (0.0183)	**0.0022 (0.0012)**	0.0168 (0.0035)
	DRO-DEEPHIT (SPLIT)	**0.0177 (0.0140)**	**0.0426 (0.0252)**	0.0030 (0.0027)	**0.0160 (0.0051)**
FLC (gender)	DeepHit	0.0330 (0.0049)	0.0152 (0.0103)	0.0330 (0.0049)	0.0152 (0.0103)
	DeepHit$_G$(R&P)	0.0297 (0.0075)	0.0132 (0.0110)	0.0330 (0.0049)	0.0152 (0.0103)
	DRO-DEEPHIT	0.0233 (0.0111)	0.0134 (0.0107)	**0.0022 (0.0012)**	0.0049 (0.0020)
	DRO-DEEPHIT (SPLIT)	**0.0177 (0.0140)**	**0.0118 (0.0118)**	0.0030 (0.0027)	**0.0031 (0.0007)**
SUPPORT (age)	DeepHit	0.0187 (0.0049)	0.0200 (0.0047)	0.0187 (0.0049)	0.0200 (0.0047)
	DeepHit$_G$(R&P)	0.0139 (0.0036)	0.0082 (0.0052)	0.0125 (0.0010)	0.0067 (0.0018)
	DRO-DEEPHIT	**0.0013 (0.0024)**	**0.0054 (0.0025)**	**0.0001 (0.0001)**	**0.0031 (0.0011)**
	DRO-DEEPHIT (SPLIT)	0.0133 (0.0064)	0.0135 (0.0081)	0.0109 (0.0083)	0.0140 (0.0084)
SUPPORT (gender)	DeepHit	0.0187 (0.0049)	0.0101 (0.0054)	0.0187 (0.0049)	0.0101 (0.0054)
	DeepHit$_G$(R&P)	0.0128 (0.0010)	0.0038 (0.0016)	0.0122 (0.0006)	0.0039 (0.0019)
	DRO-DEEPHIT	**0.0013 (0.0024)**	**0.0026 (0.0014)**	**0.0001 (0.0001)**	**0.0012 (0.0005)**
	DRO-DEEPHIT (SPLIT)	0.0133 (0.0064)	0.0077 (0.0052)	0.0109 (0.0083)	0.0069 (0.0057)
SUPPORT (race)	DeepHit	0.0187 (0.0049)	0.0108 (0.0024)	0.0187 (0.0049)	0.0108 (0.0024)
	DeepHit$_G$(R&P)	0.0125 (0.0012)	0.0051 (0.0023)	0.0120 (0.0018)	0.0059 (0.0026)
	DRO-DEEPHIT	**0.0013 (0.0024)**	**0.0042 (0.0031)**	**0.0001 (0.0001)**	**0.0025 (0.0014)**
	DRO-DEEPHIT (SPLIT)	0.0133 (0.0064)	0.0137 (0.0044)	0.0109 (0.0083)	0.0110 (0.0041)
SEER (age)	DeepHit	0.0153 (0.0023)	0.0600 (0.0076)	0.0153 (0.0023)	0.0600 (0.0076)
	DeepHit$_G$(R&P)	0.0122 (0.0051)	0.0501 (0.0166)	0.0019 (0.0002)	0.0185 (0.0010)
	DRO-DEEPHIT	**0.0055 (0.0052)**	**0.0290 (0.0207)**	**0 (0)**	**0.0008 (0.0007)**
	DRO-DEEPHIT (SPLIT)	0.0124 (0.0031)	0.0473 (0.0135)	0.0106 (0.0049)	0.0438 (0.0176)
SEER (race)	DeepHit	0.0153 (0.0023)	0.0175 (0.0053)	0.0153 (0.0023)	0.0175 (0.0053)
	DeepHit$_G$(R&P)	0.0138 (0.0049)	0.0170 (0.0059)	0.0022 (0.0036)	0.0092 (0.0022)
	DRO-DEEPHIT	**0.0055 (0.0052)**	**0.0085 (0.0066)**	**0 (0)**	**0.0008 (0.0014)**
	DRO-DEEPHIT (SPLIT)	0.0124 (0.0031)	0.0208 (0.0065)	0.0106 (0.0049)	0.0208 (0.0057)

Frank E Harrell, Robert M Califf, and David B Pryor. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Shu Hu and George H Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR, 2022.

Table H.13: SODEN test set individual and group fairness on the FLC, SUPPORT, SEER datasets when hyperparameter tuning is based on CI and F_{CG}.

Datasets	Methods	CI-based Tuning		F_{CG}-based Tuning	
		$F_I\downarrow$	$F_G\downarrow$	$F_I\downarrow$	$F_G\downarrow$
FLC (age)	SODEN	0.0019 (0.0037)	0.0101 (0.0103)	0.0019 (0.0037)	0.0101 (0.0103)
	SODEN$_G$(R&P)	0.0046 (0.0037)	0.0153 (0.0066)	0.0004 (0.0007)	0.0051 (0.0035)
	DRO-SODEN	**0.0001 (0.0003)**	**0.0046 (0.0024)**	**2.2814e-05 (5.7901e-05)**	**0.0034 (0.0017)**
FLC (gender)	SODEN	**0.0019 (0.0037)**	**0.0017 (0.0008)**	0.0019 (0.0037)	0.0017 (0.0008)
	SODEN$_G$(R&P)	0.0031 (0.0036)	0.0027 (0.0017)	0.0004 (0.0007)	0.0013 (0.0007)
	DRO-SODEN	0.0041 (0.0112)	0.0056 (0.0115)	**2.2814e-05 (5.7901e-05)**	**0.0012 (0.0006)**
SUPPORT (age)	SODEN	0.0606 (0.0051)	0.0585 (0.0035)	0.0606 (0.0051)	0.0585 (0.0035)
	SODEN$_G$(R&P)	0.0472 (0.0103)	0.0419 (0.0117)	0.0384 (0.0061)	**0.0323 (0.0071)**
	DRO-SODEN	**0.0361 (0.0201)**	**0.0373 (0.0177)**	**0.0332 (0.0193)**	0.0345 (0.0169)
SUPPORT (gender)	SODEN	0.0604 (0.0047)	0.0114 (0.0049)	0.0604 (0.0047)	0.0114 (0.0049)
	SODEN$_G$(R&P)	0.0575 (0.0061)	**0.0085 (0.0046)**	0.0387 (0.0061)	0.0058 (0.0036)
	DRO-SODEN	**0.0503 (0.0107)**	0.0087 (0.0040)	**0.0255 (0.0084)**	**0.0044 (0.0019)**
SUPPORT (race)	SODEN	0.0604 (0.0047)	0.0266 (0.0147)	0.0604 (0.0047)	0.0266 (0.0147)
	SODEN$_G$(R&P)	0.0482 (0.0107)	0.0204 (0.0077)	0.0387 (0.0061)	0.0164 (0.0055)
	DRO-SODEN	**0.0404 (0.0175)**	**0.0165 (0.0147)**	**0.0255 (0.0084)**	**0.0099 (0.0088)**
SEER (age)	SODEN	0.0714 (0.0029)	0.1725 (0.0112)	0.0714 (0.0029)	0.1725 (0.0112)
	SODEN$_G$(R&P)	0.0706 (0.0027)	0.1701 (0.0079)	0.0702 (0.0028)	0.1726 (0.0111)
	DRO-SODEN	**0.0596 (0.0126)**	**0.1294 (0.0413)**	**0.0367 (0.0187)**	**0.0736 (0.0442)**
SEER (race)	SODEN	0.0714 (0.0029)	**0.0606 (0.0134)**	0.0714 (0.0029)	0.0606 (0.0134)
	SODEN$_G$(R&P)	0.0698 (0.0028)	0.0668 (0.0140)	0.0692 (0.0030)	0.0664 (0.0126)
	DRO-SODEN	**0.0453 (0.0137)**	0.0606 (0.0139)	**0.0348 (0.0158)**	**0.0559 (0.0210)**

Table H.14: Test set scores for DRO-COX (SPLIT) on the FLC (age) dataset using $n_2 = 0.1n, 0.2n, 0.3n, 0.4n, 0.5n$ (corresponding to $n_1 = 0.9n, 0.8n, 0.7n, 0.6n, 0.5n$). The format of this table is similar to that of Table 2 although here we do not bold or highlight any cells, as our main finding here is that the scores are not dramatically different for the different choices for n_1 or n_2.

	n_2	Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS$\downarrow$	CI(%)$\downarrow$	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	$0.1n$	0.7822 (0.0183)	0.1410 (0.0056)	0.4670 (0.3846)	0 (0)	0 (0)
	$0.2n$	0.7945 (0.0069)	0.1402 (0.0029)	0.3610 (0.2667)	0 (0)	0 (0)
	$0.3n$	0.7970 (0.0037)	0.1397 (0.0025)	0.2560 (0.1559)	0 (0)	0 (0)
	$0.4n$	0.7970 (0.0043)	0.1392 (0.0015)	0.2940 (0.1387)	0 (0)	0 (0)
	$0.5n$	0.7964 (0.0045)	0.1389 (0.0008)	0.2350 (0.1277)	0 (0)	0 (0)
Nonlinear	$0.1n$	0.7583 (0.0109)	0.1907 (0.0764)	2.1490 (1.0704)	1.8664e-04 (5.5992e-04)	3.8323e-05 (1.1497e-04)
	$0.2n$	0.7712 (0.0107)	0.1622 (0.0095)	2.2640 (0.7685)	2.4905e-05 (7.4715e-05)	5.2623e-06 (1.5787e-05)
	$0.3n$	0.7709 (0.0205)	0.1650 (0.0025)	2.3830 (0.4080)	0 (0)	0 (0)
	$0.4n$	0.7731 (0.0178)	0.1633 (0.0057)	2.3860 (0.2411)	1.0570e-07 (3.1711e-07)	6.5156e-08 (1.9547e-07)
	$0.5n$	0.7784 (0.0092)	0.1647 (0.0037)	2.3210 (0.3590)	0 (0)	0 (0)

Shu Hu, Yiming Ying, and Siwei Lyu. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 2020.

Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. TkML-AP: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7649–7657, 2021.

Shu Hu, Xin Wang, and Siwei Lyu. Rank-based decomposable losses in machine learning: A survey. *arXiv preprint arXiv:2207.08768*, 2022a.

Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu. Sum of ranked range loss for supervised learning. *Journal of Machine Learning Research*, 23(112):1–44, 2022b.

John D Kalbfleisch and Ross L Prentice. *The Statistical Analysis of Failure Time Data (2nd ed)*. John Wiley & Sons, 2002.

Table H.15: Test set scores for DRO-COX (SPLIT) on the FLC (age) dataset using censoring rate imbalance ratios (abbreviated below as just "Ratio") of 0%, 20%, 40%, 60%, 80%, and 100%. The format of this table is similar to that of Table 2.

	Ratio (%)	Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS↓	CI(%)↓	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	0	0.7964 (0.0045)	0.1389 (0.0008)	0.2350 (0.1277)	0 (0)	0 (0)
	20	0.7977 (0.0053)	0.1393 (0.0009)	0.1520 (0.0846)	0 (0)	0 (0)
	40	0.7990 (0.0066)	0.1402 (0.0017)	0.2200 (0.1239)	0 (0)	0 (0)
	60	0.7965 (0.0070)	0.1410 (0.0022)	0.2810 (0.1840)	0 (0)	0 (0)
	80	0.7935 (0.0074)	0.1454 (0.0075)	0.5660 (0.3148)	0 (0)	0 (0)
	100	0.7929 (0.0102)	0.1341 (0.0006)	0.3970 (0.2665)	0 (0)	0 (0)
Nonlinear	0	0.7784 (0.0092)	0.1647 (0.0037)	2.3210 (0.3590)	0 (0)	0 (0)
	20	0.7734 (0.0188)	0.1647 (0.0048)	2.4380 (0.3981)	5.7227e-08 (1.7168e-07)	3.9546e-08 (1.1864e-07)
	40	0.7753 (0.0187)	0.1677 (0.0028)	2.1080 (0.5425)	0.0001 (0.0002)	1.3683e-05 (4.1049e-05)
	60	0.7853 (0.0120)	0.1700 (0.0002)	1.1180 (0.6116)	0 (0)	0 (0)
	80	0.7900 (0.0122)	0.1703 (0.0002)	0.6790 (0.5659)	0 (0)	0 (0)
	100	0.7577 (0.0059)	0.1646 (0.0021)	0.6390 (0.4295)	0 (0)	0 (0)

Table H.16: Test set scores of DRO-COX (SPLIT, ONE SIDE) vs DRO-COX (SPLIT) on the FLC (age) dataset. The format of this table is the same that of Table 2 except without any cells highlighted in green as we are not comparing against baselines by previous authors.

	Methods	Accuracy Metrics		Fairness Metrics		
		$C^{td}\uparrow$	IBS↓	CI(%)↓	$F_{CI}\downarrow$	$F_{CG}\downarrow$
Linear	DRO-COX (SPLIT, ONE SIDE)	0.7810 (0.0109)	**0.1330 (0.0002)**	0.4060 (0.2847)	**0 (0)**	**0 (0)**
	DRO-COX (SPLIT)	**0.7964 (0.0045)**	0.1389 (0.0008)	**0.2350 (0.1277)**	**0 (0)**	**0 (0)**
Non-linear	Deep DRO-COX (SPLIT, ONE SIDE)	0.7554 (0.0231)	**0.1332 (0.0002)**	**1.9000 (0.6850)**	1.6544e-04 (1.2172e-04)	4.0388e-05 (3.1972e-05)
	Deep DRO-COX (SPLIT)	**0.7784 (0.0092)**	0.1647 (0.0037)	2.3210 (0.3590)	**0 (0)**	**0 (0)**

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James Foulds. Equitable allocation of healthcare resources with fair survival models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 190–198. SIAM, 2021.

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

John P Klein and Melvin L Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, and Norman Desbiens. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

Havard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30, 2019.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. Evaluating model performance under worst-case subpopulations. *Advances in Neural Information Processing Systems*, 2021.

Intae Moon, Stefan Groha, and Alexander Gusev. Survlatent ode: A neural ode based time-to-event model with competing risks for longitudinal data improves cancer-associated venous thromboembolism (vte) prediction. In *Machine Learning for Healthcare Conference*, 2022.

Md Mahmudur Rahman and Sanjay Purushotham. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1452–1462, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Raphael Sonabend, Florian Pfisterer, Alan Mishler, Moritz Schauer, Lukas Burk, and Sebastian Vollmer. Flexible group fairness metrics for survival analysis. *arXiv preprint arXiv:2206.03256*, 2022.

Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. *Advances in Neural Information Processing Systems*, 2007.

Weijing Tang, Kevin He, Gongjun Xu, and Ji Zhu. Survival analysis via ordinary differential equations. *Journal of the American Statistical Association*, pages 1–16, 2022a.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *J. Mach. Learn. Res.*, 23: 34–1, 2022b.

Jing Teng. SEER breast cancer data. *IEEE Dataport*, 2019. URL `https://dx.doi.org/10.21227/a9qy-ph35`.

Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

Zhiliang Wu, Yinchong Yang, Peter A Fashing, and Volker Tresp. Uncertainty-aware time-to-event prediction using deep kernel accelerated failure time models. In *Machine Learning for Healthcare Conference*, pages 54–79. PMLR, 2021.

Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.