Distributionally Robust Learning in Survival Analysis

Yeping Jin YEPJIN@BU.EDU
Boston University, USA

Lauren Wise LWISE@BU.EDU
Boston University, USA

Ioannis Ch. Paschalidis YANNISP@BU.EDU
Boston University, USA

Abstract

We introduce an innovative approach that incorporates a *Distributionally Robust Learning (DRL)* approach into Cox regression to enhance the robustness and accuracy of survival predictions. By formulating a DRL framework with a Wasserstein distance-based ambiguity set, we develop a variant Cox model that is less sensitive to assumptions about the underlying data distribution and more resilient to model misspecification and data perturbations. By leveraging Wasserstein duality, we reformulate the original min-max DRL problem into a tractable regularized empirical risk minimization problem, which can be computed by exponential conic programming. We provide guarantees on the finite sample behavior of our DRL-Cox model. Moreover, through extensive simulations and real world case studies, we demonstrate that our regression model achieves superior performance in terms of prediction accuracy and robustness compared with traditional methods.

Data and Code Availability A portion of the data we used is publicly available and is included in the code below. The remaining data is sourced from the Pregnancy Study Online (PRESTO), a web-based preconception cohort study (Wise et al., 2015). Due to privacy considerations, these data cannot be shared publicly, as PRESTO participants did not provide informed consent for external data sharing. Our code is available at https://github.com/noc-lab/drl_cox.

Institutional Review Board (IRB) The research was conducted using de-identified data and was exempt from IRB approval.

1. Introduction

Survival analysis is a cornerstone in the field of medical research, providing insights into the time until an event of interest, such as death or disease recurrence. The Cox proportional hazards model (Cox, 1972) has become a foundational tool in survival analysis, allowing researchers to assess the impact of various covariates on survival time without the need to specify the underlying survival distribution explicitly. Despite its widespread adoption, the classical Cox model and its extensions often assume a known or fixed distribution of covariates and model parameters. This assumption may result in significant inaccuracies in survival predictions when faced with real-world data complexities, such as departure from proportional hazards, data heterogeneity, and the presence of outliers.

Recent advancements in optimization and statistical learning have introduced the concept of *Distributionally Robust Learning (DRL)*, a methodology that seeks to improve model performance under distributional uncertainty. DRL achieves this by optimizing worst-case performance across a set of possible distributions defined within an ambiguity set, thus offering a more resilient approach to modeling under uncertainty. The integration of DRL into survival analysis, particularly the Cox regression model, presents an opportunity to address some of the inherent limitations of traditional survival analysis methods by enhancing their robustness against uncertainties in the data distribution.

1.1. Related Works

In recent years, DRL has undergone extensive development and become a powerful tool for enhancing model robustness in the presence of uncertainty and distributional shifts. DRL is designed to minimize worst-case loss over a set of possible distributions, rather than relying on a single empirical distribution. Some of the earlier works (Ben-Tal et al., 2009; Delage and Ye, 2010; Goh and Sim, 2010) laid the foundation for broader applications in data-driven optimization, with extensions into machine learning (Chen and Paschalidis, 2018; Esfahani and Kuhn, 2018; Gao and Kleywegt, 2023). These works developed a comprehensive DRL framework under the Wasserstein metric and applied this framework onto various topics, such as nearest-neighbor estimation, semi-supervised learning and reinforcement learning.

More recent developments have integrated DRL into more complex settings such as deep learning and survival analysis. For example, Esfahani and Kuhn (2018) expanded DRL to settings with ambiguous data, optimizing models to handle uncertainties in real-world applications. Moreover, Sotudian et al. (2023) introduced DRL into *Learning-To-Rank (LTR)* methods and constructed a new robust LTR model which addresses the robustness issue by using the Wasserstein framework, effectively mitigating the impact of noise, adversarial perturbations, and data contamination in ranking tasks. These advancements demonstrate the growing importance of DRL in creating more reliable, robust models, which has been widely applied into fields where variability and data noise can significantly affect model performance, such as finance (Zhu and Fukushima, 2009; Glasserman and Xu, 2012) and healthcare (Mo et al., 2021; Hao et al., 2023).

On the other hand, model robustness has been a central focus in the field of survival analysis for several decades, with researchers continually developing methods to improve the resilience of survival models against challenges like outliers, censoring, and distributional shifts. Over the years, techniques such as penalized Cox (Verweij et al., 1994) and adaptive Lasso Cox (Zhang and Lu, 2007) have been introduced to enhance the robustness of Cox models, particularly in medical datasets where prediction accuracy is critical for patient outcomes. These advancements aim to ensure that survival models not only perform well under ideal conditions but also maintain reliability in the presence of real-world data variability.

However, despite the wide application of DRL in various fields, its utilization in survival analysis has been relatively rare. To our knowledge, there is only one series of existing work (Hu and Chen, 2022) in this area, which used a KL divergence framework (Duchi and Namkoong, 2021) to apply DRL on the Cox regression model. Although their approach demonstrated improvements in fairness metrics, it does not guarantee robustness under outliers and perturbations when compared to the standard Cox model. This highlights a major gap in the literature, as existing methods do not fully address the challenges posed by noisy or highly variable datasets, which are common in real-world applications such as healthcare. Therefore, there is a pressing need for further exploration of DRL-based Cox models that not only improve fairness but also ensure robustness to data irregularities, making survival predictions more reliable and applicable to broader domains. Addressing this gap could lead to substantial advancements in fields where robust survival predictions are critical, including personalized medicine and risk assessment.

1.2. Main Contribution

This paper aims to bridge the gap between robust optimization techniques and survival analysis by proposing a DRL-enhanced Cox regression model with a Wasserstein-distance induced ambiguity set. Our contributions are twofold. First, we introduce a novel approach to survival analysis that leverages the strengths of DRL to mitigate the effects of uncertain data distributions, thereby improving the reliability and accuracy of survival estimates. By adjusting the Cox loss function to adapt the DRL framework, we establish a well-defined min-max stochastic formulation for the DRL-Cox model. We then use Wasserstein duality to derive an exponential conic program as a tractable relaxation, where equality is reached if the feature space $\mathcal{X}$ coincides with $\mathbb{R}^d$. Furthermore, we carry out a stochastic analysis of the DRL-Cox model. We show that, with high probability, the optimal DRL loss acts as an upper bound for the true expected loss, highlighting the model's favorable behavior in finite sample scenarios. This suggests that the model maintains robustness even in the presence of data contamination or noise. Second, through a series of simulations and real-world applications, we empirically demonstrate the advantages of our proposed model over classical Cox regression techniques as well as penalized Cox models, highlighting its potential to provide more robust and accurate predictions in various settings. The numerical experiments demonstrate that our DRL-Cox model exceeds the two prior Cox models by more than 5% in terms of predictive accuracy. This improvement highlights the effectiveness of incorporating DRL in survival analysis, enhancing the model's ability to handle distributional shifts and outliers more effectively than standard Cox models.

The remainder of this paper is organized as follows: Sec. 2 introduces background material and presents a basic formulation of the DRL problem. Sec. 3 presents our main theoretical results, including a tractable form of the DRL-enhanced Cox model and the associated optimization strategy. After proving the main theorem, we also investigate the new model's finite sample performance. Sec. 4 details the experimental setup, data description, and presents results from our empirical analysis. Finally, Sec. 5 presents the implications of our findings, potential limitations, and directions for future research.

Notational Conventions: Boldfaced lowercase letters denote vectors, boldface upper case letters denote matrices, and ordinary lowercase letters denote scalars. Given any integer $N > 0$, we use $[\![N]\!]$ to denote $\{1, \dots, N\}$. We use $\mathbb{R}$ to represent the set of real numbers, and $\mathbb{R}^+$ the set of non-negative real numbers. All vectors are column vectors and prime denotes transpose. For space-saving reasons, we write $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. Given $p \geq 1$ and $\mathbf{x} \in \mathbb{R}^d$, $\|x\|_p := (\sum_{i\in[\![d]\!]} \mathbf{x}_i^p)^{1/p}$ denotes the ℓ_p norm. $\mathbb{E}_{\mathbb{P}}$ denotes the expectation under a probability distribution $\mathbb{P}$. For a dataset $\mathcal{D} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, we use $\hat{\mathbb{P}}_N$ to denote the empirical measure supported on $\mathcal{D}$. In particular, $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{z}_i}$, where $\delta_{\mathbf{z}_i}$ denotes the Dirac delta function at point $\mathbf{z}_i$.

2. Background Material

2.1. Cox Regression Model

The Cox regression model estimates a hazard function, representing the instantaneous risk of an event occurring at a given time, while accounting for the effects of covariates. In our setting, the training dataset is

$$\mathcal{D} = \{\mathbf{z}_i : i \in [\![N]\!]\} = \{(\mathbf{x}_i, y_i, \zeta_i) : i \in [\![N]\!]\},$$

where each data point $\mathbf{z}_i \in \mathcal{D}$ has feature vectors $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, observed duration time $y_i \in \mathcal{Y} \subset \mathbb{R}^+$, and event indicator $\zeta_i \in \{0, 1\}$. If $\zeta_i = 1$, then the event of interest occurs after time $t = y_i$. Otherwise, $\zeta_i = 0$ indicates that the event of interest did not occur, and data are censored after time $t = y_i$. In this case, the true duration time of point $\mathbf{z}_i$ is unknown, but we have a lower bound $t \geq y_i$ instead. Standard Cox regression predicts the hazard rate with the model

$$h_{\boldsymbol{\beta}}(\mathbf{x}, t) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}},$$

where one computes an optimal $\boldsymbol{\beta}$ to minimize the following loss:

$$L(\mathbf{x}, \mathbf{y}, \boldsymbol{\zeta}, \boldsymbol{\beta}) = \sum_{i=1}^N \zeta_i \left[\log\left(\sum_{j:y_j\geq y_i} e^{\boldsymbol{\beta}'\mathbf{x}_j}\right) - \boldsymbol{\beta}'\mathbf{x}_i\right], \tag{1}$$

where $\mathbf{y} = (y_1, \dots, y_N)$ and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_N)$. The above loss function sums up the individual losses of every uncensored data point. For each individual loss, we compute the log-sum-exponential function of all data points that have longer duration times. It is noteworthy that L is convex with regards to $\mathbf{x}$ and $\boldsymbol{\beta}$, but not with regards to $\mathbf{y}$, which is one of the main challenges for DRL, as we will explain in Sec. 2.3.

2.2. DRL formulation for the Cox Model

DRL considers a range of possible distributions contained within a defined uncertainty set, also known as the ambiguity set. Defining the ambiguity set is crucial, leading to the development of several distinct frameworks. This paper adopts the approach of Chen and Paschalidis (2020), wherein the ambiguity set is characterized using the Wasserstein distance. The Wasserstein distance serves as a metric between probability distributions as follows.

Definition 1 (Wasserstein Distance) *Consider two Polish (i.e., complete, separable, metric) probability spaces $(\mathcal{V}_1, \mathbb{P})$ and $(\mathcal{V}_2, \mathbb{Q})$ and a lower-semicontinuous cost function $s : \mathcal{V}_1 \times \mathcal{V}_2 \to \mathbb{R} \cup \{\infty\}$. Then, the (order-1) Wasserstein distance can be defined as:*

$$W_s(\mathbb{P}, \mathbb{Q}) = \min_{\boldsymbol{\pi}} \int_{\mathcal{V}_1\times\mathcal{V}_2} s(\mathbf{v}_1, \mathbf{v}_2) d\boldsymbol{\pi}(\mathbf{v}_1, \mathbf{v}_2),$$

where $\boldsymbol{\pi} \in \mathcal{P}(\mathcal{V}_1 \times \mathcal{V}_2)$ is a joint probability distribution of $\mathbf{v}_1, \mathbf{v}_2$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.

The basic formulation of a DRL Cox problem can be expressed as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \sup_{\mathbb{P}\in\Omega_\epsilon} \mathbb{E}_{\mathbb{P}}[l(\mathbf{x}, y, \zeta, \boldsymbol{\beta})], \tag{2}$$

where $\boldsymbol{\beta}$ are the coefficients associated with the covariates $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, $(\mathbf{x}, y, \zeta)$ represents the uncertain data point with probability distribution $\mathbb{P}$, and $l(\mathbf{x}, y, \zeta, \boldsymbol{\beta})$ is the individual loss function of a single data point in Cox regression, which will be determined in the next subsection. In (2), $\Omega_\epsilon := \{\mathbb{P} : W_s(\mathbb{P}, \mathbb{P}_0) \leq \epsilon\}$ is the set of all probability distributions that are within a Wasserstein ball (induced by the metric s) of radius ϵ, centered at a nominal distribution $\mathbb{P}_0$, which will be taken to be the empirical distribution $\hat{\mathbb{P}}_N$ induced by the training data set. By construction, (2) ensures that the optimization solution is robust against training distributions within the ambiguity set.

2.3. Choice of Loss Function

One of the principal challenges encountered in this study lies in the inherent structure of the standard Cox loss function. The DRL model requires that the loss function depends only on individual data points, but the individual loss as delineated in (1) does not adhere to this prerequisite. Sample splitting (Hu and Chen, 2022), as an alternative strategy, has been proposed to address this limitation. Despite its efficacy in enhancing fairness, this methodology does not yield superior accuracy when subjected to data contamination, in comparison to analyses performed with the original Cox model.

As a novel approach, this study proposes treating all training data as fixed constants for the computation of the individual loss terms. Consequently, the individual loss function is redefined as follows:

$$l(\mathbf{x}, y, \zeta, \boldsymbol{\beta}) = \zeta\left[\log\left(e^{\boldsymbol{\beta}'\mathbf{x}} + \sum_{j: y_j \geq y} e^{\boldsymbol{\beta}'\mathbf{x}_j}\right) - \boldsymbol{\beta}'\mathbf{x}\right]. \quad (3)$$

Compared to the standard Cox individual loss in (1), we introduce an additional exponential term in (3). This modification is intentional and aligns with the structure of the Cox proportional hazards model, where the partial likelihood inherently depends on pairwise comparisons across the dataset. By incorporating this term, we ensure that the loss function remains well-defined under the DRL framework in (2) while preserving its theoretical consistency with the proportional hazards assumption. Additionally, this adjustment guarantees the non-negativity of the loss for data points with long durations, providing stability in optimization. We acknowledge that this design deviates from the standard Cox loss; however, as demonstrated in our numerical experiments, this trade-off contributes to the robust performance of our approach. With this foundation, we now present our main result.

3. Formulation of a DRL-Cox Model

3.1. Tractable form of DRL-Cox

Given N training data points $\{(\mathbf{x}_i, y_i, \zeta_i)\}_{i \in [\![N]\!]}$, we consider the empirical distribution they induce, which will serve as $\mathbb{P}_0$, the center of ambiguity set Ω_ϵ. Plugging the individual loss (3) into (2), we obtain the following stochastic program:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \sup_{\mathbb{P} \in \Omega_\epsilon} \mathbb{E}_{(\mathbf{x}, y, \zeta) \sim \mathbb{P}}\left[\zeta\left(\log\left(e^{\boldsymbol{\beta}'\mathbf{x}} + \sum_{i: y_i \geq y} e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - \boldsymbol{\beta}'\mathbf{x}\right)\right]. \quad (4)$$

Evidently, Equation (4) presents challenges in terms of direct solvability due to its inherent complexity. Consequently, it becomes imperative to identify a tractable form that facilitates its estimation. To this end, we introduce the main theorem of this study, which posits the following convex program as an upper bound to Equation (4), and we prove the theorem in two phases. Initially, a strong dual formulation is constructed for the inner supremum problem, followed by discretizing the non-convex $y-$direction. Subsequently, through the application of the dual norm and the convex conjugate, a relaxation of the supremum problem is attained. This process culminates in the computation of the convex conjugate, which yields a tractable form of the original problem.

Theorem 2 (Relaxation of DRL-Cox) *Suppose the training data points $z_1, \ldots, z_N$ are bounded and sorted in decreasing order with regard to duration y, and the Wasserstein distance is induced by the ℓ_p norm. Let (p, q) be Hölder conjugates, so that $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then the following exponential conic program provides an upper bound for (4):*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \quad \epsilon\|(\boldsymbol{\beta}, \alpha)\|_q + \frac{1}{N}\sum_{i=1}^{N} \zeta_i s_i \quad (5)$$

$$\text{s.t. } s_i \geq \log\left(e^{\boldsymbol{\beta}'\mathbf{x}_i} + \sum_{i=1}^{k} e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - \boldsymbol{\beta}'\mathbf{x}_i - \alpha\,(y_i - y_k), \quad \forall\, 1 \leq i \leq k \leq N.$$

Moreover, if the covariate space satisfies $\mathcal{X} = \mathbb{R}^d$, then the optimal cost of (5) becomes equal to the value of the stochastic program (4).

Proof We first introduce the following corollary to construct a strong dual for the inner supremum of (4).

Corollary 3 *([Gao and Kleywegt](#), [2023](#), Cor.2) Suppose that we use the empirical distribution*

$$\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{z}_i},$$

as the center of ambiguity set Ω_ϵ and a Wasserstein distance induced by the ℓ_p norm, where $\mathbf{z}_i$, $i \in [\![N]\!]$, are the observed realizations of $\mathbf{z}$. Then the primal problem $v_P = \sup_{\mathbb{Q}\in\Omega_\epsilon} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{z},\boldsymbol{\beta})]$ has a strong dual

$$v_P = \min_{\lambda\geq 0}\left\{\lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N}\sup_{\mathbf{z}\in\mathcal{Z}}[l(\mathbf{z},\boldsymbol{\beta}) - \lambda\|\mathbf{z}-\mathbf{z}_i\|_p]\right\}. \quad (6)$$

Recall $\mathbf{z} = (\mathbf{x}, y, \zeta)$. Then the loss function becomes

$$l(\mathbf{z},\boldsymbol{\beta}) = \zeta\left(\log\left(e^{\boldsymbol{\beta}'\mathbf{x}} + \sum_{i:y_i\geq y} e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - \boldsymbol{\beta}'\mathbf{x}\right), \quad (7)$$

and let $\mathcal{Z}$ the space of $\mathbf{z}$'s. For each $i \in [\![N]\!]$, we examine the ith supremum term in the right hand side of (6):

$$\sup_{\mathbf{z}\in\mathcal{Z}}[l(\mathbf{z},\boldsymbol{\beta}) - \lambda\|\mathbf{z}-\mathbf{z}_i\|_p]. \quad (8)$$

If $\zeta_i = 0$, $l(\mathbf{z},\boldsymbol{\beta})$ becomes 0 by (7), so that the supremum term in (8) also becomes 0 at the optimal solution $\mathbf{z}^* = \mathbf{z}_i$. Therefore, we may set the ith supremum term in the right hand side of (6) to be $\zeta_i s_i$, where

$$s_i := \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \{l(\mathbf{x},y) - \lambda\|(\mathbf{x}-\mathbf{x}_i, y-y_i)\|_p\}. \quad (9)$$

Note that in (9), we simplify the loss function $l(\mathbf{x}, y, \zeta, \boldsymbol{\beta})$ to $l(\mathbf{x}, y)$ because ζ_i has been extracted, and $\boldsymbol{\beta}$ remains fixed within the supremum term:

$$\begin{aligned} l(\mathbf{x}, y) :&= l(\mathbf{x}, y, 1, \boldsymbol{\beta}) \\ &= \log\left(e^{\boldsymbol{\beta}'\mathbf{x}} + \sum_{i:y_i\geq y} e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - \boldsymbol{\beta}'\mathbf{x}. \end{aligned}$$

Our current goal is to simplify (9). Although the utilization of duality could facilitate this process for convex loss functions $l(\mathbf{x}, y)$, the loss function l, as previously mentioned, exhibits a piece-wise constant nature with respect to the y-direction, rendering it non-convex. However, note that $l(\mathbf{x},\cdot)$ is monotonically decreasing, which implies that given fixed $\mathbf{x}$, the function

$$l(\mathbf{x}, y) - \lambda\|(\mathbf{x}-\mathbf{x}_i, y-y_i)\|_p$$

will similarly decrease for $y \geq y_i$. Consequently, the supremum s_i will be attained when $y^* \leq y_i$. In fact, given the piece-wise constant characteristic of $l(\mathbf{x},\cdot)$, the supremum will be achieved within the set $\{y_i, y_{i+1}, \ldots, y_n\}$. Therefore, we proceed to discretize y in the following manner:

$$s_i = \max_{i\leq j\leq N}\sup_{\mathbf{x}\in X}[l(\mathbf{x}, y_j) - \lambda\|(\mathbf{x}-\mathbf{x}_i, y_j - y_i)\|_p]. \quad (10)$$

Since the non-convex variable y has been discretized, we introduce two notational simplifications. For $j \in [\![N]\!]$, we may write the loss function as

$$l_j(\mathbf{x}) := l(\mathbf{x}, y_j), \quad (11)$$

and we may simplify the jth supremum term in s_i as

$$s_i^j := \sup_{\mathbf{x}\in X}[l_j(\mathbf{x}) - \lambda\|(\mathbf{x}-\mathbf{x}_i, y_j - y_i)\|_p]. \quad (12)$$

In order to evaluate the supremum in (10), we introduce the dual norm and the convex conjugate.

Definition 4

1. (Dual Norm) *Given a norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ is defined as:*

$$\|\boldsymbol{\theta}\|_* := \sup_{\|\mathbf{x}\|\leq 1} \boldsymbol{\theta}'\mathbf{x}. \quad (13)$$

2. (Convex Conjugate) *Given a convex function l, its convex conjugate $l^*(\cdot)$ is:*

$$l^*(\boldsymbol{\theta}) := \sup_{\mathbf{x}\in\text{dom } l}\{\boldsymbol{\theta}'\mathbf{x} - l(\mathbf{x})\}. \quad (14)$$

In particular, we have $l(\mathbf{x}) = \sup_{\boldsymbol{\theta}\in\Theta}[\boldsymbol{\theta}'\mathbf{x} - l^*(\boldsymbol{\theta})]$, where

$$\Theta := \{\boldsymbol{\theta} : l^*(\boldsymbol{\theta}) < \infty\} \quad (15)$$

is the effective domain of the convex conjugate l^*.

According to (13), the dual norm of ℓ_p norm is the ℓ_q norm for $\frac{1}{p} + \frac{1}{q} = 1$. On the other hand, since l_j is convex, it is equal to the convex conjugate of l_j^*. Therefore, we have the following relaxation for s_i^j:

$$\begin{aligned} s_i^j &= \sup_{\mathbf{x}\in X}\sup_{\boldsymbol{\theta}\in\Theta}[\boldsymbol{\theta}'\mathbf{x} - l_j^*(\boldsymbol{\theta}) - \lambda\|(\mathbf{x}-\mathbf{x}_i, y-y_i)\|_p] \\ &= \sup_{\mathbf{x}\in X}\sup_{\boldsymbol{\theta}\in\Theta}\inf_{\|(\mathbf{r},\alpha)\|_q\leq\lambda}[\boldsymbol{\theta}'\mathbf{x} - l_j^*(\boldsymbol{\theta}) + \mathbf{r}'(\mathbf{x}-\mathbf{x}_i) + \alpha(y-y_i)] \quad (16a) \\ &= \sup_{\boldsymbol{\theta}\in\Theta}\inf_{\|(\mathbf{r},\alpha)\|_q\leq\lambda}\sup_{\mathbf{x}\in X}[(\boldsymbol{\theta}+\mathbf{r})'\mathbf{x} - l_j^*(\boldsymbol{\theta}) - \mathbf{r}'\mathbf{x}_i + \alpha(y-y_i)] \quad (16b) \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta}\inf_{\|(\mathbf{r},\alpha)\|_q\leq\lambda}\sup_{\mathbf{x}\in\mathbb{R}^d}[(\boldsymbol{\theta}+\mathbf{r})'\mathbf{x} - l_j^*(\boldsymbol{\theta}) - \mathbf{r}'\mathbf{x}_i + \alpha(y-y_i)], \quad (16c) \end{aligned}$$

where (16a) follows from the definition (13), and (16b) is a direct result of the Minimax Theorem (von Neumann, 1928). The inner supremum over $\mathbf{x} \in \mathbb{R}^d$ in (16c) reaches ∞ unless $\mathbf{r} = -\boldsymbol{\theta}$. For all $j \in [\![N]\!]$, set

$$\kappa_j := \sup\{\|\boldsymbol{\theta}\|_q \colon l_j^*(\boldsymbol{\theta}) < \infty\}. \tag{17}$$

If $\kappa_j > \lambda$, according to (15), we can pick some $\boldsymbol{\theta} \in \Theta$ such that $\|\boldsymbol{\theta}\|_q > \lambda$, which makes the inner supremum attain ∞. Otherwise, $\kappa_j \leq \lambda$, and we cam take $\mathbf{r} = -\boldsymbol{\theta}$ to achieve the inner infimum. Therefore,

$$\begin{aligned} s_i^j &\leq \inf_{\|(-\boldsymbol{\theta},\alpha)\|_q \leq \lambda} \sup_{\boldsymbol{\theta} \in \Theta} [-l_j^*(\boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{x}_i + \alpha(y - y_i)] \\ &= \inf_{\|(\boldsymbol{\theta},\alpha)\|_q \leq \lambda} l_j(\mathbf{x}_i) - \alpha(y_i - y). \end{aligned} \tag{18}$$

Take $\kappa := \max_{j \in [\![N]\!]} \kappa_j$, then according to the relaxation in (18) and the Minimax Theorem (von Neumann, 1928), for all $\lambda \geq \kappa$, we obtain the following upper bound of (10):

$$\begin{aligned} s_i = \max_{i \leq j \leq N} s_i^j &\leq \max_{i \leq j \leq N} \inf_{\|(\boldsymbol{\theta},\alpha)\|_q \leq \lambda} l_j(\mathbf{x}_i) - \alpha(y_i - y_j) \\ &= \inf_{\|(\boldsymbol{\theta},\alpha)\|_q \leq \lambda} \max_{i \leq j \leq N} l_j(\mathbf{x}_i) - \alpha(y_i - y_j). \end{aligned}$$

Therefore, by taking $\lambda = \|(\kappa, \alpha)\|_q$, we conclude the following relaxation of the primal problem (6):

$$\begin{aligned} v_P &= \min_{\lambda \geq 0} \ \lambda\epsilon + \frac{1}{N}\sum_{i=1}^N \zeta_i s_i \\ &\leq \epsilon\|(\kappa,\alpha)\|_q + \frac{1}{N}\sum_{i=1}^N \zeta_i \max_{i \leq j \leq N}[l(\mathbf{x}_i, y_j) - \alpha(y_i - y_j)]. \end{aligned} \tag{19}$$

Lastly, we aim to compute κ, which is the maximum of (17). To that end, we consider the behavior of $l_j^*(\boldsymbol{\theta})$. Given fixed $\boldsymbol{\theta} \in \Theta$ and $j \in [\![N]\!]$, we denote $f(\mathbf{x}) = \boldsymbol{\theta}'\mathbf{x} - l_j(\mathbf{x})$, which implies $l_j^*(\boldsymbol{\theta}) = \sup_{\mathbf{x} \in \text{dom } l} f(\mathbf{x})$. To maximize $f(\mathbf{x})$, we compute its gradient and Hessian matrix:

$$\begin{aligned} \nabla f(\mathbf{x}) &= \boldsymbol{\theta} - \nabla\left(\log\left(e^{\boldsymbol{\beta}'\mathbf{x}} + \sum_{i=1}^j e^{\boldsymbol{\beta}'\mathbf{x}_i}\right) - \boldsymbol{\beta}'\mathbf{x}\right) \\ &= \boldsymbol{\theta} + \frac{C}{e^{\boldsymbol{\beta}'\mathbf{x}} + C}\boldsymbol{\beta}, \\ \nabla^2 f(\mathbf{x}) &= -\frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{(e^{\boldsymbol{\beta}'\mathbf{x}} + C)^2}\boldsymbol{\beta}\boldsymbol{\beta}', \ \text{where } C = \sum_{i=1}^j e^{\boldsymbol{\beta}'\mathbf{x}_i}. \end{aligned}$$

Since the Hessian is negative semi-definite, $f(\mathbf{x})$ will reach the maximum if and only if $\nabla f(\mathbf{x})$ reaches 0. Therefore, for all $\boldsymbol{\theta} \in \mathbb{R}^d$, $l_j^*(\boldsymbol{\theta}) < \infty$ if and only if exists $c \in (-1, 0)$ such that $\boldsymbol{\theta} = c\boldsymbol{\beta}$. Therefore, we obtain $\kappa_j = \|\boldsymbol{\beta}\|_* = \|\boldsymbol{\beta}\|_q$ for all j, and so is κ. Plugging $\kappa = \|\boldsymbol{\beta}\|_q$ into (19), we reach our desired relaxation of the stochastic program in (4):

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{P} \in \Omega_\epsilon} E_{\mathbb{P}}[l(\mathbf{x}, y, \zeta, \boldsymbol{\beta})] \leq \min_{\boldsymbol{\beta},\alpha} \epsilon\|(\boldsymbol{\beta},\alpha)\|_q + \frac{1}{N}\sum_{i=1}^N \zeta_i s_i,$$

where $s_i = \max_{i \leq j \leq N}[l(\mathbf{x}_i, y_j, 1, \boldsymbol{\beta}) - \alpha(y_i - y_j)]$. ∎

3.2. Performance Analysis

Next, we analyze the performance of the DRL-Cox model. Our goal is to derive stochastic guarantees that ensure the model's robustness and generalization capability. While there exists analysis (Esfahani and Kuhn, 2018) on the asymptotic consistency of DRL, our study primarily focuses on the finite sample performance, which is directly relevant to the numerical experiments presented in the following section.

Given the empirical distribution $\hat{\mathbb{P}}_N$ of N observed data points in the support of $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}$, we denote the true measure by $\mathbb{P}^*$. Recall that $\mathbb{P}^*$ is unknown since $\hat{\mathbb{P}}_N$ is contaminated. Therefore, we compute the DRL problem (2) using $\hat{\mathbb{P}}_N$ to implicitly optimize over the true measure that is included in the ambiguity set with high confidence. Let $\hat{J}_N$ and $\hat{\boldsymbol{\beta}}_N$ be the optimal cost and optimal solution to the DRL problem (2), respectively. We thus have:

$$\begin{aligned} \hat{J}_N &= \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \sup_{\mathbb{P} \in \Omega_\epsilon} \mathbb{E}_{\mathbb{P}}[l(\mathbf{x}, y, \zeta, \boldsymbol{\beta})] \\ &= \sup_{\mathbb{P} \in \Omega_\epsilon} \mathbb{E}_{\mathbb{P}}[l(\mathbf{x}, y, \zeta, \hat{\boldsymbol{\beta}}_N)], \end{aligned}$$

where $\Omega_\epsilon = \{\mathbb{P} : W_s(\mathbb{P}, \hat{\mathbb{P}}_N) \leq \epsilon\}$ is the Wasserstein ball defining the ambiguity set. To evaluate the behavior of the optimal solution $\hat{\boldsymbol{\beta}}_N$, we check if the expected loss $\mathbb{E}_{\mathbb{P}^*}[l(\mathbf{x}, y, \zeta, \hat{\boldsymbol{\beta}}_N)]$ is bounded above by the training loss $\hat{c}_N$. If this event occurs with high probability, we can assert that $\hat{\boldsymbol{\beta}}_N$ exhibits favorable finite sample behavior. In fact, we derive the following lemma, which follows from a previous work (Zhao and Guan, 2015, Prop. 3).

Lemma 5 (Performance Guarantee) *Suppose the data space $\mathcal{X} \times \mathcal{Y}$ is bounded and finite, and the true measure $\mathbb{P}^*$ is finite. Given any significance level $\alpha \in (0, 1)$, define the following threshold value:*

$$B(\alpha) := \sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p \colon \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{X} \times \mathcal{Y}\}\sqrt{\frac{\log(1/\alpha)}{N}}.$$

Then for all ambiguity sets with radii $\epsilon \geq B(\alpha)$, we have the following robustness guarantee:

$$\mathbb{P}(\mathbb{E}_{\mathbb{P}^*}[l(\mathbf{x}, y, \zeta, \hat{\boldsymbol{\beta}}_N)] \leq \hat{J}_N) \geq 1 - \alpha. \tag{20}$$

4. Experiments

In this section, we present numerical experiments assessing the performance of our DRL-Cox model. The goal is to evaluate its robustness under the variability and imperfections of real-world data.

To ensure a comprehensive evaluation, we benchmark DRL-Cox against multiple survival analysis models, including the regular Cox model, the Ridge Cox model (Verweij et al., 1994), the Lasso Cox model (Tibshirani, 1997), the Elastic Net Cox model (Simon et al., 2011), and the Sample-Splitting Cox model (Hu and Chen, 2022). Beyond Cox-based models, we further include the Accelerated Failure Time (AFT) model (Kalbfleisch and Prentice, 2002) and the Random Survival Forest (RSF) model (Ishwaran et al., 2008), ensuring a diverse set of methodologies for comparison.

For consistency, all models are evaluated using the widely recognized concordance index (C-index) (Harrell et al., 1982) and the Integrated Time-Dependent AUC (iAUC) (Heagerty and Zheng, 2005), ensuring a standardized and comprehensive metric for performance comparison.

Since DRL-Cox introduces robustness through distributional considerations, a key factor influencing its performance is the choice of the ambiguity radius ϵ. This parameter governs the balance between robustness and predictive accuracy. A small ambiguity radius results in an ambiguity set tightly concentrated around the empirical distribution, making the model nearly equivalent to the standard Cox model but with limited robustness to distributional shifts. Conversely, a large ambiguity radius increases conservatism, optimizing for the worst-case scenario over a broader range of distributions, which may lead to underestimation of risk scores and degraded performance in well-specified settings.

Recall that the performance guarantee in (20) provides a theoretically justified lower bound for the ambiguity radius: for any significance level α, choosing $\epsilon \geq B(\alpha)$ ensures that the expected loss under the true measure $\mathbb{P}^*$ does not exceed the empirical objective $\hat{J}_N$ with probability at least $1 - \alpha$. This threshold scales as $\mathcal{O}(\sqrt{\log(1/\alpha)/N})$, revealing a fundamental trade-off between robustness and sample efficiency. To refine the selection of ϵ, we leverage the measure concentration method (Chen and Paschalidis, 2020), which bounds the deviation between empirical and true expectations to determine a statistically justified range of ambiguity radii. This prevents overconservatism while maintaining robustness guarantees. To further fine-tune ϵ, we apply cross-validation techniques, systematically evaluating different radii on validation sets to empirically select the optimal value that minimizes out-of-sample loss while preserving robustness against distributional shifts. This combined approach ensures that DRL-Cox achieves strong generalization with high probability.

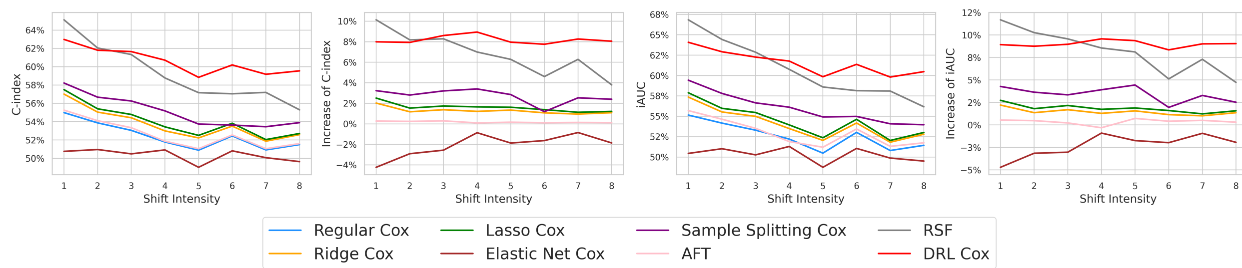
Meanwhile, it is also noteworthy that the convex program in Eq. (5) encompasses $O(N^2)$ constraints, a factor which significantly impedes computational efficiency. In practical applications, a more feasible approach involves retaining only the constraints associated with s_i in Eq. (5) for $i \leq k < i+\gamma$, thereby reducing the overall constraint count to $O(\gamma N)$. This modification not only enhances computational tractability but also maintains the integrity of the optimization process within a manageable scope. We then proceed to the following experiments with $\gamma = 3$. In our experiments, the average runtime per trial is approximately 30 seconds, indicating that while the method remains computationally demanding, it is still feasible for practical applications.

For the rest of this section, we test the model on two distinct groups of data with different contamination patterns.

Table 1: Summary Statistics of the WHAS500 Dataset (Goldberg et al., 2000).

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	y	ζ
mean	69.846	26.614	78.266	87.018	6.116	144.704	0.156	0.750	0.400	882.436	0.430
std	14.491	5.406	21.545	23.586	4.714	32.295	0.363	0.433	0.490	705.665	0.496
min	30.000	13.045	6.000	35.000	0.000	57.000	0.000	0.000	0.000	1.000	0.000
max	104.000	44.839	198.000	186.000	47.000	244.000	1.000	1.000	1.000	2358.000	1.000

Figure 1: Performance metrics under varying distributional shifts for different models.



(a) C-index

(b) C-index Increase Relative to Regular Cox

(c) iAUC

(d) iAUC Increase Relative to Regular Cox

4.1. WHAS500 Dataset with Covariate Distributional Shift

The WHAS500 dataset (Goldberg et al., 2000) contains data on 500 patients hospitalized for acute myocardial infarction (heart attack) in Worcester, Massachusetts. It is widely used in survival analysis for time-to-event modeling, with a moderate censoring rate of 57% and 14 features, ensuring balanced data utilization. The statistical summary of a subset of the features is presented in Table 1.

We now introduce a distributional shift in its covariate. We define eight levels of shift intensity, progressively replacing 1 to 8 of these features with values sampled from a normal distribution. For each intensity level, we conduct 50 trials and compute the average C-index and iAUC to evaluate the model's performance under distributional shifts. We visualize the comparison of the two metrics in Figure 1.

According to the results, DRL-Cox consistently outperforms all other models at high level of contamination, maintaining the highest C-index and iAUC values under the most intense shifts. The RSF model demonstrates moderate robustness, though it exhibits drastic decrease at intermediate and high shift levels. Meanwhile, the Sample-Splitting Cox model (Hu and Chen, 2022), despite also incorporating a DRL framework, does not achieve robustness in this scenario. This outcome is expected, as its design primarily focuses on improving fairness in survival analysis (Hu and Chen, 2024) rather than optimizing predictive performance under distributional shift.

Moreover, it is also noteworthy that neither of the two test scores exhibit a strictly proportional relationship with distributional shift severity. Unlike traditional regression models where increasing noise typically leads to a consistent decline in performance, survival analysis methods such as the Cox model rely on pairwise comparisons of relative risk rather than absolute predictions. This structure makes them more resilient to certain types of distributional shifts, as long as the relative ranking of risk scores remains stable. In certain instances, models demonstrate a transient performance recovery at intermediate shift levels, implying that moderate degrees of distributional perturbation may introduce noise that remains within a tolerable range or even facilitates enhanced generalization. Specifically, if the introduced perturbations disrupt spurious correlations or regularize the model implicitly, they may help maintain or even improve predictive stability. This non-monotonic behavior underscores the intricate interplay between feature distributional shifts and model adaptability, suggesting that the impact of shift severity on predictive performance is highly dependent on the underlying data structure and is not governed by a simple linear trend.

4.2. PRESTO Miscarriage Dataset with Outliers

In addition to the public dataset, we incorporate real-world datasets from previous studies on predictive modeling of miscarriage (Yland et al., 2022, 2024). This dataset is sourced from the *Pregnancy Study Online (PRESTO)*, a web-based preconception cohort study by Wise et al. (2015). The Miscarriage dataset spans from 2013 to 2022 and involves 8,739 females aged 21–45 in the U.S. and Canada. This dataset comprises 189 predictive variables, and the censor rate is 79.64%. We defined miscarriage as pregnancy loss before 20 completed weeks of gestation. In this

Table 2: Test scores for Miscarriage under different outlier severity levels with fixed outlier ratio (20%). The best method for each outlier level is highlighted in bold.

	Severity 1		Severity 2		Severity 3		Severity 4		Severity 5	
Method	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC
Regular Cox	0.5570	0.5727	0.5526	0.5675	0.5499	0.5657	0.5503	0.5644	0.5483	0.5646
Ridge Cox	0.5622	0.5782	0.5607	0.5770	0.5568	0.5728	0.5595	0.5752	0.5599	0.5777
Lasso Cox	0.5609	0.5713	0.5600	0.5650	0.5555	0.5637	0.5606	0.5598	0.5606	0.5609
Elastic Net Cox	0.5623	0.5703	**0.5657**	0.5748	0.5614	0.5710	0.5618	0.5664	0.5634	0.5628
Sample Splitting Cox	0.5328	0.5403	0.5359	0.5312	0.5159	0.5215	0.5085	0.5153	0.5077	0.5083
AFT	0.5565	0.5743	0.5511	0.5690	0.5483	0.5658	0.5461	0.5616	0.5464	0.5654
RSF	0.5587	0.5705	0.5545	0.5649	0.5618	0.5710	0.5662	0.5785	0.5626	0.5726
DRL-Cox	**0.5675**	**0.5819**	0.5656	**0.5812**	**0.5619**	**0.5770**	**0.5665**	**0.5808**	**0.5646**	**0.5803**

Table 3: Test scores for Miscarriage under different outlier ratios, with fixed outlier severity (level 3). The best method for each outlier ratio is highlighted in bold.

	5% Outliers		10% Outliers		15% Outliers		20% Outliers		25% Outliers		30% Outliers	
Method	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC	C-Index	iAUC
Regular Cox	0.5613	0.5789	0.5553	0.5706	0.5544	0.5683	0.5580	0.5741	0.5521	0.5690	0.5540	0.5700
Ridge Cox	0.5613	0.5770	0.5630	0.5788	0.5622	0.5771	0.5616	0.5779	0.5575	0.5751	0.5629	0.5805
Lasso Cox	0.5614	0.5680	0.5612	0.5625	0.5609	0.5650	0.5620	0.5739	0.5578	0.5674	0.5613	0.5671
Elastic Net Cox	0.5629	0.5717	0.5662	0.5764	0.5664	0.5768	0.5628	0.5719	0.5592	0.5693	0.5659	0.5724
Sample Splitting Cox	0.5351	0.5395	0.5186	0.5276	0.5361	0.5479	0.5220	0.5290	0.5204	0.5343	0.5245	0.5265
AFT	0.5548	0.5712	0.5543	0.5710	0.5538	0.5690	0.5559	0.5732	0.5512	0.5712	0.5521	0.5706
RSF	0.5639	0.5742	0.5582	0.5677	0.5648	0.5759	0.5611	0.5700	0.5615	0.5707	0.5603	0.5700
DRL-Cox	**0.5687**	**0.5834**	**0.5679**	**0.5823**	**0.5677**	**0.5815**	**0.5684**	**0.5840**	**0.5639**	**0.5800**	**0.5675**	**0.5840**

dataset, the time to event was calculated as the difference between the Exit week and the Start week (the gestational week at the time of enrollment), using weeks as the time scale.

Given the high dimensionality and censoring rate of this dataset, training is inherently challenging and prone to the curse of dimensionality (Donoho, 2000). Instead of introducing a distributional shift, we adopt a more conservative contamination approach by simulating outliers. Specifically, we contaminate the dataset by injecting outliers at varying ratios, ranging from 5% to 30%. To generate these outliers, we randomly select numerical features and perturb them with Gaussian noise of varying variance, thereby mimicking potential real-world perturbations encountered in survival analysis.

As we described above, our simulated noise depends on two parameters: the proportion of outliers in the dataset (outlier ratios) and the degree of noise introduced by the outliers (outlier severity). For each combination of these parameters, we implement five iterations and compare the averaged metrics for the benchmarks, illustrating how the model's performance is affected under different levels of outliers contamination. The results are shown in Table 2 and Table 3.

Comparing to the distributional shifted dataset, the Miscarriage dataset with outliers presents a different dynamic. While all models generally perform well, DRL-Cox exhibits a more pronounced improvement in C-index compared to the baseline Cox model, especially in scenarios involving higher outlier ratios and severity levels. The spikes in performance seen in the DRL-Cox model in these conditions indicate its ability to maintain robust performance even in the face of more irregular and complex data distributions, which are often present in miscarriage-related datasets.

Overall, the above two experiments offer a direct visualization of the relationship between varying levels of contamination and the model's predictive power, highlighting the superior robustness of the DRL-Cox model compared to its competitors. The results demonstrate that under intense level of contamination, the DRL-Cox model consistently outperforms all other (robust) Cox models, as well as the AFT Model and RSF model, offering greater resilience and predictive accuracy when handling real-world data irregularities, such as outliers and perturbations, as seen in both datasets.

5. Conclusion and Further Direction

In this article, we introduced a novel approach to enhance the robustness of Cox regression models through the application of DRL. By addressing inherent limitations in the standard Cox loss function, we have demonstrated the feasibility of deriving a more tractable form of the model. Our methodology, which leverages duality alongside the discretization of non-convex directions, culminates in a convex program that significantly enhances the model's robustness. The empirical evaluation reveals that the DRL-Cox model invariably outperforms other traditional survival methods.

In conclusion, this study's findings advocate for the integration of distributionally robust learning techniques in survival analysis, particularly in scenarios characterized by data uncertainty and noise. Despite the issue of computational complexity, the proposed DRL-Cox model represents an important step forward in developing more resilient and accurate predictive models.

On the other hand, our approach inherently relies on the convexity of the individual loss function with respect to the x-direction, making direct extensions to deep survival models non-trivial. However, there exists a potential approach to incorporate DRL-based constraints into the weight matrices of deep survival models such as DeepHit (Lee et al., 2018), ensuring that the learned representations remain stable under perturbations. Such an approach has proven effective in various deep learning applications, including classification tasks and medical imaging (Chen et al.; Hao et al., 2023), suggesting that DRL could also be integrated into deep survival learning frameworks. Future research could explore methods to adapt the DRL framework for deep survival models, potentially through structured approximations or alternative optimization techniques.

Additionally, further investigations into the bounds of ambiguity set adjustments and their implications on model performance across diverse datasets could yield deeper insights. Exploring alternative loss functions within the DRL framework may also lead to novel advancements in survival analysis and related fields. Through the advancements presented in this paper, we contribute to the broader discourse on the intersection of robust optimization and survival analysis, offering practical and theoretical insights that could pave the way for more sophisticated and resilient statistical models in medical research and beyond.

References

Ben-Tal, A., Ghaoui, L. El, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018.

Ruidi Chen and Ioannis Ch. Paschalidis. *Distributionally Robust Learning*. Now Publishers Inc., Hanover, MA, USA, 2020. doi: 10.1561/2400000026.

Ruidi Chen, Boran Hao, and Ioannis Ch. Paschalidis. Distributionally robust multiclass classification and applications in deep image classifiers. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP49357.2023.10095775.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 12 1972. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58:595–612, 06 2010. doi: 10.1287/opre.1090.0741.

David Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 01 2000.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49, 06 2021. doi: 10.1214/20-AOS2004.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1-2):115–166, 2018. doi: 10.1007/S10107-017-1172-1.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.*, 48(2):603–655, 2023. doi: 10.1287/MOOR.2022.1275.

Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Available at SSRN 2167765*, 14, 1 2012. doi: 10.2139/ssrn.2167765.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58:902–917, 08 2010. doi: 10.1287/opre.1090.0795.

Robert J. Goldberg, Jorge Yarzebski, Darleen Lessard, and Joel M. Gore. Decade-long trends and factors associated with time to hospital presentation in patients with acute myocardial infarction: The Worcester heart attack study. *Archives of Internal Medicine*, 160(21):3217–3223, 11 2000. ISSN 0003-9926. doi: 10.1001/archinte.160.21.3217.

Boran Hao, Guoyao Shen, Ruidi Chen, Chad W. Farris, Stephan W. Anderson, Xin Zhang, and Ioannis Ch. Paschalidis. Distributionally robust image classifiers for stroke diagnosis in accelerated MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 768–777, Cham, 2023. ISBN 978-3-031-43904-9.

Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030.

Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005. doi: 10.1111/j.0006-341X.2005.030814.x.

Shu Hu and George H. Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health, ML4H 2022*, volume 193, pages 62–87, 2022.

Shu Hu and George H. Chen. Fairness in survival analysis with distributionally robust optimization. *Journal of Machine Learning Research*, 25(246):1–85, 2024.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008. doi: 10.1214/08-AOAS169.

J. D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*. Wiley series in probability and statistics. J. Wiley, Hoboken, N.J, 2nd ed edition, 2002. ISBN 978-0-471-36357-6.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11842.

Mo, Weibin, Zhengling Qi, and Yufeng Liu. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674, 2021. doi: 10.1080/01621459.2020.1796359.

Noah R. Simon, Jerome H. Friedman, Trevor J. Hastie, and Robert Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39 5:1–13, 2011.

Shahabeddin Sotudian, Ruidi Chen, and Ioannis Ch Paschalidis. Distributionally robust learning-to-rank under the Wasserstein metric. *PLOS ONE*, 18 (3):e0283574, Mar 2023. doi: 10.1371/journal.pone.0283574.

Robert Tibshirani. The LASSO method for variable selection in the Cox model. *Statistics in medicine*, 16 4:385–95, 1997.

Verweij, Pierre J. M., Van Houwelingen, and Hans C. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436, 1994. doi: 10.1002/sim.4780132307.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Lauren A. Wise, Kenneth J. Rothman, Ellen M. Mikkelsen, Joseph B. Stanford, Amelia K. Wesselink, Craig McKinnon, Siobhan M. Gruschow, Casie E. Horgan, Aleta S. Wiley, Kristen A. Hahn, Henrik Toft Sorensen, and Elizabeth E. Hatch. Design and conduct of an internet-based preconception cohort study in north america: Pregnancy study online. *Paediatric and Perinatal Epidemiology*, 29 (4):360–371, 2015. doi: 10.1111/ppe.12201.

J. Yland, Wang, Taiyao, Zahra Zad, Sydney K Willis, Tanran R Wang, Amelia K Wesselink, Tammy Jiang, Elizabeth E Hatch, Lauren A Wise, and Ioannis Ch Paschalidis. Predictive models of pregnancy based on data from a preconception cohort study. *Human Reproduction*, 37(3):565–576, 01 2022. ISSN 0268-1161. doi: 10.1093/humrep/deab280.

J. Yland, Zahra Zad, Tanran R. Wang, Amelia K. Wesselink, Tammy Jiang, Elizabeth E. Hatch, Ioannis Ch. Paschalidis, and Lauren A. Wise. Predictive models of miscarriage on the basis of data from a preconception cohort study. *Fertility and Sterility*, 122(1):140–149, 2024. ISSN 0015-0282. doi: 10.1016/j.fertnstert.2024.04.007.

Hao Helen Zhang and Wenbin Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 05 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm037.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. *Available on Optimization Online*, 2015.

Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research*, 57:1155–1168, 10 2009. doi: 10.1287/opre.1080.0684.