ALIGNING DATA-DRIVEN PREDICTORS WITH ALLOCATION: A DECISION-FOCUSED APPROACH TO SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning predictors have become essential tools for guiding automated decision making. However, a major misalignment persists: predictive models are typically optimized in terms of standard statistical metrics in isolation from the algorithmic tasks they inform. We highlight this incongruity in both theory and practice in the high-stakes domain of organ allocation. We demonstrate that any algorithm relying on (even near-perfect) survival predictors optimized for standard metrics—such as the *Concordance index (C-index)*—can yield arbitrarily poor outcomes when used for allocation, failing to guarantee utility better than a uniform random selection. To bridge the gap between survival analysis and policy optimization, we introduce a *decision-focused learning* approach based on optimizing *normalized discounted cumulative gain (NDCG)*, a mainstay metric in information retrieval. We establish the utility of NDCG in survival analysis by proving that it translates to guarantees on the performance of allocation. Empirically, we develop a bootstrapping approach to optimize the NDCG of existing survival models. Unlike prior work, we also address the challenge of right censoring when evaluating NDCG. We show the aforementioned misalignment persists in historical heart transplant data from the US, and our method dramatically boosts the NDCG across a suite of survival models by 40-100%. We anticipate that our framework will find broader applications in decision making with predictions.

1 INTRODUCTION

Real-world decision making increasingly relies on algorithms powered by *machine learning (ML)* predictors trained on vast amounts of historical data. From resource allocation to automated planning and scheduling, these data-driven systems are deployed in high-stakes environments. However, an underlying disconnect persists: the development of classical algorithms for these problems is often disjoint from the design of the predictive models they leverage. ML models are typically optimized in isolation for standard statistical metrics, while the downstream algorithms using these predictions either fail to account for the predictor's performance profile or suffer because the algorithmic objective is misaligned with the model's training. The gap between predictive accuracy and algorithmic utility can have severe consequences, particularly in high-stakes applications such as organ allocation.

Organ transplantation is the treatment of choice for many terminal illnesses. Across organ types, the demand for deceased-donor organs outpaces the available supply (Cameli et al., 2022). In the US alone, thousands of patients with end-stage heart failure are waitlisted for a life-saving organ.

The current US heart transplant allocation policy places patients into rigid hierarchical tiers and allocates the organ to the highest-priority compatible patient. The policy often treats patients with heterogeneous clinical profiles as effectively identical. A major criticism of the policy is that it does not leverage finer-grained predictions of pretransplant mortality and post-transplant outcomes (Shore et al., 2020; Zhang et al., 2024). As a result, the US is transitioning to new data-driven solutions to improve the efficiency of the heart transplantation system (Papalexopoulos et al., 2024). The allocations of other organs, such as lungs (OPTN, 2025; Gottlieb et al., 2017), livers (Kamath et al., 2001; Allen et al., 2024), and kidneys (Abraham et al., 2007; Mayer & Persijn, 2006), already rely on such computational methods in the US and abroad.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

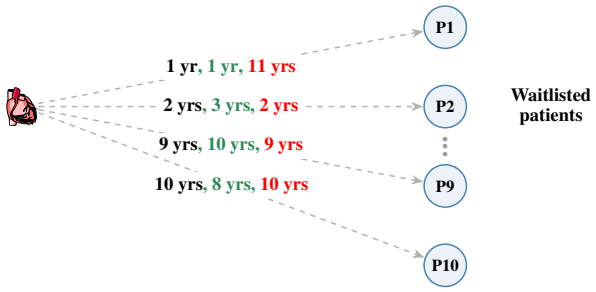


Figure 1: Illustration of heart transplant allocation with predicted outcomes. The leftmost value shows the unknown, ground-truth patient survival, the middle green value shows the predictions with NDCG@1 $= 0.9$, and the rightmost red value shows the predictions with C-index $= 0.8$.

A common data-driven approach to allocating organs relies on predictors of transplant outcomes, such as the expected life-years gained from an operation (Berrevoets et al., 2021; 2020; Zilberstein et al., 2026b; Zhang et al., 2024). The field of *survival analysis* has developed powerful statistical models for estimating such outcomes (Cox, 1972; Katzman et al., 2018; Lee et al., 2018; Wei, 1992; Nagpal et al., 2021; Luck et al., 2018). Yet, when these models are integrated into allocation mechanisms, the aforementioned disconnect surfaces. Survival models are traditionally optimized for and evaluated on metrics such as *Concordance index (C-index)* or average error, which measure aggregate performance across an entire dataset. However, when a donor heart arrives, the matching algorithm does not need perfect point-estimates of survival for all patients; rather, it requires a guarantee that it can identify the single best available match.

Our contributions As we will demonstrate, matching with a predictor that is optimized for C-index can have arbitrarily bad outcomes. We show that any deterministic algorithm relying on a predictor with near-perfect C-index can obtain a near-zero fraction of the optimal utility (Proposition 1). We then prove that no algorithm relying on a predictor with near-perfect C-index can guarantee more utility than random selection (Proposition 2), showing that C-index is a non-informative measure for allocating even a single donor. This failure is not just restricted to C-index. Most aggregate metrics, such as average error, can also lead to arbitrarily bad outcomes.

We take a step towards bridging the gap between predictive modeling in survival analysis and the requirements of downstream allocation policies. While we focus on matching with predicted edge-weights and predictors for survival analysis, our methods provide a template for evaluating and optimizing ML models whose primary purpose is to inform discrete allocation decisions.

We begin by establishing a formal link between a predictor's NDCG@k and the utility guarantee of downstream allocation (Theorem 1). We prove that the NDCG@1 of a predictor translates to a provable guarantee on the utility of greedy allocation policies (Corollary 1), a property not shared by the C-index. Figure 1 illustrates this discrepancy.

We then introduce the use of *normalized discounted cumulative gain (NDCG)* (Järvelin & Kekäläinen, 2002; Wang et al., 2013) for survival analysis, ultimately using censored-survival estimators to construct a decision-focused top-k ranking metric and training target. NDCG cannot be directly applied to survival analysis due to right censoring: many data points are only represented by a lower bound on their true survival times because the patient is still alive or their follow-up has ended. We propose two estimators of NDCG for right censored data, and prove that both provide unbiased estimates of the true *discounted cumulative gain (DCG)*. We show how such estimators can be used to select the model with superior NDCG.

Finally, we propose a method to bootstrap existing survival predictors to optimize a model for NDCG. We show using real historical heart transplant data that our estimators of NDCG can accurately identify model superiority and our bootstrapping method significantly improves the NDCG, with gains of $40\% - 100\%$ in NDCG@1, while maintaining comparable performance on standard survival analysis metrics.

Our work exposes failures in the current design of high-stakes decision-making systems for organ allocation, and we offer theoretically grounded solutions to these failures. We show, both theoretically

and empirically, that better predictive performance according to standard aggregate metrics need not translate into better downstream allocation outcomes. Predictive models of this kind are components of current mechanisms for allocating lungs, livers, and kidneys, suggesting that these systems may be misaligned with their life-saving objectives. Beyond transplantation, for ML to be safely deployed, its predictive components must be aligned to the downstream actions they inform, and our methods support this.

The mismatch between prediction and optimization is studied more broadly in the literature, and our work is the first to connect survival analysis with *decision-focused* and *end-to-end learning* (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022; Mandi et al., 2024; Capitaine et al., 2025). These lines of work focus on aligning ML models with the decision-making tasks they inform. In healthcare, this mismatch has also been recognized in the context of causal treatment effects (Vanderschueren et al., 2024; Kamran et al., 2024; Frauen et al., 2025; Fernández-Loría & Provost, 2022; Arno et al., 2026). Our paper focuses on survival analysis, which has the unique challenge of right censoring, and predicting the top-ranked candidate. We are the first to employ such techniques for organ allocation. We provide further discussion of related work in Appendix A.

2 PRELIMINARIES

We begin by reviewing standard predictive measures from information retrieval and survival analysis.

DCG and NDCG Many information retrieval settings are concerned with providing an accurate ranking of a set of datapoints (*e.g.*, recommendations). Given N inputs, we typically care about the k highest-ranked predictions rather than the entire population. Let T_i denote the relevance (*e.g.*, utility) of the point ranked ith by a prediction model where a lower rank means higher utility. In standard settings, the ground-truth relevance is known. The *discounted cumulative gain at k (DCG@k)* evaluates the quality of the top k ranked items, $\text{DCG@}k = \sum_{i=1}^{k} \frac{T_i}{\log_2(i+1)}$. To normalize DCG, we compare it to the *ideal DCG@k (IDCG@k)*, which is the maximum DCG@k achievable if the ranking were perfectly ordered by the true relevance. The *normalized discounted cumulative gain (NDCG@k)* is $\text{DCG@}k/\text{IDCG@}k$. So, $\text{NDCG@}k = 1$ represents a perfect ordering of the top k points. For further background on information retrieval, we refer to Burges (2010) and Schütze et al. (2008).

Survival analysis In the typical setting for survival analysis, we are given a dataset of individuals $i \in \{1, \ldots, N\}$. Let T_i^* denote the true unobserved survival time of person i and C_i the censoring time. The observable random variable $T_i = \min\{T_i^*, C_i\}$ and the event indicator $\delta_i = \mathbb{I}\{T_i^* \leq C_i\}$ where $\mathbb{I}$ is the binary indicator function. Let $X_i \in \mathbb{R}^d$ be the baseline covariate vector for patient i. Instead of predicting an unbounded survival time, we can also shift the target to predict survival within a fixed horizon, τ, where $T_i^{(\tau),*} = \min\{T_i^*, \tau\}$ and $\delta_i^{(\tau)} = \max\{\delta_i, \mathbb{I}\{T_i \geq \tau\}\}$. We define $S(t \mid X) = \mathbb{P}(T^* > t \mid X)$ as the true conditional survival function and $G(t \mid X) = \mathbb{P}(C > t \mid X)$ as the true conditional censoring survival function. We assume that the covariates X capture both the survival and censoring mechanisms and they are independent conditioned on X.

Assumption 1 (Conditionally independent censoring)**.** $(T^* \perp\!\!\!\perp C) \mid X$.

For heart transplantation, we aim to predict how long a new organ will sustain a patient following an operation. A ubiquitous challenge with healthcare datasets is right censoring due to patients stopping reporting. We know the last time a patient reported their condition, not the true event time.

To adapt NDCG for survival data, the relevance score becomes the true survival time (or the restricted true survival time). However, for censored patients, the true relevance remains unobservable. Therefore, we require new estimators to compute NDCG for right censored datasets.

Concordance index The standard metric for evaluating predicted rankings in survival analysis is the *concordance index (C-index)*. The C-index measures pairwise accuracy and is defined as the ratio of concordant pairs among all *comparable* pairs of patients. There are different ways of computing the C-index for right censored datasets (Gönen & Heller, 2005; Uno et al., 2011), and we adopt the commonly used Harrell's C-index (Harrell et al., 1982). These variations do not change the underlying principle of the metric. A pair of patients (i, j) is comparable if we definitively know

which patient experienced the event of interest (death, graft failure, *etc*.) first. A pair is comparable only if the patient with the shorter observed time experienced the event (*i.e.*, $T_i < T_j$ and $\delta_i = 1$).

3 Aligning machine learning predictors with policy optimization

In this section, we analyze how ML predictors interact with the allocation policies they inform. We begin by presenting a motivating example that highlights the limitations of allocating based on a predictor optimized for C-index (or other aggregate metrics). We assume that all utility is non-negative: following a transplant, the time a patient survives cannot be less than zero.

3.1 A motivating example

Consider a fully-observed dataset (*e.g.*, $\delta_i = 1 \ \forall i$), consisting of N transplant candidates and a single donor to be allocated. The goal is to allocate the donor to the patient with the best outcome (*e.g.*, maximize utility). Suppose the utility of allocating to patient i is $T_i^* = i$. An allocation algorithm does not know the true utility, but rather relies on a predictive model.

Now suppose a model correctly ranks the utility for patients $i \in [2, N]$, but incorrectly predicts that patient 1, who has the worst outcome, has the best one. When we evaluate the concordance of this model, we obtain a C-index of $1 - \frac{2}{N}$. As N grows, the C-index quickly approaches 1 despite the model recommending the patient with the worst outcome.

For $N = 100$, the above example results in a C-index of 0.98. This value is much higher than the state-of-the-art models for predicting graft survival following a heart transplant, which are around 0.6 (Aleksova et al., 2020; Lee et al., 2018; Anagnostides et al., 2025; Ayers et al., 2021). However, it is evident that despite a high C-index, a model can be arbitrarily bad at predicting the top candidate. A greedy algorithm allocating transplants based on this predictor would make catastrophic decisions. The C-index is not the right measure of a predictor that is being leveraged by a decision-making algorithm. We can formalize the failure of matching with concordance for an allocation algorithm that selects only a single candidate using a predictor.

Proposition 1. *For any deterministic algorithm selecting a single candidate based on a predictor $\hat{f}$, and for any $c \in (0, 1)$ and $\rho \in (0, 1)$, there exists a set of N candidates and a predictor $\hat{f}$ with C-index at least c, such that the algorithm achieves at most a ρ fraction of the optimal utility.*

Furthermore, not only do deterministic algorithms fail, but *no algorithm*—even a randomized one—relying on concordance can hope to do better than uniform random guessing.

Proposition 2. *For any algorithm selecting a single candidate based on a predictor $\hat{f}$ and for any $c \in (0, 1)$, there exists a set of N candidates and a predictor $\hat{f}$ with C-index at least c, such that the algorithm cannot guarantee more expected utility than that of a uniform random selection.*

This failure is not unique to the C-index. Many aggregate metrics, including standard average errors, also fail to provide utility guarantees. However, if we compute the NDCG@1 of the above predictor with $N = 100$, we would indeed see the catastrophic performance; the NDCG@1 $= 0.01$.

3.2 Allocating with NDCG

The failure of C-index arises from its inability to identify high-utility candidates. NDCG, on the other hand, captures exactly this quantity, and we can indeed bound the worst-case utility an allocation algorithm can achieve against the optimal allocation, $U(\mathrm{OPT})$. Specifically, given a predictor $\hat{f}$ with NDCG@k at least α, we can randomly select one of the top k ranked candidates proportional to the logarithmic discount in the DCG@k definition. We refer to this algorithm as the *position-weighted allocation algorithm* (PWA) which we present in Algorithm 1.

Theorem 1. *Given N candidates and a predictor $\hat{f}$ with NDCG@k at least α, algorithm* PWA *(Algorithm 1) selecting a single candidate based on $\hat{f}$ achieves expected utility at least $\alpha/W_k \cdot U(\mathrm{OPT})$ where $W_k = \sum_{i=1}^{k} 1/\log_2(i+1)$.*

The bound we obtain is a function of both α and k, and degrades approximately linearly in k. For $k = 5$, PWA guarantees roughly $\alpha/3$ of the optimal utility. NDCG@k can have higher variance at

lower values of k, so while optimizing for lower values of k is theoretically better, it may be the case that there are practical tradeoffs for robustness. When evaluating the predictor at $k = 1$, the randomized policy reduces into the greedy algorithm, yielding a direct result for greedy allocation.

Corollary 1. *Given N candidates and a predictor $\hat{f}$ with NDCG@1 at least α, algorithm* PWA *selecting a single candidate based on $\hat{f}$ achieves utility at least $\alpha \cdot U(\mathrm{OPT})$.*

These results stand in stark contrast to the performance of aggregate metrics which fail to guarantee any meaningful utility for single-item allocation. To align a data-driven predictor with the downstream allocation, these results prove that NDCG is a theoretically sound metric to optimize for.

While Theorem 1 establishes a bound for a single allocation, real-world scenarios often require sequential matching. Over multiple donor arrivals, we can apply the techniques we develop in the remainder of this paper to individual predictors for different representative donor types, consistently optimizing for the NDCG@k.

4 NDCG FOR CENSORED DATASETS

We have shown that optimizing for NDCG is a better target for predictors used by a greedy allocation algorithm than aggregate metrics like C-index. However, since the true relevance score T_i^* is unobservable for right censored data points, we cannot directly compute the standard DCG metrics in the survival analysis context. In this section, we propose two estimators of DCG that account for censoring. We show that both are unbiased whether we use the true survival time or the restricted survival time as the relevance score. It follows from the linearity of expectation that the DCG is unbiased if the relevance is conditionally unbiased given X_i. We then discuss how unbiased estimates of DCG translate to estimates of NDCG and evaluating survival models for allocation.

4.1 UNBIASED ESTIMATES OF RELEVANCE

The first method replaces the unobserved survival time for censored patients with its conditional expectation given a survival function $\hat{S}$. We define the *expected years (EY)* relevance estimator as

$$\hat{T}_i^{\mathrm{EY}} = \delta_i T_i + (1 - \delta_i)\mathbb{E}_{\hat{S}}[T_i^* \mid T_i^* > T_i, X_i].$$

To prove that the EY estimator is unbiased, we need to assume that $\hat{S}$ is unbiased.

Assumption 2 (Unbiasedness of $\hat{S}$)**.** *The conditional survival function $\hat{S}$ is conditionally unbiased such that the expected value of the estimated survival times match the true survival time. That is,* $\mathbb{E}\left[\mathbb{E}_{\hat{S}}[T_i^* \mid T_i^* > T_i, X_i]\right] = \mathbb{E}[T_i^* \mid T_i^* > T_i, X_i]$.

Proposition 3 (Unbiasedness of EY estimator)**.** *Under Assumptions 1 and 2, $\hat{T}_i^{EY}$ is a conditionally unbiased estimator of $\mathbb{E}[T_i^* \mid X_i]$.*

Using the same argument, we can also show that the estimator is unbiased in the restricted setting (Proposition 5). The theoretical unbiasedness of the EY estimator relies on the assumption that the conditional survival estimator is unbiased over the censored population. While relying on nuisance estimators[1] is common in statistics, in some ways this approach seems circular: we require a survival model to compute the DCG of another survival model. But this also highlights where the power of our bootstrapping framework stems from. If we can use a survival model as a nuisance estimator, we can also leverage it to bootstrap another model that is optimized for extremal ranking.

An alternative to imputation is the *inverse probability of censoring weighting (IPCW)* (Graf et al., 1999; Gerds & Schumacher, 2006). IPCW discards censored patients from the evaluation and re-weights the observed instances to account for the discarded population, using the censoring function $\hat{G}(t \mid X)$. The censoring survival function $\hat{G}(t \mid X)$ is the probability of remaining uncensored up to

[1] A nuisance parameter is one that is not of primary interest but must be accounted for to analyze the target parameters.

time t given X. The IPCW estimator is

$$\hat{T}_i^{\text{IPCW}} = \frac{\delta_i T_i}{\hat{G}(T_i \mid X_i)}.$$

In order to prove unbiasedness of the IPCW estimator, we need to assume that $\hat{G}$ is specified such that the inverse weights correctly recover the population expectation.

Assumption 3 (Correct specification of censoring model)**.** *The conditional censoring model $\hat{G}$ is correctly specified such that it matches the true conditional censoring distribution $G(t \mid X) = \mathbb{P}(C_i > t \mid X_i)$. We also assume that $\hat{G}(t \mid X) > 0$ for all t, X.*

Proposition 4 (Unbiasedness of IPCW estimator)**.** *Under Assumptions 1 and 3, $\hat{T}_i^{IPCW}$ is a conditionally unbiased estimator of $\mathbb{E}[T_i^* \mid X_i]$.*

As is the case for the EY estimator, the IPCW estimator is also unbiased in the restricted setting (Proposition 6). Restricting the horizon τ bounds the maximum weight at $1/\hat{G}(\tau \mid X)$ and is a practical solution for stability since IPCW weights can have high variance. Appendix C discusses the estimators and the restricted setting further.

4.2 NDCG FROM UNBIASED RELEVANCE

Given unbiased estimates of relevance, linearity of expectation ensures that $\widehat{\text{DCG}}$ and $\widehat{\text{DCG@}k}$ are unbiased. However, the IDCG requires sorting these relevances. Because the maximum operator is convex, sorting noisy estimates amplifies positive errors, leading to a positive bias. As a result, the estimate of NDCG is generally negatively biased. However, the bias does not affect the relative comparisons of two survival models. For a given dataset, $\widehat{\text{IDCG}}$ acts as a normalizer determined by the ground-truth scores and any nuisance estimators. It is independent of the models we evaluate. Therefore, given two survival models, A and B, if $\widehat{\text{NDCG}}_A > \widehat{\text{NDCG}}_B$, then $\widehat{\text{DCG}}_A > \widehat{\text{DCG}}_B$. The evaluation preserves the relative ranking of the models, allowing us to determine whether model A is superior to model B. In addition, $\widehat{\text{NDCG}}$ is scale-consistent with respect to multiplication. If $\widehat{\text{NDCG}}_A = 2\widehat{\text{NDCG}}_B$, it implies model A achieves twice the discounted gain as model B on the dataset since $\widehat{\text{IDCG}}$ cancels out in the ratio.

5 OPTIMIZING FOR NDCG VIA BOOTSTRAPPING

In this section, we briefly describe some of the most commonly deployed models for survival prediction, and then present our approach for bootstrapping a survival predictor for superior NDCG.

Baseline models We include in our evaluation a suite of common predictors used for survival analysis. These include the non-parametric *Kaplan-Meier (KM)* estimator (Kaplan & Meier, 1958) and *Random Survival Forest (RSF)* (Ishwaran et al., 2008), the semi-parametric *Cox regression (Cox)* model (Cox, 1972), the fully parametric *accelerated failure time (AFT)* model (Wei, 1992), and the deep neural network models *DeepSurv* (Katzman et al., 2018) and *DeepHit* (Lee et al., 2018). More details of these predictors can be found in Appendix D. These models are typically evaluated on their C-index. We now show how to leverage a conditional survival predictor $\hat{S}(t \mid X)$ to bootstrap another conditional survival predictor that is optimized specifically for NDCG.

Architecture Our bootstrapping framework operates using a two-stage approach. First, a base-survival model (*e.g.*, a model from above) is trained on the censored survival data to produce a conditional survival function $\hat{S}(t \mid X)$. Using this baseline, we compute the imputed label given the covariates X_i, and a restriction to the horizon, τ, $\hat{y}(X_i, \tau) = T_i + \int_{T_i}^{\tau} \hat{S}(t|X_i) / \hat{S}(T_i|X_i) dt$, or if necessary, we default to the restricted mean $\hat{y}(X_i, \tau) = \int_0^{\tau} \hat{S}(t \mid X_i) dt$. To construct the training labels for the second stage, we create a pseudo-label y_i^* that blends the observed outcomes with the model's conditional expectations to handle censoring. For a patient with observed time T_i and event indicator δ_i, we set y_i^* as τ if $T_i \geq \tau$, T_i if $T_i < \tau$ and $\delta_i = 1$, and $\hat{y}(X_i, \tau)$ if $T_i < \tau$ and $\delta_i = 0$.

Finally, we train a model, denoted $f_{\hat{S}}(X_i)$, to predict y_i^* using an objective function designed to optimize the ranking quality and the prediction error. We utilize a gradient-boosted decision tree as the underlying architecture. Other architectures are possible as well, such as a deep neural network.

Loss function We train the second model using a hybrid objective function to balance the identification of the top patients with the accuracy of the expected survival times. The total loss $\mathcal{L}_{\text{hybrid}}$ is a convex combination of the *mean squared error (MSE)* and a pairwise ranking penalty, motivated by LambdaRank (Burges, 2010). We combine the two loss functions using a hyperparameter $\alpha \in [0, 1]$, $\mathcal{L}_{\text{hybrid}} = \alpha\mathcal{L}_{\text{MSE}} + (1-\alpha)\mathcal{L}_{\text{rank}}$ and evaluate over a resampled mini-batch (*i.e.*, query group) of patients B. The regression component, $\mathcal{L}_{\text{MSE}} = \frac{1}{|B|}\sum_i (f_{\hat{S}}(X_i) - y_i^*)^2$, provides stability by ensuring the predicted survival time does not deviate too much from the imputed labels. The ranking component, $\mathcal{L}_{\text{rank}}$, optimizes for NDCG through pairwise losses scaled by the change in NDCG. We consider all pairs (i, j) where patient i outlived patient j. That is, $y_i^* > y_j^*$. If the predicted difference $f_{\hat{S}}(X_i) - f_{\hat{S}}(X_j)$ is less than a margin m, we apply a penalty that is scaled by $|\Delta\text{NDCG}_{i,j}|$, representing the absolute change in the batch's overall NDCG if patient i and j were to swap ranks, $\mathcal{L}_{\text{rank}} = \sum_{i,j\in B | y_i^* > y_j^*} \frac{1}{2} \max\left\{0, m - \left(f_{\hat{S}}(X_i) - f_{\hat{S}}(X_j)\right)\right\}^2 \cdot |\Delta\text{NDCG}_{i,j}|$. Scaling by the batch-wide $|\Delta\text{NDCG}_{i,j}|$ ensures the model prioritizes a correct ordering of the top-ranked candidates, effectively maximizing the utility of the resulting allocation. We leverage the batch-wide metric rather than $\Delta\text{NDCG@}k_{i,j}$ to provide a smoother gradient. We show in Appendix H.1 that the hybrid loss function achieves better NDCG@1 and concordance than either loss individually.

6 Experiments

We utilize the United Network for Organ Sharing (UNOS) patient registry containing clinical data for adult heart transplants in the US dating back to 1987. We provide a summary of the dataset's characteristics in Table A4. Our learning objective is to estimate *graft survival*, defined as the time elapsed from transplant to organ failure or recipient death. To ensure stability, we restrict to predicting up to $\tau = 20$ years, aligning with the 95th percentile of censoring times in the data (Table A5).

An alternative goal to graft survival prediction would be to predict *life-years gained (LYG)* of a transplant, which is the difference in years of life with and without a transplant. While LYG is often used for allocation, it relies on the graft survival prediction. For heart allocation, graft survival is also the driving factor of LYG: since conditioned on being waitlisted, a patient's survival without a transplant is typically short, the post-transplant outcome dominates LYG (Colvin et al., 2025).

6.1 Warm-up: Artificial censoring

We begin by evaluating our bootstrapping framework and our NDCG estimators under *artificial censoring*. We restrict the initial cohort to only patients with observed events, obtaining a ground-truth dataset where the exact survival time T_i^* is known for every individual. We then introduce artificial censoring to simulate the information loss present in real-world clinical registries. We provide further details on the experimental setup in Appendix F.

Results of bootstrapping Table 1 presents the comparative performance between baseline survival predictors and their bootstrapped counterparts, and we show the percent gain in NDCG in Figure 2. We report ground-truth metrics calculated using the hidden true event times. Across all baseline models, our bootstrapping framework consistently yields substantial improvements in NDCG@k. We see that the standard survival models achieve poor NDCG@k, with the values consistently below 0.3 for $k = 1$. In contrast, our approach improves upon this to over 0.4 and as high as 0.51 for NDCG@1. The increase for NDCG@1 is substantial, with the largest improvement over 0.2 when we bootstrap, and our method consistently yielding at least a 40% increase.

The gain in NDCG is achieved without compromising other metrics; our framework maintains nearly identical (in fact, marginally superior) C-index and AUC scores compared to the standard models. We observe that the full NDCG is around 0.9 across all models. This is largely due to the density of survival times in the cohort, where the cumulative sum of relevance is dominated by the mass of long-term survivors, making the metric less sensitive to individual swaps.

Predictor	Framework	C-Index	AUC	NDCG@1	NDCG@5	NDCG@10	NDCG@50	NDCG@100	NDCG
KM	Standard	.500 ±.000	.500 ±.000	.230 ±.185	.213 ±.082	.229 ±.064	.253 ±.032	.270 ±.027	.850 ±.003
	+ Ours	**.636** ±.003**	**.754** ±.011**	**.437** ±.261**	**.413** ±.123**	**.424** ±.092**	**.427** ±.040**	**.446** ±.027**	**.902** ±.003**
Cox	Standard	.623 ±.004	.763 ±.011	.288 ±.211	.343 ±.114	.360 ±.082	.411 ±.045	.424 ±.032	.898 ±.003
	+ Ours	**.637** ±.003**	**.770** ±.010**	**.494** ±.256**	**.425** ±.117**	**.425** ±.095**	**.426** ±.045**	**.442** ±.034**	**.903** ±.003**
AFT	Standard	.631 ±.003	.744 ±.009	.264 ±.228	.246 ±.107	.272 ±.093	.352 ±.045	.389 ±.032	.895 ±.003
	+ Ours	**.639** ±.003**	**.758** ±.012**	**.401** ±.246**	**.420** ±.114**	**.421** ±.084**	**.421** ±.041**	**.443** ±.025**	**.903** ±.003**
RSF	Standard	.614 ±.004	.777 ±.009	.360 ±.256	.409 ±.145	.396 ±.112	.413 ±.052	.422 ±.033	.896 ±.004
	+ Ours	**.634** ±.003**	.776 ±.010	**.513** ±.286*	**.455** ±.122	**.441** ±.095*	**.436** ±.050*	**.448** ±.031**	**.903** ±.003**
DeepSurv	Standard	.626 ±.003	.766 ±.011	.272 ±.234	.276 ±.144	.330 ±.101	.375 ±.042	.405 ±.030	.897 ±.003
	+ Ours	**.637** ±.003**	**.769** ±.011**	**.406** ±.254**	**.400** ±.154**	**.403** ±.114**	**.423** ±.052**	**.439** ±.037**	**.903** ±.003**
DeepHit	Standard	.626 ±.003	.748 ±.012	.216 ±.220	.283 ±.128	.318 ±.107	.371 ±.055	.395 ±.034	.895 ±.004
	+ Ours	**.637** ±.003**	**.765** ±.012**	**.450** ±.272**	**.376** ±.123**	**.387** ±.090**	**.422** ±.049**	**.435** ±.032**	**.902** ±.003**

*$p < .05$,**$p < .01$ (Wilcoxon signed-rank test vs. standard model)

Table 1: Performance of models on ground-truth metrics under artificial censoring.

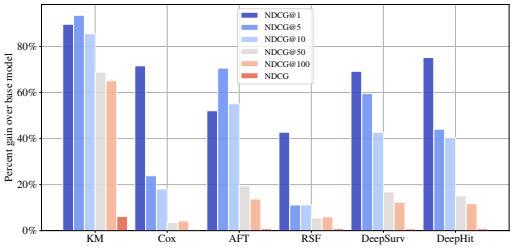


Figure 2: Average % gain in NDCG@k of bootstrapped model over baseline predictor.

Accuracy of NDCG estimators We evaluate the fidelity of the EY and IPCW estimators for NDCG by comparing their outputs against the ground-truth NDCG available from our artificial censoring. For the EY estimator, we utilize all baseline survival predictors as nuisance models barring RSF due to its computational overhead. For the IPCW estimator, we employ the KM and Cox nuisance models. Figure A1 displays the *mean absolute error (MAE)* for each estimator across different values of k. We observe that the estimation error generally decreases as k increases, approaching near-zero error for the full NDCG estimation. This trend is expected, as larger values of k aggregate more points, which averages out individual estimation errors. For NDCG@1, the MAE exceeds 0.1. The IPCW estimator has the lowest average error as k increases. In general, we expect some error, particularly due to the bias detailed in Section 4.2.

Despite the moderate absolute error, the estimators demonstrate strong correlation when compared to the ground-truth NDCG (Figure A2). The EY estimator, aggregated across all nuisance models, exhibits a high degree of correlation with the true NDCG, achieving a Spearman rank correlation coefficient of 0.8 for $k \in \{1, 5, 10, 50\}$. In contrast, the IPCW estimator displays higher variance; while individual estimates are noisier, the mean of the distribution tracks the ground-truth effectively. We observe some positive bias in the EY estimator likely due to it over-estimating population survival.

Finally, we evaluate the *relative* accuracy of our estimators (*i.e.*, their reliability in performing model selection). We report the pairwise concordance: for all pairs of survival models, (A, B), we determine the frequency with which the estimator correctly identifies the superior model. That is, whether $\widehat{\text{NDCG}}_A > \widehat{\text{NDCG}}_B$ given that $\text{NDCG}_A > \text{NDCG}_B$. As shown in Table 2, our estimators are consistent at predicting the relative ranking of models. We also evaluate an ensemble estimation,

Method	Nuisance model	NDCG@1	NDCG@5	NDCG@10	NDCG@50	NDCG@100	NDCG
IPCW	KM	77.8%	83.1%	78.7%	81.3%	85.8%	91.1%
IPCW	Cox	80.9%	73.8%	76.9%	72.9%	74.7%	83.1%
IPCW	Ensemble (Avg)	80.9%	79.6%	77.8%	79.1%	80.0%	87.1%
EY	KM	83.6%	83.2%	84.5%	85.5%	84.6%	86.1%
EY	Cox	82.3%	81.1%	81.7%	83.4%	81.7%	83.7%
EY	AFT	83.2%	82.6%	83.5%	84.6%	83.7%	85.7%
EY	DeepSurv	83.1%	82.2%	82.8%	84.1%	83.2%	85.4%
EY	DeepHit	81.7%	79.3%	80.7%	82.5%	81.5%	82.3%
EY	Ensemble (Avg)	80.9%	83.1%	89.3%	91.1%	92.9%	91.1%

Table 2: Model pairwise accuracy by estimator.

Predictor	Framework	C-Index	AUC@5	EY NDCG@k		IPCW NDCG@k	
				1	5	1	5
KM	Standard	.500 ±.000	.500 ±.000	.647 ±.288	.610 ±.158	.430 ±.388	.350 ±.189
	+ Ours	**.618** ±.004**	**.643** ±.005**	**.888** ±.184**	**.833** ±.108**	**.758** ±.523**	**.688** ±.284**
Cox	Standard	.601 ±.004	.627 ±.005	.699 ±.271	.729 ±.100	.155 ±.414	.198 ±.202
	+ Ours	**.618** ±.004**	**.652** ±.005**	**.834** ±.066**	**.800** ±.056**	**.370** ±.468**	**.427** ±.260**
AFT	Standard	.599 ±.004	.636 ±.004	.600 ±.312	.682 ±.153	.068 ±.129	.081 ±.068
	+ Ours	**.617** ±.004**	**.652** ±.005**	**.810** ±.194**	**.778** ±.105**	**.128** ±.179	**.178** ±.130**
RSF	Standard	.592 ±.004	.634 ±.005	.720 ±.167	.694 ±.106	.064 ±.068	.073 ±.048
	+ Ours	**.618** ±.004**	**.656** ±.005**	**.882** ±.157**	**.822** ±.075**	**.564** ±.468**	**.513** ±.229**
DeepSurv	Standard	.601 ±.004	.627 ±.005	.600 ±.333	.622 ±.142	.183 ±.302	.112 ±.113
	+ Ours	**.617** ±.004**	**.650** ±.005**	**.759** ±.169	**.756** ±.094**	**.269** ±.521	**.355** ±.334**
DeepHit	Standard	.590 ±.004	.582 ±.008	.649 ±.284	.650 ±.130	.192 ±.458	.175 ±.178
	+ Ours	**.613** ±.003**	**.620** ±.006**	**.730** ±.306*	**.769** ±.102**	**.601** ±.659**	**.487** ±.257**

*$p < .05$,** $p < .01$ (Wilcoxon signed-rank test vs standard model)

Table 3: Performance of models on metrics and estimated NDCG (avg over all nuisance models).

which averages the estimates over all nuisance models. The EY estimator demonstrates superior reliability, correctly identifying the better-performing model in over 80% of cases, and reaching 90% accuracy in many instances. While the IPCW estimator is more volatile, it still achieves over 77% pairwise accuracy for $k = 1$. In Appendix I, we further validate the estimators under increasingly covariate-dependent censoring. Although the absolute NDCG estimates can be biased, the estimators, particularly EY, reliably preserve model rankings and identify the model with superior ground-truth NDCG. These results establish our estimators as a reliable tool for model selection.

6.2 Evaluating on the full dataset

Having validated our estimators and our bootstrapping approach under artificial censoring, we now apply our methods to the complete UNOS registry. Unlike the artificial setup, the survival times of censored patients in this dataset are truly unknown. We rely on the NDCG estimators to assess the ranking performance. We summarize the results in Table 3 and visually show the percent gain in the NDCG estimates in Figure A3. The results show a consistent performance gain across all baseline predictors when integrated into our bootstrapping framework. The C-index and *AUC at 5 years (AUC@5)* again are very slightly improved for every baseline model tested when using bootstrapping. The AUC@5 measures the model's ability to distinguish between graft failure and survival at five years post-transplant. These results suggest that the architecture we use for the bootstrapped model may be an effective model for this predictive task.

When evaluating the estimated NDCG@k, we see the power of bootstrapping. Across both EY and IPCW estimators, our framework outperforms standard survival models, and by substantial margins at small values of k, regardless of the underlying imputer. While the NDCG@1 is comparable to the NDCG@5 for our approach, the variance of the estimation decreases as k increases. This is one reason why we may use a larger k when selecting a predictor and invoking Algorithm 1. The IPCW trends in the same direction as the EY estimates, confirming that our approach is effective at improving the identification of high-utility candidates. Despite the variance, we consistently observe a massive statistically significant improvement in the expected values of the NDCG across every survival model.

7 Conclusions

We addressed the misalignment between predictive components for survival analysis and the requirements of decision making. We demonstrated that predictors with standard aggregate metrics cannot guarantee any utility when used for allocation. To bridge this gap, we established NDCG@1 as a theoretically grounded measure that directly translates to allocation performance.

In addition, we developed novel estimators of NDCG for right censored datasets and proposed a bootstrapping approach to optimize survival models for NDCG. Our empirical results on real heart transplant data showed the effectiveness of our estimators at determining model superiority and the substantial increase in NDCG when using bootstrapping. By aligning the objectives of survival modeling with allocation optimization, our framework provides a scalable template for improving outcomes in organ transplantation and other domains of decision-making under uncertainty.

AI USE STATEMENT

In this work, we used generative AI tools for developing the software used in the experiments and editing the manuscript for clarity and consistency. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Relying on real-world historical data introduces inherent limitations. The UNOS registry, like many medical databases, contains erroneous entries and missing data. We preprocess the dataset to impute missing entries and filter inconsistent values. However, there may be unobserved confounders absent from the registry which influence outcomes. Systemic biases also present in historical clinical decisions can inadvertently be translated to the predictors. Despite this limitation, our bootstrapping approach is complementary to future advancements in the underlying survival predictors.

The unbiasedness of our DCG estimators relies on strong assumptions regarding the nuisance models. While theoretically necessary, the assumptions are not always achievable in practice. If the underlying survival model suffers from miscalibration, the EY estimator will inherit this bias. The IPCW estimator similarly requires that the censoring distribution is correctly specified. If censoring is highly informative or the probability of being uncensored approaches zero, the IPCW weights lead to high-variance estimates. This motivates the use of restricted horizons. However, restricted horizons come at the cost of expressivity. For our application, this restriction is not a major limitation (since the median graft survival is well under 20 years), but it could be a barrier to other applications. Finally, even with unbiased DCG estimators, the resulting NDCG estimate is not unbiased. Our framework reliably determines relative performance and aligns the learning with the allocation, but cannot determine the exact ground-truth NDCG. These limitations underscore the difficulty of predictive tasks under right censoring and promote future work in bias mitigation and model calibration.

REPRODUCIBILITY STATEMENT

All proofs of claims can be found in Appendix B. We present all necessary details for reproducibility in Appendices E and F including hyperparameter settings, software libraries, and compute details. We also provide the code used for the experiments as part of the submission. The UNOS patient registry is not publicly available due to sensitive information.

REFERENCES

David J Abraham, Avrim Blum, and Tuomas Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *ACM Conference on Electronic Commerce (EC)*, 2007.

Natasha Aleksova, Ana C Alba, Victoria M Molinero, Katherine Connolly, Ani Orchanian-Cheff, Mitesh Badiwala, Heather J Ross, and Juan G Duero Posada. Risk prediction models for survival after heart transplantation: a systematic review. *American Journal of Transplantation*, 20(4): 1137–1151, 2020.

Elisa Allen, Rhiannon Taylor, Alexander Gimson, and Douglas Thorburn. Transplant benefit-based offering of deceased donor livers in the united kingdom. *Journal of Hepatology*, 81(3):471–478, 2024.

Ioannis Anagnostides, Zachary Sollie, Arman Kilic, and Tuomas Sandholm. Policy optimization for dynamic heart transplant allocation. *Circulation*, 152(Suppl_3):A4369427–A4369427, 2025.

Ioannis Anagnostides, Itai Zilberstein, Zachary W Sollie, Arman Kilic, and Tuomas Sandholm. Position: Machine learning for heart transplant allocation policy optimization should account for incentives. In *International Conference on Machine Learning (ICML)*, 2026.

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Rank-learner: Orthogonal ranking of treatment effects. In *International Conference on Machine Learning (ICML)*, 2026.

Pranjal Awasthi and Tuomas Sandholm. Online stochastic optimization in the large: Application to kidney exchange. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

Brian Ayers, Tuomas Sandholm, Igor Gosev, Sunil Prasad, and Arman Kilic. Using machine learning to improve survival prediction after heart transplantation. *Journal of Cardiac Surgery*, 36(11): 4113–4120, 2021.

Jeroen Berrevoets, James Jordon, Ioana Bica, and Mihaela van der Schaar. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Jeroen Berrevoets, Ahmed M. Alaa, Zhaozhi Qian, James Jordon, Alexander E. S. Gimson, and Mihaela van der Schaar. Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In *International Conference on Machine Learning (ICML)*, 2021.

Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pp. 45–57, 1975.

Christopher JC Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81, 2010.

Robin Burke, Alexander Felfernig, and Mehmet H Göker. Recommender systems: An overview. *AI Magazine*, 32(3):13–18, 2011.

Matteo Cameli, Maria Concetta Pastore, Alessandro Campora, Matteo Lisi, and Giulia Elena Mandoli. Donor shortage in heart transplantation: How can we overcome this challenge? *Frontiers in Cardiovascular Medicine*, 9:1001002, 2022.

Zhaohui Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference on Machine Learning (ICML)*, 2007.

Aymeric Capitaine, Maxime Haddouche, Eric Moulines, Michael I Jordan, Etienne Boursier, and Alain Durmus. Online decision-focused learning. *arXiv:2505.13564*, 2025.

Monica M Colvin, Jodi M Smith, Yoon Son Ahn, Kelsi A Lindblad, Dzhuliyana Handarova, Ajay K Israni, and Jon J Snyder. OPTN/SRTR 2023 annual data report: Heart. *American Journal of Transplantation*, 25(2):S329–S421, 2025.

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

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

John P. Dickerson and Tuomas Sandholm. FutureMatch: Combining human value judgments and machine learning to match in dynamic environments. In *Conference on Artificial Intelligence (AAAI)*, 2015.

John P Dickerson, David F Manlove, Benjamin Plaut, Tuomas Sandholm, and James Trimble. Position-indexed formulations for kidney exchange. In *ACM Conference on Economics and Computation (EC)*, 2016.

Priya L. Donti, J. Zico Kolter, and Brandon Amos. Task-based end-to-end model learning in stochastic optimization. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022.

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. *arXiv preprint arXiv:2505.13092*, 2025.

Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

Jens Gottlieb, Jacqueline Smits, Rene Schramm, Frank Langer, Roland Buhl, Christian Witt, Martin Strueber, and Hermann Reichenspurner. Lung transplantation in Germany since the introduction of the lung allocation score: a retrospective analysis. *Deutsches Arzteblatt International*, 114(11): 179, 2017.

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.

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

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.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D'Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.

Fahad Kamran, Maggie Makar, and Jenna Wiens. Learning to rank for optimal treatment allocation under resource constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

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):24, 2018.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Conference on Artificial Intelligence (AAAI)*, 2018.

Margaux Luck, Tristan Sylvain, Joseph Paul Cohen, Heloise Cardinal, Andrea Lodi, and Yoshua Bengio. Learning to rank for censored survival data. *arXiv preprint arXiv:1806.01984*, 2018.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research (JAIR)*, 80:1623–1701, 2024.

Gert Mayer and Guido G Persijn. Eurotransplant kidney allocation system (ETKAS): rationale and implementation. *Nephrology Dialysis Transplantation*, 21(1):2–3, 2006.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *Communications of the ACM*, 65(7):33–35, 2022.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep cox mixtures for survival regression. In *Machine Learning for Healthcare Conference*, pp. 674–708, 2021.

OPTN. Continuous distribution. `https://www.hrsa.gov/optn/policies-bylaws/policy-issues/continuous-distribution`, January 2025.

Theodore Papalexopoulos, James Alcorn, Dimitris Bertsimas, Rebecca Goff, Darren Stewart, and Nikolaos Trichakis. Reshaping national organ allocation policy. *Operations Research*, 72(4): 1475–1486, 2024.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, 2020.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *International Conference on World Wide Web (WWW)*, 2007.

Marco Rossetti, Fabio Stella, and Markus Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *ACM Conference on Recommender Systems*, pp. 31–34, 2016.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Judy Hanwen Shen, Ellen Vitercik, and Anders Wikum. Algorithms with calibrated machine learning predictions. In *International Conference on Machine Learning (ICML)*, 2025.

Parshin Shojaee, Xiaoyu Chen, and Ran Jin. Adaptively weighted top-N recommendation for organ matching. *ACM Transactions on Computing for Healthcare*, 3(1):1–29, 2021.

Supriya Shore, Jessica R Golbus, Keith D Aaronson, and Brahmajee K Nallamothu. Changes in the United States adult heart allocation policy: challenges and opportunities. *Circulation: Cardiovascular Quality and Outcomes*, 13(10):e005795, 2020.

Xuanming Su and Stefanos Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing & Service Operations Management*, 6(4):280–301, 2004.

Yuan Tao and Huifu Xu. Necessary optimality conditions for integrated learning and optimization problem in contextual optimization. *arXiv:2601.16581*, 2026.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

Toon Vanderschueren, Wouter Verbeke, Felipe Moraes, and Hugo Manuel Proença. Metalearners for ranking treatment effects. *arXiv preprint arXiv:2405.02183*, 2024.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv:2308.01222*, 2023.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of NDCG type ranking measures. In *Conference on Learning Theory (COLT)*, 2013.

Lee-Jen Wei. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Conference on Artificial Intelligence (AAAI)*, 2019.

Kevin C Zhang, Nikhil Narang, Carine Jasseron, Richard Dorent, Kevin A Lazenby, Mark N Belkin, Jonathan Grinstein, Anoop Mayampurath, Matthew M Churpek, Kiran K Khush, and William F Parker. Development and validation of a risk score predicting death without transplant in adult heart transplant candidates. *Journal of the American Medical Association (JAMA)*, 331(6):500–509, 2024.

Itai Zilberstein, Ioannis Anagnostides, Zachary W Sollie, Arman Kilic, and Tuomas Sandholm. Learning potentials for dynamic matching and application to heart transplantation. *arXiv preprint arXiv:2602.08878*, 2026a.

Itai Zilberstein, Ioannis Anagnostides, Zachary W Sollie, Arman Kilic, and Tuomas Sandholm. Near-optimal dynamic matching via coarsening with application to heart transplantation. In *International Conference on Machine Learning (ICML)*, 2026b.

A Related work

Decision-focused learning and end-to-end learning A closely related line of work is *decision-focused learning (DFL)*. As in our paper, the DFL framework is motivated by the fact that, in machine learning, optimization is typically based on estimators. These two pieces are often treated in isolation by typical approaches (Wilder et al., 2019). Specifically, a predictive model is first trained through some measure of accuracy, for example, mean squared error. Then that model's predictions are given as input to an optimization algorithm in order to make a decision. While this two-stage approach is justified when the predictive model is highly accurate, in complex tasks state-of-the-art models will inevitably be imperfect. The training process involves tradeoffs as to the nature of such errors. When prediction and optimization are treated in isolation, it creates a critical misalignment.

To address this misalignment, Donti et al. (2017) proposed an end-to-end approach for learning ML models in a way that directly captures the final task-based objectives for which the models will be used; this approach is referred to as *end-to-end model learning*. Similarly, decision-focused learning endeavors to align decision and learning (Wilder et al., 2019; Elmachtoub & Grigas, 2022; Mandi et al., 2024; Capitaine et al., 2025). This was pioneered by Wilder et al. (2019), who studied certain classes of combinatorial optimization problems, and is also known as integrated decision learning (Tao & Xu, 2026). Closer to our work is the recent paper of Capitaine et al. (2025), which examines DFL in dynamic environments where the objective and data distribution evolve over time. For a survey on this rapidly growing line of work, we refer to Mandi et al. (2024).

There are two key challenges in applying those prior frameworks in our setting: i) the presence of right censoring, which significantly complicates statistical estimation, and ii) the objective we want to optimize, namely NDCG@1, is highly sensitive to distribution shifts.

Algorithms with predictions There is a flourishing line of work on algorithm design through the use of *predictions* (Mitzenmacher & Vassilvitskii, 2022), where the goal is to improve performance under reliable estimators and revert back to worst-case performance when the predictors are inaccurate. Heart transplant allocation can also be viewed from that perspective. As motivated above, the interplay between prediction quality and algorithm design is at the core of our approach, driven by the observation that the estimation part needs to be informed by the policy optimization component. These two are typically treated in isolation in prior papers in the line of work on algorithms with predictions.

Calibration Prediction that aligns with downstream decision tasks can also be accomplished through *calibration* (Dawid, 1982; Wang, 2023; Foster & Vohra, 1998), as it effectively allows treating the estimated quantities as probabilities; exploring how calibration can be used in cases where there is right censoring is an interesting direction for future work. To connect with the line of work on algorithms with predictions, recent work by Shen et al. (2025) studied the ski rental problem under calibrated predictors.

Moreover, the interaction between estimation and downstream decisions is also central in the framework of performative prediction (Perdomo et al., 2020), where the underlying distribution in the estimation part shifts due to the strategic decisions of the population.

Recommender systems Another closely related line of research is that of recommender systems (Burke et al., 2011; Wang et al., 2013; Rossetti et al., 2016; Järvelin & Kekäläinen, 2002), which focuses on identifying the optimal items to present to a user. Learning-to-rank (Cao et al., 2007; Burges, 2010) is a foundational technique for training a recommender system, and is the underlying learning objective of our bootstrapping approach. While the connection between learning-to-rank and allocation has been explored in domains such as healthcare resource management (Kamran et al., 2024) and ad auctions (Richardson et al., 2007), the formal link between ranking metrics and downstream allocation utility has not been previously established.

Survival prediction and organ allocation Survival analysis, a specialized field of time-to-event prediction, focuses on modeling the distribution of event times under conditions of right censoring (Kaplan & Meier, 1958; Wei, 1992). Unlike standard regression tasks, the presence of censoring means that for many subjects, we only possess a lower bound on the true event time, rather than an exact observation. Cox regression (Cox, 1972) is one of the foundational statistical models for survival analysis. Cox models, which are often evaluated via concordance, are commonly used in organ allocation (Papalexopoulos et al., 2024; OPTN, 2025; Dickerson & Sandholm, 2015). More recently, deep learning solutions have been developed to extend beyond semi-parametric models for survival analysis (Lee et al., 2018; Nagpal et al., 2021; Katzman et al., 2018).

The application of matching donor organs to patients for transplantation is a high-stakes application of machine learning and matching algorithms (Su & Zenios, 2004; Abraham et al., 2007; Awasthi & Sandholm, 2009; Dickerson & Sandholm, 2015; Dickerson et al., 2016; Berrevoets et al., 2020; 2021; Anagnostides et al., 2026; Shojaee et al., 2021). Most data-driven approaches rely on *predictions* of transplant outcomes (Berrevoets et al., 2020; 2021; Dickerson & Sandholm, 2015; Papalexopoulos et al., 2024; Shojaee et al., 2021; Zilberstein et al., 2026a). These existing methods treat the predictive model as a black-box optimized for aggregate statistical accuracy. While some work has focused on making the predictions more robust (Berrevoets et al., 2021), the predictive objectives still remain decoupled from the combinatorial requirements of the matching algorithm. Because these predictors are trained in isolation from the downstream matching objective, the resulting allocation mechanisms lack provable bounds on solution quality relative to the true underlying utility.

Conditional average treatment effect estimation One of the most similar lines of work to ours concerns estimating *conditional average treatment effect (CATE)*, a causal inference problem. Learning-to-rank has also been identified as a core task for CATE when allocating treatments (Kamran et al., 2024; Fernández-Loría & Provost, 2022). Similar to our bootstrapping approach, re-training a ranker based on an underlying predictive model is a common technique (Vanderschueren et al., 2024; Frauen et al., 2025; Arno et al., 2026). Our work differentiates itself from this line of work in multiple ways. First, survival analysis revolves around a different predictive task from the causal treatment effects, which is more akin to predicting the life-years gained. Prior work on CATE has also not addressed right censoring, a distinguishing challenge of survival analysis. Finally, in general, these prior works focus on ranking an entire set of patients, whereas our allocation is most concerned with predicting the top-k candidates correctly for very small values of k (*i.e.*, $k = 1$).

B Omitted algorithms and proofs

We begin by proving our claims regarding allocation using a predictor to determine edge weights with only a guarantee on the concordance index of the predictor.

Proposition 1. *For any deterministic algorithm selecting a single candidate based on a predictor $\hat{f}$, and for any $c \in (0, 1)$ and $\rho \in (0, 1)$, there exists a set of N candidates and a predictor $\hat{f}$ with C-index at least c, such that the algorithm achieves at most a ρ fraction of the optimal utility.*

Proof. Let $N \geq 2$ and consider an arbitrary ranking of the N candidates determined by $\hat{f}$. Let $\pi(i) \in \{1, \ldots, N\}$ denote the predicted rank of candidate i, where a lower $\pi(i)$ corresponds to a higher utility.

Given this predicted ranking, the deterministic policy will select a specific candidate. We will adversarially construct the true utilities based on the policy's selection.

For a small constant $\epsilon > 0$, we assign the true utilities, T_i, as follows.

- For the selected candidate i, $T_i = \rho - \epsilon \cdot \pi(i)$
- For the $N-1$ non-selected candidates j, $T_j = 1 + \epsilon \cdot (N - \pi(j))$

The true utility of the selected candidate is less than the utility of any candidate not selected. The policy achieves a total utility of

$$U(\mathsf{ALG}) = \rho - \epsilon \cdot \pi(i) < \rho.$$

Because $N \geq 2$, the optimal policy will select a candidate not selected by the algorithm, achieving a utility of

$$U(\mathrm{OPT}) = 1 + \epsilon \cdot (N - \pi(j)) > 1.$$

The ratio of the optimal utility achieved by the policy is therefore upper bounded by ρ.

We now calculate the C-index of the predictor $\hat{f}$. The total number of comparable pairs is $\binom{N}{2}$. All $\binom{N-1}{2}$ pairs excluding candidate i are perfectly ranked. The incorrectly ranked pairs occur when comparing the selected candidate to the rest. There are at most $(N-1)$ such pairs. The C-index is bounded below by

$$1 - \frac{(N-1)}{\binom{N}{2}} = 1 - \frac{2}{N}.$$

As $N \to \infty$, the C-index approaches 1. We can choose a sufficiently large $N > 2/(1-c)$. For such an N, the predictor achieves C-index at least c while the allocation policy is restricted to a ρ fraction of the optimal utility. □

We continue with the proof of Proposition 2.

Proposition 2. *For any algorithm selecting a single candidate based on a predictor $\hat{f}$ and for any $c \in (0,1)$, there exists a set of N candidates and a predictor $\hat{f}$ with C-index at least c, such that the algorithm cannot guarantee more expected utility than that of a uniform random selection.*

Proof. Consider any algorithm that assigns a probability $p_i \in [0,1]$ to each candidate $i \in \{1, \ldots, N\}$ based on the utility determined by $\hat{f}$. We assume $\sum_{i=1}^{N} p_i = 1$. Let Δ be the total probability mass on the candidate with the highest probability. Let $\pi(i) \in \{1, \ldots, N\}$ denote the predicted rank of candidate i, where a lower $\pi(i)$ corresponds to a higher utility.

We assign to the candidate with the highest probability a true, unknown utility of $\rho - \epsilon \cdot \pi(i)$ and a utility of $1 + \epsilon \cdot (N - \pi(i))$ to the remaining $N-1$ candidates for any $\rho \in (0,1)$ and $\epsilon > 0$.

As seen in the proof of Proposition 1, if the predictor correctly ranks all the patients except the patient with the largest probability of selection, it achieves a C-index at least c for sufficiently large N as

$$1 - \frac{(N-1)}{\binom{N}{2}} = 1 - \frac{2}{N}.$$

The expected utility of the algorithm is

$$\mathbb{E}[U(\mathsf{ALG})] = \sum_{i=1}^{N} p_i T_i \leq \Delta \cdot \rho + (1-\Delta) \cdot (1 + \epsilon N) = 1 - \Delta(1-\rho) + (1-\Delta)\epsilon N.$$

The optimal offline utility is $U(\mathrm{OPT}) \geq 1$. The expected fraction of the optimal utility is then

$$\frac{\mathbb{E}[U(\mathsf{ALG})]}{\mathbb{E}[U(\mathrm{OPT})]} \leq \frac{1 - \Delta(1-\rho) + (1-\Delta)\epsilon N}{1} \leq 1 - \Delta(1-\rho)$$

as ϵ tends to 0 for a fixed N.

Since $\rho \in (0,1)$, this expected ratio is maximized when Δ is minimized. The minimum value of the largest probability is achieved when the distribution is uniform. Any non-uniform policy necessarily has a larger Δ and can obtain a strictly lower expected ratio of the optimal utility as N grows. □

Algorithm 1 Position-weighted allocation (PWA)

Require: N candidates, predictor $\hat{f}$ with NDCG@$k \geq \alpha$.

1: Order the N candidates according to $\hat{f}$.
2: Set $w_i = 1/\log_2(i+1)$ and $W_k = \sum_{i=1}^{k} w_i$.
3: Randomly allocate to one of the top k ranked candidates proportional to w_i/W_k.

We move on to the proof of Theorem 1, and provide the pseudocode for Algorithm 1. Recall that we assume non-negative weights ($T_i \geq 0 \ \forall_i$) as in our application patients cannot have passed away in the past.

Theorem 1. *Given N candidates and a predictor $\hat{f}$ with NDCG@k at least α, algorithm* PWA *(Algorithm 1) selecting a single candidate based on $\hat{f}$ achieves expected utility at least $\alpha/W_k \cdot U(\mathrm{OPT})$ where $W_k = \sum_{i=1}^{k} 1/\log_2(i+1)$.*

Proof. Let $\pi(i) \in \{1, \ldots, N\}$ denote the index of the candidate at rank i ordered by $\hat{f}$. We denote the true utility of candidate i as T_i and assume without loss of generality that $T_i \geq T_j$ for all $i \leq j$. PWA selects rank $i \in [1, k]$ with probability w_i/W_k. The expected utility of PWA is

$$\mathbb{E}[\mathsf{PWA}] = \sum_{i=1}^{k} \frac{w_i}{W_k} T_{\pi(i)} = \frac{\mathrm{DCG@}k}{W_k}.$$

Since NDCG@$k \geq \alpha$, DCG@$k \geq \alpha$IDCG@k, so

$$\mathbb{E}[\mathsf{PWA}] \geq \frac{\alpha \mathrm{IDCG@}k}{W_k}.$$

The worst case occurs when IDCG@k is minimized. Since T_1 is the utility of the best candidate (and also the utility of OPT),

$$\mathrm{IDCG@}k = T_1 + \sum_{i=2}^{k} T_i w_i.$$

This quantity is minimized exactly when $T_i = 0$ for all $i \in [2, k]$ since the weights are non-negative. This completes the proof as

$$\mathbb{E}[\mathsf{PWA}] \geq \frac{\alpha T_1}{W_k} \geq \frac{\alpha}{W_k} U(\mathrm{OPT}).$$

□

The proof of Corollary 1 follows immediately since $W_1 = 1$. Next, we prove the unbiasedness of our NDCG estimators.

Proposition 3 (Unbiasedness of EY estimator)**.** *Under Assumptions 1 and 2, $\hat{T}_i^{EY}$ is a conditionally unbiased estimator of $\mathbb{E}[T_i^* \mid X_i]$.*

Proof. We can write the expected value of $\hat{T}_i^{\mathrm{EY}}$ based on the event indicator δ_i,

$$\mathbb{E}[\hat{T}_i^{\mathrm{EY}} \mid X_i] = \mathbb{E}[\delta_i T_i \mid X_i] + \mathbb{E}\left[(1-\delta_i)\mathbb{E}_{\hat{S}}[T_i^* \mid T_i^* > T_i, X_i] \mid X_i\right].$$

We consider the two terms and values of δ_i.

1. $\delta_i = 1$: The event is fully observed, so $T_i = T_i^*$ which implies $\mathbb{E}[\delta_i T_i \mid X_i] = \mathbb{E}[\delta_i T_i^* \mid X_i]$.
2. $\delta_i = 0$: The observation is censored, so $T_i^* > T_i$. Under Assumption 1, $\delta_i = 0$ does not influence the survival time T_i^* other than the survival up to the censoring time. It follows from Assumption 2 and the Law of Iterated Expectation, that

$$\mathbb{E}\left[(1-\delta_i)\mathbb{E}_{\hat{S}}[T_i^* \mid T_i^* > T_i, X_i] \mid X_i\right] = \mathbb{E}[(1-\delta_i)T_i^* \mid X_i].$$

Combining both cases,

$$\mathbb{E}[\delta_i T_i^* \mid X_i] + \mathbb{E}[(1-\delta_i)T_i^* \mid X_i] = \mathbb{E}\left[(\delta_i + 1 - \delta_i)T_i^* \mid X_i\right] = \mathbb{E}[T_i^* \mid X_i].$$

□

Continuing with the IPCW estimator.

Proposition 4 (Unbiasedness of IPCW estimator)**.** *Under Assumptions 1 and 3, $\hat{T}_i^{IPCW}$ is a conditionally unbiased estimator of $\mathbb{E}[T_i^* \mid X_i]$.*

Proof. Note that $\delta_i = \mathbb{I}\{C_i \geq T_i^*\}$ and $T_i = T_i^*$ when $\delta_i = 1$. Since the numerator evaluates to zero when $\delta_i = 0$, we can substitute T_i^* for T_i inside the expectation without changing the expected value. Under Assumption 1 and the Law of Iterated Expectation, we get

$$\mathbb{E}_{T^*,C,X}\left[\frac{\mathbb{I}\{C_i \geq T_i^*\}T_i^*}{\hat{G}(T_i^* \mid X_i)} \,\middle|\, X_i\right] = \mathbb{E}_{T^*,X}\left[\mathbb{E}_C\left[\frac{\mathbb{I}\{C_i \geq T_i^*\}T_i^*}{\hat{G}(T_i^* \mid X_i)} \,\middle|\, T_i^*, X_i\right] \,\middle|\, X_i\right].$$

Since T_i^* is constant relative to the inner expectation over C, we factor it out, giving

$$\mathbb{E}_{T^*,X}\left[T_i^* \cdot \mathbb{E}_C\left[\frac{\mathbb{I}\{C_i \geq T_i^*\}}{\hat{G}(T_i^* \mid X_i)} \,\middle|\, T_i^*, X_i\right] \,\middle|\, X_i\right].$$

By Assumption 3, the inner expectation equals 1. Therefore, $\mathbb{E}_{T^*,X}[T_i^* \cdot 1 \mid X_i] = \mathbb{E}[T_i^* \mid X_i]$.

□

C Further discussion of NDCG estimators

The restricted EY estimator is $\hat{T}_i^{(\tau),\text{EY}} = \delta_i^{(\tau)} T_i^{(\tau)} + \left(1 - \delta_i^{(\tau)}\right) \mathbb{E}_{\hat{S}}\left[T_i^{(\tau),*} \mid T_i^* > T_i, X_i\right]$.

If we again assume the conditional survival function is conditionally unbiased, we obtain that the restricted estimator provides conditionally unbiased estimates of the relevance.

Assumption 4 (Restricted unbiasedness of $\hat{S}$)**.** *The conditional survival function $\hat{S}$ is conditionally unbiased such that the expected value of the restricted estimated survival times match the true restricted survival time. That is, $\mathbb{E}\left[\mathbb{E}_{\hat{S}}[T_i^{(\tau),*} \mid T_i^* > T_i, X_i]\right] = \mathbb{E}[T^{(\tau),*} \mid T_i^* > T_i, X_i]$.*

Proposition 5 (Unbiasedness of restricted EY estimator)**.** *Under Assumptions 1 and 4, $\hat{T}_i^{(\tau),EY}$ is a conditionally unbiased estimator of $\mathbb{E}\left[T_i^{(\tau),*} \mid X_i\right]$.*

An identical argument to the proof of Proposition 3 holds for the proof of Proposition 5.

The restricted IPCW estimator, $\hat{T}_i^{(\tau),\text{IPCW}} = \frac{\delta_i^{(\tau)} T_i^{(\tau)}}{\hat{G}(T_i^{(\tau)} \mid X_i)}$, is also unbiased by an identical argument to the unrestricted one.

Proposition 6 (Unbiasedness of restricted IPCW estimator)**.** *Under Assumptions 1 and 3, $\hat{T}_i^{(\tau),IPCW}$ is a conditionally unbiased estimator of $\mathbb{E}\left[T_i^{(\tau),*} \mid X_i\right]$.*

As with the restricted case for the EY estimator, an identical argument holds for the proof of Proposition 6.

Since $\hat{G}(t \mid X) \to 0$ in the tail, IPCW weights can grow arbitrarily large. Restricting the horizon τ bounds the maximum weight at $1/\hat{G}(\tau \mid X)$ and is a practical solution for stability. Another advantage of the restricted case of IPCW is that we only require calibration over the horizon τ. In Assumption 3, we also require that $\hat{G}(t \mid X) > 0$ to avoid division by 0. In practice, this is not guaranteed in the unrestricted setting, but can be guaranteed in the restricted case.

IPCW has a number of drawbacks. It assigns a weight of zero to any data point censored before τ, discarding a significant portion of the dataset. In clinical settings with heavy censoring, this reduces

the sample size of the evaluation metric. Although the horizon τ theoretically bounds the maximum weight, patients with a high probability of being censored still receive large inverse weights. This introduces high variance into the evaluation metric, meaning a single heavily weighted patient can entirely dominate the DCG score.

D FURTHER DETAILS OF BASELINE SURVIVAL PREDICTORS

We present each baseline survival predictor in more detail.

- *Kaplan-Meier estimator (KM)* (Kaplan & Meier, 1958) is a non-parametric model that predicts the survival function directly from empirical data. It computes the survival probability at time t as the product of conditional survival probabilities at all observed event times $t_i \leq t$, defined as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

  where d_i is the number of events occurring at time t_i, and n_i is the number of patients at risk prior to time t_i. The KM estimator does not use any covariates in its prediction and therefore does not explicitly satisfy the assumptions required to achieve an unbiased estimate of relevance. We still include it as a baseline as it is commonly used in the clinical setting.

- *Cox regression (Cox)* (Cox, 1972) uses a semi-parametric model relying on the proportional hazards assumption which state that the hazard function can be expressed as

$$h(t \mid x) = h_0(t) \exp(\theta^\top x)$$

  where $h_0(t)$ is the baseline hazard function. The baseline hazard is typically fit using the Breslow estimator (Breslow, 1975). The survival function, $S(t \mid x)$ is computed as

$$S(t \mid x) = \exp\left(-\exp(\theta^\top x) \int_{z=0}^{t} h_0(z) dz\right).$$

- *Accelerated Failure Time (AFT)* (Wei, 1992) is a fully parametric model that assumes the effect of covariates is to accelerate (or decelerate) the time to an event. It models the logarithm of the survival time T as a linear function of the covariates x,

$$\log(T) = \theta^\top x + \sigma\epsilon$$

  where θ is the vector of coefficients, σ is a scale parameter, and ϵ is an error term following an underlying assumed distribution (typically the Weibull distribution). The survival function $S(t \mid x)$ is related to the baseline survival function $S_0(t)$ by

$$S(t \mid x) = S_0(t \exp(-\theta^\top x)).$$

- *Random Survival Forest (RSF)* (Ishwaran et al., 2008) is a non-parametric ensemble model that extends random forests to right-censored data. Each tree is trained on a bootstrap sample, and nodes are recursively split using the log-rank statistic. The ensemble cumulative hazard is obtained by averaging across the trees in the forest.

- *DeepSurv* (Katzman et al., 2018) is a non-linear extension of Cox regression that uses deep neural networks to model the proportional hazards. DeepSurv replaces the linear combination of covariates with the scalar output of a deep neural network, $f_\theta(x)$. The hazard function is modeled as

$$h(t \mid x) = h_0(t) \exp(f_\theta(x)).$$

- *DeepHit* (Lee et al., 2018) is a discrete-time survival model that relaxes the proportional hazards assumption. DeepHit discretizes the time horizon into intervals and uses a deep neural network to estimate the probability of the event occurring within each interval.

E MODEL AND TRAINING CONFIGURATIONS

We detail the hyperparameters and architectural configurations for all models used in our experiments. All models are implemented in Python 3.11, leveraging common implementations from the `scikit-survival`, `lifelines`, and `pycox` libraries. We preprocess the dataset in the same way for all models. We impute missing features using median imputation, scale features using a standard scaler, and leverage a one-hot encoding of categorical features. All experiments are conducted on an M4 Pro processor with 24GB unified memory, and terminate within hours on this processor. We report the final hyperparameters used in experiments following tuning.

E.1 OUR BOOTSTRAPPING APPROACH

Our approach utilizes a *Gradient Boosted Decision Tree (GBDT)* implemented via the LightGBM library, optimized with a custom objective detailed in Section 5. We present the necessary hyperparameters for reproducibility in Table A1. Note training also uses early stopping using a validation set.

Hyperparameter	**Value**
α	0.3
Margin m	1.0
Batch size	256
Learning rate	0.05
Number of estimators	100
Number of leaves	31
Min. child samples	40
Subsample ratio	0.8
Colsample by tree	0.8
Early stopping	50 rounds

Table A1: Bootstrapped model hyperparameters.

E.2 DEEP NEURAL NETWORKS

DeepSurv and DeepHit are implemented using the `pycox` library configured using the values in Table A2. Training also uses early stopping based on the performance on the validation set.

Hyperparameter	**DeepSurv**	**DeepHit**
Hidden layers	[64, 32]	[64, 32]
Activation	ReLU	ReLU
Dropout	0.4	0.5
Batch normalization	True	True
Optimizer	Adam	Adam
Learning rate	1e-3	1e-3
Batch size	256	256
Training Epochs	75	75
Specific Params	Weight decay: 1e-4	Bins: 30, α: 0.3, σ: 0.3

Table A2: DeepSurv and DeepHit configurations.

E.3 STATISTICAL BASELINES

The statistical baselines are implemented from standard libraries with L_2 regularization to prevent overfitting shown in Table A3.

Model	Library	Parameter(s)
Cox	`scikit-survival`	$\alpha = 0.1$
AFT	`lifelines`	penalizer$= 0.1$
KM	`lifelines`	N/A (Non-parametric)
RSF	`scikit-survival`	estimators$= 50$, max depth$= 15$, min sample splits$= 50$

Table A3: Statistical model parameters.

F Further details of the experimental setup

In the artificial censoring experiment, for patients selected for censoring, we draw a censoring time $C_i \in [0, T_i^*]$ uniformly at random. We conduct 5 iterations of 5-fold cross-validation. In each iteration, the dataset is independently re-censored. We cross-fit the nuisance models within each test fold to avoid overfitting to the observed outcomes and to ensure that the nuisance estimates used for evaluation are generated independently of the data used to train the predictors.

For the full experiment on the UNOS registry, we again conduct 5 iterations of 5-fold cross-validation using the same cross-fitting setup.

G Omitted tables and figures

Total Patients (N)	Censored (%)	Total Features	Categorical Features
60,055	29,707 (49.47%)	92	46

Table A4: Summary of UNOS dataset characteristics and feature composition.

Metric (Years)	Entire Cohort	Observed	Censored
Median	5.60	5.71	5.25
75th Percentile	10.83	11.46	10.01
90th Percentile	16.32	17.01	15.82
95th Percentile	20.01	20.37	19.64
99th Percentile	26.76	26.42	27.01
Maximum	36.02	35.25	36.02

Table A5: Distribution of follow-up times in years across the overall cohort, observed events, and censored patients.

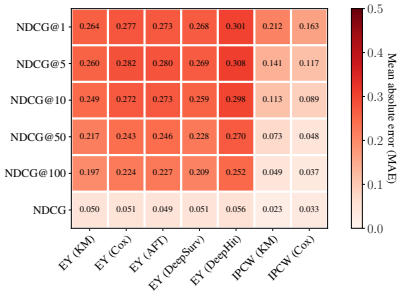
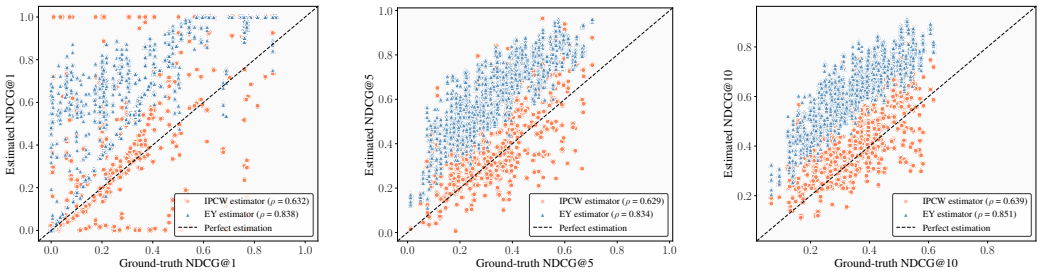
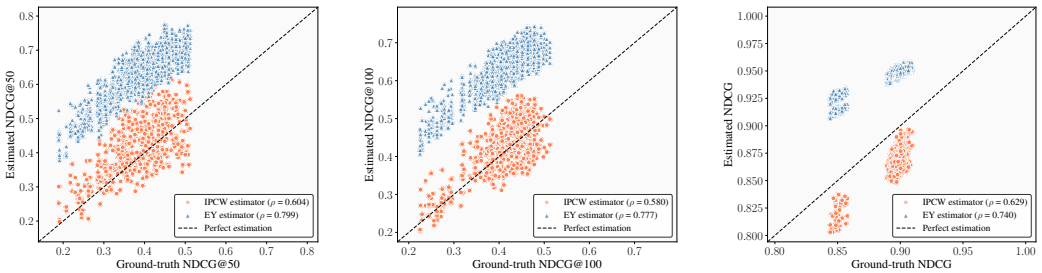


Figure A1: Mean absolute error in NDCG estimation.

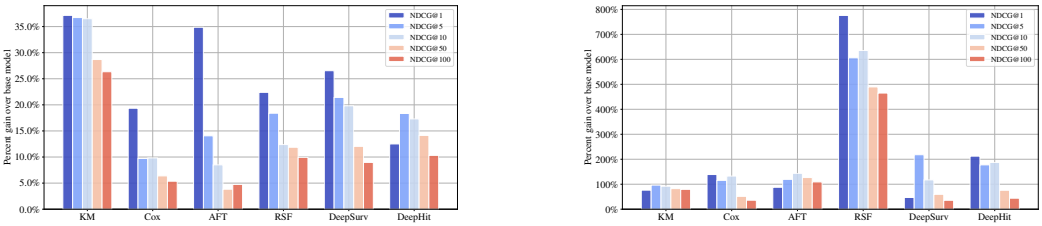


(a) Estimation of NDCG@1. (b) Estimation of NDCG@5. (c) Estimation of NDCG@10.



(d) Estimation of NDCG@50. (e) Estimation of NDCG@100. (f) Estimation of NDCG.

Figure A2: Ground-truth versus estimated NDCG@k. Each data point is an individual estimate for a model and estimator. We report the Spearman rank correlation coefficient, ρ, for each estimator.



(a) Average % gain of EY estimation. (b) Average % gain of IPCW estimation.

Figure A3: Average % gain in NDCG@k estimations of bootstrapped model over baseline predictor.

Predictor	**Framework**	**EY NDCG@k**				
		1	**5**	**10**	**50**	**100**
KM	Standard	.647 ±.288	.610 ±.158	.604 ±.094	.614 ±.040	.612 ±.027
	+ Ours	**.888** ±.184**	**.833** ±.108**	**.824** ±.075**	**.790** ±.036**	**.773** ±.029**
Cox	Standard	.699 ±.271	.729 ±.100	.729 ±.065	.729 ±.036	.730 ±.022
	+ Ours	**.834** ±.066**	**.800** ±.056**	**.801** ±.045**	**.776** ±.033**	**.769** ±.024**
AFT	Standard	.600 ±.312	.682 ±.153	.707 ±.105	.732 ±.042	.718 ±.028
	+ Ours	**.810** ±.194**	**.778** ±.105**	**.767** ±.081**	**.760** ±.032**	**.752** ±.027**
RSF	Standard	.720 ±.167	.694 ±.106	.715 ±.069	.701 ±.037	.698 ±.028
	+ Ours	**.882** ±.157**	**.822** ±.075**	**.804** ±.068**	**.784** ±.034**	**.767** ±.029**
DeepSurv	Standard	.600 ±.333	.622 ±.142	.631 ±.106	.676 ±.047	.690 ±.034
	+ Ours	**.759** ±.169	**.756** ±.094**	**.757** ±.065**	**.757** ±.039**	**.752** ±.032**
DeepHit	Standard	.649 ±.284	.650 ±.130	.647 ±.088	.669 ±.043	.685 ±.039
	+ Ours	**.730** ±.306*	**.769** ±.102**	**.759** ±.093**	**.763** ±.046**	**.755** ±.035**

*$p < .05$,** $p < .01$ (Wilcoxon signed-rank test vs standard model)

Table A6: Performance of models from the EY NDCG estimator. We report the average over all nuisance models for each estimator.

Predictor	Framework	IPCW NDCG@k				
		1	**5**	**10**	**50**	**100**
KM	Standard	.430 $\pm$.388	.350 $\pm$.189	.322 $\pm$.138	.296 $\pm$.080	.287 $\pm$.064
	+ Ours	**.758** $\pm$.523**	**.688** $\pm$.284**	**.619** $\pm$.221**	**.539** $\pm$.111**	**.516** $\pm$.078**
Cox	Standard	.155 $\pm$.414	.198 $\pm$.202	.177 $\pm$.131	.213 $\pm$.070	.244 $\pm$.056
	+ Ours	**.370** $\pm$.468**	**.427** $\pm$.260**	**.412** $\pm$.212**	**.323** $\pm$.089**	**.332** $\pm$.074**
AFT	Standard	.068 $\pm$.129	.081 $\pm$.068	.087 $\pm$.052	.094 $\pm$.033	.107 $\pm$.032
	+ Ours	**.128** $\pm$.179	**.178** $\pm$.130**	**.213** $\pm$.113**	**.212** $\pm$.062**	**.223** $\pm$.048**
RSF	Standard	.064 $\pm$.068	.073 $\pm$.048	.064 $\pm$.034	.064 $\pm$.027	.065 $\pm$.020
	+ Ours	**.564** $\pm$.468**	**.513** $\pm$.229**	**.468** $\pm$.185**	**.376** $\pm$.075**	**.370** $\pm$.069**
DeepSurv	Standard	.183 $\pm$.302	.112 $\pm$.113	.148 $\pm$.099	.183 $\pm$.062	.223 $\pm$.057
	+ Ours	**.269** $\pm$.521	**.355** $\pm$.334**	**.322** $\pm$.244**	**.292** $\pm$.124**	**.302** $\pm$.088**
DeepHit	Standard	.192 $\pm$.458	.175 $\pm$.178	.153 $\pm$.120	.239 $\pm$.071	.292 $\pm$.064
	+ Ours	**.601** $\pm$.659**	**.487** $\pm$.257**	**.439** $\pm$.205**	**.421** $\pm$.110**	**.421** $\pm$.079**

$^{*}p < .05, ^{**}p < .01$ (Wilcoxon signed-rank test vs standard model)

Table A7: Performance of models from the IPCW NDCG estimator. We report the average over all nuisance models for each estimator.

H Ablation studies

H.1 Hybrid loss function

We test how different values of α in the hybrid objective affect the performance of our bootstrapping approach. We fix the underlying predictor as Cox regression and show the result using the artificial censoring setup in Table A8. Using $\alpha = 0.3$ consistently outperforms using either the MSE loss or ranking loss alone, confirming that both components are beneficial.

Framework	C-Index	NDCG@k			
		1	**5**	**10**	**50**
Cox (Standard)	0.623	0.288	0.343	0.360	0.411
+ Ours ($\alpha = 0$)	0.580	0.305	0.237	0.265	0.296
+ Ours ($\alpha = 0.3$)	0.637	0.494	0.425	0.425	0.426
+ Ours ($\alpha = 0.6$)	0.636	0.368	0.409	0.417	0.418
+ Ours ($\alpha = 1$)	0.636	0.468	0.429	0.432	0.431

Table A8: Effect of varying α in the hybrid loss function.

I Validation of NDCG estimators under covariate-dependent censoring

We further validate the NDCG estimators under increasingly covariate-dependent censoring. The censoring mechanism here depends only on observed baseline covariates and is conditionally independent of the event time.

For each patient in the fully observed cohort, we independently sample a censoring time according to

$$C_i \mid X_i \sim \text{Exponential}\left(\lambda_0 e^{\gamma \eta_i}\right),$$

and observe

$$\widetilde{T}_i = \min\{T_i^*, C_i\}, \text{ and } \delta_i = \mathbb{I}\{T_i^* \leq C_i\}.$$

We define η_i as a fixed deterministic function of a small set of baseline prognostic covariates,

$$\eta_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} s_j z_{ij}, \text{ and } \sum_i \eta_i = 0,$$

where z_{ij} is the standardized value of covariate j for patient i, and $s_j \in \{-1, +1\}$ orients each covariate so that larger η_i corresponds to worse prognosis (e.g., older age or higher BMI). Because η_i

is a deterministic function of X_i and the censoring time is sampled independently of T_i^* conditional on X_i, this construction satisfies $T_i^* \perp C_i \mid X_i$ for every value of γ.

The parameter γ controls the strength of the dependence between censoring and the observed covariates. At $\gamma = 0$, the censoring rate is constant across patients and censoring is independent of X_i. As γ increases, higher-risk patients are increasingly likely to be censored earlier, making conditional nuisance models more important. For each value of γ, we recalibrate λ_0 so that the overall artificial censoring rate remains fixed at 49.47%, matching the censoring rate in the full UNOS cohort. The amount of censoring is held fixed while only its dependence on the covariates changes. We evaluate $\gamma \in \{0, 0.5, 1\}$ using 5×5-fold cross-validation, with an independent realization of the censoring process in each repetition.

Table A9 reports the bias of the resulting NDCG estimates by estimator, nuisance model, and γ. The absolute NDCG estimates remain biased even though the relevance and DCG estimators are unbiased. This is expected since computing IDCG requires sorting noisy relevance estimates, and this nonlinear normalization introduces bias into NDCG.

For the full NDCG, the EY estimator exhibits only a small positive bias, between $+0.031$ and $+0.041$, and is remarkably stable across γ. The bias is larger for small values of k, particularly for NDCG@1, where estimating the highest-ranked candidate is inherently more sensitive to noise in the relevance estimates. The IPCW estimator exhibits a substantial negative NDCG bias. This effect is consistent with the higher variance of IPCW relevance estimates: the subsequent sorting and normalization used to construct IDCG can amplify this noise. Neither estimator provides uniformly accurate NDCG estimates across all values of k.

		NDCG@1			**NDCG@10**			**NDCG**		
Method	**Nuisance**	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1
IPCW	KM	-0.264	-0.152	-0.097	-0.278	-0.216	-0.125	-0.257	-0.201	-0.148
IPCW	Cox	-0.249	-0.099	-0.006	-0.272	-0.226	-0.123	-0.254	-0.194	-0.140
EY	KM	0.133	0.143	0.148	0.104	0.122	0.139	0.031	0.031	0.032
EY	Cox	0.191	0.190	0.184	0.162	0.169	0.178	0.035	0.035	0.034
EY	AFT	0.259	0.241	0.228	0.229	0.229	0.224	0.038	0.037	0.036
EY	DeepSurv	0.148	0.165	0.164	0.125	0.146	0.155	0.033	0.033	0.033
EY	DeepHit	0.232	0.228	0.229	0.207	0.215	0.223	0.041	0.041	0.041
EY	Ensemble	0.193	0.193	0.191	0.165	0.176	0.184	0.036	0.035	0.035

Table A9: Bias of the estimated NDCG under increasingly covariate-dependent censoring. Positive values indicate overestimation relative to the ground-truth NDCG.

Despite the bias, the estimators remain useful for their primary purpose in our framework: identifying which predictive model has superior NDCG. Table A10 reports the percentage of pairwise model comparisons for which each estimator recovers the correct order induced by the ground-truth NDCG.

The EY estimator is consistently reliable across all censoring mechanisms. For NDCG@1, its pairwise accuracy ranges from approximately 75% to 83%, while for the full NDCG it ranges from approximately 83% to 91%. Increasing the dependence of censoring on the covariates does not systematically degrade model-selection accuracy. The performance remains stable or improves across most configurations.

IPCW is less reliable at small k, particularly for NDCG@1, where its pairwise accuracy is only slightly above random guessing at $\gamma = 0$. Nevertheless, its performance improves as γ increases, reaching above 60% accuracy for NDCG@1 at $\gamma = 1$ and substantially higher accuracy for larger k. These results reinforce the distinction between absolute NDCG estimation and model selection: although the estimated NDCG values can be biased, particularly at small k, their relative ordering is more stable, especially for the EY estimator.

		NDCG@1			NDCG@10			NDCG		
Method	**Nuisance**	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1
IPCW	KM	55.1%	58.9%	62.3%	60.4%	66.6%	68.2%	75.4%	77.8%	80.0%
IPCW	Cox	55.5%	58.9%	61.5%	60.7%	63.0%	64.5%	75.8%	75.9%	74.4%
EY	KM	74.8%	78.8%	79.6%	75.6%	79.4%	80.0%	86.6%	89.0%	89.6%
EY	Cox	75.5%	79.0%	81.0%	77.0%	80.8%	80.1%	82.9%	85.7%	88.5%
EY	AFT	75.6%	78.7%	81.5%	**79.0%**	**81.8%**	81.2%	86.0%	87.8%	88.3%
EY	DeepSurv	75.9%	**79.9%**	**83.0%**	78.2%	81.4%	80.4%	**87.2%**	**90.1%**	89.6%
EY	DeepHit	76.1%	79.9%	81.3%	77.1%	80.5%	80.8%	86.2%	87.9%	88.1%
EY	Ensemble	**77.1%**	79.8%	81.9%	77.8%	81.6%	**81.6%**	86.8%	88.8%	**90.7%**

Table A10: Model pairwise accuracy by estimator under increasingly covariate-dependent censoring.