Subgroup Discovery with the Cox Model

Zachary Izzo
NEC Labs America
zach@nec-labs.com

Iain Melvin
NEC Labs America
iain@nec-labs.com

Abstract

We study the problem of subgroup discovery with Cox regression models and introduce a method for finding an interpretable subset of the data on which a Cox model is highly accurate. Our method relies on two technical innovations: the *expected prediction entropy*, a novel metric for evaluating survival models which predict a hazard function; and the *conditional rank distribution*, a statistical object which quantifies the deviation of an individual point to the distribution of survival times in an existing subgroup. Because of the interpretability of the discovered subgroups, in addition to improving the predictive accuracy of the model, they can also form meaningful, data-driven patient cohorts for further study in a clinical setting.

1 Introduction

Cox regression is a popular approach for survival analysis, where the goal is to model the distribution of the time until a "unit" (e.g., patient) experiences a "failure" (an event of interest, e.g., death or onset of a disease) conditional on relevant covariates. While the Cox model is appealing for its simplicity and ease of interpretability, in practice, the modeling assumptions can be violated leading to inaccurate predictions [13]. Neural network-based methods for survival analysis have gained popularity in the machine learning community in recent years, and these methods are more flexible and capable of modeling more complex relationships in the data than the Cox model [23, 19, 15]. However, due to their black-box, uninterpretable nature, these methods have not been widely employed in practice. In this paper, we address the problem of using interpretable methods to accurately model survival data. Rather than trying to model the entire dataset simultaneously, we instead find a subset of the data (referred to as a subgroup) on which an interpretable Cox model is highly accurate. The subgroup itself is defined via easily interpretable criteria, namely, by thresholding the covariate values. Thus, in addition to improving the predictive accuracy of our model, the discovered subgroups can also be used to define meaningful patient cohorts for future clinical study.

Our Contributions In summary, this paper makes the following contributions:

1. We introduce an algorithm for finding an interpretable region of the feature space in which a Cox model makes confident, accurate predictions. The region is defined by thresholding covariate values, which aligns with common clinical research practice and makes the discovered subgroups useful for further clinical study.
2. Our method relies primarily on two technical innovations. First, we introduce the expected prediction entropy (EPE), a novel metric for evaluating survival models. The EPE gives a more fine-grained evaluation of survival models by taking into account the confidence of the model's prediction. Second, we introduce the *conditional rank distribution*, a statistical object which quantifies the deviation of an individual point to the distribution of survival times in an existing subgroup.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

2 Related Work

Survival Analysis: Models The Cox model [6, 7] is a standard method for survival analysis, and it has found widespread use in practice due to its ease of interpretation. Nevertheless, the interpretability comes at the cost of strong modeling assumptions which may be violated in practice [13]. Within the machine learning community, there has been a great deal of effort to apply modern ML techniques to survival data and provide more powerful and flexible models. One class of approaches estimates the risk or hazard function using flexible models like neural networks, [19, 15, 43, 3], trees [2], Gaussian processes [21], or kernels [4]. Another common approach seeks to directly model the survival distribution [23, 34]. The authors of [40] forego these standard modeling targets and instead try to directly predict the failure order.

Survival Analysis: Evaluations In addition to improving modeling flexibility, there has been a great deal of work on proper evaluation metrics for survival models. One of the most common evaluation metrics is Harrell's concordance index (C-index) [12], which evaluates a survival model according to how well the predicted failure order of the units matches the data. [45] studied proper scoring rules for survival analysis (i.e., metrics which are minimized by the true survival function) for use both as loss functions for training models, and as evaluation metrics. [11] introduce D-calibration as a calibration metric for models which predict a survival distribution, and which is similar in spirit to the expected calibration error used to evaluate classification models. [33] proposed a method for estimating the mean absolute error of a survival model with censored observations, which can then be used as an evaluation metric.

Subgroup Discovery Our work sits at the intersection of two orthogonal topics, survival analysis and subgroup discovery. At a high level, subgroup discovery refers to mining datasets for subsets or regions in which the data distribution is in some sense "interesting," usually quantified by a numerical score function taking an extreme value when evaluated on the subgroup [10, 1, 24, 17, 44]. While subgroup discovery is a general problem, it has found a great deal of applications in biostatistics [28, 30]. Many methods have been proposed to study heterogeneous treatment effects in patient populations, in particular to find patient groups which experience enhanced benefit from a treatment [20, 27, 9, 26, 29, 36, 35, 25]; or for purposes of patient stratification [31, 5, 16]. [42] studied the subgroup discovery problem for the Cox model but defined subgroups via the two sides of a hyperplane rather than an axis-aligned box.

3 Problem Setup

Our high-level goal is to find an interpretable region R^* of the feature space in which a Cox model makes confident and accurate predictions.[1] For the interpretability of the discovered region, we follow the setting of [17] and restrict ourselves to axis-aligned boxes. Such regions correspond to thresholding individual features. It remains to give a precise definition of "confident and accurate predictions."

The Cox model is inherently relative in nature: because the baseline hazard function is not estimated, the Cox model by itself does not provide information about the survival distribution of an individual unit. Rather, given two (or more) units, the Cox model predicts the probability that one unit will fail before another. This fact is the motivation behind one of the most common evaluation metrics for Cox models, the C-index [12], which measures the fraction of pairs of units in the data for which the predicted failure order matched the true failure order.

The problem with using the C-index as our accuracy measure is that it does not take model confidence into account. If the model predicts a 99.9% chance that unit 1 fails before unit 2, but in fact unit 2 fails before unit 1, this is penalized equally as if the model had predicted a 50.1% chance of unit 1 failing before unit 2 when evaluating the C-index. For detecting more subtle deviations from the Cox model, a more sensitive metric is needed.

A natural alternative to consider is the (log) partial likelihood, the quantity used as a loss function for fitting the Cox model. While the partial likelihood does take model confidence into account, it is not

[1] For readers who are unfamiliar with survival analysis in general or the Cox model in particular, we provide the necessary background information in Appendix A.

suitable for *comparing* different groups of data. This is because the value of the partial likelihood depends heavily on the size of the risk sets R_i. For instance, the first unit to fail out of 1000 units was given a predicted 10% chance of being the first to fail by the model, this could reasonably be considered a very confident and accurate prediction (a $100\times$ improvement over a random guess). On the other hand, if only two units were at risk and the model assigned a 10% chance of failure to the unit which actually failed first, this would constitute a confident but inaccurate prediction. However, these two scenarios contribute equally to the value of the partial likelihood.

The inadequacy of the existing metrics discussed above motivate definition of our accuracy metric, the *expected prediction entropy (EPE)*.

3.1 Expected Prediction Entropy

Let $\lambda(t, x)$ be the true hazard function. Conditional on a failure occurring at time t among two units with features x_1 and x_2, the probability that x_1 experiences failure is

$$\lambda(t, x_1)/(\lambda(t, x_1) + \lambda(t, x_2)). \tag{1}$$

Given a survival model which predicts an instantaneous hazard rate $\hat{\lambda}(t, x)$, we can evaluate the goodness of fit of our model to the data by measuring its ability to discriminate between which of two units at risk will fail.

Definition 3.1 (Expected Prediction Entropy)**.** Let P be a probability distributions over $\mathbb{R}^d \times \mathbb{R}_{\geq 0}$ which denotes the joint distribution of a (feature, survival time) pair. Let (X, T), $(X', T') \sim P$ be two i.i.d. draws from P, and define $Y = \mathbb{1}\{T \leq T'\}$. Let $\hat{\lambda}$ be an estimate for the hazard function which defines the distribution of T conditional on X, and let $R \subseteq \mathbb{R}^d$ be a sub-region of the feature space. We define the *expected prediction entropy (EPE)* as $\mathrm{EPE}(\hat{\lambda}, R) =$

$$\mathbb{E}\left[-Y \log \frac{\hat{\lambda}(T, X)}{\hat{\lambda}(T, X) + \hat{\lambda}(T, X')} - (1 - Y) \log \frac{\hat{\lambda}(T', X')}{\hat{\lambda}(T', X) + \hat{\lambda}(T', X')} \;\middle|\; X, X' \in R\right]. \tag{2}$$

By equation (1), Y is a Bernoulli random variable with parameter $p = \lambda(T, X_1)/(\lambda(T, X) + \lambda(T, X'))$. Since the cross entropy loss is a proper scoring rule, it follows that the minimum of (2) occurs when ratio of the estimated hazard functions equals its true value, i.e., when

$$\frac{\hat{\lambda}(T, X)}{\hat{\lambda}(T, X) + \hat{\lambda}(T, X')} = \frac{\lambda(T, X)}{\lambda(T, X) + \lambda(T, X')}.$$

In particular, the true hazard function λ (along with any positive scalar multiple of λ) minimizes this loss function. In general, a lower EPE indicates a more accurate survival model.

Estimating EPE Empirically Let $\{(x_i, t_i, \delta_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}_{\geq 0} \times \{0, 1\}$ be a survival dataset with features x_i, event times t_i, and censoring indicators δ_i. An empirical estimate of the EPE is given by

$$-\frac{1}{N} \sum_{i\,:\,\delta_i = 1} \sum_{j \in R_i} \log \frac{\hat{\lambda}(t_i, x_i)}{\hat{\lambda}(t_i, x_i) + \hat{\lambda}(t_i, x_j)}, \tag{3}$$

where $R_i = \{j \;:\; t_j > t_i\}$ is the risk set at time t_i (minus the i-th datapoint itself) and $N = \sum_{i\,:\,\delta_i = 1} |R_j|$ is the total number of comparable event times. In the case that there is no censoring (i.e., $\delta_i = 1$ for all i), (3) gives an unbiased estimate for (2). In the presence of censoring, the fact that we can only compare two datapoints when the first event time was uncensored may introduce a bias.

Specialization to the Cox Model The EPE has a particularly interesting interpretation when $\hat{\lambda}$ is given by a Cox model, i.e., $\hat{\lambda}(t, x) = \lambda_0(t) e^{\beta^\top x}$. In this case, (2) reduces to

$$\mathbb{E}\left[-Y \log \frac{1}{1 + e^{-\beta^\top (X - X')}} - (1 - Y) \log \frac{1}{1 + e^{\beta^\top (X - X')}} \;\middle|\; X, X' \in R\right]. \tag{4}$$

Observe that this is the standard cross entropy loss for a logistic model trained to predict the label Y from the feature differences $X_1 - X_2$.

We remark that the expression (4) appeared in [38] as a lower bound for the C-index. The authors use this lower bound directly to train a Cox model, instead of the standard partial likelihood. [22] used the same expression as an approximation to the partial likelihood, using a risk set of size 1 to avoid memory constraints during model training. [40] also explored this expression in the context of ranking losses, which are again used to train relative risk models. To the best of our knowledge, we are the first to explore the usefulness and properties of the EPE as an *evaluation metric*, not merely as a loss function.

3.2 Mathematical Problem Statement

The EPE allows us to quantify when our model is making confident and accurate predictions. Thus, we can state the precise goal for our method as follows. Let $\mathcal{A}$ be the set of all axis-aligned boxes in the feature space of the data. Define $\mathcal{R}_{\text{EPE}} = \text{argmin}_{R \in \mathcal{A}} \text{EPE}(R)$ to be the set of all axis-aligned boxes which minimize the conditional EPE. Finally, let $\mathcal{R}^{\max}_{\text{EPE}} = \text{argmax}_{R \in \mathcal{R}_{\text{EPE}}} \text{vol}(R)$ be the regions of maximal volume which minimize the EPE. Our goal is to find (or approximate) a region $R^* \in \mathcal{R}^{\max}_{\text{EPE}}$. Equivalently:

Find the largest possible region which minimizes the conditional EPE.

4 Method

We follow the general algorithmic framework used by [17]. Namely, we first find a small "core" group of points which we are confident belong to the desired region R^*, and we use these points to fit a coarse model (Section 4.1). After the core group is selected, we examine each other point in the dataset and "reject" points which could not feasibly follow the same model as the points in the core group (Section 4.2). In the final step, we select a region which is as large as possible but contains no rejected points. For this final step, we use the same "growing box" procedure as [17].

4.1 Core Group Selection: Minimize EPE

We select a core group which minimizes the training EPE. Specifically, for each point in the dataset, we consider all points that lie within a small ℓ_∞ ball centered at that point and fit a Cox model to this group. We then compute the empirical EPE via (3), where $\hat{\lambda}$ is given by the Cox model fit to the points in this group. The core group is chosen to be whichever group had minimal empirical EPE.

4.2 Rejection Criterion: Conditional Rank Statistics

We restrict our attention to the case of uncensored data for now. For the Cox model with unconstrained baseline hazard function, all of the information is contained in the order of failures. Thus, we examine the probability of the rank statistics of the observed points, conditional on the estimated Cox coefficients and the observed failure order of the core group.

Specifically, let β be the fitted model coefficients and $x_1, \ldots, x_n$ be the feature vectors in the core group, labeled such that $t_1 < t_2 < \cdots < t_n$. For a "test" point with features x^* and failure time t^*, we wish to compute the probability that the rank of x^* is at least as extreme (high or low) as its observed value, conditional on the other observed failure times and assuming that x^* follows the same Cox model as the core group. To do this, we work with the *conditional rank distribution* of x^*, defined as:

$$r^c_k(x^*) = \mathbb{P}(t_{k-1} < t^* < t_k \mid x^*, x_1, \ldots, x_n;\ t_1 < \cdots < t_n), \tag{5}$$

where the probability is computed assuming each pair (x, t) follows the same Cox model with fixed (unknown) baseline hazard function $\lambda_0(t)$ and Cox coefficients β. It will also be convenient to define the *unconditional rank probabilities* of x^* as

$$r_k(x^*) = \mathbb{P}(t_1 < \cdots < t_{k-1} < t^* < t_k < \cdots < t_n \mid x^*, x_1, \ldots, x_n). \tag{6}$$

By Bayes' rule, we have that $r^c_k(x^*) = r_k(x^*)/(\sum_{j=1}^n r_j(x^*))$. It thus suffices to compute the unconditional rank probabilities of x^*. When the data are generated according to the Cox model, we have

$$r_k(x^*) = \prod_{i=1}^{n+1} \frac{\exp(\beta^\top x_i^{(k)})}{\sum_{j=i}^{n+1} \exp(\beta^\top x_j^{(k)})}, \tag{7}$$

where we have defined $x_i^{(k)} = x_i$ if $i < k$, $x_i^{(k)} = x^*$ if $i = k$, and $x_i^{(k)} = x_{i-1}$ if $i > k$ (i.e., the i-th feature vector when x^* has been "inserted" in the k-th position). Using the expression from Bayes' rule, we can then compute the conditional rank distribution for x^*.

Finally, let $\mathrm{rank}(x^*)$ denote the random variable whose value is the rank of the "test" unit with features x^*, and let k^* be its observed value (i.e., the rank of t^* among $t_1, \ldots, t_n$). We define the *rank tail probability*

$$\tau^* = \min\left\{\sum_{k=1}^{k^*} r_k^c(x^*),\ \sum_{k=k^*}^{n+1} r_k^c(x^*)\right\}$$

and check whether $\tau^* < \alpha/2$.

Generalization to Censored Data The conditional rank distribution has a straightforward generalization to the partial likelihood and censored data. We again consider the distribution of possible failure times for x^* among all of the events (failure or censoring) experienced by the other points. If we know the actual rank of x^* (i.e., if it failed), then we can conduct a two-tailed test after computing the distribution. If x^* was censored, then we can only form a test based on its right tail.

Let $y_1 < \cdots < y_n$ be the event times for the points with features $x_1, \ldots, x_n$ in the core group, and let δ_i be the corresponding failure indicators ($\delta_i = \mathbb{1}\{x_i$ failed (was not censored) at time $y_i\}$). The partial likelihood that x^* fails with event rank k is

$$r_k(x^*) = \prod_{i=1}^{n+1}\left(\frac{\exp(\beta^\top x_i^{(k)})}{\sum_{j=i}^{n+1}\exp(\beta^\top x_j^{(k)})}\right)^{\delta_i} = \prod_{i\,:\,\delta_i=1}\left(\frac{\exp(\beta^\top x_i^{(k)})}{\sum_{j=i}^{n+1}\exp(\beta^\top x_j^{(k)})}\right), \tag{8}$$

where $x_i^{(k)}$ are defined as before. Note that this is simply the standard Cox partial likelihood if x^* fails as the k-th event. The conditional failure "likelihoods" $r_k^c(x^*)$ are then defined analogously to equation (5), though we note that these are no longer actually probabilities or proper likelihoods in the presence of censoring.

Fast Implementation Computing the conditional rank probabilities naively is inefficient on large datasets, scaling as $\Omega(n^3)$. Using some recursive relationships between the unconditional rank probabilities, we can drastically reduce this runtime down to $O(n)$ which also led to marked practical efficiency gains. Details can be found in Appendix B.

4.3 Determining the Final Region

Once we have determined a core group and rejected points which cannot feasibly follow the same model as the core group, we can directly apply the "growing box" procedure in Algorithm 2 of [17]. Intuitively, starting from the average of the features in the core group, we allow each side of the region to expand (potentially at different speeds) until it collides with a rejected point. This procedure continues until all sides of the region have collided with a rejected point or reached a predetermined maximum value.

5 Theoretical Results

In this section, we will show theoretically that in a well-specified setting, our method recovers the correct region. Our proof relies on several assumptions and simplifications:

1. The hazard function for the entire dataset has the form $\lambda(t; x) = \lambda_0(t)e^{h(x)}$ for some unknown risk function h.
2. There is no censoring in the data.
3. There is a unique largest region R^* which minimizes the EPE, and R^* is an axis-aligned box. Conditional on $x \in R^*$, we have $h(x) = \beta^\top x$ for some β, i.e., the Cox model is well-specified.
4. The core group selection procedure (Section 4.1) finds a group of points which belong to R^*, and the Cox model fit to these points recovers the true parameters β.

5. The conditional rank distribution converges to its expected value. We use this limiting distribution for the analysis, rather than the finite sample version described in Section 4.2.

Under these assumptions, our main theorem shows that our method can approximately recover the ground truth region R^* with high probability, given a large enough effect size. The full proof of this theorem can be found in Appendix C.

Theorem 5.1. *Let $\hat{R}_N$ be the region output by our method on a dataset of N i.i.d. points satisfying the above assumptions. For any constant $\varepsilon > 0$, there is a constant C_ε (depending on ε) such that if $|h(x) - \beta^\top x| \geq C_\varepsilon$ outside of R^*, then with probability at least 0.99, we have that $R^* \subseteq \hat{R}_N$ and $\mathrm{vol}(\hat{R}_N \setminus R^*) \leq C'\varepsilon$ for another constant C' as $N \to \infty$.*

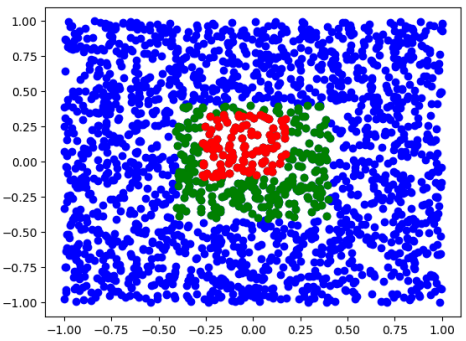
6 Experiments

We compared our method against three baselines: the *vanilla Cox model*, i.e. fitting a Cox model to the entire dataset; *random*, in which we construct the bounding boxes of random subsets of the training data and select the box with the best validation EPE; and an ablation of our own method without the growing box procedure *(no expansion)*. Full descriptions are given in Appendix D.1.

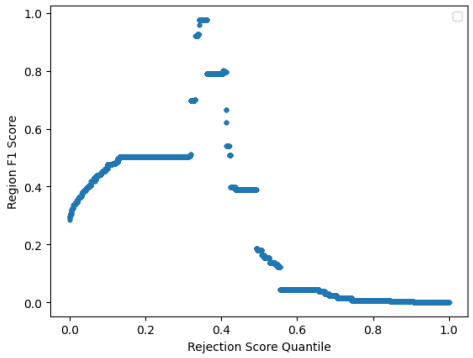
We evaluate each method according to several metrics. Our primary goal is to minimize the *expected prediction entropy (EPE)*. We report the empirical estimate of $\mathrm{EPE}(R)$ on the test set for the region R discovered by each method. When R^* is known, we can also compute the F1 score of an estimated region by defining $\text{Precision} = \mathrm{vol}(\hat{R} \cap R^*)/\mathrm{vol}(\hat{R})$ and $\text{Recall} = \mathrm{vol}(\hat{R} \cap R^*)/\mathrm{vol}(R^*)$.

We created two synthetic datasets to test our method's ability to recover a ground truth region R^* (Synth 1 & Synth 2). A full description of their construction is given in Appendix D.2. We also test our method on several standard, publicly available survival analysis benchmarks included in the sksurv package [32]. These include Breast Cancer [8], GBSG2 [37], Lung Cancer [18], and AIDS [14].

Figure 1 shows the results of the two components of the method on synthetic data. Figure 1a shows the results of the core group finding procedure. The green points belong to R^*, the blue points are outside R^*, and the red points constitute the core group. The method correctly identifies a core group belonging to R^*, as desired. Figure 1b shows the F1 score for the estimated region vs. the threshold set for the hyperparameter τ^* in Section 4.2. When the rejection threshold is set properly, it is possible for the method to exactly recover R^* (which is equivalent to reaching an F1 score of 1).



(a) Core group selection results. By choosing the core group with minimum EPE, we successfully obtain a subset of R^*.



(b) Rejection/growing box results. When tuned correctly, the method can correctly recover the ground truth R^*.

Figure 1: Recovering the ground truth region R^* on synthetic data.

Table 1 shows the EPE of the regions discovered by each method. The results are averaged over 10 training/test splits, and the best test EPE for each method is selected in each run. On 4/6 datasets, our method finds the best region among all of the methods, and on 5/6 datasets, it reduced the EPE significantly as compared to the procedure of fitting the Cox model to the entire dataset. This indicates that meaningful subgroups may exist in real survival data.

	Synth 1	Synth 2	Lung Cancer	Breast Cancer	GBSG2	AIDS
Cox	0.70	0.70	**0.52**	0.67	0.59	0.44
Random	0.70	0.70	**0.52**	0.67	0.58	0.44
No Expansion	0.68	0.47	0.69	0.23	0.23	**0.01**
Full Method	**0.35**	**0.29**	0.56	**0.18**	**0.14**	0.15

Table 1: Mean EPE results on synthetic and real data averaged over 10 training/test splits (lower is better). The best test EPE is selected for each method in each run.

7 Conclusion

In this work, we introduced subgroup discovery method which finds interpretable subsets of the feature space in which a Cox model makes confident and accurate predictions. Our method relies on two components: the expected prediction entropy (EPE), which quantifies the ability of a survival model to discriminate between the relative risk of failure for two units; and the conditional rank distribution, a statistical object which can be used to measure the deviation of an individual datapoint to the distribution of survival times in an existing subgroup. We gave asymptotic convergence guarantees for our method in a well-specified setting and confirmed its effectiveness empirically on synthetic and real datasets.

Limitations & Future Work While the convergence of the conditional rank distribution to its population-level analog is intuitive and supported empirically, it remains to provide a rigorous proof. It also remains to prove that the core group selection procedure correctly selects a subset of R^*. Expanding our understanding of the CRD in the presence of censoring is also of interest.

The EPE may be useful more broadly as an evaluation metric for survival models which predict a hazard function. However, as the EPE depends not only on the accuracy of the model, but also on the distribution of patient covariates and the intrinsic difficulty of distinguishing between units, a more complete and quantitative understanding of these factors is necessary for it to be maximally useful as an evaluation tool.

Lastly, the experiments we conducted serve as an initial proof of concept for the efficacy of our method. Moving forward, more extensive empirical evaluation should be undertaken. It is especially interesting to undertake case studies where an interpretable modeling outcome is desired, but where the Cox model is known to give a poor fit to all of the data (e.g., in the scenario discussed by [13]).

References

[1] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[2] Dimitris Bertsimas, Jack Dunn, Emma Gibson, and Agni Orfanoudaki. Optimal survival trees. *Machine learning*, 111(8):2951–3023, 2022.

[3] Linus Bleistein, Van Tuan NGUYEN, Adeline Fermanian, and Agathe Guilloux. Dynamic survival analysis with controlled latent states. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=xGlVkBSDdt`.

[4] George H Chen. Survival kernets: Scalable and interpretable deep kernel survival analysis with an accuracy guarantee. *Journal of Machine Learning Research*, 25(40):1–78, 2024.

[5] Gong Chen, Hua Zhong, Anton Belousov, and Viswanath Devanarayan. A prim approach to predictive-signature development for patient stratification. *Statistics in medicine*, 34(2):317–342, 2015.

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

[7] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[8] Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Françoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, Mahasti Saghatchian d'Assignies, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast

cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214, 2007.

[9] Elise Dusseldorp and Iven Van Mechelen. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237, 2014.

[10] Jerome H Friedman and Nicholas I Fisher. Bump hunting in high-dimensional data. *Statistics and computing*, 9(2):123–143, 1999.

[11] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

[12] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[13] Miguel A Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

[14] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2008.

[15] Bingqing Hu and Bin Nan. Conditional distribution function estimation using neural networks for censored and uncensored data. *Journal of Machine Learning Research*, 24(223):1–26, 2023.

[16] Xin Huang, Yan Sun, Paul Trow, Saptarshi Chatterjee, Arunava Chakravartty, Lu Tian, and Viswanath Devanarayan. Patient subgroup identification for clinical drug development. *Statistics in medicine*, 36(9):1414–1428, 2017.

[17] Zachary Izzo, Ruishan Liu, and James Zou. Data-driven subgroup identification for linear regression. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL `https://proceedings.mlr.press/v202/izzo23a.html`.

[18] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[19] 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*, 18:1–12, 2018.

[20] Victoria Kehl and Kurt Ulm. Responder identification in clinical trials with censored data. *Computational statistics & data analysis*, 50(5):1338–1355, 2006.

[21] Hideaki Kim. Survival permanental processes for survival analysis with time-varying covariates. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=CYCzfXn6cZ`.

[22] Havard Kvamme, Ornulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

[23] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[24] Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*, pages 1–16. Springer, 2008.

[25] Michael Lingzhi Li and Kosuke Imai. Statistical performance guarantee for selecting those predicted to benefit most from treatment. *arXiv preprint arXiv:2310.07973*, 2023.

[26] Ilya Lipkovich and Alex Dmitrienko. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of biopharmaceutical statistics*, 24(1):130–153, 2014.

[27] Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621, 2011.

[28] Ilya Lipkovich, Alex Dmitrienko, and Ralph B D'Agostino Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1): 136–196, 2017.

[29] Ilya Lipkovich, Alex Dmitrienko, Kaushik Patra, Bohdana Ratitch, and Erik Pulkstenis. Subgroup identification in clinical trials by stochastic sidescreen methods. *Statistics in Biopharmaceutical Research*, 9(4):368–378, 2017.

[30] Ilya Lipkovich, David Svensson, Bohdana Ratitch, and Alex Dmitrienko. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *arXiv preprint arXiv:2311.14889*, 2023.

[31] Wolfgang Polonik and Zailong Wang. Prim analysis. *Journal of Multivariate Analysis*, 101(3): 525–540, 2010.

[32] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL `http://jmlr.org/papers/v21/20-729.html`.

[33] Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An effective meaningful way to evaluate survival models. *arXiv preprint arXiv:2306.01196*, 2023.

[34] David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1190–1205. PMLR, 2022.

[35] Patrick M Schnell. Monte carlo approaches to frequentist multiplicity-adjusted benefiting subgroup identification. *Statistical Methods in Medical Research*, 30(4):1026–1041, 2021.

[36] Patrick M Schnell, Peter Müller, Qi Tang, and Bradley P Carlin. Multiplicity-adjusted semi-parametric benefiting subgroup identification in clinical trials. *Clinical Trials*, 15(1):75–86, 2018.

[37] M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, RL Neumann, and HF Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

[38] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20, 2007.

[39] Lu Tian. Survival analysis (stat331) lecture notes, 2015. URL `https://web.stanford.edu/~lutian/coursepdf/unit1.pdf`.

[40] Andre Vauvelle, Benjamin Wild, Roland Eils, and Spiros Denaxas. Differentiable sorting for censored time-to-event data. *Advances in Neural Information Processing Systems*, 37, 2023.

[41] Rasmus Waagepetersen. Cox's proportional hazards model and cox's partial likelihood. `https://people.math.aau.dk/~rw/Undervisning/DurationAnalysis/Slides/lektion3.pdf`, 2022.

[42] Susan Wei and Michael R Kosorok. The change-plane cox model. *Biometrika*, 105(4):891–903, 2018.

[43] Ruofan Wu, Jiawei Qiao, Mingzhe Wu, Wen Yu, Ming Zheng, Tengfei LIU, Tianyi Zhang, and Weiqiang Wang. Neural frailty machine: Beyond proportional hazard assumption in neural survival regressions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=3Fc9gnR0fa`.

[44] Sascha Xu, Nils Philipp Walter, Janis Kalofolias, and Jilles Vreeken. Learning exceptional subgroups by end-to-end maximizing KL-divergence. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 55267–55285. PMLR, 2024.

[45] Hiroki Yanagisawa. Proper scoring rules for survival analysis. In *International Conference on Machine Learning*, pages 39165–39182. PMLR, 2023.

A Background on Survival Analysis and the Cox Model

A *survival time* is a non-negative random variable T which describes the amount of time until an event of interest. Examples of commonly modeled events include the onset of a disease, the death of a patient, the time at which a customer stops using a product or platform, or the failure of a mechanical component. The arbitrary event to be modeled is referred to as a *failure*. Unlike more typical regression tasks in machine learning where the goal is to give a point estimate of a continuous-valued target, the goal of survival analysis is usually to model the *distribution* of T conditional on some associated covariates $X \in \mathbb{R}^d$.

Natural modeling targets for describing the distribution of T include standard probabilistic quantities such as the probability density function (pdf) or cumulative distribution function (cdf) of T, conditional on the features X, and indeed some survival analysis methods take this approach. A more common target, however, is the *hazard function*, defined as

$$\lambda(t; x) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T \leq t + dt \mid T \geq t, X = x)}{dt}. \tag{9}$$

The hazard function can be thought of as an instantaneous rate of failure in the infinitesimal time interval $[t, t + dt)$, conditional on surviving up to time t and on the features $X = x$. The hazard function is related to more standard quantities like the pdf or cdf. Specifically, letting $F(t, x) = \mathbb{P}(T \geq t \mid X = x)$ be the cdf and $f(t, x)$ the associated pdf (assuming one exists), we have the following identities:

$$S(t, x) := 1 - F(t, x) = \exp\left\{-\int_0^t \lambda(u, x)\, du\right\}, \qquad f(t, x) = \lambda(t, x) S(t, x).$$

The complement $S(t, x)$ of the cdf is referred to as the *survival function*. The existence of these formulas shows that determining the hazard function completely specifies the distribution of $T|X$, as it completely specifies the pdf or cdf. In a biomedical context, the hazard function has several advantageous properties which make it a natural modeling target, including but not limited to interpretability. For instance, a patient in remission from cancer would naturally be more interested in knowing the conditional probability of a recurrence given that they have not experienced one yet, rather than an absolute probability which is more easily described by the cdf [39].

The Cox model posits a particular semiparametric form for the hazard function which implies that a unit change in each covariate has a multiplicative effect on the hazard function, i.e.,

$$\lambda(t; z) = \lambda_0(t) \exp(\beta^\top z) \tag{10}$$

for some coefficients β.

A.1 Fitting the Cox Model with the Partial Likelihood

This subsection follows the derivation of [41]. Suppose that the failure times are given by $t_1 < \cdots < t_m$. Let L_i denote the index of the individual who fails at time t_i. Let T_ℓ denote the random failure time for the ℓ-th individual, and define $R(t)$ to be the *risk set* at time t, i.e. the set of individuals $R(t) = \{\ell \,:\, T_\ell \geq t\}$ who have not failed before time t.

We begin by computing the probability of an *individual* failure event, given the risk set at that time and the parameters β. That is, we wish to compute

$$\mathbb{P}(L_i = \ell \mid T_{L_i} = t_i,\ R(t_i) = R_i). \tag{11}$$

When the failure times are continuous random variables, the probability that $T_{L_i} = t_i$ is zero. Thus we will instead consider

$$\mathbb{P}(L_i = \ell \mid T_{L_i} \in [t_i, t_i + dt),\ R(t_i) = R_i) \tag{12}$$

and let $dt \to 0$. First, observe that we have

$$\begin{aligned}
&\mathbb{P}(L_i = \ell, T_{L_i} \in [t_i, t_i + dt) \mid R(t_i) = R_i) \\
&\quad = \mathbb{P}(T_\ell \in [t_i, t_i + dt),\ T_k > T_\ell\ \forall k \in R_i \setminus \{\ell\} \mid R(t_i) = R_i) \\
&\quad = \mathbb{P}(T_\ell \in [t_i, t_i + dt),\ T_k > t_i + dt\ \forall k \in R_i \setminus \{\ell\} \mid R(t_i) = R_i) + O((dt)^2) \\
&\quad = (\lambda(t_i; z_\ell)dt) \prod_{k \in R_i \setminus \{\ell\}} (1 - \lambda(t_i; z_k)dt) + O((dt)^2) \\
&\quad = \lambda_0(t_i) \exp(z_\ell^\top \beta)dt + O((dt)^2). \qquad (13)
\end{aligned}$$

Using equation (13), we also have that

$$\begin{aligned}
\mathbb{P}(T_{L_i} \in [t_i, t_i + dt) \mid R(t_i) = R_i) &= \sum_{j \in R_i} \mathbb{P}(L_i = j,\ T_{L_i} \in [t_i, t_i + dt) \mid R(t_i) = R_i) \\
&= \sum_{j \in R_i} \lambda_0(t_i) \exp(z_j^\top \beta)dt + O((dt)^2). \qquad (14)
\end{aligned}$$

Combining (13) and (14), we find that

$$\mathbb{P}(L_i = \ell \mid T_{L_i} = t_i, R(t_i) = R_i) = \frac{\exp(z_\ell^\top \beta)}{\sum_{j \in R_i} \exp(z_j^\top \beta)}. \qquad (15)$$

Given the failure times t_i and associated risk sets R_i (which account for both previous failures and censoring), [6] then proposed estimating β by maximizing the log partial likelihood

$$\mathcal{L}(\beta) := \sum_{i=1}^{m} z_{\ell_i}^\top \beta - \sum_{i=1}^{m} \log \left(\sum_{j \in R_i} \exp(z_j^\top \beta) \right), \qquad (16)$$

where ℓ_i is the index of the individual which failed at time t_i. While each term (15) is a likelihood in the traditional sense, [7] showed that $\exp(\mathcal{L}(\beta))$ is *not* a marginal or conditional likelihood (unless one makes restrictive assumptions on the censoring patterns/failure times). Nevertheless, maximizing (16) still enjoys many of the same properties as traditional MLE, such as asymptotic normality and consistency [7].

B Runtime Improvements

A naive implementation of the conditional rank tail probability took over 20 seconds to evaluate on a single point in some early experiments. Thus, a faster implementation is necessary. To avoid cumbersome notation, we will use the abbreviation $r_k = r_k(x^*; X, \delta, \beta)$.

First, we observe that the naive computation of a single r_k from equation (7) will require $\Omega(n^2)$ time. This can easily be reduced to $O(n)$ by updating the partial sum contained in the denominator as each term in the product is computed, rather than recomputing it from scratch each time. With this modification, we can compute r_1 in $O(n)$ time.

We can obtain another speedup by computing the remaining r_k recursively, rather than repeatedly using the procedure above from scratch for each r_k. A direct calculation using the formula (8) shows that

$$r_{k+1} = \frac{(1 - \delta_k)e^{\beta^\top x^*} + S_k}{e^{\beta^\top x^*} - e^{\beta^\top x_k} + S_k} \cdot r_k, \qquad (17)$$

where we have defined $S_k = \sum_{i=k}^{n} e^{\beta^\top x_i}$. Again using the running partial sum trick to quickly compute S_k (rather than computing from scratch each time), we can compute the next r_{k+1} in constant time using the previous one. This means that $r_1, \ldots, r_{n+1}$ can *all* be computed using only $O(n)$ time total.

The pseudocode for the resulting procedure is given in Algorithm 1. We have replaced the rank probabilities r_k with the logarithms since when working with large datasets, working directly with the product of many probabilities (even when each is individually of "reasonable" size) can lead to numerical issues. Given the set of $\log r_k$, the conditional probability distribution r^c can then be computed by taking a softmax.

Algorithm 1 Fast computation of the log rank probabilities with censoring

$S \leftarrow \sum_{i=1}^{n} e^{\beta^\top x_i}$
log_prod $\leftarrow \beta^\top x^* - \log(S + e^{\beta^\top x^*})$
for $i = 1, \ldots, n$ **do**
  log_prod $\leftarrow$ log_prod $+ \delta_i(\beta^\top x_i - \log S)$
  $S \leftarrow S - e^{\beta^\top x_i}$
end for
$\log r_1 \leftarrow$ log_prod

$S \leftarrow \sum_{i=1}^{n} e^{\beta^\top x_i}$
for $k = 1, \ldots, n$ **do**
  $\log r_{k+1} \leftarrow \log r_k + \log(S + (1 - \delta_k)e^{\beta^\top x^*}) - \log(S + e^{\beta^\top x^*} - e^{\beta^\top x_k})$
  $S \leftarrow S - e^{\beta^\top x_k}$
end for

return $\log r_1, \ldots, \log r_{n+1}$

C Proof of Theorem 5.1

In this section, we provide the proof of Theorem 5.1. We first restate our assumptions and give more precise conditions when needed, as well as define the notation used in the proof.

Assumption C.1. We assume that the hazard function has the form $\lambda(t; x) = \lambda_0(t)e^{h(x)}$ where $\lambda_0(t) > 0$ for all t. Conditional on $x \in R^*$, we have $h(x) = \beta^\top x$ for some fixed β, i.e., the Cox model is well-specified in R^*.

We define the following quantities:

- $T(r)$ denotes a random survival time with hazard function $\lambda_0(t)e^r$, i.e. the survival time of a unit where the log relative hazard $h(x) = r$. In particular, the data are generated according to $T(h(X))$, where X is sampled from the feature distribution.
- $G(t)$ denotes the marginal CDF for survival times sampled from points belonging to R^*. That is, $G(t) = \mathbb{P}(T(h(X)) \leq t \mid X \in R^*)$, where the probability is computed with respect to both the randomness in $X \in R^*$ and $T(h(X))$.
- $G^{-1} : [0, 1) \to \mathbb{R}_{\geq 0}$ denotes the inverse CDF.
- $F(q, r)$ denotes the probability that a unit with log-hazard r fails before the q-th quantile of the marginal failure time distribution from R^*, i.e., $F(q, r) = \mathbb{P}(T(r) \leq G^{-1}(q))$.
- p_{I} denotes the type I error rate, i.e., the probability that a point which belongs to R^* is rejected.
- p_{II} denotes the type II error rate, i.e., the probability that we fail to reject a point close to each face of R^*.
- q denotes the q-th quantile of the limiting rank quantile distribution G.
- N denotes the total number of datapoints in the training dataset.
- m denotes the number of points which, if rejected, will constitute a satisfactory detection of some face of R^*. (These correspond to points in $R_{\varepsilon,j,\pm}$ from [17].)
- n denotes the number of points which belong to R^* and shouldn't be rejected. We can expect $n = \Omega(N)$.

- We will use η to denote a small probability which needs to be lower bounded, and p to denote a small probability which needs to be upper bounded.

Assumption C.2. No censoring occurs in the data.

Assumption C.3. The core group selected by the first stage of the algorithm belongs entirely to R^*. Furthermore, the error in the Cox model fit to this core group is negligible, i.e. we have $\hat{\beta} \approx \beta$ where $\hat{\beta}$ are the Cox coefficients fit on the core group.

Assumption C.4. The conditional rank distribution for each point converges to its expectation, i.e., for each feature vector x in the dataset we have

$$\sum_{k \le pn} r_k^c(x; X, \beta) \approx F(p, \beta^\top x).$$

We conjecture that Assumptions C.3 and C.4 can be proven to hold under reasonable assumptions on the data generation process and for a sufficiently large sample size. Both assumptions hold empirically on our synthetic datasets.

We now proceed with the proof of Theorem 5.1, which is broken down into five steps.

Step 1: It suffices to assume $\lambda_0(t) \equiv 1$.

Consider the random variable $\tilde{T} = \int_0^T \lambda_0(s)\, ds$. Note that this is a monotonic change of variables since $\lambda_0(s) > 0$. In addition, note that the survival function $\tilde{S}(t, x)$ of $\tilde{T}$ conditional on features $X = x$ is given by

$$\begin{aligned}
\tilde{S}(t, x) &= \mathbb{P}(\tilde{T} \ge t \mid X = x) \\
&= \mathbb{P}\left(\int_0^T \lambda_0(s)\, ds \ge t \mid X = x \right) \\
&= \mathbb{P}(T \ge \Lambda_0^{-1}(t) \mid X = x) \\
&= \exp\left(-e^{h(x)} \Lambda_0(\Lambda_0^{-1}(t)) \right) \\
&= \exp(-e^{h(x)} t). \qquad (18)
\end{aligned}$$

This implies that $\tilde{T}$ has hazard function $\tilde{\lambda}(t; x) = e^{h(x)}$, which in particular means that the baseline hazard function under this transformation is $\tilde{\lambda}_0(t) \equiv 1$. Since the transformation is monotonic, all of the ranks will be preserved, so all of the results which hold for T hold also for $\tilde{T}$ and vice-versa.

Step 2: How low does the individual false positive rate need to be to guarantee an overall FPR of at most p_{I}? Let the resulting maximum individual FPR be p.

Let p be an upper bound on the probability that an individual point in R^* is rejected. A union bound implies that the overall false positive rate is at most pn. In particular, if $\boxed{p \le p_{\mathrm{I}}/n}$ then the overall false positive rate is at most p_{I}.

Step 3: How large can we make the lower rejection region for a point x, such that its individual FPR is at most p from Step 2? That is, what is the quantile $q(x)$ of the conditional quantile distribution such that $\mathbb{P}(T(x) \le q(x)) \le p$, conditional on $T(x)$ following the same Cox model?

Given x, we wish to determine the maximum possible q such that

$$F(q, \beta^\top x) = \mathbb{P}(T(x) \le G^{-1}(q)) \le p \quad \Leftrightarrow \quad \mathbb{P}(T(x) > G^{-1}(q)) \ge 1 - p.$$

From Step 1, it suffices to consider the case where $\lambda_0(t) \equiv 1$. In this case, $T(x)$ is an exponential random variable with rate $e^{\beta^\top x}$ and we can compute the tail probability explicitly. In particular, we

have

$$\mathbb{P}(T(x) > G^{-1}(q)) = \exp(-e^{\beta^\top x} G^{-1}(q)) \geq 1 - p \tag{19}$$

$$\Leftrightarrow \quad -e^{\beta^\top x} G^{-1}(q) \geq \log(1-p) \tag{20}$$

$$\Leftrightarrow \quad G^{-1}(q) \leq -\log(1-p) e^{-\beta^\top x} \tag{21}$$

$$\Leftrightarrow \quad q \leq G\left(-\log(1-p) e^{-\beta^\top x}\right) \tag{22}$$

$$= \mathbb{P}_{T \sim \text{Core}}\left(T \leq -\log(1-p) e^{-\beta^\top x}\right) \tag{23}$$

$$= \mathbb{E}_{X \sim \text{Core}}\left[1 - \exp\left(-e^{\beta^\top X} \cdot (-\log(1-p) e^{-\beta^\top x})\right)\right] \tag{24}$$

$$= 1 - \mathbb{E}_{X \sim \text{Core}}\left[(1-p)^{e^{\beta^\top (X-x)}}\right]. \tag{25}$$

Thus we can take $\boxed{q(x) = 1 - \mathbb{E}_{X \sim \text{Core}}\left[(1-p)^{e^{\beta^\top (X-x)}}\right]}$.

Step 4: How high does the individual true positive rate need to be to guarantee an overall false negative rate of at most p_{II}?

Let η be a lower bound on the individual TPR. Then the overall false negative rate is at most

$$(1-\eta)^m \leq p_{\text{II}} \quad \Leftrightarrow \quad 1 - \eta \leq p_{\text{II}}^{1/m} \quad \Leftrightarrow \quad \boxed{\eta \geq 1 - p_{\text{II}}^{1/m}}.$$

Step 5: Lower bound the hazard rate for "good rejections" in terms of the type II error and conditional quantile distribution. Equivalently, when is the hazard large enough for the individual good failure probability to be at least what is required by Step 4?

WLOG we will focus on rejecting points with the lower tail of the conditional rank distribution. The same argument works with the inequalities reversed for the upper tail.

To obtain maximum power for the test at a fixed false positive rate, we will reject points with features x falling below the quantile $q(x)$ defined in Step 3. The question is equivalent to determining conditions on $h(x)$ such that

$$F(q(x), h(x)) = \mathbb{P}(T(x) \leq G^{-1}(q(x))) \geq \eta.$$

Again from Step 1, it suffices to consider $\lambda_0(t) \equiv 1$ so that $T(x)$ is exponential with rate $e^{h(x)}$. Let $q = q(x)$. Then we have

$$\mathbb{P}(T(x) \leq G^{-1}(q)) \geq \eta \quad \Leftrightarrow \quad \mathbb{P}(T(x) > G^{-1}(q)) \leq 1 - \eta \tag{26}$$

$$\Leftrightarrow \quad \exp(-e^{h(x)} G^{-1}(q)) \leq p_{\text{II}}^{1/m} \tag{27}$$

$$\Leftrightarrow \quad -e^{h(x)} G^{-1}(q) \leq \frac{1}{m} \log p_{\text{II}} \tag{28}$$

$$\Leftrightarrow \quad h(x) \geq \log^2 p_{\text{II}}^{-1} - \log m - \log G^{-1}(q). \tag{29}$$

It remains to upper bound $-\log G^{-1}(q)$, or equivalently to lower bound $G^{-1}(q)$. This is again equivalent to finding a lower bound on t such that $\mathbb{P}_{T \sim \text{Core}}(T \leq t) \leq q$. Using the expression for

$q = q(x)$ from Step 3, we have

$$\mathbb{P}_{T\sim\text{Core}}(T \le t) \le q \tag{30}$$

$$\Leftrightarrow \quad \mathbb{E}_{X\sim\text{Core}}\left[1 - \exp(-e^{\beta^\top X} t)\right] \le 1 - \mathbb{E}_{X\sim\text{Core}}\left[(1-p)^{e^{\beta^\top (X-x)}}\right] \tag{31}$$

$$\Leftrightarrow \quad \mathbb{E}_{X\sim\text{Core}}\left[\exp\left(\log(1-p)e^{-\beta^\top x}e^{\beta^\top X}\right)\right] \le \mathbb{E}_{X\sim\text{Core}}\left[\exp(-te^{\beta^\top X})\right]. \tag{32}$$

As long as $t \le -\log(1-p)e^{-\beta^\top x}$, the integrand on the LHS of (32) is pointwise less than or equal to the integrand on the RHS. In particular, we can take t equal to this upper bound, or equivalently

$$G^{-1}(q) \ge -\log(1-p)e^{-\beta^\top x}.$$

Plugging this into (29), we can sufficiently bound the false negative rate provided that

$$h(x) \ge \log^2 p_{\text{II}}^{-1} - \log m - \log^2(1-p)^{-1} + \beta^\top x. \tag{33}$$

Finally, substituting $p = p_{\text{I}}/n$, we require that

$$h(x) - \beta^\top x \ge \log^2 p_{\text{II}}^{-1} - \log m - \log^2(1 - p_{\text{I}}/n)^{-1}. \tag{34}$$

Observe that

$$\log\log\frac{1}{1-p_{\text{I}}/n} \ge \log\log(1+p_{\text{I}}/n) \ge \log\frac{p_{\text{I}}}{2n}.$$

Thus, it suffices to have

$$h(x) - \beta^\top x \ge \log^2\frac{1}{p_{\text{II}}} + \log\frac{2n}{p_{\text{I}}} - \log m.$$

In order for $\text{vol}(\hat{R}_N \setminus R^*) \le \varepsilon$, the m points must lie within an $O(\varepsilon)$ distance of each of the faces of R^*. With high probability, there will be at least $\Omega(\varepsilon N)$ such points, so we may assume that $m \ge c\varepsilon N$ for some constant $c > 0$. Thus, the final inequality simplifies to

$$\boxed{h(x) - \beta^\top x \ge \log^2\frac{1}{p_{\text{II}}} + \log\frac{2}{cp_{\text{I}}\varepsilon}}.$$

Thus we can take the constant C_ε to be equal to the RHS of this inequality.

To prove the desired result, we can now directly apply logic from the analogous proof in [17]. Specifically, since we have assumed that the core group lies within R^* and we have shown that no points in R^* will be rejected (Step 2), the logic of [17] shows that $R^* \subseteq \hat{R}_N$ given correct settings for the expansion speed $s_j^{\pm}$ of each side of the box. Similarly, the choice of C_ε implies that there will be a rejected point within an $O(\varepsilon)$ distance from each face of R^* (Steps 4 & 5), meaning that each face of $\hat{R}_N$ will stop expanding within $O(\varepsilon)$ distance of the corresponding face of R^* and yielding $\text{vol}(\hat{R}_N \setminus R^*) = O(\varepsilon)$. This completes the proof.

D Experiment Details

D.1 Baselines

Vanilla Cox model In some cases, subgroup discovery may not be necessary and fitting a single model to the entire dataset (as is standard) may suffice. Thus, we include as a baseline a Cox model fit to the whole dataset.

Random At its core, our method selects a group of points from the dataset and uses the resulting bounding box for these points to define the region. Thus, the natural weakest baseline for comparison would be to select several *random* groups of points from the dataset, form their bounding box, and select the region with the best score on the validation set.

No Expansion We test the efficacy of the box expansion part of our procedure by comparing against the results when we use the same computational budget to just select and validate a larger number of core groups according to lowest training EPE.

D.2 Synthetic Data

We created two synthetic datasets to test our method's ability to recover a ground truth region R^*. The features are drawn $X \sim \text{Unif}(B)$, where $B = [-1, 1]^d$ is a bounding box for the data. We set the ground truth region $R^* = [-(1/6)^{1/d}, (1/6)^{1/d}]^d$ so that it occupies $1/6$ of the total area of the feature space, regardless of the dimension. Conditional on $X \in R^*$, the survival time T is drawn according to the Cox model with $\lambda_0(t) \equiv 1$ and $\beta = c\mathbf{1}$, where c is a scalar which varies the effect size and $\mathbf{1}$ is the d-dimensional 1s vector. Note that in this case, the survival times are exponentially distributed. Conditional on $X \notin R^*$, we have two different settings. In one setting (Synth 1), we draw T according to the Cox model with the same baseline hazard $\lambda_0(t) \equiv 1$, but with different Cox coefficients. This matches the setting of the theory. To test the case where our assumptions on the data outside R^* are not met, we also tried generated $T \sim \text{Unif}([0, \tau])$ conditional on $X \notin R^*$ (Synth 2).