DISCOVERING SUBGROUPS WITH EXCEPTIONAL SURVIVAL CHARACTERISTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

In many applications, it is important to identify subpopulations that survive longer or shorter than the rest of the population. In medicine, for example, it allows determining which patients benefit from treatment, and in predictive maintenance, which components are more likely to fail. Existing methods for discovering subgroups with exceptional survival characteristics rely on restrictive assumptions about the survival model (e.g. proportional hazards), require pre-discretized features, and, as they compare average statistics, tend to overlook individual heterogeneity. In this paper, we propose SYSURV, a non-parametric, fully differentiable method that discovers human-readable rules selecting subgroups with exceptional survival characteristics. Empirical evaluation on a wide range of datasets and settings, including a case study on cancer data, shows that SYSURV reveals insightful and actionable survival subgroups, outperforming the state of the art.

1 INTRODUCTION

Survival analysis traditionally focuses on estimating whether a *given* group of individuals has different survival characteristics than a reference population. This has applications in many fields, but most obviously in precision medicine, as it allows characterizing patients who benefit from a treatment. What if the groups are not yet known? Can we *learn* subgroups that stand out in terms of survival characteristics? Can we learn *easily interpretable* rules that select such subgroups? Can we learn these without restrictive assumptions or pre-discretizing features, allowing even heavy censoring, while keeping individual heterogeneity in mind? That is exactly the topic of this paper.

We consider time-to-event data, where each sample $(\mathbf{x}, \delta, t)$ consists of features $\mathbf{x}$, an event (e.g. death) indicator δ, and an observed time t. We are interested in learning conjunctive rules composed of conditions on the features (e.g. $x_1 \in [0.2, 1.0]$ **and** $x_2 \in [0.8, 0.9]$). The rule specifies a subgroup with an exceptional survival characteristic.

For example, on the left in Figure 1, we show the result of our method, SYSURV, on neck cancer treatment data. It learns that "patients whose tumors have altered gene expressions related to DNA repair" tend to respond very poorly to treatment, which is in line with medical understanding (Barker et al., 2015). It is also actionable, as we can test, identify, and treat those patients differently. On the right-hand side, we consider unemployment data. SYSURV finds that "people older than 58 *with* unemployment insurance" tend to have exceptionally long durations until re-employment (group 1), whereas "people who previously earned a low wage and *without* unemployment insurance" tend to find new jobs much more quickly (group 2).

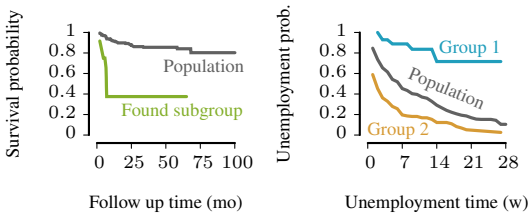


Figure 1: *Survival subgroups*. SYSURV finds patients suffering from a therapy-resistant tumor **(Left)**, and people with exceptionally long (1) resp. short (2) durations until re-employment **(Right)**.

To find these subgroups, we use a *non-parametric* measure of exceptionality based on *individual* survival curves. That is, unlike the state of the art, we do *not* rely on the proportional hazards model, which would restrict us to finding subgroups that have the same functional shape as the overall

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.

population. Using individual survival curves also gives valuable information during optimization, i.e., which subjects to include or exclude. We show how to learn rules that single out subgroups with exceptional survival characteristics in a differentiable manner, automatically finding relevant features and intervals in a way that allows easy optimization, yet crisp logical interpretation after learning.

Through an extensive set of experiments on synthetic and real-world data, we show that our method outperforms state-of-the-art methods – both in terms of recovering the ground truth, and in terms of exceptionality of the found subgroups. Through a case study on treatment of head and neck squamous cell carcinoma, we confirm that SYSURV rediscovers biomarker associations with good, respectively, poor response to treatment, while also identifying novel subgroups that warrant further investigation.

Our main contributions are as follows

1. We propose a differentiable, non-parametric measure of subgroup exceptionality in right-censored data based on individual survival estimates.
2. We ensure a stable, non-moving optimization target by decoupling survival modeling from rule learning, and provide a test for the statistical significance of the discovered subgroups.
3. Our method, SYSURV, empirically outperforms the state-of-the-art on extensive synthetic and real-world benchmarks.

All our code is available in the supplementary material.

2 RELATED WORK

Survival Analysis. Most ML research on time-to-event data focuses on individual outcome prediction, where tree-based (ensemble) methods—including Random Survival Forests—represent the state of the art (Ishwaran et al., 2008; Barnwal et al., 2022; Bertsimas et al., 2022; Zhang et al., 2024; Huisman et al., 2024). Alternatively, neural networks and multimodal architectures have also been explored for estimating survival curves from high-dimensional covariates, sequences, or images (Yu et al., 2011; Fotso, 2018; Cheerla & Gevaert, 2019; Vale-Silva & Rohr, 2021; Meng et al., 2022; 2023; Saeed et al., 2024; Farooq et al., 2025; Li et al., 2025). SYSURV does not compete with these outcome predictors, but rather leverages an individual estimator to obtain its optimization targets.

Survival Clustering aims to partition subjects into groups with distinct survival profiles (Manduchi et al., 2022; Nagpal et al., 2021; Cui et al., 2024; Hou et al., 2024; de Boer et al., 2024). Recent methods parameterize latent mixtures via variational autoencoders (Manduchi et al., 2022), MLPs (Hou et al., 2024), proportional hazards (Nagpal et al., 2021), or vector quantization (de Boer et al., 2024). While related, clustering assigns discrete memberships without descriptive rules explaining *how* a patient belongs to a cluster, often relying on restrictive parametric assumptions.

Subgroup Discovery (Klösgen, 1996) finds and describes subpopulations that exhibit exceptional behavior on a target property (Friedman & Fisher, 1999; Lavrač et al., 2004; Kalofolias & Vreeken, 2022). Exceptional model mining (Leman et al., 2008; Duivesteijn et al., 2016) extends this to discrepancies between local and population models. While traditional methods rely on discrete search heuristics, Xu et al. (2024) showed that gradient-based optimization can learn subgroups on continuous features without pre-discretization. However, their formulation cannot handle right-censoring without inducing bias, collapses under event-time overlap, and ignores directionality.

Survival Subgroup Discovery. Existing methods typically optimize the logrank statistic (Mantel, 1966) using combinatorial heuristics: RULEKIT (Gudyś et al., 2020) employs beam search, FIBERS (Urbanowicz et al., 2023) uses evolutionary algorithms, and ESMAMDS (Vimieiro et al., 2025) relies on ant colony optimization. Relator et al. (2018) find significant subgroups by controlling family-wise error rates, but at high computational cost. In contrast, SYSURV introduces a differentiable rule-learning framework for censored data with a non-parametric objective capturing deviations, with post-hoc statistical guarantees.

3 DISCOVERING SURVIVAL SUBGROUPS

We aim to learn rules that select subgroups with survival trends that are exceptional compared to those of the general population. For that, we need 3 interdependent ingredients: a way to model

subgroup survival, a measure for subgroup exceptionality, and a way to learn these rules. We start with notation, and then present each of these ingredients in turn.

3.1 Notation

We consider time-to-event (survival) data, where we have a dataset $D = \{(\mathbf{x}^{(i)}, t^{(i)}, \delta^{(i)})\}_{i=1}^{n}$ consisting of n i.i.d. realizations, called *subjects*, from a joint distribution $\mathbb{P}(\mathbf{X}, T, \Delta)$. Here, $\mathbf{x} \in \mathbb{R}^p$ is a vector of p covariates that describes the subject, $t \in \mathbb{R}_{\geq 0}$ measures time since the start of its observation, and δ indicates whether the event of interest (e.g. relapse) has occurred yet. If the event did occur ($\delta = 1$), t is the time at which it was observed. If it did not occur, t is the latest time we observed the subject to still be fine, and so the outcome is said to be *censored*. A survival function $S(t)$ gives the probability of a subject surviving beyond time t, i.e. the complementary cumulative distribution function $\mathbb{P}(T > t)$. To ease notation, we denote the domains of $\mathbf{x}$ by $\mathcal{X}$ and t by $\mathcal{T}$.

A subgroup Q is a set of subjects selected from D using a rule $\sigma_Q \colon \mathcal{X} \to \{0, 1\}$ that indicates whether a subject belongs to Q or not, i.e. $Q = \{(\mathbf{x}, t, \delta) \in D \mid \sigma_Q(\mathbf{x}) = 1\}$. The survival function of subgroup Q is denoted by $\hat{S}_Q(t)$.

3.2 Subgroup Survival Model

We model subgroup survival by estimating individual-level survival functions, reflecting the precision medicine principle that patient heterogeneity requires personalized outcome profiles rather than population aggregates (Feuerriegel et al., 2024). Individual survival functions $\hat{S}(t \mid \mathbf{x})$ leverage covariates $\mathbf{x}$ to yield more accurate survival estimates than marginal estimators (Cox, 1972; Kaplan & Meier, 1958), providing fine-grained signal during optimization to guide subgroup inclusion. Given individual estimates, the survival function for a subgroup Q selected by rule σ_Q is defined as

$$\hat{S}_Q(t) \coloneqq \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{x}}}[\hat{S}(t \mid \mathbf{x}) \mid \sigma_Q(\mathbf{x}) = 1] \,.$$

To obtain individual survival functions, we fit a non-parametric conditional estimator, or population model, M over the entire dataset. For our instantiation, we use the random survival forest (RSF) (Ishwaran et al., 2008). This is a fully non-parametric continuous-time model that extends the random forest (Breiman, 2001) to time-to-event data. In contrast to well-known parametric models, it does not restrict us to specific survival distributions (e.g. Weibull) or structures (e.g. Cox) and allows us to find subgroups with exceptional survival characteristics in general. Decoupling population survival estimation from the subgroup search allows us to ensure a stable, non-moving optimization target across candidate subgroups, and to perform a one-time precomputation for computational efficiency.

3.3 Subgroup Exceptionality Measure

Having specified a way to model subgroup survival, we need to define a way to measure the exceptionality of a given subgroup versus a reference group. In traditional survival analysis, the groups of interest are assumed to be *given* and exceptionality is measured on average, at the group level. As we will detail below, this has drawbacks in terms of sensitivity to per-subject deviations when we aim to *learn* the subgroup.

First, we show in Fig. 2 (left) how individual estimates can *reveal* differences from the reference survival that group-level estimates *obscure*. The dark green line shows the group-level average over four individuals (light green lines). The average curve obscures that the curves of subjects 1 and 2 are very different, i.e. delayed, compared to those of subjects 3 and 4, effectively underestimating the true exceptionality.

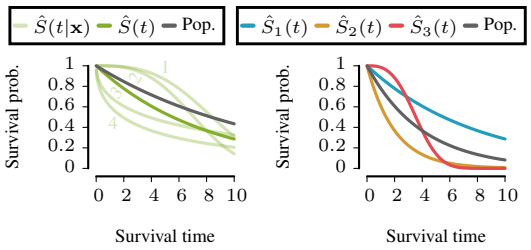


Figure 2: **(Left)** Individual survival functions $\hat{S}(t|\mathbf{x})$ are more informative than their group-level estimate $\hat{S}(t)$. **(Right)** The absolute difference between group and reference survival (Pop.) is effective in the absence of crossing ($\hat{S}_1$ and $\hat{S}_2$) and in its presence ($\hat{S}_3$), unlike the signed difference.

In general, individual-level exceptionality measures offer increased sensitivity to deviations of individual survival functions, such as those crossing the reference, making them particularly suited for optimization of subgroup membership rules.

Independent of whether we measure exceptionality at the group or individual level, it is important to consider the absolute difference to the reference. We give an example in Fig. 2 (right). While it is trivial to identify groups 1 and 2 as exceptional even when we measure relative to the reference, the large exceptionality of group 3 only becomes clear when we take the absolute difference.

Next, we formalize these intuitions, and show that considering absolute differences of individual estimates to the reference gives us more signal than the absolute difference in group means.

Proposition 3.1. *Given two groups A and B, selectable by rule σ_A and σ_B, respectively, for which the expected group-level survival at any time t are $\hat{S}_A(t)$ and $\hat{S}_B(t)$, and for which individual-level survival is denoted by $\hat{S}(t \mid \mathbf{x})$. The expected absolute difference in survival of the subjects selectable by σ_A is as, or more, sensitive than the absolute difference of $\hat{S}_A(t)$ from $\hat{S}_B(t)$,*

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ell_t^1(s(\mathbf{x}), s_B) \mid \sigma_A(\mathbf{x}) = 1] \geq \ell_t^1(s_A, s_B) \ ,$$

where $\ell_t^1(\cdot,\cdot)$ is an absolute difference measure, and we write $s_\circ$ for $\hat{S}_\circ(t)$, and $s(\mathbf{x})$ for $\hat{S}(t \mid \mathbf{x})$.

We provide the proof, as well as an informative Corollary, in Appx. A. For our purposes, s_A would typically belong to a group of interest, e.g. a subgroup, and s_B to a reference group, e.g. the entire dataset. To measure the exceptionality between two survival estimates $s_A, s_B \geq 0$ at some time t, we consider the L^1 distance $\ell_t^1(s_A, s_B) \coloneqq |s_A - s_B|$.

To capture the survival interplay between groups, we leverage individual estimates over the entire time domain $\mathcal{T} = [0, t_{\max}]$ by extending the absolute difference measure $\ell_t^1(\cdot,\cdot)$ for two survival functions S_A, S_B as $\ell_{\mathcal{T}}^1(S_A, S_B) \coloneqq \int_{t\in\mathcal{T}} |S_A(t) - S_B(t)|\, dt$. Finally, we define our exceptionality measure as the expected difference in survival of the subjects of a subgroup selectable by σ from the estimated survival in the population $\hat{S}_D(t)$ throughout $\mathcal{T}$ as

$$\phi(\sigma, \sigma_D) \coloneqq \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ell_{\mathcal{T}}^1(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)) \mid \sigma(\mathbf{x}) = 1] \ . \tag{1}$$

No Blind Spots. Placing the absolute value inside the expectation introduces a crucial blind spot. Under group-level difference ($|\mathbb{E}[\hat{S}(t \mid \mathbf{x})] - \hat{S}_D(t)|$), individual deviations in opposing directions cancel out. In contrast, expected individual-level deviations ($\mathbb{E}[|\hat{S}(t \mid \mathbf{x}) - \hat{S}_D(t)|]$) reward any departure from the reference population without mutual cancellation, providing better gradients throughout training. While this formulation prioritizes individual exceptionality over strictly enforced homogeneity, discovered subgroups empirically maintain strong internal cohesion (Appx. K.2).

In our experiments, we will compare to methods that rely on the logrank for measuring exceptionality, a signed, group-level statistic, widely used in survival analysis.

3.4 Rule Learner

Armed with our subgroup survival model and exceptionality measure, we now present SYSURV for learning rules that select subgroups with exceptional survival characteristics, via gradient-descent.

Learnable Rules. Subgroups are selected via rules, so we first formalize the language of these rules. Traditionally, a subgroup is defined by a hard rule $\sigma\colon \mathbf{x} \to \bigwedge_{j=1}^{p} \pi(x_j; \alpha_j, \beta_j)$, where each π is a Boolean condition evaluating to true (1) if a covariate x falls within the interval defined by lower and upper bounds $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}$, e.g. "$18 <$ age < 32". To integrate rule induction into a gradient-based optimization pipeline, we employ a continuous relaxation of these logical expressions.

To facilitate differentiable rule induction, we use soft rules $\hat{\sigma}\colon \mathcal{X} \to [0,1]$ consisting of soft conditions $\hat{\pi}\colon \mathbb{R} \to [0,1]$. These soft conditions model the probability of a covariate being inside the specified interval via a composition of two opposing sigmoids located at the learnable bounds α and β. A temperature hyperparameter $\tau > 0$ controls the strictness of the bounds; as $\tau \to 0$, the soft condition $\hat{\pi}$ converges to a Boolean interval (Fig. 3a).

We adopt the following soft condition and differentiable rule learner definitions

$$\hat{\pi}(x_j; \alpha_j, \beta_j, \tau) := \frac{e^{\frac{1}{\tau}(2x_j - \alpha_j)}}{e^{\frac{1}{\tau}x_j} + e^{\frac{1}{\tau}(2x_j - \alpha_j)} + e^{\frac{1}{\tau}(3x_j - \alpha_j - \beta_j)}} \quad \text{and} \quad \hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) := \frac{\sum_{j=1}^{p} w_j}{\sum_{j=1}^{p} w_j \, \hat{\pi}_j^{-1}} \,,$$

respectively, where $\hat{\pi}_j$ denotes $\hat{\pi}(x_j; \alpha_j, \beta_j, \tau)$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w})$, and $\mathbf{w} \in \mathbb{R}^p$ represents learnable feature weights (Xu et al., 2024). These weights allow the model to perform feature selection, i.e. a covariate x is actively included in the conjunction when $w > 0$ and effectively ignored as $w \to 0$. The harmonic mean structure is chosen because it serves as a smooth approximation of the logical-and operator; if any single active condition $\hat{\pi}$ approaches zero, the entire rule output $\hat{\sigma}$ tends toward zero. Henceforth, the output $\hat{\sigma}(\mathbf{x})$ can be interpreted as the probability $\mathbb{P}(\hat{\sigma}(\mathbf{X}) = 1 \mid \mathbf{X} = \mathbf{x})$ that a subject belongs to the subgroup (yellow box in Fig. 3b).

Objective. Our goal is to learn those rules to select the subgroups that maximize our objective. For rule updates, we need to take the gradient of our objective w.r.t the parameters of the soft rule $\hat{\sigma}$. Hence, we rewrite the expectation in Eq. (1) to derive our objective where $\hat{\sigma}$ explicitly appears as

$$\phi(\hat{\sigma}, \hat{\sigma}_D) = \int_{\mathbf{x} \in \mathcal{X}} \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)) \, \frac{\mathbb{P}(\mathbf{X} = \mathbf{x})}{\mathbb{P}(\hat{\sigma}(\mathbf{X}) = 1)} \, \hat{\sigma}(\mathbf{x}) \, d\mathbf{x} \,, \tag{2}$$

by using the expected value definition, marginalization rule and Bayes' theorem. We estimate the integral over $\mathcal{X}$ and $\mathcal{T}$ in Eq. (2) using the standard Monte Carlo and trapezoidal methods, respectively. Our (core) objective is defined as

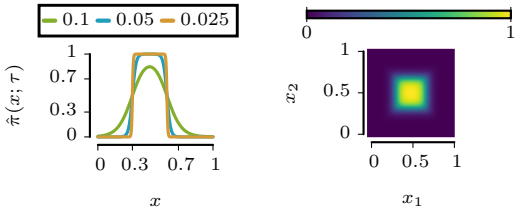
$$\hat{\phi}(\hat{\sigma}, \hat{\sigma}_D) := \frac{1}{n_{\hat{\sigma}}} \sum_{i=1}^{n} \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}^{(i)}), \hat{S}_D(t)) \, \hat{\sigma}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \,,$$

where the subgroup size $n_{\hat{\sigma}} = \sum_{i=1}^{n} \hat{\sigma}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$.

To avoid learning overly specific subgroups, and as is standard in subgroup discovery, we introduce a size penalty n_σ^γ to our objective. This allows us to control the trade-off between exceptionality and subgroup size via a hyperparameter $\gamma \in [0, 1]$. To learn non-redundant subgroups, we sequentially learn consecutive rules that describe subgroups whose survival is simultaneously exceptional w.r.t the population and the set of q preceding subgroups. For that, we add a regularizing term to our objective that ensures that we get a diverse set of subgroups. Thus,

$$\arg\max_{\boldsymbol{\theta}} \left[n_{\hat{\sigma}}^\gamma \, \hat{\phi}(\hat{\sigma}, \hat{\sigma}_D) + \sum_{g=1}^{q} n_{\hat{\sigma}}^\gamma \, \hat{\phi}(\hat{\sigma}, \hat{\sigma}_g) \right],$$

gives our diversity, size-aware (full) objective.



(a) Soft condition (b) Soft rule

Figure 3: *Soft rules.* The soft condition approaches a hard interval with decreasing temperature $\tau \to 0$ **(a)**. Multiple soft conditions combine to form a soft rule depicted as a hyper box in the covariate space **(b)**.

Optimization. Gradient-based optimization allows us to efficiently learn the parameters of our rules, and therewith scale SYSURV to large datasets. For every subject, each feature is subjected to its corresponding soft condition according to the learned bounds, which are then weighted and combined into a soft rule that predicts a membership probability. We iteratively learn the weights and bounds using standard first-order gradient-based optimization techniques (Kingma & Ba, 2015) for a number of epochs, while annealing the temperature to arrive at crisp discretizations. For the pseudocode and the annealing schedule see Appx. B.

Rule Sparsity. In correlated data, there often exist different ways (i.e. different conditions over different features) to select the same subgroup. Some of these features may be easier to measure, others may be causally more relevant. As the goal of SYSURV is exploratory, the default is to not sparsify the rule but rather let the user decide which features (terms) are most informative. In specific cases, such as when dealing with high-dimensional data, the user may instead wish a small, non-redundant set of terms. To this end, we propose a post-hoc strategy, in which we iteratively remove that condition from the rule that minimally changes which subjects are selected, and stop when a minimal Jaccard similarity is reached. We present the pseudocode in Appx. C.

3.5 Statistical & Theoretical Guarantees

In critical applications, it is vital to statistically validate the results. We provide post-hoc statistical guarantees via permutation-tests. Following Duivesteijn & Knobbe (2011), we create independent copies of the input dataset and break the dependency between the covariates and the outcomes by random permutation. Running SYSURV on each of these copies enables building a distribution of false exceptionalities, by applying the method of moments. Then, we can determine the significance of the exceptionality of a proper subgroup from its Z-score. We use an alpha cutoff of 0.05, and apply Bonferroni correction when discovering multiple subgroups (Dunn, 1961). Details in Appx. D.

In Appx. E, we prove Lipschitz stability to population model errors, uniform consistency over the rule space, asymptotic subgroup recovery under a margin, and vanishing relaxation bias.

4 Experiments

Next, we empirically evaluate SYSURV on synthetic and real-world data. We compare it to state-of-the-art methods RuleKit (Gudyś et al., 2020), FIBERS (Urbanowicz et al., 2023), and EsmamDs (Vimieiro et al., 2025). All baselines optimize the logrank for exceptionality, but do so employing different search strategies. We use the implementations of the respective authors, and allow each method, including SYSURV, up to 48 hours per experiment. We tune hyperparameters for all methods using grid search (see Appx. F). We make all code available (see supplementary), and give the pseudocode for our data generator and additional results in Appx. H and K.

4.1 Synthetic Data

We evaluate on synthetic data with known ground truth ($n = 10\,000$, $p = 10$). We sample background survival times from a Weibull distribution, $Weibull(1.5, 5)$, subject to 10% uniform censoring. We randomly plant a conjunctive rule over $k = 2$ features covering 20% of the population, rescaling its event times so that subjects experience their event $5\times$ earlier than the population on expectation ($Weibull(1.5, 1)$). We provide complete generator details and pseudocode in Appx. H.

Scalability. We assess the performance of SYSURV and its competitors on data where we vary the number of features. We report the average F_1-scores over 10 runs in Fig. 4 (left). We see that all methods are stable in terms of F1-scores as the dimensionality increases, and that RuleKit does not finish in time for more than 500 features. SYSURV outperforms the closest competitor, EsmamDs, by a wide margin. We provide the runtimes in Fig. 9 (left) in Appx. K.

Censoring. We assess robustness to censoring when increasing the percentage of censored subjects. We give the results in Fig. 4 (center); we see that SYSURV performs very well, with F_1 of 0.8 and above for up to 60% censoring. We check whether having the subgroup outsurvive the population, and vice versa, affects performance by varying the subgroup and fixing the population hazards, and give the results in Fig. 9 (center).

Subgroup Size. We assess retrieval quality as the percentage of subjects that belong to the planted subgroup increases. We see in Fig. 4 (right) that all methods improve as the planted subgroup size increases, with SYSURV performing the best. We attribute the improvement in the baselines to them generally preferring larger subgroups, according to our observations (Appx. K). In Fig. 9 (right), we also provide results on sample sufficiency by varying dataset sizes.

Model Sensitivity. We also evaluate sensitivity to the underlying population survival model across four distinct regressors: RSF, Gradient Boosting, ExtraTrees, and single Decision Trees (Appx. K.3). SYSURV is not sensitive to the specific model family, but rather correlates directly with the population model's goodness-of-fit as measured by the Integrated Brier Score (IBS) on held-out samples, demonstrating that SYSURV reliably leverages faithful survival estimates regardless of architecture.

Ablation. To examine the drivers of retrieval improvement over the baselines, we ablate discretization first, then compound with search strategy (Appx. K.3). Directly optimizing continuous intervals outperforms pre-discretization ($F_1 = 0.87$ vs. 0.76). Evaluating SYSURV with discrete heuristic search from EsmamDs drops retrieval to $F_1 = 0.37$, confirming continuous optimization as the primary driver. Under 10-fold cross-validation ($n = 5000$), SYSURV remains stable from train to test ($F_1 = 0.89 \rightarrow 0.88$), whereas baselines overfit.

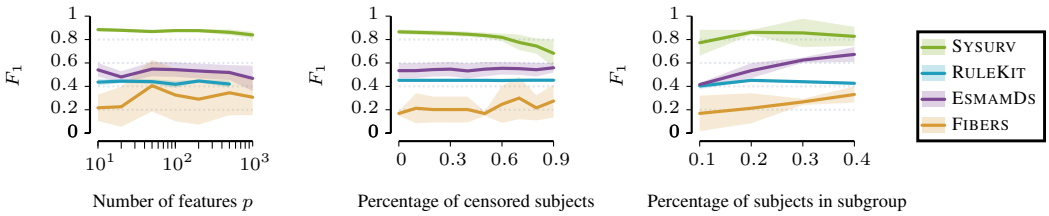


Figure 4: *Synthetic setting.* Comparison of SYSURV and each of RULEKIT, ESMAMDS, and FIBERS in terms of F_1-scores recovering planted subgroups with increasingly large dataset dimensionalities **(Left)**, increasingly many censored subjects **(Center)**, and increasingly large planted subgroups **(Right)**. ESMAMDS is the closest competitor to SYSURV closely followed by RULEKIT. Higher is better. The shaded areas show ±1 standard error over 10 runs.

Overall, SYSURV is stable with average $F_1 \approx 0.8$, up to very high rates of censoring ($\geq 80\%$). ESMAMDS is the closest competitor, hovering around 0.55, closely followed by RULEKIT consistently at 0.45, while FIBERS follows in all experiments, with average $F_1 \approx 0.35$ and large standard errors.

4.2 REAL-WORLD DATA

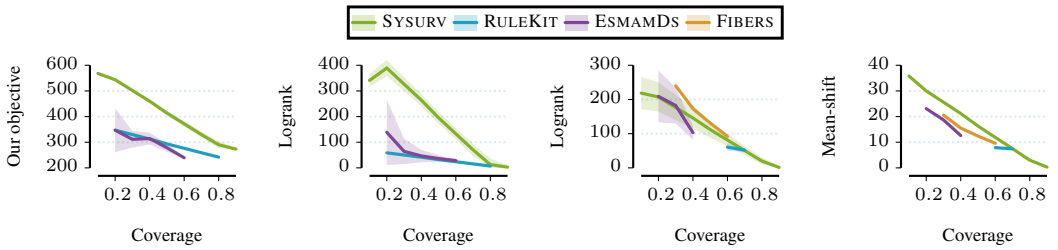
Next, we evaluate how well SYSURV performs on real-world data across 13 time-to-event datasets from the SurvSet repository (Drysdale, 2022) that span different domains and have at least 1000 subjects and 7 covariates. Because the ground truth is unknown, we evaluate discovered subgroups along two axes: out-of-sample generalization and size-controlled exceptionality frontiers.

Out-of-Sample Exceptionality. First, to evaluate generalization to unseen subjects, we evaluate out-of-sample exceptionality on held-out test sets using 10-fold cross-validation across 3 metrics: logrank, mean-shift, and our objective, defined in Appx. I. As before, we compare SYSURV to RULEKIT, ESMAMDS, and FIBERS, where we consider the best of the top-5 subgroups found by each baseline, and the first subgroup found by SYSURV. We report the results in Table 1. For mean-shift, we are only interested in the magnitude of the shift, and hence report absolute values. We provide full-dataset in-sample results, qualitative curves, and detailed diagnostics in Appx. K.

We see that SYSURV finds subgroups that stay exceptional on the test set, outperforming competitors. It achieves the best average rank on our objective (1.31 ± 0.72) and mean-shift (1.54 ± 1.01), and ties for first on logrank (2.08 ± 1.14). For datasets such as *Actg*, *DataDIVAT2*, *Grace*, *ProstateSurvival*, *Rdata*, and *Scania*, it scores highest on each metric.

Table 1: *Real-world setting.* Exceptionality on held-out test sets of subgroups discovered by SYSURV (*Ours*), RULEKIT (RK), ESMAMDS (ED), and FIBERS (FIB), measured using 3 metrics (our objective, logrank, and mean-shift). Greater is better. Highest values marked in bold.

	Our objective				Logrank				Mean-shift			
	Ours	RK	ED	FIB	*Ours*	RK	ED	FIB	*Ours*	RK	ED	FIB
UnempDur	**4.7**	3.7	3.7	3.7	7.1	**23.0**	**23.0**	**23.0**	**2.2**	2.1	2.1	2.1
Nwtco	**1691.5**	859.2	1575.9	384.1	21.5	6.3	**23.5**	2.6	**806.1**	248.4	578.8	86.3
Rott2	**53.9**	35.0	50.1	50.1	19.3	10.6	**30.3**	**30.3**	25.2	11.4	**25.2**	**25.2**
Rdata	**1502.7**	1248.4	1278.5	893.7	**12.3**	10.4	11.0	6.3	**1154.8**	895.4	955.9	535.2
Aids2	**230.5**	128.7	111.6	27.1	0.8	**1.3**	0.7	0.0	**72.6**	58.9	41.8	0.2
Dialysis	**13.5**	9.6	6.2	5.9	11.1	10.7	18.9	**20.7**	**8.8**	7.2	3.0	3.2
TRACE	1.7	1.6	**2.1**	1.9	3.1	5.6	16.7	**19.4**	0.6	0.7	**1.9**	1.6
Support2	205.1	230.1	**314.7**	171.6	5.3	11.9	**26.3**	5.0	197.8	219.9	**368.5**	105.1
DataDIVAT2	**3.1**	2.3	2.4	1.9	**4.2**	2.3	3.2	1.6	**2.4**	1.5	1.7	1.2
ProstateSurv.	**16.9**	7.5	14.5	9.9	**32.1**	5.3	29.5	18.5	**4.2**	2.9	3.4	2.0
Actg	**35.3**	30.4	24.0	6.6	**4.8**	2.9	2.8	0.2	**67.1**	14.5	14.4	1.3
Scania	**3.1**	2.8	2.5	2.0	**1.8**	1.6	1.2	1.3	**2.5**	1.7	1.3	1.0
Grace	**43.7**	27.7	29.8	20.2	**3.9**	3.4	3.3	0.6	**36.6**	19.9	20.5	5.7
Avg. rank	**1.31**	2.77	2.15	3.46	**2.08**	2.62	**2.08**	2.92	**1.54**	2.62	2.15	3.38
± std	±0.72	±0.80	±0.66	±0.84	±1.14	±1.00	±0.92	±1.33	±1.01	±0.62	±0.86	±1.00



(a) *Support2* (b) *Support2* (c) *Rott2* (d) *Rott2*

Figure 5: *Coverage vs. exceptionality frontiers.* Exceptionality across coverage levels (10% to 90%) with bootstrapped LOWESS error bands (200 resamples) on *Support2* for our objective (**a**) and logrank (**b**), and on *Rott2* for logrank (**c**) and mean-shift (**d**). SYSURV discovers more exceptional subgroups than baselines across matched coverage levels. Additional per-dataset frontiers in Appx. K.2.

Size-Controlled Exceptionality. A potential confound in subgroup discovery is that average deviation metrics naturally reward smaller, more extreme cohorts, whereas logrank scales with sample size. To verify whether SYSURV's advantage stems from genuine contrast rather than differing group sizes, we evaluate the coverage vs. exceptionality frontier using bootstrapped LOWESS fits (200 resamples) across coverage levels from 10% to 90% on each dataset (Appx. K.2). In Fig. 5, we show the resulting frontiers on *Support2* (our objective and logrank) and *Rott2* (logrank and mean-shift).

We see that SYSURV discovers more exceptional subgroups than all baselines across matched coverage levels. Across all benchmark datasets, SYSURV ranks first across all coverage levels on our objective (51 out of 51 available comparisons), achieves the top rank in 34/51 combinations on mean-shift, and remains competitive on logrank (13/51), confirming that its lead holds at matched subgroup sizes.

Crucially, while baselines cluster between 20% and 70% with wide coverage gaps, SYSURV populates the entire axis via its tunable size control. Further, SYSURV frequently discovers subgroups with under 10% coverage (median 6.5%; Table 3 in Appx. K.2) that are almost always sufficiently large and statistically acceptable. Beyond our objective, SYSURV retains an edge on mean-shift across most benchmarks, with baselines leading on both external metrics on only 3 of 10 datasets.

Overall, these real-world experiments show that SYSURV performs well on a wide range of settings and on data from different domains. Learned rules consistently generalize to unseen subjects on held-out test data while outperforming competitors across multiple exceptionality metrics. Importantly, coverage-matched frontiers confirm that this advantage reflects genuine survival contrast rather than subgroup size artifacts, yielding more exceptional subgroups across user-adjustable coverage levels.

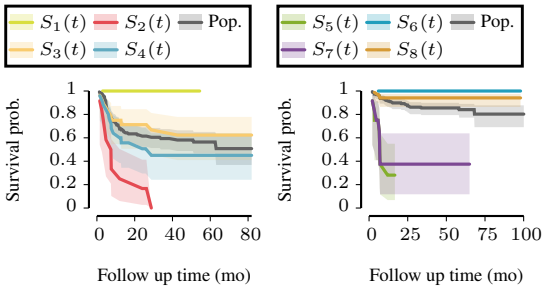
4.3 CASE STUDY: NECK CANCER

Last, in collaboration with domain experts, we qualitatively evaluate SYSURV via a case study. We retrospectively consider locally advanced head and neck squamous cell carcinoma (HNSCC) patients undergoing radiochemotherapy (Schmidt et al., 2020). The treatment data consists of two cohorts. The first, or primary, cohort are patients who only received radiochemotherapy ($n = 136$). The second, or postoperative, cohort are patients who received radiochemotherapy after a tumor removal operation ($n = 190$). The covariates are tumor gene expressions ($p = 158$) and the outcome is the time until tumor recurrence. Censoring is 60% for the first and 85% for the second cohort. From a clinical perspective, we are both interested in subgroups of subjects that respond better, and those that respond worse to treatment, as this allows more targeted (personalized) treatment.

We run SYSURV on both datasets and discover four subgroups each. We show the survival curves in Fig. 6 and give the rules in Tables 14 and 15 in Appx. K. For both cohorts, SYSURV finds two subgroups that respond better, and two that respond worse, to the respective treatments than the overall populations. Undergoing the tumor removal operation expectedly improves the overall survival, nevertheless, subgroup discovery is perfectly suited, when randomized control trials may fall short, for pointing out when this is not the case, and when it is even detrimental.

First, consider the primary cohort. We see in Fig. 6a that approximately half of the population are still alive after 80 months. Subgroups S_1 (yellow, n=6) and S_3 (orange, n=28) identify subjects that respond better, while subgroups S_2 (red, n=12) and S_4 (blue, n=25) respond worse. Although the effect sizes are not significant under our test, it is highly encouraging that the rules select on meaningful biomarkers. Schmidt et al. (2020) recently reported on the survival profile when stratifying primary radiochemotherapy patients based on hypoxia 15- and 26-gene expressions. Notably, multi-gene signatures of comparable complexity are well-established clinical precedents in HNSCC, where validated prognostic classifiers routinely comprise 15 or 26 genes (Toustrup et al., 2011; Eustace et al., 2013; Toustrup et al., 2016). Subgroup S_4 not only shows a similar profile, but like subgroup S_3 also mainly selects on hypoxia-associated genes.

Next, we investigate the results on the postoperative cohort. Here, almost 80% of patients survive longer than 80 months (Fig. 6b). We are specifically interested in subgroups S_5 (green, n=12) and S_7 (purple, n=12) as these identify subjects responding very poorly. Both are significant under our statistical model. As before, the corresponding rules make biological sense: they both select on genes relating to cellular communication, metabolism, and DNA repair, which are related to highly aggressive tumor subtypes (Lendahl et al., 2009; Toustrup et al., 2011) for which conventional treatment not only fails to improve patient condition but even promotes recurrence (Barker et al., 2015). The learned rules allow clinicians to pre-screen these subjects and offer them alternative treatment.



(a) Primary cohort (b) Postoperative cohort

Figure 6: *Case study.* SYSURV discovers, overall, sets of diverse subgroups in the primary cohort (**a**), and two subgroups of poor responders in the post-operative cohort (**b**) of HNSCC data. The shaded areas show 95% confidence intervals.

5 CONCLUSION

We proposed SYSURV, a differentiable framework for survival subgroup discovery. By quantifying subgroup exceptionality through individual-level survival deviations from non-parametric estimates, SYSURV avoids restrictive parametric assumptions and overcomes effect cancellation inherent in aggregate statistics. Through gradient-based optimization of soft feature intervals, it learns crisp, human-interpretable conjunctive rules without pre-discretization. Extensive experiments on synthetic and real-world datasets demonstrate that SYSURV often outperforms state-of-the-art baselines across held-out test sets and coverage-matched frontiers, while discovering clinically meaningful biomarker signatures in neck cancer patients.

Limitations. SYSURV provides a novel method to uncover, possibly novel, subgroups with exceptional survival characteristics. SYSURV is strictly associational, and while its results may indicate (unknown) causal mechanisms, it does not provide any guarantees in this regard. Before putting any of its results into (clinical) practice, randomized controlled trials, or other ways to verify causality should be employed. Moreover, SYSURV relies on the population model to obtain individual survival functions, which dictates the quality of the results. In particular, our objective may become overestimated, but there exist asymptotic methods for debiasing it (Ito et al., 2018).

Future Work. In SYSURV, we consider tabular data; however, many applications involve other data modalities such as images or sequences. We intend to extend SYSURV to incorporate such structured data alongside tabular covariates in a meaningful way to discover survival subgroups characterized by both types of data. This may involve anchoring the discovered subgroups in interpretable tabular features as well as emergent visual features, or sequential patterns. Furthermore, although we do not constrain SYSURV to finding compact, low-variance subgroups so as to keep it as sensitive as possible, empirical evaluation shows that discovered cohorts exhibit strong internal cohesion in practice (see Appx. K.2). Nevertheless, intra-subgroup cohesion can be a favorable inductive bias to enforce for some applications (Boley et al., 2017), which can be achieved by adding a tunable regularization term to our objective that encourages low intra-subgroup variance in survival, at a computational cost.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in drafting paragraphs around results in the appendix, polishing sentences here and there, grammar, spelling and punctuation checking, tightening and sketching auxiliary proofs, and understanding case study results. Additionally, we used generative AI tools for editing scientific tables and software code.

We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, and cleaning and reformatting datasets, while assisting with translation is not applicable to this work.

We have reviewed all AI-assisted work. We checked LLM-generated text, math, and code for potential mistakes and verified/tested them for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Given these potentially interesting cohorts, researchers may reproduce previous findings or make new discoveries that allow for more informed decisions about treatment and intervention strategies. We underscore the importance of responsible use of our method in critical settings: the rules SYSURV learns are based on correlations extracted from potentially selection-biased observational data. Thus, any decisions based on these rules must be made with caution and in conjunction with domain experts as misuse may lead to unfair treatment of specific demographics. Further, we declare that this work does not involve crowdsourcing nor research with human subjects.

REPRODUCIBILITY STATEMENT

The main text, appendices and source code together provide the means necessary to reproduce the results in the paper. These include the method, post hoc routines, mathematical proofs, algorithms, experimental setup, synthetic data generation, real-world datasets or their references, hyperparameters, seeds, compute, and training details.

REFERENCES

Holly E Barker, James TE Paget, Aadil A Khan, and Kevin J Harrington. The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. *Nature Reviews Cancer*, 15(7): 409–425, 2015.

Avinash Barnwal, Hyunsu Cho, and Toby Hocking. Survival Regression with Accelerated Failure Time Model in XGBoost. *Journal of Computational and Graphical Statistics*, 31(4):1292–1302, 2022.

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

Mario Boley, Bryan R Goldsmith, Luca M Ghiringhelli, and Jilles Vreeken. Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery. *Data Mining and Knowledge Discovery*, 31(5):1391–1418, 2017.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.

David R Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–202, 1972.

Chang Cui, Yongqiang Tang, and Wensheng Zhang. Deep Survival Analysis With Latent Clustering and Contrastive Learning. *IEEE Journal of Biomedical and Health Informatics*, 28(5):3090–3101, 2024.

Jasper de Boer, Klest Dedja, and Celine Vens. SurvivalLVQ: Interpretable supervised clustering and prediction in survival analysis via Learning Vector Quantization. *Pattern Recognition*, 153:110497, 2024.

Erik Drysdale. SurvSet: An open-source time-to-event dataset repository. *arXiv preprint arXiv:2203.03094*, 2022.

Wouter Duivesteijn and Arno Knobbe. Exploiting False Discoveries–Statistical Validation of Patterns and Quality Measures in Subgroup Discovery. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 151–160. IEEE, 2011.

Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional Model Mining: Supervised descriptive local pattern mining with complex target concepts. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

Anthony Eustace, Natalie Mani, Paul N Span, Jillian J Irlam, Jacqueline Taylor, Gillian N Betts, Helen Denley, Crispin J Miller, Jarrod J Homer, Madan Sen, et al. A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer. *Clinical Cancer Research*, 19(17):4879–4888, 2013.

Aiman Farooq, Deepak Mishra, and Santanu Chaudhury. Survival Prediction in Lung Cancer through Multi-Modal Representation Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3907–3915. IEEE, 2025.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Stephane Fotso. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv preprint arXiv:1801.05512*, 2018.

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

Adam Gudyś, Marek Sikora, and Łukasz Wróbel. RuleKit: A comprehensive suite for rule-based learning. *Knowledge-Based Systems*, 194:105480, 2020.

Bojian Hou, Zixuan Wen, Jingxuan Bao, Richard Zhang, Boning Tong, Shu Yang, Junhao Wen, Yuhan Cui, Jason H Moore, Andrew J Saykin, Heng Huang, Paul M. Thompson, Marylyn D. Ritchie, Christos Davatzikos, and Li Shen. Interpretable deep clustering survival machines for Alzheimer's disease subtype discovery. *Medical Image Analysis*, 97:103231, 2024.

Tim Huisman, Jacobus GM van der Linden, and Emir Demirović. Optimal Survival Trees: A Dynamic Programming Approach. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12680–12688, 2024.

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

Shinji Ito, Akihiro Yabe, and Ryohei Fujimaki. Unbiased Objective Estimation in Predictive Optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2176–2185. PMLR, 2018.

Janis Kalofolias and Jilles Vreeken. Naming the Most Anomalous Cluster in Hilbert Space for Structures with Attribute Information. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4057–4064, 2022.

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

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Willi Klösgen. Explora: a multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pp. 249–271. American Association for Artificial Intelligence, 1996.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5(Feb):153–188, 2004.

Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional Model Mining. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pp. 1–16. Springer, 2008.

Urban Lendahl, Kian Leong Lee, Henry Yang, and Lorenz Poellinger. Generating specificity and diversity in the transcriptional response to hypoxia. *Nature Reviews Genetics*, 10(12):821–832, 2009.

Yingxue Li, Xiaolu Zhang, Jun Hu, Wenwen Xia, Ziyi Liu, Xiaobo Qin, Binjie Fei, and Jun Zhou. SurvFormer: A Transformer Based Framework for Survival Analysis in Insurance Underwriting. In *Proceedings of the International World Wide Web Conference (WWW)*, pp. 325–333, 2025.

Laura Manduchi, Ričards Marcinkevičs, Michela C. Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C. Neidert, Marc Pfister, Bram Stieltjes, and Julia E Vogt. A Deep Variational Approach to Clustering Survival Data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, 1966.

Mingyuan Meng, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim. DeepMTS: Deep Multi-Task Learning for Survival Prediction in Patients With Advanced Nasopharyngeal Carcinoma Using Pretreatment PET/CT. *IEEE Journal of Biomedical and Health Informatics*, 26 (9):4497–4507, 2022.

Mingyuan Meng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. Merging-Diverging Hybrid Transformer Networks for Survival Prediction in Head and Neck Cancer. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 400–410. Springer, 2023.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep Cox Mixtures for Survival Regression. In *Machine Learning for Healthcare Conference*, pp. 674–708. PMLR, 2021.

Raissa T Relator, Aika Terada, and Jun Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Medical Genomics*, 11(Suppl 2):31, 2018.

Numan Saeed, Muhammad Ridzuan, Fadillah Adamsyah Maani, Hussain Alasmawi, Karthik Nandakumar, and Mohammad Yaqub. SurvRNC: Learning Ordered Representations for Survival Prediction Using Rank-N-Contrast. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 659–669. Springer, 2024.

Stefan Schmidt, Annett Linge, Marianne Grosser, Fabian Lohaus, Volker Gudziol, Alexander Nowak, Ingeborg Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Henning Schäfer, Anca-Ligia Grosu, Amir Abdollahi, Jürgen Debus, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Gustavo B. Baretton, Frank Buchholz, Michael Baumann, Mechthild Krause, and Steffen Löck. Comparison of GeneChip, nCounter, and Real-Time PCR–Based Gene Expressions Predicting Locoregional Tumor Control after Primary and Postoperative Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *The Journal of Molecular Diagnostics*, 22(6):801–810, 2020.

Kasper Toustrup, Brita Singers Sørensen, Marianne Nordsmark, Morten Busk, Carsten Wiuf, Jan Alsner, and Jens Overgaard. Development of a Hypoxia Gene Expression Classifier with Predictive Impact for Hypoxic Modification of Radiotherapy in Head and Neck Cancer. *Cancer Research*, 71 (17):5923–5931, 2011.

Kasper Toustrup, Brita Singers Sørensen, Mohamed A. Hassan Metwally, Trine Tramm, Lise Saksø Mortensen, Jens Overgaard, and Jan Alsner. Validation of a 15-gene hypoxia classifier in head and neck cancer for prospective use in clinical trials. *Acta Oncologica*, 55(9-10):1091–1098, 2016.

Ryan Urbanowicz, Harsh Bandhey, Malek Kamoun, Nolan Fogarty, and Yi-An Hsieh. Scikit-FIBERS: An 'OR'-Rule Discovery Evolutionary Algorithm for Risk Stratification in Right-Censored Survival Analyses. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pp. 1846–1854, 2023.

Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021.

Renato Vimieiro, Juliana Barcellos Mattos, and Paulo SG de Mattos Neto. EsmamDS: A more diverse exceptional survival model mining approach. *Information Sciences*, 690:121549, 2025.

Sascha Xu, Nils Philipp Walter, Janis Kalofolias, and Jilles Vreeken. Learning Exceptional Subgroups by End-to-End Maximizing KL-Divergence. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 55267–55285. PMLR, 2024.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2011.

Rui Zhang, Rui Xin, Margo Seltzer, and Cynthia Rudin. Optimal Sparse Survival Trees. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 352–360. PMLR, 2024.

Appendix: Table of Contents

Below is an index of all supplementary materials and appendices accompanying this paper:

- **Appendix A: Proof of Proposition 3.1** 15
- **Appendix B: The SYSURV Algorithm** 16
- **Appendix C: Rule Pruning** 17
- **Appendix D: Statistical Validation** 18
- **Appendix E: Statistical Properties and Theoretical Guarantees** 19
- **Appendix F: Hyperparameters of SYSURV and Baselines** 26
- **Appendix G: Computational Complexity of SYSURV** 26
- **Appendix H: Survival Data Generation** 27
- **Appendix I: Exceptionality Metrics** 28
- **Appendix J: Connection Between Our Measure and the Logrank** 28
- **Appendix K: Additional Experimental Results and Ablations** 32
  - **K.1: Synthetic Data** 32
  - **K.2: Real-World Data** 32
    - In-Sample Exceptionality 32
    - Intra-Subgroup Cohesion 35
    - Coverage vs. Exceptionality Frontiers 36
  - **K.3: Ablation Studies and Model Sensitivity** 40
  - **K.4: Case Study** 41

A Proof of Proposition 3.1

Proposition 3.1. *Given two groups A and B, selectable by rule σ_A and σ_B, respectively, for which the expected group-level survival at any time t are $\hat{S}_A(t)$ and $\hat{S}_B(t)$, and for which individual-level survival is denoted by $\hat{S}(t \mid \mathbf{x})$. The expected absolute difference in survival of the subjects selectable by σ_A is as, or more, sensitive than the absolute difference of $\hat{S}_A(t)$ from $\hat{S}_B(t)$,*

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ell_t^1(s(\mathbf{x}), s_B) \mid \sigma_A(\mathbf{x}) = 1] \geq \ell_t^1(s_A, s_B) \ ,$$

where $\ell_t^1(\cdot,\cdot)$ is an absolute difference measure, and we write $s_\circ$ for $\hat{S}_\circ(t)$, and $s(\mathbf{x})$ for $\hat{S}(t \mid \mathbf{x})$.

Proof. Let $Z = s(\mathbf{X}) - s_B$ be a random variable. The triangle inequality property of the absolute value function (Jensen's inequality applied to the convex absolute value function) states that

$$\mathbb{E}[\ |Z|\ ] \geq |\ \mathbb{E}[Z]\ | \ .$$

It also holds to condition the expectations using the indicator $\sigma_A(\mathbf{x}) = 1$ to get

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ |Z|\ \mid \sigma_A(\mathbf{x}) = 1\ ] \geq |\ \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[Z \mid \sigma_A(\mathbf{x}) = 1]\ | \ .$$

We substitute the definition of Z into the inequality to get

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ |s(\mathbf{x}) - s_B|\ \mid \sigma_A(\mathbf{x}) = 1] \geq |\ \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[s(\mathbf{x}) - s_B \mid \sigma_A(\mathbf{x}) = 1]\ | \ .$$

We use the linearity of the expectation, $\mathbb{E}[Z + W] = \mathbb{E}[Z] + \mathbb{E}[W]$, and note that the expectation of the positive constant s_B is the constant itself to get

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ |s(\mathbf{x}) - s_B|\ \mid \sigma_A(\mathbf{x}) = 1\ ] \geq |\ \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[s(\mathbf{x}) \mid \sigma_A(\mathbf{x}) = 1] - s_B\ | \ .$$

Lastly, we substitute the definitions of $\ell_t^1(\cdot,\cdot)$ and s_A to arrive at

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ \ell_t^1(s(\mathbf{x}), s_B) \mid \sigma_A(\mathbf{x}) = 1\ ] \geq \ell_t^1(s_A, s_B) \ .$$

□

Corollary A.1. *Under the assumptions of Prop. 3.1, if there exist $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathcal{X}$ selectable by σ_A with a strictly positive probability measure such that*

$$s(\mathbf{x}^{(1)}) > s_B \wedge s(\mathbf{x}^{(2)}) < s_B \ ,$$

then the inequality in Prop. 3.1 is strict, i.e.

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\ell_t^1(s(\mathbf{x}), s_B) \mid \sigma_A(\mathbf{x}) = 1] > \ell_t^1(s_A, s_B) \ .$$

Proof. Proposition 3.1 follows from the triangle inequality property of the absolute value function. Equality holds iff all arguments are aligned, i.e. either $s(\mathbf{x}) \geq s_B$ or $s(\mathbf{x}) \leq s_B$ for all $\mathbf{x}$ selectable by σ_A. The existence of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ contradicts this, hence the inequality is strict. □

B THE SYSURV ALGORITHM

In this section, we provide the pseudocode for SYSURV in Alg. 1. Prior to executing SYSURV, we fit a global non-parametric population model to the entire dataset $RSF(\mathbf{X}, T, \Delta)$. This gives per-subject survival functions over the discrete domain of unique event times $\hat{S}(t^* \mid \mathbf{X} = \mathbf{x}^{(i)})$ as the matrix M. The rule set R is initialized as $\emptyset$ and is populated with subsequently discovered subgroups, up to a user-defined limit. On the subject level $\mathbf{x}^{(i)}$, the following steps are applied. First, each feature $x_j^{(i)}$ is binned with learned cut points α_j and β_j using the soft condition to obtain $\hat{\pi}(\mathbf{x}^{(i)}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau)$. Next, by means of the weights w_j, we combine per-feature conditions $\hat{\pi}(x_j^{(i)})$ into a conjunction $\hat{\sigma}(\mathbf{x}^{(i)}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \tau)$. Then, we estimate the expected per-subject deviation in survival within the subgroup w.r.t. the overall population according to our objective. In a backwards pass, the soft rule parameters are updated to select a more exceptional subgroup for the remainder of the epochs. We anneal the temperature τ once halfway through the epochs and once three quarters of the way. Note that $\#events$ denotes the number of unique event times t^* in D, where $t^* \in \text{uniq}(\{t_i \sim P_T \mid \delta_i = 1\})$. The MEASUREEXCEPTIONALITY algorithm, used in Alg. 1, computes exceptionality using the trapezoidal rule over the absolute differences in survival between each subject and the reference population survival. See Appx. G for a discussion on computational complexity and optimizations.

Algorithm 1 LEARNSUBGROUP

Input: design matrix $\mathbf{X}$, population model M, rules R, size penalty γ, initial temperature τ
Output: a rule $\hat{\sigma}$ that selects an exceptional subgroup
$\alpha_j \leftarrow \min \mathbf{X}_{:,j}$
$\beta_j \leftarrow \max \mathbf{X}_{:,j}$
$w_j \leftarrow 1$
Initialize soft rule $\hat{\sigma}(\cdot; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \tau)$
$S_D(t^*) \leftarrow \frac{1}{n} \sum_{i=1}^{n} S(t^* \mid \mathbf{x}^{(i)})$
$Exceptionality_D(\mathbf{x}) \leftarrow$ MEASUREEXCEPTIONALITY$(\mathbf{x}, M, S_D(t^*))$
for $e \leftarrow 1$ **to** $\#epochs$ **do**
  $Membership(\mathbf{x}) \leftarrow \hat{\sigma}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \tau)$
  $Size \leftarrow \sum_{i=1}^{n} Membership(\mathbf{x}^{(i)})$
  $\hat{\phi} \leftarrow \sum_{i=1}^{n} Membership(\mathbf{x}^{(i)}) \cdot Exceptionality_D(\mathbf{x}^{(\mathbf{i})})$
  $Weighted_\hat{\phi} \leftarrow \hat{\phi} \cdot Size^{\gamma-1}$
  $Regularizer \leftarrow 0$
  for all $\hat{\sigma}_g \in R$ **do**
    $Size_g \leftarrow \sum_{i=1}^{n} \hat{\sigma}_g(\mathbf{x}^{(i)})$
    $S_g(t^*) \leftarrow \frac{1}{Size_g} \sum_{i=1}^{n} S(t^* \mid \mathbf{x}^{(i)}) \cdot \hat{\sigma}_g(\mathbf{x}^{(i)})$
    $Exceptionality_g(\mathbf{x}) \leftarrow$ MEASUREEXCEPTIONALITY$(\mathbf{x}, M, S_g(t^*))$
    $\hat{\phi}_g \leftarrow \sum_{i=1}^{n} Membership(\mathbf{x}^{(i)}) \cdot Exceptionality_g(\mathbf{x}^{(\mathbf{i})})$
    $Regularizer \leftarrow Regularizer + \hat{\phi}_g$
  end for
  $Regularizer \leftarrow Regularizer \cdot Size^{\gamma-1}$
  $loss \leftarrow -Weighted_\hat{\phi} - Regularizer$
  Update the parameters of rule $\hat{\sigma}$ to minimize the loss
  if $(e = \#epochs/2) \vee (e = \#epochs \cdot 3/4)$ **then**
    $\tau \leftarrow \tau/2$
  end if
end for
return $\hat{\sigma}$

Algorithm 2 MEASUREEXCEPTIONALITY

Input: subject covariates $\mathbf{x}$, population model M, reference survival at the event times $S_{ref}(t^*)$
Output: subject survival exceptionality value w.r.t reference
$abs_diff \leftarrow \mid S(t^* \mid \mathbf{x}) - S_{ref}(t^*) \mid$
return $\sum_{u=0}^{\#events-1}(t^*_{u+1} - t^*_u)/2 \cdot (abs_diff[u] + abs_diff[u+1])$

C RULE PRUNING

We provide the pseudocode for our post hoc pruning algorithm that simplifies rules in the presence of collinearity in Alg. 3. The PRUNERULE algorithm is a greedy procedure for reducing the complexity of learned rules without significantly altering its subject memberships. It begins with a full rule and iteratively identifies the least influential condition and attempts to remove it by setting its weight to zero. The algorithm then evaluates each candidate rule by computing the Jaccard similarity between its membership indicators and those of the full rule. This process continues as long as the similarity remains above a specified threshold, ensuring it still describes, more or less, the same subset of the population. It is to (optionally) be executed after learning a rule using Alg. 1. We report empirical pruning results across all real-world benchmark datasets in Table 4 (Appx. K), and demonstrate its application to high-dimensional clinical cohorts in our case study (Tables 14 and 15).

Algorithm 3 PRUNERULE

Input: design matrix $\mathbf{X}$, learned rule $\hat{\sigma}$, Jaccard similarity $threshold$ (default: 0.95)
Output: pruned rule $\hat{\sigma}_{prun}$
$\hat{\sigma}_{prun} \leftarrow \hat{\sigma}$
$mask \leftarrow \hat{\sigma}(\mathbf{X}) = 1$
while $Jaccard(mask, \hat{\sigma}_{prun}(\mathbf{X}) = 1) \geq threshold$ **do**
  $indices \leftarrow \{\ j \mid ReLU(w_j^{\hat{\sigma}_{prun}}) > 0.1\ \}$
  $score^\star \leftarrow 0$
  $idx^\star \leftarrow null$
  for each $j \in indices$ **do**
    $\hat{\sigma}' \leftarrow \hat{\sigma}_{prun}$
    $w_j^{\hat{\sigma}'} \leftarrow 0$
    $score \leftarrow Jaccard(mask, \hat{\sigma}'(\mathbf{X}) = 1)$
    if $score > score^\star$ **then**
      $score^\star \leftarrow score$
      $idx^\star \leftarrow j$
    end if
  end for
  if $score^\star < threshold$ **then**
    break
  end if
  $w_{idx^\star}^{\hat{\sigma}_{prun}} \leftarrow 0$
end while
return $\hat{\sigma}_{prun}$

D STATISTICAL VALIDATION

We provide the pseudocode of our instantiation for building the model of false discoveries (DFD), proposed by Duivesteijn & Knobbe (2011), in Alg. 4. The BUILDDFD algorithm quantifies the exceptionality that can be expected by chance. By repeatedly shuffling the survival outcomes relative to the design matrix, the algorithm creates a series of null datasets where no true relationship exists. In each iteration, it runs SYSURV while recording the exceptionality scores (line 12 of Alg. 1, divided by the subgroup size from line 11) obtained from these null datasets. It returns the mean and standard deviation that parametrize a normal distribution of false exceptionalities. This distribution serves as a baseline (null model) for judging the significance of discovered subgroups in terms of their exceptionalities, subject to multiple hypothesis testing. For that, we use the Z-test to obtain one-tailed p-values with a nominal alpha cutoff of 0.05. BUILDDFD is to be executed before running Alg. 1. *$\#runs$ should be no less than 1000.*

In Fig. 7, we show an example of the distribution of false discoveries obtained from 1000 runs of BUILDDFD on the postoperative cohort from our case study, with a fitted normal distribution. The mean and standard deviation of the fitted normal distribution are $\mu = 16.9$ and $\eta = 2.77$, respectively.

Algorithm 4 BUILDDFD

Input: design matrix $\mathbf{X}$, time-to-event T, event indicator Δ, size penalty γ, initial temperature τ, number of independent samples $\#runs$ (default: 1000)
Output: parameters of the Gaussian representing the distribution of false discoveries, mean μ, and standard deviation η
$scores \leftarrow \emptyset$
$indices \leftarrow \{1, \ldots, n\}$
$permutations \leftarrow \{\ P_r \mid P_r$ is a random permutation of $indices, r = 1 \ldots \#runs\ \}$
for each $P_r \in permutations$ **do**
    $T' \leftarrow T[P_r]$
    $\Delta' \leftarrow \Delta[P_r]$
    $M \leftarrow RSF(\mathbf{X}, T', \Delta')$
    $\hat{\sigma} \leftarrow$ LEARNSUBGROUP$(\mathbf{X}, M, \emptyset, \gamma, \tau)$
    $scores \leftarrow scores \cup \{\hat{\phi} \cdot Size^{-1} \mid \hat{\phi}$ is the exceptionality of the rule $\hat{\sigma}\}$
end for
$\mu \leftarrow$ mean$(scores)$
$\eta \leftarrow$ std$(scores)$
return μ, η

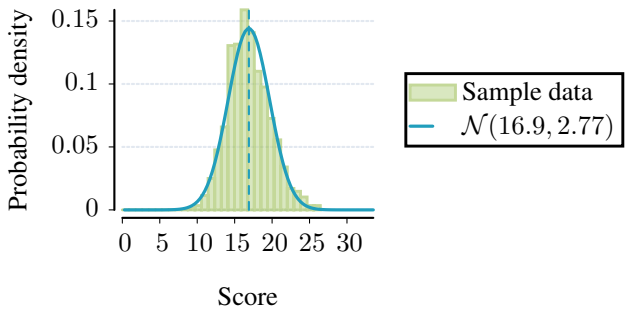
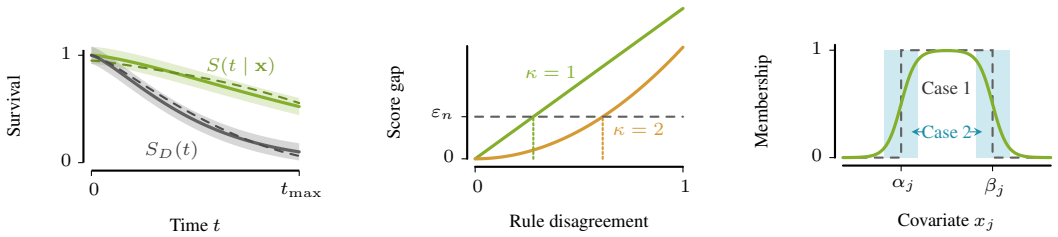


Figure 7: Histogram of the exceptionality scores obtained from 1000 runs of BUILDDFD on the postoperative cohort from our case study, with a fitted normal distribution. The mean and standard deviation of the fitted normal distribution are $\mu = 16.9$ and $\eta = 2.77$, respectively.



(a) Error stability (Prop. E.2) (b) Margin recovery (Thm. E.7) (c) Soft relaxation (Prop. E.9)

Figure 8: *Overview of theoretical guarantees.* **(a)** Bounded estimation errors (shaded bands) around the subject survival curve $S(t \mid \mathbf{x})$ and population curve $S_D(t)$ perturb the area between them at most linearly. **(b)** Bounding the vertical score gap by ε_n forces the horizontal rule disagreement toward zero. **(c)** The smooth condition (solid) matches the crisp interval (dashed) in Case 1 and only differs inside two thin boundary strips around α_j and β_j in Case 2.

E Statistical Properties and Theoretical Guarantees

In a Nutshell. In practice, SYSURV operates under three approximations relative to an ideal population-level search: it relies on survival curves estimated by a population model M, it learns rules from a finite sample of n subjects, and it replaces hard Boolean intervals with smooth, differentiable boundaries. In this section, we establish four theoretical guarantees for SYSURV—Lipschitz stability of the exceptionality score to survival estimation error (Prop. E.2), uniform consistency over the hypothesis space with a finite-sample bound (Thm. E.4), asymptotic subgroup recovery under a margin condition (Thm. E.7), and vanishing continuous relaxation bias as the temperature is annealed (Prop. E.9)—with the core geometric intuition for each illustrated in Fig. 8.

To formalize these guarantees, let $S(t \mid \mathbf{x}) \coloneqq \mathbb{P}(T > t \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[\mathbb{I}(T > t) \mid \mathbf{X} = \mathbf{x}]$ denote the true individual-level survival function for a subject with covariates $\mathbf{x} \in \mathcal{X}$, and let $S_D(t) \coloneqq \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[S(t \mid \mathbf{x})]$ denote the true population survival function. Given the population model's individual survival estimates $\hat{S}(t \mid \mathbf{x})$, let $\hat{S}_D(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{S}(t \mid \mathbf{x}^{(i)})$ denote the empirical reference survival curve (Alg. 1, line 7), and for any hard rule $\sigma\colon \mathcal{X} \to \{0,1\}$, let $\hat{\phi}(\sigma, \sigma_D) \coloneqq \frac{1}{n_\sigma}\sum_{i=1}^{n} \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}^{(i)}), \hat{S}_D(t))\, \sigma(\mathbf{x}^{(i)})$ with $n_\sigma \coloneqq \sum_{i=1}^{n} \sigma(\mathbf{x}^{(i)})$ denote its empirical exceptionality on D. Under the ground-truth survival curves, the true population counterpart to our expected exceptionality measure $\phi(\sigma, \sigma_D)$ in Eq. (1) and empirical objective $\hat{\phi}(\sigma, \sigma_D)$ is

$$\phi^*(\sigma, \sigma_D) \coloneqq \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\left[\ell^1_{\mathcal{T}}\big(S(t \mid \mathbf{x}), S_D(t)\big) \;\middle|\; \sigma(\mathbf{x}) = 1\right] ,$$

where $\ell^1_{\mathcal{T}}(S(t \mid \mathbf{x}), S_D(t)) \in [0, t_{\max}]$ since survival probabilities lie in $[0,1]$ (the baseline curve used in each variant is indicated by the decoration: $\hat{\phi}$ uses the empirical reference $\hat{S}_D$, ϕ uses the population expectation $\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\hat{S}(t \mid \mathbf{x})]$, and ϕ^* uses the true population curve S_D). Throughout this section, all probabilistic statements condition on a fixed population model M, treating $\hat{S}(t \mid \mathbf{x})$ as a deterministic function of $\mathbf{x}$; the only remaining randomness is the i.i.d. draw of the dataset $D = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ used for subgroup optimization. This conditioning is natural because SYSURV first trains M and then freezes it before the subgroup search (Alg. 1, line 6). The asymptotic results (Thms. E.4 and E.7) additionally let $\epsilon_M \to 0$ as the model improves with growing data, which can be formalized by sample splitting or by assuming a separate training set for M. Let $\mathcal{H} \coloneqq \{\mathbf{x} \mapsto \bigwedge_{j=1}^{p} \pi(x_j; \alpha_j, \beta_j) \mid \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^p, \alpha_j \le \beta_j\}$ denote the hypothesis class of hard conjunctive rules. To ensure well-posed conditional expectations and prevent degenerate subgroups, we impose a minimum coverage condition on candidate rules.

Assumption E.1 (Minimum Subgroup Coverage)**.** The covariate domain $\mathcal{X} \subseteq \mathbb{R}^p$ and time horizon $\mathcal{T} = [0, t_{\max}]$ are compact, and candidate subgroup rules are restricted to the hypothesis class $\mathcal{H}_{\mu_0} \coloneqq \{\sigma \in \mathcal{H} \mid \mathbb{P}(\sigma(\mathbf{x}) = 1) \ge \mu_0\}$ for a minimum population coverage threshold $\mu_0 \in (0,1)$.

Proposition E.2. *Under Assump. E.1, let $\hat{S}(t \mid \mathbf{x})$ be the individual survival estimator from population model M for $S(t \mid \mathbf{x})$, with uniform estimation error $\epsilon_M \coloneqq \sup_{\mathbf{x}\in\mathcal{X},\, t\in\mathcal{T}} \left|\hat{S}(t \mid \mathbf{x}) - S(t \mid \mathbf{x})\right|$.*

For any candidate subgroup rule $\sigma \in \mathcal{H}_{\mu_0}$, the deviation of the expected exceptionality $\phi(\sigma, \sigma_D)$ from the true population exceptionality $\phi^*(\sigma, \sigma_D)$ is bounded such that

$$|\phi(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D)| \leq 2\, t_{\max}\, \epsilon_M + \mathcal{O}_p(n^{-1/2}) \ .$$

We provide the proof at the end of this section.

Stability to Survival Estimation Error. Prop. E.2 confirms that inaccuracies in $\hat{S}(t \mid \mathbf{x})$ cannot amplify non-linearly during rule evaluation: the perturbation to any candidate subgroup's score scales at most linearly with ϵ_M (with factor $2\, t_{\max}$ accounting for both the individual curve and the reference curve $\hat{S}_D(t)$), up to a standard $\mathcal{O}_p(n^{-1/2})$ empirical averaging fluctuation that vanishes identically if $\hat{S}_D(t)$ is taken as the population expectation $\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\hat{S}(t \mid \mathbf{x})]$. We note that if the uniform supremum ϵ_M is replaced by an $L^1(P_{\mathbf{X}})$ average estimation error, an analogous bound holds with an additional $1/\mu_0$ factor, since the proof only requires a bound on $\mathbb{E}_{\mathbf{x}}[|\hat{S}(t \mid \mathbf{x}) - S(t \mid \mathbf{x})| \mid \sigma(\mathbf{x}) = 1]$, which is at most $\mathbb{E}_{\mathbf{x}}[|\hat{S}(t \mid \mathbf{x}) - S(t \mid \mathbf{x})|]/\mu_0$ by the coverage assumption. To guarantee that the empirical objective converges as sample size grows, we next assume that the population model M achieves vanishing uniform error. While L^2 or pointwise consistency is standard for many non-parametric survival estimators, uniform convergence additionally requires regularity of the survival function and sufficient smoothness; in the finite-sample bounds (Thm. E.4), ϵ_M enters as a deterministic bias term for any fixed model, so practitioners may simply report or bound ϵ_M empirically.

Assumption E.3 (Consistency of the Population Survival Model)**.** Under conditionally independent censoring and positivity over $\mathcal{T} = [0, t_{\max}]$, the survival estimation error of the population model M vanishes in probability, $\epsilon_M \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$. (This is formalized by assuming M is trained on an independent sample of the same size, or equivalently via sample splitting; the finite-sample bounds in Thms. E.4 and E.7 hold for any fixed M with ϵ_M treated as a deterministic quantity.)

Theorem E.4. *The hypothesis class $\mathcal{H}$ of conjunctive rules in $\mathbb{R}^p$ has Vapnik-Chervonenkis dimension $\mathrm{VC}(\mathcal{H}) = 2p$. Under Assump. E.1, for any sample size $n \geq p$ and failure probability $\xi \in (0, 1)$, with probability at least $1 - \xi$ over the random draw of D, the empirical exceptionality $\hat{\phi}(\sigma, \sigma_D)$ converges uniformly to the true population exceptionality $\phi^*(\sigma, \sigma_D)$ across all rules $\sigma \in \mathcal{H}_{\mu_0}$ s.t.*

$$\sup_{\sigma \in \mathcal{H}_{\mu_0}} \left| \hat{\phi}(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D) \right| \leq 2\, t_{\max}\, \epsilon_M + \frac{C\, t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}} \ ,$$

where $C > 0$ is a universal numerical constant. Consequently, if Assump. E.3 also holds, the empirical maximizer $\hat{\sigma}_n \coloneqq \arg\max_{\sigma \in \mathcal{H}_{\mu_0}} \hat{\phi}(\sigma, \sigma_D)$ is asymptotically consistent,

$$\phi^*(\hat{\sigma}_n, \sigma_D) \xrightarrow{\mathbb{P}} \max_{\sigma \in \mathcal{H}_{\mu_0}} \phi^*(\sigma, \sigma_D) \quad \text{as } n \to \infty \ .$$

We provide the proof at the end of this section.

Uniform Generalization and Consistency. Theorem E.4 shows that axis-aligned conjunctive rules have a finite VC dimension of exactly $2p$ (two threshold degrees of freedom per covariate), ensuring that the generalization gap decays uniformly at the parametric rate $\mathcal{O}(\sqrt{p \log(en/p)/n})$, inversely proportional to the minimum coverage μ_0. When paired with a consistent survival estimator satisfying Assump. E.3, the true exceptionality of the empirically learned rule $\hat{\sigma}_n$ converges in probability to the global optimum.

Assumption E.5 (Exceptionality Margin Condition)**.** There exists an optimal subgroup rule $\sigma^* \in \arg\max_{\sigma \in \mathcal{H}_{\mu_0}} \phi^*(\sigma, \sigma_D)$ s.t. for all candidate rules $\sigma \in \mathcal{H}_{\mu_0}$, the true exceptionality gap satisfies

$$\phi^*(\sigma^*, \sigma_D) - \phi^*(\sigma, \sigma_D) \geq c \cdot \mathbb{P}_{\mathbf{x}\sim P_{\mathbf{X}}}\left(\sigma(\mathbf{x}) \neq \sigma^*(\mathbf{x})\right)^{\kappa} \ ,$$

for some margin constant $c > 0$ and sharpness exponent $\kappa \geq 1$.

Remark E.6 (When the margin holds). Assump. E.5 is a standard structural condition in statistical learning theory, analogous to Tsybakov's margin condition in classification. It holds with $\kappa = 1$ whenever the true subgroup boundary separates regions with a uniform survival contrast: if subjects

inside σ^* have survival curves that differ from S_D by at least $\Delta > 0$ in $\ell^1_{\mathcal{T}}$ norm, while subjects outside differ by at most $\Delta' < \Delta$, then the exceptionality gap is lower-bounded by $(\Delta - \Delta') \cdot \mathbb{P}(\sigma(\mathbf{x}) \neq \sigma^*(\mathbf{x}))$ with $c = \Delta - \Delta'$. More generally, $\kappa > 1$ arises when the survival contrast decays smoothly near the optimal boundary. In subgroup discovery, Assump. E.5 is most naturally satisfied when the target subgroup has a clinically meaningful survival difference from the population—precisely the setting of practical interest.

Theorem E.7. *Under Assumps. E.1 and E.5, for any $n \geq p$ and $\xi \in (0, 1)$, with probability at least $1 - \xi$ over the random draw of D, the probability that the empirical maximizer $\hat{\sigma}_n$ disagrees with the ground-truth optimal rule σ^* on a randomly drawn subject $\mathbf{x} \sim P_{\mathbf{X}}$ is bounded s.t.*

$$\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x}) \right) \leq \left(\frac{4\, t_{\max}\, \epsilon_M + \frac{2C t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}}}{c} \right)^{1/\kappa},$$

which, if additionally Assump. E.3 holds, vanishes in probability, $\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}}(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x})) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.

We provide the proof at the end of this section.

Subgroup Recovery Under a Margin. While Thm. E.4 guarantees score optimality, Thm. E.7 links this to structural subpopulation recovery via Assump. E.5. Crucially, the margin is formulated in terms of subject membership disagreement $\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}}(\sigma(\mathbf{x}) \neq \sigma^*(\mathbf{x}))$ rather than parameter distance $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$, making it invariant to uninformative or collinear covariates that induce identical subject partitions (where $\kappa = 1$ corresponds to a sharp jump in survival behaviour across the true subgroup boundary, and $\kappa = 2$ corresponds to a smooth quadratic peak). When a population contains multiple distinct exceptional subgroups, Assump. E.5 can in principle be applied to each isolated mode of the exceptionality landscape. Extending the formal guarantees to the sequential discovery procedure with the diversity regularizer is beyond the scope of this work, though the single-rule bounds apply to each step conditional on the previously discovered rules.

Assumption E.8 (Regularity of Covariate Marginals and Active Weights)**.** *(Data regularity.)* Under $P_{\mathbf{X}}$, each continuous covariate x_j admits a marginal density f_{X_j} bounded above by $M_f < \infty$ on $\mathcal{X}$, while any discrete covariate has interval endpoints α_j, β_j separated from its support atoms by at least $\delta_0 > 0$. *(Algorithmic regularity.)* The weights of the active features in the soft rule satisfy $w_j \geq w_{\min} > 0$ for all $j \in S_{\text{act}}$ (with $w_j = 0$ for inactive features).

Proposition E.9. *Under Assumps. E.1 and E.8, for a soft rule $\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)$ with temperature $\tau > 0$ relaxing a hard rule $\sigma \in \mathcal{H}_{\mu_0}$ over the active feature set $S_{\text{act}} \coloneqq \{j \in \{1, \ldots, p\} \mid w_j > 0\}$, the expected rule discrepancy satisfies $\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|] \leq \mathcal{O}(p \cdot \tau \log(1/\tau))$, and the resulting relaxation bias in the survival exceptionality objective $\phi(\hat{\sigma}, \hat{\sigma}_D)$ in Eq. (2) is bounded s.t.*

$$|\phi(\hat{\sigma}, \hat{\sigma}_D) - \phi(\sigma, \sigma_D)| \leq \mathcal{O}\left(\frac{p\, t_{\max}}{\mu_0} \tau \log(1/\tau) \right).$$

We provide the proof at the end of this section.

Continuous Rule Relaxation Bias. While the preceding results analyze the discrete rule space $\mathcal{H}_{\mu_0}$, SYSURV optimizes the differentiable soft objective using sigmoidal conditions $\hat{\pi}_j$ and a weighted harmonic mean $\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)$. Pointwise convergence of $\hat{\sigma}$ to σ away from the interval endpoints is not sufficient on its own, because $\hat{\pi}_j = 1/2$ at the boundary regardless of τ. Under Assump. E.8, Prop. E.9 shows that the discrepancy is confined to a boundary layer of width $\mathcal{O}(\tau \log(1/\tau))$ around each active endpoint (vanishing at an even faster $\mathcal{O}(\tau)$ rate for discrete covariates once $\tau \log(1/\tau) < \delta_0$), confirming that annealing $\tau \to 0$ drives the relaxation bias to zero.

Remark E.10 (End-to-end error decomposition). Combining Props. E.2 and E.9 with Thm. E.4 via the triangle inequality, the total gap between the soft empirical objective and the true population exceptionality decomposes as

$$\left| \hat{\phi}(\hat{\sigma}, \hat{\sigma}_D) - \phi^*(\sigma, \sigma_D) \right| \leq \underbrace{\mathcal{O}\left(\frac{p\, t_{\max}}{\mu_0} \tau \log(1/\tau) \right)}_{\text{relaxation bias}} + \underbrace{2\, t_{\max}\, \epsilon_M}_{\text{model bias}} + \underbrace{\mathcal{O}\left(\frac{t_{\max}}{\mu_0} \sqrt{\frac{p \log n}{n}} \right)}_{\text{finite-sample}},$$

showing that all three sources vanish independently as $\tau \to 0$, $\epsilon_M \to 0$, and $n \to \infty$.

PROOFS OF PROPOSITIONS E.2 AND E.9 AND THEOREMS E.4 AND E.7

Proposition E.2. *Under Assump. E.1, let $\hat{S}(t \mid \mathbf{x})$ be the individual survival estimator from population model M for $S(t \mid \mathbf{x})$, with uniform estimation error $\epsilon_M \coloneqq \sup_{\mathbf{x}\in\mathcal{X},\, t\in\mathcal{T}} \left|\hat{S}(t \mid \mathbf{x}) - S(t \mid \mathbf{x})\right|$. For any candidate subgroup rule $\sigma \in \mathcal{H}_{\mu_0}$, the deviation of the expected exceptionality $\phi(\sigma, \sigma_D)$ from the true population exceptionality $\phi^*(\sigma, \sigma_D)$ is bounded such that*

$$|\phi(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D)| \leq 2\, t_{\max}\, \epsilon_M + \mathcal{O}_p(n^{-1/2}) \,.$$

Proof. We first bound the integrated deviation $\ell^1_{\mathcal{T}}(\hat{S}_D(t), S_D(t)) = \int_{\mathcal{T}} \ell^1_t(\hat{S}_D(t), S_D(t))\, dt$ between the empirical reference survival curve $\hat{S}_D(t)$ and the true population survival function $S_D(t)$. At any time $t \in \mathcal{T}$, adding and subtracting the empirical average of the true conditional survival functions $\frac{1}{n}\sum_{i=1}^n S(t \mid \mathbf{x}^{(i)})$ and applying the triangle inequality separates the model estimation error from the finite-sample fluctuation as

$$\begin{aligned}
\ell^1_t\big(\hat{S}_D(t), S_D(t)\big) &= \left|\frac{1}{n}\sum_{i=1}^n \Big(\hat{S}(t \mid \mathbf{x}^{(i)}) - S(t \mid \mathbf{x}^{(i)})\Big) + \left(\frac{1}{n}\sum_{i=1}^n S(t \mid \mathbf{x}^{(i)}) - S_D(t)\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \ell^1_t\big(\hat{S}(t \mid \mathbf{x}^{(i)}), S(t \mid \mathbf{x}^{(i)})\big) + \left|\frac{1}{n}\sum_{i=1}^n S(t \mid \mathbf{x}^{(i)}) - S_D(t)\right| \,.
\end{aligned}$$

By definition of the uniform estimation error ϵ_M, each term in the first sum is at most ϵ_M. Integrating both sides over the compact horizon $\mathcal{T} = [0, t_{\max}]$ from Assump. E.1 yields

$$\ell^1_{\mathcal{T}}\big(\hat{S}_D(t), S_D(t)\big) \leq t_{\max}\, \epsilon_M + Z_n \,, \quad \text{where } Z_n \coloneqq \int_0^{t_{\max}} \left|\frac{1}{n}\sum_{i=1}^n S(t \mid \mathbf{x}^{(i)}) - S_D(t)\right| dt \,. \tag{3}$$

Because the random variables $S(t \mid \mathbf{x}^{(i)}) \in [0,1]$ are i.i.d. with mean $S_D(t)$ and variance $\mathrm{Var}(S(t \mid \mathbf{x})) \leq 1/4$ at every $t \in \mathcal{T}$, Tonelli's theorem and Jensen's inequality imply $\mathbb{E}[Z_n] \leq \int_0^{t_{\max}} \sqrt{\mathrm{Var}(S(t \mid \mathbf{x}))/n}\, dt \leq \frac{t_{\max}}{2\sqrt{n}}$, and hence $Z_n = \mathcal{O}_p(n^{-1/2})$ by Markov's inequality.

Next, for any subject with covariates $\mathbf{x} \in \mathcal{X}$, we bound the error in their integrated survival deviation $\ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t))$ relative to $\ell^1_{\mathcal{T}}(S(t \mid \mathbf{x}), S_D(t))$ (illustrated in Fig. 8a). Applying the integral triangle inequality $|\int f - \int g| \leq \int |f - g|$ over $\mathcal{T}$ gives

$$\Big|\ell^1_{\mathcal{T}}\big(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)\big) - \ell^1_{\mathcal{T}}\big(S(t \mid \mathbf{x}), S_D(t)\big)\Big| \leq \int_{\mathcal{T}} \Big|\big|\hat{S}(t \mid \mathbf{x}) - \hat{S}_D(t)\big| - \big|S(t \mid \mathbf{x}) - S_D(t)\big|\Big|\, dt \,.$$

By the reverse triangle inequality $||a| - |b|| \leq |a - b|$ applied to $a = \hat{S}(t \mid \mathbf{x}) - \hat{S}_D(t)$ and $b = S(t \mid \mathbf{x}) - S_D(t)$, followed by the standard triangle inequality, the integrand is bounded at every time $t \in \mathcal{T}$ as

$$\begin{aligned}
\Big|\big|\hat{S}(t \mid \mathbf{x}) - \hat{S}_D(t)\big| - |S(t \mid \mathbf{x}) - S_D(t)|\Big| &\leq \Big|\Big(\hat{S}(t \mid \mathbf{x}) - S(t \mid \mathbf{x})\Big) - \Big(\hat{S}_D(t) - S_D(t)\Big)\Big| \\
&\leq \ell^1_t\big(\hat{S}(t \mid \mathbf{x}), S(t \mid \mathbf{x})\big) + \ell^1_t\big(\hat{S}_D(t), S_D(t)\big) \,.
\end{aligned}$$

Integrating over $\mathcal{T} = [0, t_{\max}]$ and substituting ϵ_M together with Eq. (3) yields the uniform per-subject bound

$$\Big|\ell^1_{\mathcal{T}}\big(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)\big) - \ell^1_{\mathcal{T}}\big(S(t \mid \mathbf{x}), S_D(t)\big)\Big| \leq 2\, t_{\max}\, \epsilon_M + Z_n \,. \tag{4}$$

Finally, we substitute this per-subject bound into the subgroup exceptionality measure. Under Assump. E.1, $\mathbb{P}(\sigma(\mathbf{x}) = 1) \geq \mu_0 > 0$, so the conditional expectation is well-defined. Using the linearity of conditional expectation and Jensen's inequality for the absolute value function, we get

$$\begin{aligned}
|\phi(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D)| &= \Big|\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\Big[\ell^1_{\mathcal{T}}\big(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)\big) - \ell^1_{\mathcal{T}}\big(S(t \mid \mathbf{x}), S_D(t)\big) \;\Big|\; \sigma(\mathbf{x}) = 1\Big]\Big| \\
&\leq \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\Big[\Big|\ell^1_{\mathcal{T}}\big(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)\big) - \ell^1_{\mathcal{T}}\big(S(t \mid \mathbf{x}), S_D(t)\big)\Big| \;\Big|\; \sigma(\mathbf{x}) = 1\Big] \,.
\end{aligned}$$

Because Eq. (4) holds uniformly for all $\mathbf{x} \in \mathcal{X}$ and $Z_n = \mathcal{O}_p(n^{-1/2})$, taking the conditional expectation preserves the upper bound and yields

$$|\phi(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D)| \leq 2\, t_{\max}\, \epsilon_M + \mathcal{O}_p(n^{-1/2}) \,.$$

□

Theorem E.4. *The hypothesis class $\mathcal{H}$ of conjunctive rules in $\mathbb{R}^p$ has Vapnik-Chervonenkis dimension $\mathrm{VC}(\mathcal{H}) = 2p$. Under Assump. E.1, for any sample size $n \geq p$ and failure probability $\xi \in (0,1)$, with probability at least $1-\xi$ over the random draw of D, the empirical exceptionality $\hat{\phi}(\sigma, \sigma_D)$ converges uniformly to the true population exceptionality $\phi^*(\sigma, \sigma_D)$ across all rules $\sigma \in \mathcal{H}_{\mu_0}$ s.t.*

$$\sup_{\sigma \in \mathcal{H}_{\mu_0}} \left| \hat{\phi}(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D) \right| \leq 2\, t_{\max}\, \epsilon_M + \frac{C\, t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}} \,,$$

where $C > 0$ is a universal numerical constant. Consequently, if Assump. E.3 also holds, the empirical maximizer $\hat{\sigma}_n \coloneqq \arg\max_{\sigma \in \mathcal{H}_{\mu_0}} \hat{\phi}(\sigma, \sigma_D)$ is asymptotically consistent,

$$\phi^*(\hat{\sigma}_n, \sigma_D) \xrightarrow{\mathbb{P}} \max_{\sigma \in \mathcal{H}_{\mu_0}} \phi^*(\sigma, \sigma_D) \quad \text{as } n \to \infty \,.$$

Proof. Each conjunctive rule $\sigma \in \mathcal{H}$ is the product of p interval conditions $\pi(x_j; \alpha_j, \beta_j)$, where each interval $[\alpha_j, \beta_j]$ is the intersection of two half-spaces in $\mathbb{R}$. By a standard result in statistical learning theory, the class of axis-aligned hyper-rectangles in $\mathbb{R}^p$ shatters at most $2p$ points and cannot shatter any set of $2p+1$ points, which establishes $\mathrm{VC}(\mathcal{H}) = 2p$.

To bound the uniform deviation between $\hat{\phi}(\sigma, \sigma_D)$ and $\phi^*(\sigma, \sigma_D)$, we write $d^*(\mathbf{x}) \coloneqq \ell^1_{\mathcal{T}}(S(t \mid \mathbf{x}), S_D(t))$ and $\hat{d}(\mathbf{x}) \coloneqq \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t))$ for the true and estimated per-subject survival deviations. Expressing the population objective as the ratio $\phi^*(\sigma, \sigma_D) = A(\sigma)/p_\sigma$ with numerator $A(\sigma) \coloneqq \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{x}}}[d^*(\mathbf{x})\, \sigma(\mathbf{x})]$ and coverage $p_\sigma \coloneqq \mathbb{P}(\sigma(\mathbf{x}) = 1)$, and the empirical objective as $\hat{\phi}(\sigma, \sigma_D) = \hat{A}(\sigma)/\hat{p}_\sigma$ with $\hat{A}(\sigma) \coloneqq \frac{1}{n}\sum_{i=1}^n \hat{d}(\mathbf{x}^{(i)})\, \sigma(\mathbf{x}^{(i)})$ and $\hat{p}_\sigma \coloneqq \frac{n_\sigma}{n}$, we apply the algebraic identity $\frac{u}{v} - \frac{u^*}{v^*} = \frac{u-u^*}{v} - \frac{u^*(v-v^*)}{v \cdot v^*}$ and the triangle inequality to obtain

$$\left| \hat{\phi}(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D) \right| \leq \frac{\left| \hat{A}(\sigma) - A(\sigma) \right|}{\hat{p}_\sigma} + \frac{A(\sigma)\, |\hat{p}_\sigma - p_\sigma|}{\hat{p}_\sigma\, p_\sigma} \,. \tag{5}$$

Bounding the numerator and denominator. For the numerator error $|\hat{A}(\sigma) - A(\sigma)|$, we add and subtract the empirical average under true deviations $\frac{1}{n}\sum_{i=1}^n d^*(\mathbf{x}^{(i)})\, \sigma(\mathbf{x}^{(i)})$ to get

$$\left| \hat{A}(\sigma) - A(\sigma) \right| \leq \frac{1}{n} \sum_{i=1}^n \left| \hat{d}(\mathbf{x}^{(i)}) - d^*(\mathbf{x}^{(i)}) \right| \sigma(\mathbf{x}^{(i)}) + \left| \frac{1}{n} \sum_{i=1}^n d^*(\mathbf{x}^{(i)})\, \sigma(\mathbf{x}^{(i)}) - A(\sigma) \right| \,.$$

By Eq. (4) from the proof of Prop. E.2, the first term is bounded above by $\hat{p}_\sigma(2t_{\max}\epsilon_M + Z_n)$ (this factor of $\hat{p}_\sigma$ cancels with the denominator $\hat{p}_\sigma$ in the first term of Eq. (5)). Because replacing any single observation $\mathbf{x}^{(i)}$ in D changes $\frac{1}{n}\sum_{i=1}^n S(t \mid \mathbf{x}^{(i)})$ by at most $1/n$ at each $t \in [0, t_{\max}]$, Z_n satisfies the bounded differences property with constant $t_{\max}/n$; since $\mathbb{E}[Z_n] \leq \frac{t_{\max}}{2\sqrt{n}}$, McDiarmid's inequality implies that with probability at least $1 - \xi/3$, $Z_n \leq t_{\max}\big(\frac{1}{2\sqrt{n}} + \sqrt{\frac{\log(3/\xi)}{2n}}\big)$. For the second term, the function class $\mathcal{F} \coloneqq \{\mathbf{x} \mapsto d^*(\mathbf{x})\, \sigma(\mathbf{x}) \mid \sigma \in \mathcal{H}\}$ consists of non-negative functions bounded by the envelope $t_{\max}$ and indexed by a VC-subgraph class of dimension $2p$. Applying Dudley's entropy integral bound and McDiarmid's inequality for VC classes guarantees that, with probability at least $1 - \xi/3$,

$$\sup_{\sigma \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n d^*(\mathbf{x}^{(i)})\, \sigma(\mathbf{x}^{(i)}) - A(\sigma) \right| \leq C_1\, t_{\max} \sqrt{\frac{2p \log(en/p) + \log(6/\xi)}{n}} \,. \tag{6}$$

Similarly, applying the standard Vapnik-Chervonenkis uniform convergence inequality to the indicator class $\mathcal{H}$ guarantees that, with probability at least $1 - \xi/3$,

$$\sup_{\sigma \in \mathcal{H}} |\hat{p}_\sigma - p_\sigma| \leq C_2 \sqrt{\frac{2p \log(en/p) + \log(6/\xi)}{n}} \,. \tag{7}$$

Combining the bounds and establishing consistency. By the union bound, the high-probability bound on Z_n and Eqs. (6) and (7) hold simultaneously with probability at least $1-\xi$. On this event, for any $\sigma \in \mathcal{H}_{\mu_0}$, Assump. E.1 ensures $p_\sigma \geq \mu_0$. If the right-hand side of Eq. (7) is at most $\mu_0/2$, then $\hat{p}_\sigma \geq p_\sigma - \mu_0/2 \geq \mu_0/2$; noting also that $A(\sigma) \leq t_{\max}\, p_\sigma$ and $\log(6/\xi) \leq 2\log(4/\xi)$ for $\xi \in (0,1)$, substituting the bounds on Z_n, Eq. (6), and Eq. (7) into Eq. (5) yields

$$\sup_{\sigma \in \mathcal{H}_{\mu_0}} \left| \hat{\phi}(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D) \right| \leq 2\, t_{\max}\, \epsilon_M + \frac{C\, t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}} \,,$$

where $C > 0$ absorbs the universal numerical constants (conversely, if the right-hand side of Eq. (7) exceeds $\mu_0/2$, the second term above exceeds $t_{\max}$ for $C \geq 4C_2$, so the inequality holds trivially since $\hat{\phi}(\sigma, \sigma_D), \phi^*(\sigma, \sigma_D) \in [0, t_{\max}]$). Finally, let $\sigma^* \coloneqq \arg\max_{\sigma \in \mathcal{H}_{\mu_0}} \phi^*(\sigma, \sigma_D)$ be a population optimal rule. Because $\hat{\sigma}_n$ maximizes $\hat{\phi}(\sigma, \sigma_D)$, we have $\hat{\phi}(\sigma^*, \sigma_D) - \hat{\phi}(\hat{\sigma}_n, \sigma_D) \leq 0$, so the optimality gap satisfies

$$\begin{aligned} \phi^*(\sigma^*, \sigma_D) - \phi^*(\hat{\sigma}_n, \sigma_D) &\leq \left(\phi^*(\sigma^*, \sigma_D) - \hat{\phi}(\sigma^*, \sigma_D)\right) + \left(\hat{\phi}(\hat{\sigma}_n, \sigma_D) - \phi^*(\hat{\sigma}_n, \sigma_D)\right) \\ &\leq 2 \sup_{\sigma \in \mathcal{H}_{\mu_0}} \left| \hat{\phi}(\sigma, \sigma_D) - \phi^*(\sigma, \sigma_D) \right| . \end{aligned} \tag{8}$$

Under Assump. E.3, $\epsilon_M \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, so the uniform bound on the right-hand side vanishes, proving that $\phi^*(\hat{\sigma}_n, \sigma_D) \xrightarrow{\mathbb{P}} \phi^*(\sigma^*, \sigma_D)$. □

Theorem E.7. *Under Assumps. E.1 and E.5, for any $n \geq p$ and $\xi \in (0,1)$, with probability at least $1-\xi$ over the random draw of D, the probability that the empirical maximizer $\hat{\sigma}_n$ disagrees with the ground-truth optimal rule σ^* on a randomly drawn subject $\mathbf{x} \sim P_{\mathbf{X}}$ is bounded s.t.*

$$\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x}) \right) \leq \left(\frac{4\, t_{\max}\, \epsilon_M + \frac{2C t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}}}{c} \right)^{1/\kappa} ,$$

which, if additionally Assump. E.3 holds, vanishes in probability, $\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}}(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x})) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.

Proof. Evaluating the margin condition from Assump. E.5 at the empirical maximizer $\sigma = \hat{\sigma}_n \in \mathcal{H}_{\mu_0}$ relates the subject membership disagreement probability directly to the excess exceptionality gap (illustrated in Fig. 8b) as

$$c \cdot \mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x}) \right)^{\kappa} \leq \phi^*(\sigma^*, \sigma_D) - \phi^*(\hat{\sigma}_n, \sigma_D) \,.$$

Under Assump. E.1, substituting the uniform convergence bound from Thm. E.4 into the excess gap inequality in Eq. (8) guarantees that, with probability at least $1-\xi$,

$$c \cdot \mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x}) \right)^{\kappa} \leq 4\, t_{\max}\, \epsilon_M + \frac{2C\, t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}} \,.$$

Dividing both sides by $c > 0$ and applying the strictly increasing power function $u \mapsto u^{1/\kappa}$ for $u \geq 0$ and $\kappa \geq 1$ yields the stated finite-sample recovery bound

$$\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x}) \right) \leq \left(\frac{4\, t_{\max}\, \epsilon_M + \frac{2C\, t_{\max}}{\mu_0} \sqrt{\frac{p \log(en/p) + \log(4/\xi)}{n}}}{c} \right)^{1/\kappa} .$$

Taking the limit as $n \to \infty$ under Assump. E.3 drives $\epsilon_M \xrightarrow{\mathbb{P}} 0$ and the right-hand side to zero, establishing $\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}}(\hat{\sigma}_n(\mathbf{x}) \neq \sigma^*(\mathbf{x})) \xrightarrow{\mathbb{P}} 0$. □

Proposition E.9. *Under Assumps. E.1 and E.8, for a soft rule $\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)$ with temperature $\tau > 0$ relaxing a hard rule $\sigma \in \mathcal{H}_{\mu_0}$ over the active feature set $S_{\text{act}} \coloneqq \{j \in \{1, \ldots, p\} \mid w_j > 0\}$, the expected rule discrepancy satisfies $\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|] \leq \mathcal{O}(p \cdot \tau \log(1/\tau))$, and the resulting relaxation bias in the survival exceptionality objective $\phi(\hat{\sigma}, \hat{\sigma}_D)$ in Eq. (2) is bounded s.t.*

$$|\phi(\hat{\sigma}, \hat{\sigma}_D) - \phi(\sigma, \sigma_D)| \leq \mathcal{O}\left(\frac{p\, t_{\max}}{\mu_0}\, \tau \log(1/\tau) \right) .$$

Proof. Dividing both the numerator and denominator of the soft condition $\hat{\pi}_j$ by $e^{\frac{1}{\tau}(2x_j-\alpha_j)}$ simplifies its expression to the reciprocal form

$$\hat{\pi}_j = \frac{1}{1 + e^{-\frac{1}{\tau}(x_j-\alpha_j)} + e^{\frac{1}{\tau}(x_j-\beta_j)}} \ .$$

By Assump. E.8, inactive features ($j \notin S_{\text{act}}$) have weight $w_j = 0$ and do not enter $\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)$, which therefore relaxes the hard conjunctive rule $\sigma(\mathbf{x}) = \bigwedge_{j \in S_{\text{act}}} \pi(x_j; \alpha_j, \beta_j) \in \mathcal{H}_{\mu_0}$. For any temperature $\tau \in (0, 1/2)$ and active feature $j \in S_{\text{act}}$, define the boundary margin $\eta_\tau \coloneqq \tau \log(1/\tau)$ around α_j and β_j, and let $B_j(\tau) \coloneqq [\alpha_j - \eta_\tau, \alpha_j + \eta_\tau] \cup [\beta_j - \eta_\tau, \beta_j + \eta_\tau]$ denote the boundary layer for covariate j (illustrated in Fig. 8c).

Case 1: Covariates outside the boundary layer ($\mathbf{x} \notin \bigcup_{j \in S_{\text{act}}} \{\mathbf{x} \mid x_j \in B_j(\tau)\}$). Note first that if an active interval has width $\beta_j - \alpha_j \leq 2\eta_\tau$, then the entire interval $[\alpha_j, \beta_j]$ is contained in $B_j(\tau)$ (covered in Case 2). Otherwise, if $x_j \in (\alpha_j + \eta_\tau, \beta_j - \eta_\tau)$ lies strictly inside the interval, then $\pi(x_j; \alpha_j, \beta_j) = 1$, while both $x_j - \alpha_j > \tau \log(1/\tau)$ and $\beta_j - x_j > \tau \log(1/\tau)$. Consequently, the two exponential terms in the denominator of $\hat{\pi}_j$ are each strictly less than $e^{-\log(1/\tau)} = \tau$, which implies $\hat{\pi}_j^{-1} = 1 + e^{-\frac{1}{\tau}(x_j-\alpha_j)} + e^{\frac{1}{\tau}(x_j-\beta_j)} \in [1, 1+2\tau]$. If $\sigma(\mathbf{x}) = 1$, then $x_j \in (\alpha_j + \eta_\tau, \beta_j - \eta_\tau)$ for all $j \in S_{\text{act}}$, so the weighted harmonic mean is bounded below as

$$\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) = \frac{\sum_{j \in S_{\text{act}}} w_j}{\sum_{j \in S_{\text{act}}} w_j \, \hat{\pi}_j^{-1}} \geq \frac{1}{1 + 2\tau} \geq 1 - 2\tau \ ,$$

which bounds the pointwise approximation error by $|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})| \leq 2\tau$. Conversely, if $\sigma(\mathbf{x}) = 0$, there exists at least one active feature $k \in S_{\text{act}}$ for which $x_k \notin [\alpha_k, \beta_k]$ and $x_k \notin B_k(\tau)$, meaning either $x_k < \alpha_k - \eta_\tau$ or $x_k > \beta_k + \eta_\tau$. In either case, $\hat{\pi}_k^{-1} > 1 + \tau^{-1}$, so by Assump. E.8 ($w_k \geq w_{\min} > 0$), the harmonic mean is bounded above by

$$\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) \leq \frac{\sum_{j \in S_{\text{act}}} w_j}{w_k \, \hat{\pi}_k^{-1}} \leq \frac{\sum_{j \in S_{\text{act}}} w_j}{w_{\min}} \, \tau \ ,$$

yielding $|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})| \leq \frac{\sum_{j \in S_{\text{act}}} w_j}{w_{\min}} \, \tau$, which is $\mathcal{O}(\tau)$ for fixed weights and is dominated by the $\mathcal{O}(p \cdot \tau \log(1/\tau))$ boundary-layer term in Case 2.

Case 2: Covariates inside the boundary layer ($\mathbf{x} \in \bigcup_{j \in S_{\text{act}}} \{\mathbf{x} \mid x_j \in B_j(\tau)\}$). Inside the union of the boundary layers, the difference $|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|$ is bounded by 1. By Assump. E.8, each continuous marginal density f_{X_j} is bounded above by $M_f < \infty$, so the probability that x_j falls within the two intervals of width $2\eta_\tau$ in $B_j(\tau)$ is at most $4M_f\eta_\tau = 4M_f\tau \log(1/\tau)$ (while for any discrete covariate, $\mathbb{P}(x_j \in B_j(\tau)) = 0$ once $\eta_\tau < \delta_0$). Applying the union bound over all active covariates $|S_{\text{act}}| \leq p$ bounds the total probability mass of the boundary region by

$$\mathbb{P}_{\mathbf{x} \sim P_{\mathbf{X}}} \left(\bigcup_{j \in S_{\text{act}}} \{x_j \in B_j(\tau)\} \right) \leq \sum_{j \in S_{\text{act}}} \mathbb{P}\big(x_j \in B_j(\tau)\big) \leq 4pM_f \, \tau \log(1/\tau) \ .$$

Decomposing the expectation $\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|]$ over the interior and boundary regions and combining both bounds yields $\mathcal{O}(\tau) + \mathcal{O}(p \cdot \tau \log(1/\tau)) = \mathcal{O}(p \cdot \tau \log(1/\tau))$. Since $\hat{d}(\mathbf{x}) = \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}), \hat{S}_D(t)) \in [0, t_{\max}]$ and $p_\sigma \geq \mu_0$ under Assump. E.1, both the numerator error $|\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[\hat{d}(\mathbf{x})(\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x}))]| \leq t_{\max} \, \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|]$ and coverage error $|\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)] - p_\sigma| \leq \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|]$ in Eq. (2) are $\mathcal{O}(p \cdot \tau \log(1/\tau))$. Thus, for τ sufficiently small that $\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[|\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau) - \sigma(\mathbf{x})|] \leq \mu_0/2$ (and trivially bounded by $t_{\max}$ for larger τ), $\mathbb{E}_{\mathbf{x} \sim P_{\mathbf{X}}}[\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)] \geq \mu_0/2$, and applying the algebraic identity $\frac{u}{v} - \frac{u^*}{v^*} = \frac{u - u^*}{v} - \frac{u^*(v - v^*)}{v \, v^*}$ with $u = \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x}) \, \hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)]$, $u^* = \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x}) \, \sigma(\mathbf{x})]$, $v = \mathbb{E}_{\mathbf{x}}[\hat{\sigma}(\mathbf{x}; \boldsymbol{\theta}, \tau)]$, and $v^* = p_\sigma$ (analogous to Eq. (5)) gives $|\phi(\hat{\sigma}, \hat{\sigma}_D) - \phi(\sigma, \sigma_D)| \leq \mathcal{O}\big((p \, t_{\max}/\mu_0) \, \tau \log(1/\tau)\big)$. □

F HYPERPARAMETERS OF SYSURV AND BASELINES

In this section, we discuss the hyperparameter choices for SYSURV and the baselines.

SYSURV

We have two core hyperparameters (γ and τ) in SYSURV, and some more dictated by the choice of the regression model. The most important parameter is γ. Tuning γ is crucial for steering the learning towards rules that select subgroups of a desired size range. Values of γ range from 0.0 (very small subgroup) to 1.0 (a very large subgroup) at 0.1 increments. For our experiments, we optimize γ and found that $\gamma \in [0, 0.2]$ was almost always selected. As for τ, we found that $\tau = 0.2$ worked well across settings and datasets. We have secondary parameters, namely, the number of epochs and the learning rate used for stochastic optimization. Generally, 1000 epochs paired with a learning rate of 0.01 ensured stable convergence. Since we always start with a full rule, i.e. a rule that selects all subjects by spanning the entire feature space, initialization does not affect the optimization, and because we anneal the temperature, this ensures convergence to a (local) minimum.

In this work, we instantiated SYSURV using Random Survival Forests, but in theory it can be done using any non-parametric continuous-time regression model. We do not sample features for each tree, despite the computational cost, to avoid missing important features. However, we do sample subjects, especially when n is very large. To train the RSF, we used 100 trees (300 for the case study, to account for its rather small dataset sizes), a maximum depth equal to double the number of features, and a maximum of 2000 subjects per tree. A rule of thumb for getting a reliable splitting is to have at least 10–20 events in a cohort. As such, we set the minimum number of subjects to split to 40 and, correspondingly, the minimum number of subjects in a leaf to 20.

BASELINES

We use the implementations of the baselines as provided by their authors. We take the top-5 subgroups discovered by each baseline for evaluation. We preprocess the data in the same way for each of the baselines. This only involves the use of a k-bin discretizer for continuous features using $k = 5$ and the *quantile* strategy.

G COMPUTATIONAL COMPLEXITY OF SYSURV

Modern implementations of RSFs are efficient. Nevertheless, fitting the population model requires the most time. During training, we have a computational complexity of $\mathcal{O}(\#trees \cdot N \cdot V \cdot d/\#cores)$. $\#trees$ is fixed, N is the number of subjects assigned to a tree, V is the number of features considered for splitting, d is the tree depth, and $\#cores$ is also fixed. After training, the computational complexity of predicting on the entire dataset is $\mathcal{O}(\#trees \cdot N \cdot d/\#cores)$, which happens only once, and the results are saved.

During learning, the computational complexity is $\mathcal{O}(\#epochs \cdot N \cdot V)$, where $\#epochs$ is fixed. We omit this in Alg. 1, but we save unchanging data structures for efficiency. All the computations are done using highly optimized matrix operations through specialized libraries.

Hardware. For our experiments, we used machines equipped with 2x AMD Epyc 7773x 2.2GHz (base), 3.5GHz (max boost) with 128 real cores and 2TB of memory.

H SURVIVAL DATA GENERATION

In Alg. 5, we provide the pseudocode for MAKESURVIVALDATA, the routine we use for generating synthetic survival data with known ground truth, allowing independent configuration of n, p, k, censoring fraction, and subgroup size. We sample p feature variables $\mathbf{X}_{:,j} \sim \mathcal{U}(0,1)$ and randomly choose $k \leq p$ conditional features to define a hyper-box $[\alpha_j, \alpha_j + \epsilon]$ covering a target fraction $ratio_{target}$ (default 20%) of the feature space, with side length $\epsilon = \sqrt[k]{ratio_{target}}$. Subjects are assigned via $\sigma \sim Bernoulli(ratio_{target})^n$, ensuring that non-subgroup samples ($\sigma = 0$) lie strictly outside the rule bounds across conditional features, while subgroup samples ($\sigma = 1$) satisfy all k bounds. Survival times are drawn from Weibull distributions, chosen for their parametric configurability. To introduce intra-group heterogeneity, scales incorporate a linear covariate effect $\mathcal{N}_0 = \sum_{j=1}^{k} \mathbf{X}[:, v_j] - k/2$ scaled by ψ, which is iteratively shrunk to guarantee positive, non-overlapping scale supports. In our default benchmark ($n = 10\,000, p = 10, k = 2$), both groups share shape 1.5 but distinct base scales (1 vs. 5); thus, subgroup subjects experience events $5\times$ earlier on expectation. Finally, a fraction $ratio_{cens}$ (default 10%) of subjects undergo uniform right-censoring, where censored times are drawn uniformly from $[0, Y]$.

Algorithm 5 MAKESURVIVALDATA

Input: number of samples n, number of covariates p, number of true conditions k, Weibull scale of non-subgroup subjects $scale_{nsg}$, Weibull shape of non-subgroup subjects $shape_{nsg}$, Weibull scale of subgroup subjects $scale_{sg}$, Weibull shape of subgroup subjects $shape_{sg}$, percentage of subjects in the subgroup $ratio_{target}$, percentage of censored subject outcomes $ratio_{cens}$
Output: synthetic design matrix $\mathbf{X}$, synthetic time-to-event T, synthetic event indicator Δ
$\mathcal{V} \leftarrow Sample(p, k,$ without replacement$)$
$\epsilon \leftarrow \sqrt[k]{ratio_{target}}$
for $j \leftarrow 1$ **to** k **do**
  $\alpha_j \sim Uniform(0, 1 - \epsilon)$
end for
$\mathbf{X} \sim Uniform(0,1)^{n \times p}$
for $j \leftarrow 1$ **to** k **do**
  $\mathbf{X}_{sg}[:, j] \sim Uniform(\alpha_j, \alpha_j + \epsilon)^n$
  $\mathbf{X}_{nsg}[:, j] \sim Uniform(0,1)^n$
  while $\exists\ \mathbf{x}_{nsg}[j] \mid \alpha_j \leq \mathbf{x}_{nsg}[j] \leq \alpha_j + \epsilon$ **do**
    $\mathbf{X}_{nsg}[:, j] \sim Uniform(0,1)^n$
  end while
end for
$\sigma \sim Bernoulli(ratio_{target})^n$
for $j \leftarrow 1$ **to** k **do**
  $\mathbf{X}[\sigma = 1, v_j] \leftarrow \mathbf{X}_{sg}[\sigma = 1, j]$
  $\mathbf{X}[\sigma = 0, v_j] \leftarrow \mathbf{X}_{nsg}[\sigma = 0, j]$
end for
$\mathcal{N}_0 \leftarrow \sum_{j=1}^{k} \mathbf{X}[:, v_j] - k/2$
$\psi \leftarrow 1$
$scale_{sg} \leftarrow scale_{sg} + \mathcal{N}_0 \cdot \psi$
$scale_{nsg} \leftarrow scale_{nsg} + \mathcal{N}_0 \cdot \psi$
while $min(scale_{sg}) < 0 \vee min(scale_{nsg}) < 0 \vee |Supp(scale_{nsg}) \cap Supp(scale_{sg})| > 0$ **do**
  $\psi \leftarrow \psi \cdot 0.9$
end while
$Y_{sg} \leftarrow scale_{sg} \cdot Weibull(shape_{sg}, 1)$
$Y_{nsg} \leftarrow scale_{nsg} \cdot Weibull(shape_{nsg}, 1)$
$Y[\sigma = 1] \leftarrow Y_{sg}[\sigma = 1]$
$Y[\sigma = 0] \leftarrow Y_{nsg}[\sigma = 0]$
$\Delta \sim Bernoulli(1 - ratio_{cens})$
$T[\Delta = 0] \sim Uniform(0, Y[\Delta = 0])$
$T[\Delta = 1] \leftarrow Y[\Delta = 1]$
return $\mathbf{X}$, T, Δ

I EXCEPTIONALITY METRICS

To evaluate the discovered subgroups on real-world data across methods, we use our (discrete-time) exceptionality measure (Φ), the two-sample logrank statistic (LR), and the absolute mean-shift (AMS), between the subgroup Q and the population D. Those are formally defined as

$$\Phi(Q, D) = \frac{1}{|Q|} \sum_{i \in Q} \sum_{t \in \mathcal{T}}^{\text{trapz}} \left| \hat{S}(t \mid \mathbf{x}^{(i)}) - \hat{S}_D(t) \right|$$

$$AMS(Q, D) = \left| \left(\frac{1}{|Q|} \sum_{i \in Q} t^{(i)} \right) - \left(\frac{1}{|D|} \sum_{i \in D} t^{(i)} \right) \right|$$

$$LR(Q, D) = \frac{\sum_{\textit{event times}} (d_t^Q - e_t^Q)}{\sqrt{\sum_{\textit{event times}} e_t^Q \cdot (\frac{r_t - d_t}{r_t}) \cdot (\frac{r_t - r_t^Q}{r_t - 1})}} ,$$

where $\sum^{\text{trapz}}$ denotes numerical integration via the trapezoidal rule over unique event times in $\mathcal{T}$ (Alg. 2), at-risks $r_t^\circ = \left|\{i \in \circ \mid (t^{(i)} > t \wedge \delta^{(i)} = 1) \vee \delta^{(i)} = 0\}\right|$, had-events $d_t^\circ = \left|\{i \in \circ \mid t^{(i)} = t \wedge \delta^{(i)} = 1\}\right|$, total at-risks $r_t = r_t^Q + r_t^D$, total had-events $d_t = d_t^Q + d_t^D$, and expected-events $e_t^\circ = d_t \cdot r_t^\circ / r_t$. We compute LR via a standard library.

J CONNECTION BETWEEN OUR MEASURE AND THE LOGRANK

We can trace a transition from the continuous-time one-sample logrank to our proposed integral measure in Eq. 1. The one-sample logrank essentially compares the observed number of events d to the expected number of events e at the group level across all event times

$$\begin{aligned}
\textit{Logrank} &\propto \sum_{\textit{event times}} (d_t^{grp} - e_t^{grp}) \\
&= \sum_{\textit{event times}} r_t^{grp} \left(\frac{d_t^{grp}}{r_t^{grp}} - \frac{d_t^{ref}}{r_t^{ref}} \right) \\
&= \sum_{\textit{event times}} r_t^{grp} \left(\hat{\lambda}_{grp}(t) - \hat{\lambda}_{ref}(t) \right) \\
&= \sum_{\textit{event times}} r_t^{grp} \Delta \left(\hat{\Lambda}_{grp}(t) - \hat{\Lambda}_{ref}(t) \right) ,
\end{aligned}$$

where $e_t^{grp} = d_t^{ref} \cdot r_t^{grp} / r_t^{ref}$, r refers to the number of individuals at risk, "grp" and "ref" refer to the group and reference populations, respectively, $\hat{\lambda}_\circ(t)$ is the hazard function, $\hat{\Lambda}_\circ(t)$ is the cumulative hazard function, and $\Delta(\cdot)$ denotes the discrete step-wise increment.

From Statistic to Distance. The full version divides this score (the numerator) by the pooled standard error to yield a symmetric test statistic. However, because variance is inversely proportional to sample size, this division dynamically scales the output, converting it from a spatial distance into a statistic. To build a valid L^1 geometric metric, we must isolate the spatial numerator. Letting $\rho_A(t)$ be the continuous at-risk process for group A, this numerator is natively written as a continuous Stieltjes integral over the study duration $\mathcal{T} = [0, t_{\max}]$,

$$\textit{Logrank} \propto \int_{\mathcal{T}} \rho_A(t)\, d\big(\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)\big) .$$

While isolating the numerator preserves geometric magnitude, it exposes four restrictive characteristics, which we sequentially dismantle to arrive at our integral-based distance metric.

Proposition J.1. *Let $\Sigma_\Lambda(t) = \hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)$ be the continuous cumulative hazard state difference over the time domain $\mathcal{T} = [0, t_{\max}]$. Integrating the state directly over the domain is equivalent to integrating its differentials scaled by a temporal persistence weight, such that*

$$\int_{\mathcal{T}} \Sigma_\Lambda(t)\, dt = t_{\max} \Sigma_\Lambda(0) + \int_{\mathcal{T}} (t_{\max} - t)\, d\Sigma_\Lambda(t) .$$

We provide the proof at the end of this section.

From Changes to States. The logrank integrates the hazard differentials $d(\cdot)$, overlooking prior hazard increments. By Prop. J.1, evaluating the integral of the accumulated *state* directly is better. It inherently applies a temporal persistence weight $(t_{\max} - t)$ to past deviation and naturally captures any initial baseline gap $(t_{\max}\Sigma_\Lambda(0))$, circumventing the logrank's memoryless aggregation.

Proposition J.2. *Given two groups A and B for which the expected group-level survival probabilities at any fixed time t are $\hat{S}_A(t)$ and $\hat{S}_B(t)$, respectively, and for which $\hat{\Lambda}_A(t)$ and $\hat{\Lambda}_B(t)$ are their corresponding cumulative hazard functions. The difference in their cumulative hazards is proportional to the difference in survival $\hat{S}_A(t) - \hat{S}_B(t)$, such that*

$$\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t) = -\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\hat{S}^*(t)} \ ,$$

where $\hat{S}^*(t)$ is exactly the logarithmic mean of the two survival probabilities, formally defined to account for potential curve crossing as

$$\hat{S}^*(t) = \begin{cases} \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\ln(\hat{S}_A(t)) - \ln(\hat{S}_B(t))} & \text{if } \hat{S}_A(t) \neq \hat{S}_B(t) \\ \hat{S}_A(t) & \text{if } \hat{S}_A(t) = \hat{S}_B(t) \end{cases} \ ,$$

which is strictly bounded by their geometric and arithmetic means, $\sqrt{\hat{S}_A(t)\hat{S}_B(t)} < \hat{S}^*(t) < \frac{\hat{S}_A(t)+\hat{S}_B(t)}{2}$ (for $\hat{S}_A(t) \neq \hat{S}_B(t)$).

We provide the proof at the end of this section.

From Hazards to Survival. If we apply the logrank's implicit continuous risk weighting $\rho_A(t) \approx N_A\hat{S}_A(t)$, while ignoring the effect of censoring, to the accumulated hazard state $\Sigma_\Lambda(t)$ adopted in the previous step, Prop. J.2 reveals a geometric scaling behavior

$$\rho_A(t)\Sigma_\Lambda(t) = -\rho_A(t)\frac{\Sigma_S(t)}{\hat{S}^*(t)} \approx -N_A \underbrace{\frac{\hat{S}_A(t)}{\hat{S}^*(t)}}_{\text{Distortion}} \Sigma_S(t) \ .$$

The raw spatial distance $\Sigma_S(t)$ is directly multiplied by the ratio $\hat{S}_A(t)/\hat{S}^*(t)$. Because $\hat{S}^*(t)$ is tightly bounded between the two survival curves, this ratio acts as a distortion strictly constrained by $\min(1, \hat{S}_A(t)/\hat{S}_B(t)) < \hat{S}_A(t)/\hat{S}^*(t) < \max(1, \hat{S}_A(t)/\hat{S}_B(t))$, which becomes unbounded as $\hat{S}_B(t)$ approaches zero. Crucially, this multiplier is asymmetric, i.e. weighting by $\hat{S}_A(t)/\hat{S}^*(t)$ evaluates the geometric deviation differently than if the reference groups were swapped to weight by $\hat{S}_B(t)/\hat{S}^*(t)$.

The full logrank test relies on its denominator to force symmetry back into the numerator. Because we discarded the denominator to preserve spatial geometry, we are left with this exposed, asymmetric distortion. Because a rigorous distance metric must satisfy spatial symmetry ($d(A, B) = d(B, A)$), we strip this proportional weight. Doing so geometrically resolves the numerator's asymmetry and isolates the pure spatial distance $\Sigma_S(t) = \hat{S}_A(t) - \hat{S}_B(t)$.

From Signed to Absolute. The logrank relies on the Cox proportional hazards condition for power, which states that the rates at which subgroup and reference survival decay (i.e. hazards) are constantly proportional over time. This assumption is often violated in practice, for example, when the survival functions for groups of interest cross one another. The logrank test inherently integrates a signed difference. If the survival curves cross, their relative hazard rates invert, causing the logrank's native differentials to flip signs over the domain $\mathcal{T}$. Consequently, the logrank inherently accumulates positive and negative areas that cancel each other out, severely reducing statistical power under crossing hazards. To resolve this, we apply an absolute value to the isolated state, yielding a robust spatial distance measure $|\Sigma_S(t)| = |\hat{S}_A(t) - \hat{S}_B(t)|$. Evaluating this absolute state measures true net geometric distance without over-penalizing high-frequency estimation noise, recovering exactly $\ell_t^1(s_A, s_B)$ from the right-hand side of the inequality in Prop. 3.1.

From Group to Individuals. Despite the temporal and spatial modifications in Steps 1 through 3, the measure thus far still operates entirely on expected group-level estimates. By definition, the group survival curve is simply the expectation of individual survival probabilities over all subjects selectable by the subgroup rule σ_A: $\hat{S}_A(t) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}[\hat{S}(t \mid \mathbf{x}) \mid \sigma_A(\mathbf{x}) = 1]$. Evaluating the deviation at this group level inherently averages away and obscures individual heterogeneity. We resolve this by taking the expectation *outside* the distance function.

Crucially, this final step strictly depends on the absolute value applied in Step 3. If we had retained the logrank's signed difference, moving the expectation outside would have no effect due to the linearity of expectation

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\left[\hat{S}(t \mid \mathbf{x}) - \hat{S}_B(t) \;\middle|\; \sigma_A(\mathbf{x}) = 1\right] = \hat{S}_A(t) - \hat{S}_B(t) \; .$$

However, because the absolute value function is convex, applying it *before* the expectation actively captures the internal heterogeneity that the group average cancels out. By Jensen's inequality we get the left-hand side of the inequality in Prop. 3.1.

Applying these four modifications in sequence, namely, integrating the accumulated state rather than differentials, mapping to survival while removing the logrank numerator's native asymmetric distortion, applying the absolute value function to prevent native cancellation, and moving the expectation outside the absolute difference to capture individual heterogeneity, exactly yields our proposed exceptionality measure from Eq. 1

$$\phi(\sigma_A, \sigma_B) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\left[\ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}), \hat{S}_B(t)) \;\middle|\; \sigma_A(\mathbf{x}) = 1\right] \; ,$$

where σ_A simplifies to σ, and B stands for the reference population/dataset D.

PROOFS OF PROPOSITIONS J.1 AND J.2

Proposition J.1. *Let $\Sigma_\Lambda(t) = \hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)$ be the continuous cumulative hazard state difference over the time domain $\mathcal{T} = [0, t_{\max}]$. Integrating the state directly over the domain is equivalent to integrating its differentials scaled by a temporal persistence weight, such that*

$$\int_{\mathcal{T}} \Sigma_\Lambda(t)\, dt = t_{\max}\Sigma_\Lambda(0) + \int_{\mathcal{T}} (t_{\max} - t)\, d\Sigma_\Lambda(t) \; .$$

Proof. We apply the standard rule of continuous integration by parts, $\int u\, dv = uv - \int v\, du$, to the integral of the state over $[0, t_{\max}]$. Let $u = \Sigma_\Lambda(t)$ and $dv = dt$. We select the valid antiderivative $v = (t - t_{\max})$. Therefore, $du = d\Sigma_\Lambda(t)$. Substituting these into the integration by parts formula yields

$$\int_0^{t_{\max}} \Sigma_\Lambda(t)\, dt = \Big[\Sigma_\Lambda(t)(t - t_{\max})\Big]_0^{t_{\max}} - \int_0^{t_{\max}} (t - t_{\max})\, d\Sigma_\Lambda(t) \; .$$

Evaluating the boundary conditions for the first term

$$\Big(\Sigma_\Lambda(t_{\max}) \cdot 0\Big) - \Big(\Sigma_\Lambda(0) \cdot (-t_{\max})\Big) = t_{\max}\Sigma_\Lambda(0) \; .$$

Substituting this back and absorbing the negative sign into the integral gives the stated equivalence

$$\int_0^{t_{\max}} \Sigma_\Lambda(t)\, dt = t_{\max}\Sigma_\Lambda(0) + \int_0^{t_{\max}} (t_{\max} - t)\, d\Sigma_\Lambda(t) \; .$$

□

Proposition J.2. *Given two groups A and B for which the expected group-level survival probabilities at any fixed time t are $\hat{S}_A(t)$ and $\hat{S}_B(t)$, respectively, and for which $\hat{\Lambda}_A(t)$ and $\hat{\Lambda}_B(t)$ are their corresponding cumulative hazard functions. The difference in their cumulative hazards is proportional to the difference in survival $\hat{S}_A(t) - \hat{S}_B(t)$, such that*

$$\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t) = -\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\hat{S}^*(t)} \; ,$$

where $\hat{S}^*(t)$ is exactly the logarithmic mean of the two survival probabilities, formally defined to account for potential curve crossing as

$$\hat{S}^*(t) = \begin{cases} \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\ln(\hat{S}_A(t)) - \ln(\hat{S}_B(t))} & \text{if } \hat{S}_A(t) \neq \hat{S}_B(t) \\ \hat{S}_A(t) & \text{if } \hat{S}_A(t) = \hat{S}_B(t) \end{cases},$$

which is strictly bounded by their geometric and arithmetic means, $\sqrt{\hat{S}_A(t)\hat{S}_B(t)} < \hat{S}^*(t) < \frac{\hat{S}_A(t)+\hat{S}_B(t)}{2}$ (for $\hat{S}_A(t) \neq \hat{S}_B(t)$).

Proof. In survival analysis, the (group-level) survival function is related to its corresponding cumulative hazard function thus

$$\hat{S}_\circ(t) = e^{-\hat{\Lambda}_\circ(t)} \implies \hat{\Lambda}_\circ(t) = -\ln(\hat{S}_\circ(t)) .$$

Consider the real-valued function $f(v) = e^{-v}$. This function is continuous and differentiable everywhere on $\mathbb{R}$, with its first derivative being $f'(v) = -e^{-v}$. We evaluate the difference in survival based on the cumulative hazards.

Case 1: The survival functions do not cross ($\hat{S}_A(t) \neq \hat{S}_B(t)$). Since $\hat{S}_A(t) \neq \hat{S}_B(t)$, it follows that $\hat{\Lambda}_A(t) \neq \hat{\Lambda}_B(t)$. We apply the Mean Value Theorem. There exists an intermediate cumulative hazard value, $\hat{\Lambda}^*(t)$, strictly between $\hat{\Lambda}_A(t)$ and $\hat{\Lambda}_B(t)$ such that

$$f(\hat{\Lambda}_A(t)) - f(\hat{\Lambda}_B(t)) = f'(\hat{\Lambda}^*(t))(\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)) .$$

Evaluating f and f' yields

$$e^{-\hat{\Lambda}_A(t)} - e^{-\hat{\Lambda}_B(t)} = -e^{-\hat{\Lambda}^*(t)}(\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)) .$$

Substituting the group-level survivals back into the left side gives

$$\hat{S}_A(t) - \hat{S}_B(t) = -e^{-\hat{\Lambda}^*(t)}(\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)) .$$

Let us define the intermediate term $\hat{S}^*(t) = e^{-\hat{\Lambda}^*(t)}$. Because $f(v) = e^{-v}$ is strictly monotonically decreasing, $\hat{S}^*(t)$ is an intermediate survival probability strictly between $\hat{S}_A(t)$ and $\hat{S}_B(t)$. Rearranging the equation to solve for $\hat{S}^*(t)$ and substituting $\hat{\Lambda}_\circ(t) = -\ln(\hat{S}_\circ(t))$ reveals its exact analytical form

$$\hat{S}^*(t) = -\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t)} = -\frac{\hat{S}_A(t) - \hat{S}_B(t)}{-\ln(\hat{S}_A(t)) - (-\ln(\hat{S}_B(t)))} = \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\ln(\hat{S}_A(t)) - \ln(\hat{S}_B(t))} .$$

This confirms $\hat{S}^*(t)$ is exactly the logarithmic mean. By standard properties of the logarithmic mean for any two distinct positive real numbers, $\hat{S}^*(t)$ is strictly bounded below by the geometric mean $\sqrt{\hat{S}_A(t)\hat{S}_B(t)}$ and above by the arithmetic mean $\frac{\hat{S}_A(t)+\hat{S}_B(t)}{2}$, fulfilling the proposition's bounds. Finally, rearranging the equation to isolate the difference in cumulative hazards yields

$$\hat{\Lambda}_A(t) - \hat{\Lambda}_B(t) = -\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\hat{S}^*(t)} .$$

Case 2: The survival functions cross ($\hat{S}_A(t) = \hat{S}_B(t)$). At the exact point of crossing, the cumulative hazards must also be equal ($\hat{\Lambda}_A(t) = \hat{\Lambda}_B(t)$), meaning the left side of our proposition evaluates to 0. However, evaluating the standard logarithmic mean on the right side yields an undefined $\frac{0}{0}$ singularity. To maintain continuity in the geometric space, we evaluate the limit of the logarithmic mean as $\hat{S}_B(t)$ approaches $\hat{S}_A(t)$. Applying L'Hôpital's rule with respect to $\hat{S}_B(t)$ yields

$$\lim_{\hat{S}_B(t)\to\hat{S}_A(t)} \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\ln(\hat{S}_A(t)) - \ln(\hat{S}_B(t))} = \lim_{\hat{S}_B(t)\to\hat{S}_A(t)} \frac{-1}{-1/\hat{S}_B(t)} = \lim_{\hat{S}_B(t)\to\hat{S}_A(t)} \hat{S}_B(t) = \hat{S}_A(t) .$$

This limit necessitates the piecewise definition $\hat{S}^*(t) = \hat{S}_A(t)$ at the crossing point. Substituting this continuous limit into the right side yields $-(0)/\hat{S}_A(t) = 0$. The equality holds trivially, and the mapping remains perfectly continuous.

Thus, the proposition holds continuously for all $t \in \mathcal{T}$. □

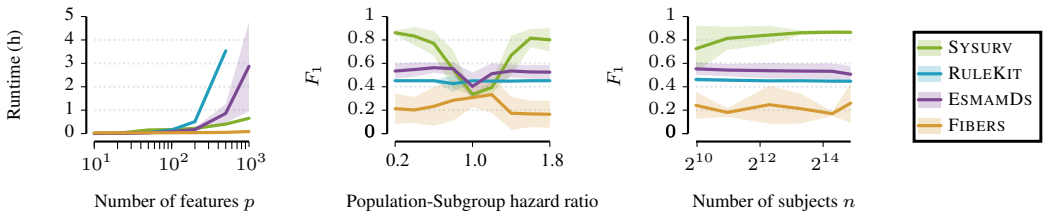


Figure 9: *Synthetic setting.* Comparison of SYSURV and each of RULEKIT, ESMAMDS, and FIBERS in terms of runtime recovering planted subgroups with increasingly large dataset dimensionalities **(Left)**. Lower is better. Also, in terms of F_1-score, comparisons between SYSURV and baselines under varying population-subgroup hazard ratios **(Center)**, where the hazard ratio indicates the rate by which the population outsurvives the subgroup, or vice versa, and increasingly large datasets **(Right)**. Higher is better. ESMAMDS is the closest competitor to SYSURV closely followed by RULEKIT. The shaded areas show ±1 standard error over 10 runs.

K ADDITIONAL EXPERIMENTAL RESULTS AND ABLATIONS

In this section, we provide additional experimental evaluations and ablation studies. In Appx. K.1, we evaluate synthetic scalability runtimes, hazard ratios, and sample sufficiency. In Appx. K.2, we provide real-world in-sample results, qualitative Kaplan-Meier curves, subgroup sizes, rule lengths before and after pruning, computational runtimes, coverage vs. exceptionality frontiers, and intra-subgroup cohesion. In Appx. K.3, we present ablation studies on feature handling, compound search strategy and feature handling, 10-fold cross-validation stability, and population model sensitivity. Lastly, in Appx. K.4, we provide the full and pruned rules discovered in our clinical case study.

K.1 SYNTHETIC DATA

We assess the runtimes of SYSURV and each of the baselines from our scalability experiment in Fig. 4 (left). See Appx. G for details regarding computational complexity. In Fig. 9 (left), we see that all the methods have consistent runtimes except ESMAMDS, which varies quite a lot as the number of dimensions increases. Expectedly, ESMAMDS and RULEKIT require increasingly more time unlike SYSURV and FIBERS. We attribute some of the increase in runtime of SYSURV to learning the population model, which varies with the choice of the regressor. Overall, SYSURV is the fastest method when factoring in retrieval performance. ESMAMDS is the closest competitor but needs around 4 times more time for 1000 features.

Next, we assess the retrieval performance of each method under varying population-subgroup hazard ratios in Fig. 9 (center). The hazard ratio indicates the rate by which the population outsurvives the subgroup, or vice versa. We can see that SYSURV consistently outperforms all baselines across all hazard ratios. As the hazard ratio approaches 1.0, the retrieval performance of all methods degrades since the survival distributions of the subgroup and population become increasingly similar.

Lastly, we assess the retrieval performance of each method under increasingly large datasets in Fig. 9 (right). We see that SYSURV consistently outperforms all baselines across all dataset sizes. As the dataset size increases, the stability of SYSURV improves, since more data is available to learn from.

K.2 REAL-WORLD DATA

In this subsection, we provide extended empirical evaluations, diagnostics, and coverage frontiers across the 13 real-world benchmark datasets.

IN-SAMPLE EXCEPTIONALITY

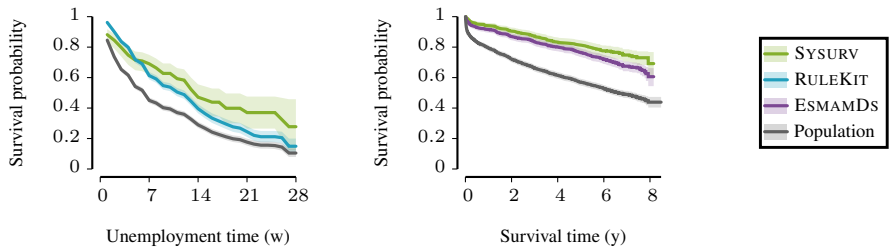
While Table 1 in the main text reports out-of-sample test exceptionality under 10-fold cross-validation, in Table 2 we provide the full-dataset (in-sample) exceptionality results across all 13 real-world datasets from our initial exploratory benchmark. In Fig. 10, we illustrate exemplary qualitative Kaplan-Meier survival curves of subgroups discovered on *Unemployment Duration* and *TRACE*.

On *Unemployment Duration* (Fig. 10a), both SYSURV and RULEKIT find subgroups with substantially longer time until re-employment than the overall population. The rule found by RULEKIT, *"unemployment insured"*, is succinct, while the rule found by SYSURV ($\gamma = 0.2$), *"age* $> 47.43 \wedge$ *replacement rate* $\in [0.11, 1.95] \wedge$ *disregard rate* $< 0.96 \wedge$ *tenure* $> 6.14 \wedge$ *log(wage)* $\in [4.20, 7.36]$*"*, provides a more detailed description of older workers who previously earned a higher wage.

On the heart attack dataset, *TRACE* (Fig. 10b), setting γ to 0.025 yields a compact subgroup of 470 subjects that covers more than half of the 896-subject cohort found by ESMAMDS. The rule found by SYSURV, *"wall motion index* $> 1.37 \wedge$ *female* $\wedge \neg$*clinical heart failure"*, provides a richer clinical characterization than ESMAMDS (*"*$\neg$*clinical heart failure"*).

Table 2: *Real-world setting.* Exceptionality of subgroups discovered on the full dataset without cross-validation by SYSURV (*Ours*), RULEKIT (RK), ESMAMDS (ED), and FIBERS (FIB), measured using 3 metrics (our objective, logrank, and mean-shift). SYSURV ranks first across metrics. Greater is better. Highest values marked in bold.

	Our objective				Logrank				Mean-shift			
	Ours	RK	ED	FIB	*Ours*	RK	ED	FIB	*Ours*	RK	ED	FIB
UnempDur	**5.1**	3.8	3.8	3.8	83.1	**226.5**	125.0	125.0	**2.4**	2.1	2.1	2.1
Nwtco	**2509.1**	1751.4	1751.4	556.1	**428.5**	281.3	31.4	0.0	**1160.1**	688.2	688.2	0.6
Rott2	**68.2**	40.3	59.3	59.3	**267.0**	93.0	71.6	74.2	**31.5**	10.0	25.2	25.2
Rdata	**1597.2**	822.7	1341.2	978.7	**120.8**	12.5	22.4	43.3	**1144.5**	240.9	967.3	705.5
Aids2	**194.6**	141.3	154.7	123.9	0.5	1.7	**4.3**	0.0	34.2	**61.2**	56.7	0.1
Dialysis	**18.3**	6.6	7.2	5.0	79.4	68.4	70.9	**110.3**	**11.3**	5.7	3.0	1.9
TRACE	1.9	1.8	**2.5**	2.2	97.6	94.6	93.2	**178.6**	1.4	0.6	**2.1**	1.7
Support2	**564.7**	347.5	**564.7**	206.1	**434.6**	59.9	43.9	11.9	**405.5**	227.6	**405.5**	71.0
DataDIVAT2	**4.1**	2.1	2.1	2.1	**33.5**	31.6	10.9	10.9	**3.2**	1.6	1.6	1.6
ProstateSurv.	**20.4**	7.5	14.1	5.3	**425.7**	64.7	82.5	82.5	**6.1**	3.7	3.4	1.0
Actg	**66.2**	51.2	31.7	19.9	**53.4**	1.1	8.0	0.0	**52.8**	27.3	14.4	0.4
Scania	3.1	**3.3**	3.1	2.9	0.3	14.7	**15.8**	9.9	0.5	1.6	**2.4**	0.0
Grace	**66.0**	28.1	35.0	23.6	**35.5**	33.3	8.2	13.8	**62.6**	16.0	31.8	11.1
Avg. rank	**1.23**	2.77	2.08	3.46	**1.77**	2.54	2.69	2.77	**1.46**	2.54	2.00	3.31
$\pm$ std	$\pm$0.58	$\pm$0.89	$\pm$0.62	$\pm$0.75	$\pm$1.12	$\pm$0.93	$\pm$0.99	$\pm$1.05	$\pm$0.84	$\pm$0.93	$\pm$0.68	$\pm$0.91



(a) Unemployment dataset (b) Heart attack dataset

Figure 10: *Qualitative real-world survival curves.* Survival subgroups discovered in the unemployment (**a**) and heart attack (**b**) datasets using SYSURV and RULEKIT resp. ESMAMDS in the full-dataset in-sample setting. The shaded areas show 95% confidence intervals.

In Table 3, we present the subgroup sizes (with coverage percentages) discovered by each method across real-world datasets. SYSURV consistently discovers smaller, more compact subgroups than all baselines. This directly reflects the effect of our size penalty γ, which prevents the discovery of overly broad cohorts. In contrast, the baselines lack direct mechanisms for penalizing cohort size, yielding substantially larger subgroups.

In Table 4, we report rule lengths (number of conditions) across real-world datasets and evaluate the effect of post hoc condition pruning (PRUNERULE, Alg. 3) using a Jaccard similarity threshold of 0.95. Unpruned SYSURV rules have a median length of 3 conditions (Q1: 2, Q3: 4), while baselines have medians of 1 (RULEKIT), 1 (ESMAMDS), and 2 (FIBERS). Post hoc pruning simplifies SYSURV rules to a median of 3 conditions (Q1: 2, Q3: 4) while preserving $\geq 95\%$ of patient memberships. Notably, on high-dimensional datasets with strong collinearity such as *Support2* ($p = 76$), pruning drastically simplifies the rule from 13 conditions down to just 4 highly actionable conditions.

In Table 5, we report the runtimes across all real-world datasets. SYSURV completes between 9 seconds and a little under 9 minutes with a median runtime of 18.6 seconds. RULEKIT is the fastest with a median of 2 seconds, while ESMAMDS is the second fastest with a median of 8 seconds but has several slow runs on larger datasets. SYSURV comes in third with one slower run (9 minutes), while FIBERS consistently takes around a minute across datasets. Overall, SYSURV requires competitive time when applied to benchmark real-world datasets.

Table 3: Subgroup sizes (with coverage percentages) per method across real-world datasets.

Dataset	n	SYSURV	RULEKIT	ESMAMDS	FIBERS
UnempDur	3241	209 (6.4%)	1446 (44.6%)	1446 (44.6%)	1446 (44.6%)
Nwtco	4028	212 (5.3%)	459 (11.4%)	459 (11.4%)	4027 (100.0%)
Rott2	2982	320 (10.7%)	1546 (51.8%)	646 (21.7%)	646 (21.7%)
Rdata	1040	144 (13.8%)	770 (74.0%)	231 (22.2%)	436 (41.9%)
Aids2	2839	262 (9.2%)	368 (13.0%)	391 (13.8%)	2838 (100.0%)
Dialysis	6805	101 (1.5%)	241 (3.5%)	1473 (21.6%)	4465 (65.6%)
TRACE	1878	470 (25.0%)	1074 (57.2%)	279 (14.9%)	576 (30.7%)
Support2	9105	79 (0.9%)	97 (1.1%)	79 (0.9%)	451 (5.0%)
DataDIVAT2	1837	63 (3.4%)	677 (36.9%)	677 (36.9%)	677 (36.9%)
ProstateSurv.	14294	1840 (12.9%)	5606 (39.2%)	3306 (23.1%)	11220 (78.5%)
Actg	1151	37 (3.2%)	122 (10.6%)	221 (19.2%)	1149 (99.8%)
Scania	1931	1630 (84.4%)	1159 (60.0%)	229 (11.9%)	1294 (67.0%)
Grace	1000	17 (1.7%)	417 (41.7%)	186 (18.6%)	726 (72.6%)
Q1 (25%) Cov.	–	3.20%	11.40%	13.77%	36.85%
Median (Q2) Cov.	–	6.45%	39.22%	19.20%	65.61%
Q3 (75%) Cov.	–	12.90%	51.84%	22.21%	78.49%

Table 4: Rule lengths (number of conditions) across real-world datasets before and after post hoc pruning (PRUNERULE, Alg. 3, Jaccard threshold 0.95), compared to baseline methods.

Dataset	p	SYSURV (Unpruned)	SYSURV (Pruned)	RULEKIT	ESMAMDS	FIBERS
UnempDur	7	5	4	1	1	1
Nwtco	14	3	2	1	1	6
Rott2	18	1	1	1	1	1
Rdata	8	2	2	1	1	2
Aids2	15	3	3	3	4	3
Dialysis	74	2	2	1	1	40
TRACE	10	3	3	1	1	2
Support2	76	13	4	1	1	1
DataDIVAT2	7	2	2	1	1	1
ProstateSurvival	9	5	4	1	2	2
Actg	24	4	4	2	2	3
Scania	12	1	1	2	1	5
Grace	7	4	4	2	1	5
Q1 (25%)	–	2	2	1	1	1
Median (Q2)	–	3	3	1	1	2
Q3 (75%)	–	4	4	2	1	5

Table 5: Runtime (seconds) per method across real-world datasets.

	SYSURV	RULEKIT	ESMAMDS	FIBERS
UnempDur	15.24	1.93	0.82	65.07
Nwtco	75.22	2.59	81.97	69.65
Rott2	20.52	3.20	7.82	70.07
Rdata	11.17	0.46	1.76	60.83
Aids2	18.64	2.55	758.92	61.93
Dialysis	525.30	65.11	4826.16	72.71
TRACE	17.32	1.08	2.30	62.28
Support2	34.42	3.86	20.84	65.12
DataDIVAT2	13.00	0.61	5.12	57.25
ProstateSurv.	28.12	12.93	7.18	71.34
Actg	14.35	1.47	21.44	60.96
Scania	19.75	1.16	51.05	63.21
Grace	9.09	0.39	1.99	61.49
Q1 (25%)	14.35	1.08	2.30	61.49
Median (Q2)	18.64	1.93	7.82	63.21
Q3 (75%)	28.12	3.20	51.05	69.65

INTRA-SUBGROUP COHESION

To investigate whether maximizing per-subject average survival deviation produces heterogeneous subgroups, we evaluate intra-subgroup cohesion across all benchmark datasets using two metrics. First, we compute the **Average IQR Ratio**, defined as the ratio of the interquartile range (IQR) of survival probabilities within the discovered subgroup to that of the population, averaged across all time points $t \in \mathcal{T}$; a ratio < 1.0 indicates that subgroup survival trajectories are more compact than the population. Second, we compute the **Silhouette Score**, which evaluates within-cluster cohesion relative to separation from the remaining population based on pairwise L^1 survival distances, with higher values indicating more cohesive and well-separated subgroups.

In Table 6, we report the results across all 13 real-world benchmark datasets. SYSURV achieves the best Average IQR Ratio (0.65 vs. 0.71–1.08) and ties for best average silhouette score (0.42, tied with ESMAMDS at 0.42), demonstrating that our objective naturally discovers compact, cohesive cohorts rather than diffuse outliers.

Table 6: Intra-subgroup cohesion metrics across real-world datasets: Average IQR Ratio (lower is better, < 1 indicates higher compactness than population) and Silhouette Score (higher is better).

	Avg. IQR Ratio				Silhouette Score			
Dataset	SYSURV	RULEKIT	ESMAMDS	FIBERS	SYSURV	RULEKIT	ESMAMDS	FIBERS
UnempDur	**0.2**	0.4	0.4	0.4	0.1	**0.6**	**0.6**	**0.6**
Nwtco	1.9	2.9	2.9	**1.0**	**0.7**	**0.7**	**0.7**	0.0
Rott2	**0.4**	1.0	0.5	0.5	0.4	0.3	**0.5**	**0.5**
Rdata	**0.3**	1.0	0.5	0.7	0.4	0.2	**0.5**	**0.5**
Aids2	1.8	1.1	1.2	**1.0**	**0.3**	**0.3**	0.1	0.0
Dialysis	**0.4**	1.0	0.7	0.8	**0.6**	0.1	0.4	0.2
TRACE	**0.5**	1.0	**0.5**	**0.5**	0.2	0.1	0.3	**0.4**
Support2	**0.0**	1.4	**0.0**	0.6	**0.7**	0.3	**0.7**	**0.7**
DataDIVAT2	**0.3**	0.7	0.7	0.7	**0.5**	0.4	0.4	0.4
ProstateSurvival	**0.3**	0.8	1.3	0.5	**0.7**	0.2	0.6	0.6
Actg	1.0	**0.8**	1.5	1.0	**0.6**	**0.6**	0.4	0.5
Scania	1.1	1.1	**0.5**	0.7	-0.1	0.1	0.1	**0.3**
Grace	**0.3**	0.9	**0.3**	0.8	**0.4**	0.3	0.2	0.3
Average	**0.65**	1.08	0.85	0.71	**0.42**	0.32	**0.42**	0.38

COVERAGE VS. EXCEPTIONALITY FRONTIERS

To verify whether SYSURV's advantage stems from genuine contrast rather than differing subgroup sizes, we evaluate the coverage vs. exceptionality frontier across coverage levels from 10% to 90%. For each dataset, we sample 200 bootstrapped sets of data points from the discovered subgroups and fit a LOWESS per set. We average these LOWESS fits and compute the error bands. In Tables 7, 8, and 9, we report the frontier values for our objective, logrank, and mean-shift, respectively. Each cell contains the metric value with error for all 4 methods formatted across four lines (one line per method): SYSURV (line 1), RULEKIT (line 2), ESMAMDS (line 3), and FIBERS (line 4). A dash (–) indicates that the coverage level lies outside the empirical domain of discovered subgroups. Across these frontier evaluations, we only report the datasets where SYSURV has entries within the evaluated coverage window; *Nwtco*, *ProstateSurvival*, and *TRACE* are omitted because SYSURV has no entries in this coverage range.

On our objective (Table 7), SYSURV ranks first across all coverage levels and datasets (51 out of 51 combinations with ≥ 2 values available). On logrank (Table 8), RULEKIT ranks first most often with 22 out of 51 combinations, followed by SYSURV with 13, ESMAMDS with 11, and FIBERS with 10. On mean-shift (Table 9), SYSURV achieves the top rank in 34 out of the 51 combinations. This confirms that holding coverage strictly equal, SYSURV consistently discovers subgroups with stronger survival divergence. Figure 11 visually plots these complete coverage vs. exceptionality frontiers across all 10 real-world datasets for our objective, logrank, and mean-shift metrics, displaying the bootstrapped LOWESS curves with standard error bands.

Table 7: Coverage vs. exceptionality frontier on **our objective**. Each cell reports four lines (one per method): SYSURV (line 1), RULEKIT (line 2), ESMAMDS (line 3), and FIBERS (line 4) with bootstrapped standard errors. Dashes (–) indicate coverage levels outside the empirical domain.

Dataset	10%	20%	30%	40%	50%	60%	70%	80%	90%
UnempDur	6.2 ± 0.7	5.8 ± 0.6	5.5 ± 0.5	5.2 ± 0.5	4.9 ± 0.4	4.6 ± 0.3	4.3 ± 0.2	4.0 ± 0.1	3.7 ± 0.0
	–	–	–	–	3.5 ± 0.0	–	–	–	–
	–	–	–	–	3.5 ± 0.0	–	–	–	–
	–	–	–	–	3.5 ± 0.0	–	–	–	–
Rott2	71.3 ± 3.1	65.8 ± 2.8	61.7 ± 2.4	57.5 ± 2.0	52.9 ± 1.6	48.7 ± 1.3	44.5 ± 1.1	40.4 ± 1.1	37.4 ± 0.5
	–	–	–	–	–	35.4 ± 0.8	32.4 ± 0.3	–	–
	–	54.6 ± 4.5	48.7 ± 3.2	40.0 ± 1.2	–	–	–	–	–
	–	–	51.6 ± 0.5	44.5 ± 0.7	39.3 ± 0.5	34.5 ± 0.1	–	–	–
Rdata	–	1515.1 ± 5.4	1391.0 ± 13.4	1266.8 ± 21.6	1130.3 ± 30.7	1065.9 ± 19.2	1003.6 ± 6.6	–	–
	–	–	–	–	–	–	–	–	–
	–	–	1213.3 ± 7.0	1041.9 ± 16.3	823.5 ± 1.8	–	–	–	–
	–	–	–	–	866.6 ± 0.0	–	–	–	–
Aids2	197.4 ± 0.0	189.0 ± 0.0	180.6 ± 0.0	172.2 ± 0.0	162.9 ± 0.0	154.6 ± 0.0	146.2 ± 0.0	137.8 ± 0.0	129.4 ± 0.0
	–	143.7 ± 0.4	138.0 ± 0.7	132.1 ± 1.2	125.6 ± 1.8	121.4 ± 1.3	118.0 ± 1.1	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Dialysis	15.0 ± 0.3	13.1 ± 0.7	11.1 ± 1.0	9.2 ± 1.3	6.8 ± 1.5	6.3 ± 1.1	5.8 ± 0.7	5.6 ± 0.4	5.4 ± 0.1
	6.7 ± 0.0	6.5 ± 0.0	6.3 ± 0.1	6.0 ± 0.1	5.8 ± 0.1	5.5 ± 0.2	5.3 ± 0.2	5.1 ± 0.2	5.1 ± 0.1
	–	6.0 ± 0.7	6.1 ± 0.4	5.5 ± 0.2	5.3 ± 0.1	–	–	–	–
	–	–	–	–	–	–	4.9 ± 0.0	–	–
Support2	568.1 ± 3.7	543.6 ± 4.7	501.7 ± 5.3	459.8 ± 6.7	413.7 ± 8.7	371.8 ± 10.7	329.8 ± 12.8	290.2 ± 13.6	273.2 ± 4.7
	–	346.9 ± 0.0	329.7 ± 0.0	312.5 ± 0.0	293.5 ± 0.0	276.3 ± 0.0	259.0 ± 0.0	241.8 ± 0.0	–
	–	346.9 ± 85.6	311.4 ± 32.5	314.3 ± 23.0	278.6 ± 16.0	239.3 ± 1.6	–	–	–
	–	–	–	–	–	–	–	–	–
DataDIVAT2	3.8 ± 0.0	3.6 ± 0.0	3.4 ± 0.0	3.1 ± 0.0	2.9 ± 0.0	2.6 ± 0.0	2.4 ± 0.0	2.2 ± 0.0	1.9 ± 0.0
	–	–	–	2.1 ± 0.0	1.8 ± 0.0	1.6 ± 0.0	–	–	–
	–	–	–	1.8 ± 0.2	1.7 ± 0.1	1.6 ± 0.0	–	–	–
	–	–	–	2.1 ± 0.0	1.8 ± 0.0	1.6 ± 0.0	–	–	–
Actg	56.0 ± 4.1	42.4 ± 9.8	39.6 ± 8.6	36.8 ± 7.4	33.7 ± 6.0	30.9 ± 4.8	28.1 ± 3.6	25.3 ± 2.4	22.5 ± 1.1
	–	33.0 ± 3.0	26.8 ± 3.5	24.2 ± 2.8	21.3 ± 2.5	18.7 ± 2.9	19.6 ± 1.7	–	–
	–	–	23.0 ± 2.5	18.3 ± 2.1	16.9 ± 0.5	–	–	–	–
	–	–	–	–	–	–	–	–	–
Scania	5.4 ± 0.0	5.1 ± 0.0	4.8 ± 0.0	4.5 ± 0.0	4.2 ± 0.0	3.9 ± 0.0	3.6 ± 0.0	3.3 ± 0.0	3.1 ± 0.0
	–	2.1 ± 0.4	2.4 ± 0.3	2.6 ± 0.3	2.9 ± 0.2	3.1 ± 0.2	3.2 ± 0.1	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Grace	63.3 ± 2.3	59.1 ± 2.0	55.0 ± 1.8	50.8 ± 1.5	46.2 ± 1.2	42.0 ± 1.0	37.9 ± 0.7	33.7 ± 0.5	29.5 ± 0.2
	–	–	–	–	26.0 ± 0.0	–	–	–	–
	–	33.3 ± 1.2	32.1 ± 1.1	29.6 ± 1.1	24.7 ± 0.7	–	–	–	–
	–	–	–	–	–	–	–	–	–

Table 8: Coverage vs. exceptionality frontier on **logrank**. Each cell reports four lines (one per method): SYSURV (line 1), RULEKIT (line 2), ESMAMDS (line 3), and FIBERS (line 4) with bootstrapped standard errors. Dashes (–) indicate coverage levels outside the empirical domain.

Dataset	10%	20%	30%	40%	50%	60%	70%	80%	90%
UnempDur	51.5 ± 18.8	45.1 ± 16.5	38.8 ± 14.1	32.4 ± 11.8	25.5 ± 9.3	19.1 ± 6.9	12.8 ± 4.6	6.5 ± 2.4	0.4 ± 0.7
	–	–	–	–	171.1 ± 0.0	–	–	–	–
	–	–	–	–	171.1 ± 0.0	–	–	–	–
	–	–	–	–	171.1 ± 0.0	–	–	–	–
Rott2	218.7 ± 47.0	207.2 ± 42.8	176.5 ± 36.6	145.8 ± 30.5	112.0 ± 23.9	81.3 ± 18.2	50.6 ± 13.3	19.9 ± 10.3	1.2 ± 4.6
	–	–	–	–	–	60.6 ± 11.2	51.9 ± 5.0	–	–
	–	209.0 ± 75.9	182.4 ± 52.6	102.9 ± 20.7	–	–	–	–	–
	–	–	240.1 ± 3.7	173.3 ± 8.6	131.1 ± 5.5	92.1 ± 1.7	–	–	–
Rdata	–	100.0 ± 4.0	70.1 ± 6.6	40.1 ± 9.9	7.1 ± 13.8	4.2 ± 8.5	2.4 ± 2.9	–	–
	–	–	–	–	–	–	–	–	–
	–	–	97.6 ± 1.5	78.8 ± 3.4	50.7 ± 0.4	–	–	–	–
	–	–	–	–	62.6 ± 0.0	–	–	–	–
Aids2	0.5 ± 0.0	0.5 ± 0.0	0.4 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
	–	11.4 ± 0.3	9.7 ± 0.3	8.3 ± 0.2	6.7 ± 0.2	5.1 ± 0.2	3.4 ± 0.1	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Dialysis	67.4 ± 2.9	53.8 ± 5.9	40.1 ± 8.8	26.5 ± 11.8	9.3 ± 13.7	8.4 ± 9.0	6.9 ± 5.8	4.3 ± 3.5	1.8 ± 1.2
	63.9 ± 0.1	57.0 ± 0.3	50.1 ± 0.6	43.2 ± 0.8	35.7 ± 1.0	28.8 ± 1.2	21.9 ± 1.4	15.0 ± 1.6	6.4 ± 0.7
	–	102.3 ± 24.0	108.3 ± 25.7	77.7 ± 13.7	72.4 ± 6.7	–	–	–	–
	–	–	–	–	–	–	89.7 ± 0.6	–	–
Support2	341.0 ± 25.2	389.1 ± 31.1	326.6 ± 27.0	264.1 ± 23.9	195.3 ± 21.9	132.8 ± 21.9	70.3 ± 23.6	12.8 ± 23.9	2.7 ± 7.9
	–	58.6 ± 0.0	50.1 ± 0.0	41.7 ± 0.0	32.4 ± 0.0	23.9 ± 0.0	15.5 ± 0.0	7.0 ± 0.0	–
	–	138.5 ± 127.3	64.4 ± 49.7	46.5 ± 24.4	36.3 ± 13.4	28.5 ± 1.4	–	–	–
	–	–	–	–	–	–	–	–	–
DataDIVAT2	31.2 ± 0.0	27.6 ± 0.0	24.0 ± 0.0	20.4 ± 0.0	16.5 ± 0.1	12.9 ± 0.1	9.3 ± 0.1	5.7 ± 0.1	2.2 ± 0.1
	–	–	–	29.4 ± 0.0	20.9 ± 0.0	13.1 ± 0.0	–	–	–
	–	–	–	20.8 ± 5.8	16.2 ± 3.1	12.1 ± 0.7	–	–	–
	–	–	–	29.4 ± 0.0	20.9 ± 0.0	13.1 ± 0.0	–	–	–
Actg	36.8 ± 7.3	29.8 ± 7.2	26.1 ± 6.3	22.4 ± 5.4	18.3 ± 4.4	14.6 ± 3.5	10.9 ± 2.6	7.2 ± 1.7	3.5 ± 0.8
	–	31.4 ± 2.8	24.2 ± 3.7	20.4 ± 2.8	16.2 ± 2.0	12.4 ± 1.7	6.6 ± 1.0	–	–
	–	–	18.3 ± 2.5	15.3 ± 2.2	15.4 ± 0.5	–	–	–	–
	–	–	–	–	–	–	–	–	–
Scania	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.1	0.3 ± 0.1	0.1 ± 0.0
	–	15.1 ± 0.5	14.2 ± 0.8	13.3 ± 1.2	12.3 ± 1.7	10.9 ± 2.3	7.6 ± 1.1	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Grace	26.0 ± 10.7	23.1 ± 9.5	20.2 ± 8.3	17.3 ± 7.1	14.2 ± 5.8	11.3 ± 4.6	8.4 ± 3.4	5.5 ± 2.2	2.6 ± 1.1
	–	–	–	–	28.5 ± 0.0	–	–	–	–
	–	27.1 ± 9.9	23.5 ± 5.7	18.6 ± 2.8	13.6 ± 0.3	–	–	–	–
	–	–	–	–	–	–	–	–	–

Table 9: Coverage vs. exceptionality frontier on **mean-shift**. Each cell reports four lines (one per method): SYSURV (line 1), RULEKIT (line 2), ESMAMDS (line 3), and FIBERS (line 4) with bootstrapped standard errors. Dashes (–) indicate coverage levels outside the empirical domain.

Dataset	10%	20%	30%	40%	50%	60%	70%	80%	90%
UnempDur	2.2 ± 0.1	2.0 ± 0.0	1.7 ± 0.0	1.4 ± 0.0	1.1 ± 0.0	0.8 ± 0.0	0.6 ± 0.0	0.3 ± 0.0	0.0 ± 0.0
	–	–	–	–	1.9 ± 0.0	–	–	–	–
	–	–	–	–	1.9 ± 0.0	–	–	–	–
	–	–	–	–	1.9 ± 0.0	–	–	–	–
Rott2	35.8 ± 3.0	30.0 ± 2.8	25.6 ± 2.4	21.2 ± 2.0	16.3 ± 1.7	11.9 ± 1.4	7.5 ± 1.3	3.0 ± 1.4	0.3 ± 0.7
	–	–	–	–	–	7.9 ± 0.8	7.4 ± 0.4	–	–
	–	23.1 ± 2.5	18.7 ± 1.9	12.6 ± 1.2	–	–	–	–	–
	–	–	20.6 ± 0.2	15.5 ± 0.4	12.4 ± 0.3	9.6 ± 0.1	–	–	–
Rdata	–	971.5 ± 22.8	707.4 ± 50.0	443.3 ± 79.7	152.9 ± 112.9	117.0 ± 69.4	90.7 ± 23.9	–	–
	–	–	–	–	–	–	–	–	–
	–	–	862.9 ± 0.3	725.1 ± 0.8	575.3 ± 0.1	–	–	–	–
	–	–	–	–	601.4 ± 0.0	–	–	–	–
Aids2	33.9 ± 0.0	30.3 ± 0.0	26.7 ± 0.0	23.0 ± 0.0	19.1 ± 0.0	15.4 ± 0.0	11.8 ± 0.0	8.2 ± 0.0	4.6 ± 0.0
	–	59.5 ± 2.2	53.5 ± 1.9	46.1 ± 1.3	38.1 ± 0.7	30.8 ± 0.4	23.5 ± 0.4	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Dialysis	9.6 ± 0.3	7.7 ± 0.6	5.9 ± 1.0	4.0 ± 1.3	1.6 ± 1.5	1.1 ± 1.0	0.7 ± 0.6	0.5 ± 0.4	0.3 ± 0.1
	5.3 ± 0.0	4.7 ± 0.1	4.0 ± 0.1	3.3 ± 0.2	2.6 ± 0.2	2.0 ± 0.3	1.3 ± 0.3	0.6 ± 0.3	0.4 ± 0.1
	–	1.7 ± 0.3	2.6 ± 0.4	2.4 ± 0.2	2.2 ± 0.1	–	–	–	–
	–	–	–	–	–	–	1.8 ± 0.0	–	–
Support2	407.9 ± 6.0	372.7 ± 7.6	313.9 ± 8.6	255.1 ± 10.9	190.4 ± 14.2	131.6 ± 17.5	72.8 ± 21.0	18.4 ± 22.0	6.3 ± 7.0
	–	223.4 ± 0.0	195.8 ± 0.0	168.3 ± 0.0	137.9 ± 0.0	110.3 ± 0.0	82.8 ± 0.0	55.2 ± 0.0	–
	–	254.9 ± 65.0	190.3 ± 25.0	161.6 ± 10.9	137.0 ± 6.0	116.0 ± 0.6	–	–	–
	–	–	–	–	–	–	–	–	–
DataDIVAT2	3.0 ± 0.0	2.6 ± 0.0	2.3 ± 0.0	2.0 ± 0.0	1.6 ± 0.0	1.3 ± 0.0	0.9 ± 0.0	0.6 ± 0.0	0.3 ± 0.0
	–	–	–	1.5 ± 0.0	1.2 ± 0.0	1.0 ± 0.0	–	–	–
	–	–	–	1.2 ± 0.2	1.1 ± 0.1	0.9 ± 0.0	–	–	–
	–	–	–	1.5 ± 0.0	1.2 ± 0.0	1.0 ± 0.0	–	–	–
Actg	40.7 ± 9.0	22.9 ± 13.5	20.0 ± 11.9	17.2 ± 10.2	14.1 ± 8.3	11.3 ± 6.6	8.5 ± 5.0	5.6 ± 3.3	2.8 ± 1.6
	–	11.0 ± 4.1	6.1 ± 4.3	4.9 ± 3.1	3.6 ± 1.8	2.4 ± 0.9	1.9 ± 0.5	–	–
	–	–	5.0 ± 1.2	4.6 ± 0.9	2.9 ± 0.2	–	–	–	–
	–	–	–	–	–	–	–	–	–
Scania	11.8 ± 0.0	10.3 ± 0.0	8.8 ± 0.0	7.3 ± 0.0	5.7 ± 0.1	4.2 ± 0.1	2.7 ± 0.1	1.2 ± 0.1	0.4 ± 0.0
	–	4.3 ± 0.6	3.6 ± 0.5	3.0 ± 0.4	2.3 ± 0.3	1.8 ± 0.3	1.4 ± 0.2	–	–
	–	–	–	–	–	–	–	–	–
	–	–	–	–	–	–	–	–	–
Grace	57.3 ± 0.0	50.9 ± 0.1	44.6 ± 0.1	38.2 ± 0.1	31.2 ± 0.1	24.9 ± 0.1	18.5 ± 0.1	12.1 ± 0.1	5.8 ± 0.2
	–	–	–	–	13.6 ± 0.0	–	–	–	–
	–	25.4 ± 3.6	20.2 ± 1.5	16.1 ± 0.7	11.9 ± 0.1	–	–	–	–
	–	–	–	–	–	–	–	–	–

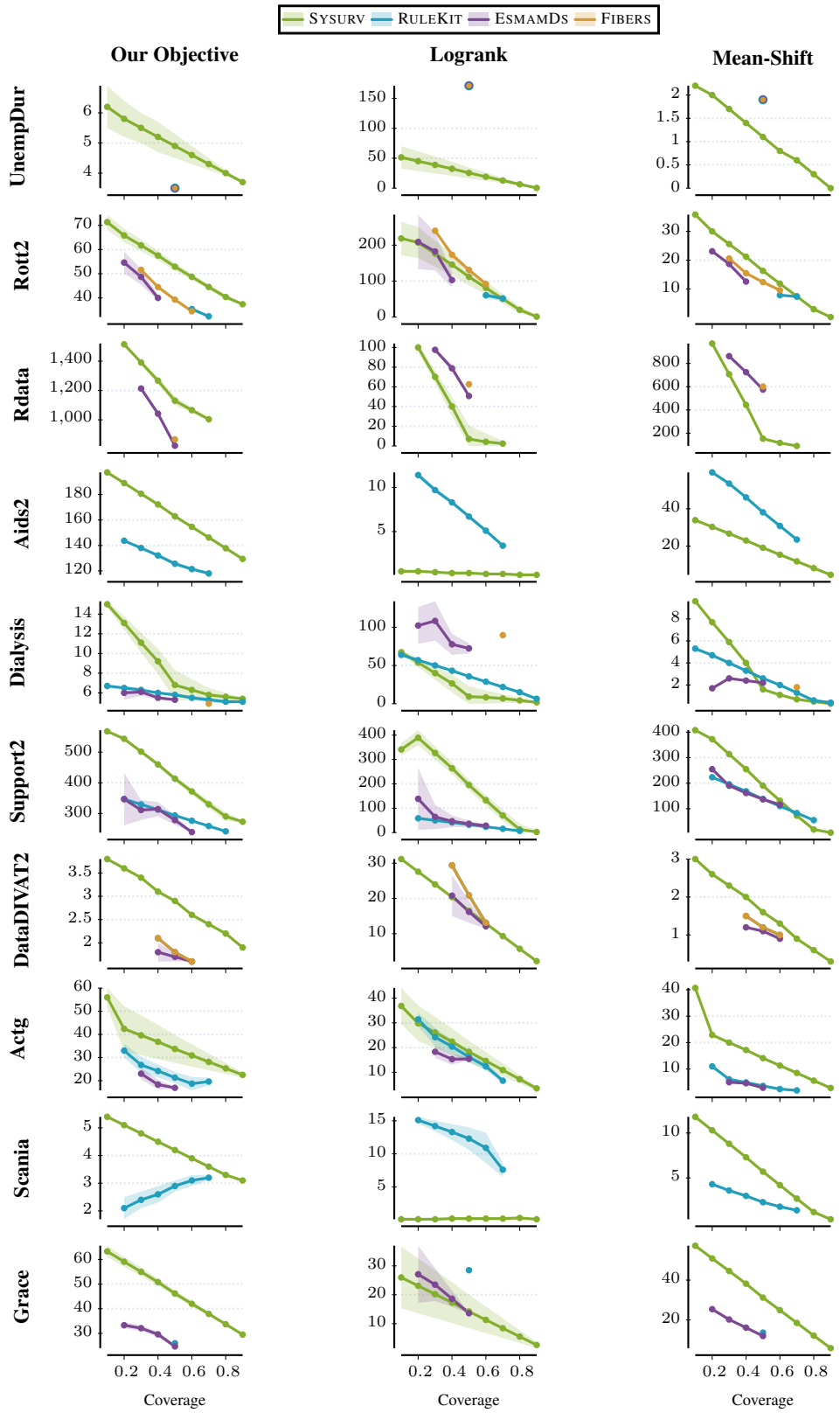


Figure 11: *Coverage vs. exceptionality frontiers.* Columns show Our Objective (left), Logrank (center), and Mean-Shift (right). Shaded bands depict bootstrapped LOWESS standard errors across coverage levels (10% to 90%). SYSURV consistently dominates on our objective and mean-shift (almost) while remaining highly competitive on logrank across matched subgroup sizes.

K.3 Ablations and Population Model Sensitivity

In this section, we present detailed ablations isolating the effects of continuous feature handling, the compound impact of search strategy and feature handling, train-test generalization, and the choice of the population model. As with the synthetic experiments, all ablation evaluations (feature handling, search strategy, and population model sensitivity) are reported over 10 random seeds.

Feature Handling. We first ablate feature handling alone by evaluating whether directly optimizing continuous feature intervals outperforms pre-discretization into bins over 10 random seeds (Table 10). Continuous feature optimization consistently achieves higher retrieval performance ($F_1 = 0.87 \pm 0.03$ vs. 0.76 ± 0.07 at $\gamma = 0.10$), showing that continuous intervals fit true subgroup boundaries more accurately than pre-computed bins while holding the optimization framework fixed.

Compound Ablation. Next, we evaluate SYSURV with discrete heuristic search from ESMAMDS to isolate the joint effect of heuristic search and pre-discretization against continuous gradient optimization (Table 11). Retrieval drops sharply to $F_1 = 0.37 \pm 0.15$ at $\gamma = 0.10$ (0.40 ± 0.14 at $\gamma = 0.05$), confirming continuous gradient optimization as the primary performance driver.

Out-Of-Sample Exceptionality. In Table 12, we evaluate the out-of-sample generalization of discovered subgroups across methods under 10-fold cross-validation ($n = 5000$). SYSURV exhibits strong stability between training and test folds ($F_1 = 0.89 \rightarrow 0.88$, precision $0.99 \rightarrow 0.98$, coverage $0.17 \rightarrow 0.17$), closely matching the true planted subgroup coverage (0.20). In contrast, baseline methods struggle to generalize: ESMAMDS suffers severe dilution on test folds (coverage expands from 0.38 to 0.92 while precision drops to 0.23), while RULEKIT and FIBERS remain trapped at much lower retrieval quality ($F_1 = 0.46$ and 0.20, respectively).

Table 10: Feature handling ablation on synthetic data (F_1 score).

γ	SYSURV (Continuous)	SYSURV (Pre-discretized)
0.05	0.80 ± 0.04	0.61 ± 0.19
0.10	$\mathbf{0.87 \pm 0.03}$	0.76 ± 0.07
0.15	0.82 ± 0.18	0.74 ± 0.09

Table 11: Compound ablation (search strategy and feature handling) on synthetic data (F_1 score).

γ	SYSURV (Cont. optimization)	SYSURV (Discrete heuristic search)
0.05	0.80 ± 0.04	0.40 ± 0.14
0.10	$\mathbf{0.87 \pm 0.03}$	0.37 ± 0.15
0.15	0.82 ± 0.18	0.37 ± 0.13

Table 12: Synthetic 10-fold cross-validation performance ($n = 5000$) comparing train and test stability across methods.

	F_1		Precision		Recall		Coverage	
Method	Train	Test	Train	Test	Train	Test	Train	Test
SYSURV	$\mathbf{0.89 \pm 0.00}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.99 \pm 0.02}$	$\mathbf{0.98 \pm 0.03}$	$\mathbf{0.82 \pm 0.02}$	0.81 ± 0.03	0.17 ± 0.01	0.17 ± 0.02
ESMAMDS	0.56 ± 0.03	0.35 ± 0.04	0.43 ± 0.01	0.23 ± 0.08	0.81 ± 0.11	$\mathbf{0.94 \pm 0.17}$	0.38 ± 0.06	0.92 ± 0.24
FIBERS	0.19 ± 0.19	0.20 ± 0.18	0.14 ± 0.15	0.14 ± 0.14	0.35 ± 0.25	0.37 ± 0.24	0.67 ± 0.14	0.66 ± 0.15
RULEKIT	0.46 ± 0.01	0.46 ± 0.06	0.46 ± 0.01	0.47 ± 0.06	0.46 ± 0.01	0.46 ± 0.05	0.20 ± 0.00	0.20 ± 0.01

Table 13: Population model stability and goodness-of-fit on synthetic data ($n = 2500$). Retrieval F_1 correlates with Integrated Brier Score (IBS) on held-out samples (lower IBS indicates better fit).

Model	γ	F_1	IBS-All	IBS-SG
RSF	0.05	0.79 ± 0.04	0.165	0.190
RSF	0.10	$\mathbf{0.87 \pm 0.02}$	0.165	0.190
RSF	0.15	0.70 ± 0.25	0.165	0.190
Gradient Boosting	0.05	0.78 ± 0.03	0.163	0.189
Gradient Boosting	0.10	0.85 ± 0.02	0.163	0.189
Gradient Boosting	0.15	0.85 ± 0.07	0.163	0.189
ExtraTrees	0.05	0.56 ± 0.09	0.165	0.200
ExtraTrees	0.10	0.70 ± 0.05	0.165	0.200
ExtraTrees	0.15	0.77 ± 0.05	0.165	0.200
Decision Tree	0.05	0.36 ± 0.04	0.263	0.305
Decision Tree	0.10	0.34 ± 0.02	0.263	0.305
Decision Tree	0.15	0.34 ± 0.02	0.263	0.305

Population Model Sensitivity. In Table 13, we evaluate SYSURV's sensitivity to the choice of the underlying population survival model over 10 random seeds using four regressors: Random Survival Forests (RSF), Gradient Boosting, ExtraTrees, and single Decision Trees. Retrieval performance does not depend on a specific architecture family, but directly tracks the population model's goodness-of-fit as measured by the Integrated Brier Score (IBS) on held-out data: models with lower IBS (better predictive calibration) yield consistently higher retrieval F_1 scores.

K.4 CASE STUDY

We present in Tables 14 and 15 examples of rules for the primary, resp., postoperative radiochemotherapy (RCT) cohorts from our case study. We also present the corresponding pruned rules, which describe subgroups whose Jaccard similarities in terms of the membership indicators to the original subgroups are no less than the threshold of 0.95. We also provide the resulting changes in subgroup size and exceptionality. For this, we use the post hoc pruning algorithm PRUNERULE (Alg. 3).

We can see that the rules are reduced to a combination of a handful of conditions after pruning. In the very first rule, we see a drastic reduction in rule size. This goes to show how the problem of collinearity is amplified when $n \approx p$, when in reality, rules having only a handful of conditions could suffice. Nevertheless, this post hoc pruning method cannot narrow down the predicates to those that, in the gene expression case, activate the rest of the genes in the same gene signature.

Table 14: Rules learned on the primary RCT cohort before and after pruning using a Jaccard similarity threshold of 0.95 for membership indicators along with the respective changes in subgroup sizes and exceptionality.

Original rule	Pruned rule	Changes
"$ACTN1 \in [-3.3, -0.9] \wedge AKT1 \in [-3.6, -1.9] \wedge ALDH3A1 \in [-8.9, -1.4] \wedge ANLN \in [-4.3, -2.2] \wedge BCL2L1 \in [-7.2, -5.3] \wedge BNIP3L \in [-3.0, -0.6] \wedge CAV1 \in [-4.4, -1.1] \wedge CDKN3 \in [-4.5, -2.5] \wedge CXCL12 \in [-8.2, -2.6] \wedge DCBLD1 \in [-6.9, -4.7] \wedge ERBB2 \in [-1.9, 0.7] \wedge ERBB3 \in [-7.0, -2.9] \wedge GNAI1 \in [-5.2, -0.0] \wedge IGF1R \in [-3.5, -0.9] \wedge LOX \in [-6.4, -3.3] \wedge MTOR \in [-3.2, -2.3] \wedge NOTCH1 \in [-4.9, -2.6] \wedge PLAU \in [-4.5, -1.2] \wedge RB1 \in [-3.8, -2.4] \wedge RELA \in [-3.2, -1.8] \wedge RFC4 \in [-4.6, -1.6] \wedge RPA2 \in [-4.0, -2.1] \wedge SNAI1 \in [-6.9, -3.3] \wedge TCF3 \in [-3.2, -1.3] \wedge TGFB1 \in [-2.6, -1.0] \wedge TPI1 \in [-7.2, -5.3] \wedge XPA \in [-4.7, -1.6] \wedge XRCC1 \in [-4.2, -2.6]$"	"$CDKN3 \in [-4.5, -2.5] \wedge ERBB2 \in [-1.9, 0.7] \wedge LOX \in [-6.4, -3.3] \wedge RB1 \in [-3.8, -2.4] \wedge RPA2 \in [-4.0, -2.1] \wedge TGFB1 \in [-2.6, -1.0]$"	Subgroup size: $6 \rightarrow 6$ Exceptionality: $19.1051 \rightarrow 19.1037$
"$AKT1 \in [-3.2, -1.2] \wedge ALDH1A1 \in [-8.1, -2.3] \wedge ASS1 \in [-4.4, -0.4] \wedge CENPK \in [-6.1, -4.4] \wedge CHEK2 \in [-5.5, -3.8] \wedge CXCR4 \in [-6.6, -2.8] \wedge CYP1B1 \in [-11.3, -4.5] \wedge ERBB3 \in [-5.5, -3.5] \wedge ERBB4 \in [-10.6, -7.6] \wedge FGFR3 \in [-7.1, -3.1] \wedge FOSL1 \in [-5.3, -0.7] \wedge GNAI1 \in [-3.2, -0.4] \wedge GPI \in [-2.0, 0.2] \wedge HIF1A \in [-6.8, -4.7] \wedge KDR \in [-5.5, -3.5] \wedge LDHA \in [-6.9, -4.3] \wedge LGALS1 \in [-2.6, 1.0] \wedge LIMD1 \in [-6.2, -4.7] \wedge LOX \in [-4.1, -0.9] \wedge MCM6 \in [-5.9, -3.7] \wedge MET \in [-5.3, -2.8] \wedge MMP7 \in [-9.2, -3.4] \wedge MYC \in [-2.8, 0.3] \wedge NBN \in [-3.7, -2.7] \wedge PDK1 \in [-4.9, -3.5] \wedge PFKFB3 \in [-3.9, -0.7] \wedge RASSF6 \in [-7.6, -5.0] \wedge RB1 \in [-4.4, -2.5] \wedge SYK \in [-6.0, -3.9] \wedge SYNGR3 \in [-8.9, -4.7] \wedge TPI1 \in [-7.2, -4.8] \wedge XRCC1 \in [-5.4, -3.4] \wedge CD44 \in [-0.1, 2.1]$"	"$ALDH1A1 \in [-8.1, -2.3] \wedge ASS1 \in [-4.4, -0.4] \wedge CYP1B1 \in [-11.3, -4.5] \wedge ERBB3 \in [-5.5, -3.5] \wedge ERBB4 \in [-10.6, -7.6] \wedge FOSL1 \in [-5.3, -0.7] \wedge GNAI1 \in [-3.2, -0.4] \wedge GPI \in [-2.0, 0.2] \wedge HIF1A \in [-6.8, -4.7] \wedge KDR \in [-5.5, -3.5] \wedge LDHA \in [-6.9, -4.3] \wedge LGALS1 \in [-2.6, 1.0] \wedge LIMD1 \in [-6.2, -4.7] \wedge LOX \in [-4.1, -0.9] \wedge MCM6 \in [-5.9, -3.7] \wedge MET \in [-5.3, -2.8] \wedge MMP7 \in [-9.2, -3.4] \wedge MYC \in [-2.8, 0.3] \wedge NBN \in [-3.7, -2.7] \wedge PDK1 \in [-4.9, -3.5] \wedge PFKFB3 \in [-3.9, -0.7] \wedge RASSF6 \in [-7.6, -5.0] \wedge RB1 \in [-4.4, -2.5] \wedge SYK \in [-6.0, -3.9] \wedge SYNGR3 \in [-8.9, -4.7] \wedge TPI1 \in [-7.2, -4.8] \wedge XRCC1 \in [-5.4, -3.4] \wedge CD44 \in [-0.1, 2.1]$"	Subgroup size: $12 \rightarrow 12$ Exceptionality: $16.6102 \rightarrow 16.6079$
"$ADM \in [-5.3, -0.8] \wedge AKT1 \in [-3.6, -1.7] \wedge ALDH3A1 \in [-8.9, -0.5] \wedge ANLN \in [-4.3, -2.8] \wedge ASS1 \in [-4.9, -0.0] \wedge ATP5G3 \in [-1.3, 1.4] \wedge ATR \in [-5.9, -3.8] \wedge BCL2L1 \in [-7.8, -4.9] \wedge BIRC5 \in [-4.7, -1.0] \wedge CAV1 \in [-4.4, 0.3] \wedge CDKN3 \in [-5.9, -2.5] \wedge CLDN4 \in [-5.0, -0.6] \wedge CXCL12 \in [-8.2, -3.2] \wedge ERBB3 \in [-6.3, -3.2] \wedge ERCC5 \in [-5.2, -3.6] \wedge FAM83B \in [-7.4, -1.7] \wedge FLT1 \in [-5.7, -3.1] \wedge GNAI1 \in [-5.2, -0.2] \wedge IGF1R \in [-3.5, -0.6] \wedge INHBA \in [-4.4, 1.9] \wedge KDR \in [-5.5, -2.8] \wedge LDHA \in [-7.5, -5.7] \wedge LGALS1 \in [-2.1, 1.0] \wedge LOX \in [-6.4, -1.4] \wedge MDM2 \in [-4.6, -2.2] \wedge MMP13 \in [-9.3, 1.3] \wedge MRE11A \in [-6.1, -3.6] \wedge MTOR \in [-3.7, -2.3] \wedge NOTCH1 \in [-5.9, -2.6] \wedge P4HA2 \in [-4.7, -1.4] \wedge PSMD9 \in [-4.2, -2.1] \wedge PTEN \in [-7.3, -5.6] \wedge RAD23B \in [-0.9, 0.8] \wedge RB1 \in [-3.8, -2.3] \wedge RPA2 \in [-4.9, -2.1] \wedge SERPINB2 \in [-6.6, -1.2] \wedge TCF3 \in [-3.5, -1.0] \wedge TGFB1 \in [-2.6, -0.3] \wedge TPI1 \in [-7.2, -5.0] \wedge XRCC1 \in [-5.3, -2.6] \wedge XRCC5 \in [-2.5, -0.8] \wedge CD44 \in [-0.8, 2.5]$"	"$ANLN \in [-4.3, -2.8] \wedge ATR \in [-5.9, -3.8] \wedge CLDN4 \in [-5.0, -0.6] \wedge CXCL12 \in [-8.2, -3.2] \wedge ERBB3 \in [-6.3, -3.2] \wedge ERCC5 \in [-5.2, -3.6] \wedge LDHA \in [-7.5, -5.7] \wedge LGALS1 \in [-2.1, 1.0] \wedge MDM2 \in [-4.6, -2.2] \wedge MTOR \in [-3.7, -2.3] \wedge PTEN \in [-7.3, -5.6] \wedge RAD23B \in [-0.9, 0.8] \wedge RB1 \in [-3.8, -2.3] \wedge SERPINB2 \in [-6.6, -1.2] \wedge TCF3 \in [-3.5, -1.0] \wedge TGFB1 \in [-2.6, -0.3] \wedge TPI1 \in [-7.2, -5.0]$"	Subgroup size: $28 \rightarrow 29$ Exceptionality: $16.0098 \rightarrow 15.7934$
"$AKT1 \in [-3.6, -1.5] \wedge ALDH3A1 \in [-8.9, -0.5] \wedge ANLN \in [-4.5, -2.6] \wedge ASS1 \in [-4.9, -0.0] \wedge ATP5G3 \in [-1.3, 1.7] \wedge ATR \in [-5.9, -3.7] \wedge BCL2L1 \in [-7.7, -5.0] \wedge BIRC5 \in [-5.3, -1.3] \wedge BNIP3L \in [-3.1, -0.6] \wedge CAV1 \in [-4.4, 0.4] \wedge CDKN3 \in [-5.8, -2.5] \wedge CLDN4 \in [-5.0, -0.7] \wedge CXCL12 \in [-8.2, -2.2] \wedge ERBB3 \in [-6.2, -3.2] \wedge ERBB4 \in [-10.6, -7.4] \wedge ERCC5 \in [-5.2, -3.6] \wedge FAM83B \in [-6.8, -1.6] \wedge GNAI1 \in [-5.2, -0.2] \wedge HIF1A \in [-7.7, -4.7] \wedge IGF1R \in [-3.5, -0.6] \wedge INHBA \in [-4.8, 1.8] \wedge KDR \in [-5.5, -3.6] \wedge LDHA \in [-7.5, -5.3] \wedge LGALS1 \in [-2.6, 1.0] \wedge LOX \in [-6.4, -1.7] \wedge MCM6 \in [-5.9, -3.0] \wedge MMP13 \in [-9.3, 1.3] \wedge MRGBP \in [-5.0, -2.8] \wedge MTOR \in [-3.8, -2.3] \wedge P4HA2 \in [-4.7, -1.4] \wedge PDK1 \in [-4.9, -2.3] \wedge PGAM1 \in [-2.6, -0.7] \wedge PSMD9 \in [-4.2, -2.0] \wedge PTEN \in [-7.3, -5.5] \wedge RAD23B \in [-0.8, 0.8] \wedge RB1 \in [-3.9, -2.3] \wedge TCF3 \in [-3.4, -1.4] \wedge TGFB1 \in [-2.6, -0.0] \wedge TP53 \in [-4.3, -1.0] \wedge TPI1 \in [-7.2, -4.9] \wedge CD44 \in [-0.4, 2.1]$"	"$ALDH3A1 \in [-8.9, -0.5] \wedge CLDN4 \in [-5.0, -0.7] \wedge ERBB3 \in [-6.2, -3.2] \wedge ERBB4 \in [-10.6, -7.4] \wedge HIF1A \in [-7.7, -4.7] \wedge LDHA \in [-7.5, -5.3] \wedge MMP13 \in [-9.3, 1.3] \wedge MRGBP \in [-5.0, -2.8] \wedge P4HA2 \in [-4.7, -1.4] \wedge PDK1 \in [-4.9, -2.3] \wedge PGAM1 \in [-2.6, -0.7] \wedge PTEN \in [-7.3, -5.5] \wedge RAD23B \in [-0.8, 0.8] \wedge TP53 \in [-4.3, -1.0] \wedge TPI1 \in [-7.2, -4.9]$"	Subgroup size: $25 \rightarrow 26$ Exceptionality: $15.8073 \rightarrow 15.2069$

Table 15: Rules learned on the postoperative RCT cohort before and after pruning using a Jaccard similarity threshold of 0.95 for membership indicators along with the respective changes in subgroup sizes and exceptionality.

Original rule	Pruned rule	Changes
"$ALDH1A1 \in [-7.0, -3.2] \wedge BCL2L1 \in [-6.9, -5.5] \wedge BIRC5 \in [-4.6, -1.6] \wedge BNIP3L \in [-2.5, -1.1] \wedge CA9 \in [-9.0, -4.4] \wedge CBX4 \in [-5.3, -3.7] \wedge CLDN4 \in [-5.3, -1.3] \wedge CXCL12 \in [-5.2, -1.1] \wedge EGLN3 \in [-4.4, -0.9] \wedge ENO1 \in [-0.2, 1.2] \wedge ENO2 \in [-5.8, -2.4] \wedge EPHA1 \in [-7.2, -4.6] \wedge GPI \in [-1.7, -0.2] \wedge HSPB1 \in [-4.0, -2.0] \wedge ITGB2 \in [-4.7, -2.3] \wedge KCTD11 \in [-2.3, 0.1] \wedge MAP2K2 \in [-4.3, -3.3] \wedge MCM6 \in [-5.8, -4.0] \wedge MMP13 \in [-1.9, 1.7] \wedge MMP9 \in [-2.7, 0.6] \wedge MPRS17 \in [-5.0, -1.2] \wedge PFKFB3 \in [-2.7, -0.2] \wedge PTEN \in [-6.6, -5.1] \wedge RAD23B \in [-1.0, 0.9] \wedge RB1 \in [-5.0, -2.5] \wedge RELA \in [-2.9, -1.6] \wedge SERPINB2 \in [-7.6, -2.6] \wedge SLC5A1 \in [-9.9, -3.6] \wedge TCF3 \in [-3.8, -2.3] \wedge XRCC1 \in [-5.2, -3.2] \wedge XRCC4 \in [-6.1, -4.2] \wedge XRCC5 \in [-2.1, -0.6]$"	"$EPHA1 \in [-7.2, -4.6] \wedge GPI \in [-1.7, -0.2] \wedge KCTD11 \in [-2.3, 0.1] \wedge MAP2K2 \in [-4.3, -3.3] \wedge MMP13 \in [-1.9, 1.7] \wedge MMP9 \in [-2.7, 0.6] \wedge XRCC5 \in [-2.1, -0.6]$"	Subgroup size: $12 \rightarrow 12$ Exceptionality: $27.8189 \rightarrow 27.7435$
"$ANXA5 \in [-3.0, -0.9] \wedge ATM \in [-3.8, -1.6] \wedge ATP5G3 \in [-1.1, 1.0] \wedge BNIP3 \in [-6.1, -2.6] \wedge BSG \in [-9.4, -5.4] \wedge CD24 \in [-8.8, -4.5] \wedge CXCL12 \in [-6.2, -1.3] \wedge DCBLD1 \in [-7.5, -4.5] \wedge DKK3 \in [-5.3, -1.4] \wedge FANCA \in [-6.7, -3.1] \wedge FN1 \in [-4.6, 1.8] \wedge HIF1A \in [-7.1, -4.6] \wedge HK2 \in [-6.7, -2.6] \wedge HSPA4 \in [-3.9, -2.2] \wedge INHBA \in [-7.9, -1.3] \wedge ITGB1 \in [-2.5, -0.1] \wedge KRT17 \in [-3.2, 2.8] \wedge LGALS1 \in [-3.4, -0.2] \wedge LOXL2 \in [-6.4, -3.0] \wedge MCM6 \in [-5.3, -2.5] \wedge MDM2 \in [-4.3, -0.7] \wedge MMP10 \in [-9.7, -1.8] \wedge MMP13 \in [-10.4, -2.8] \wedge MMP2 \in [-5.0, -0.5] \wedge MTOR \in [-3.9, -2.3] \wedge MYC \in [-3.7, -0.6] \wedge MYNN \in [-5.1, -3.0] \wedge NR1D2 \in [-4.9, -2.3] \wedge P4HA2 \in [-5.1, -2.4] \wedge PSMD9 \in [-4.3, -2.2] \wedge RMI2 \in [-5.1, -2.9] \wedge RPA2 \in [-4.5, -2.1] \wedge SFN \in [-1.6, 3.1] \wedge SLC3A2 \in [-4.3, -2.3] \wedge SMDT1 \in [-3.9, -2.2] \wedge SNAI1 \in [-7.6, -3.8] \wedge SPP1 \in [-9.0, -3.9] \wedge TCF3 \in [-4.0, -2.1] \wedge TP53 \in [-3.8, -1.0] \wedge XRCC4 \in [-5.2, -3.5] \wedge XRCC5 \in [-2.8, -0.9] \wedge YAP1 \in [-3.6, -0.2]$"	"$ATM \in [-3.8, -1.6] \wedge HIF1A \in [-7.1, -4.6] \wedge KRT17 \in [-3.2, 2.8] \wedge MMP13 \in [-10.4, -2.8] \wedge MMP2 \in [-5.0, -0.5] \wedge PSMD9 \in [-4.3, -2.2] \wedge RMI2 \in [-5.1, -2.9] \wedge SNAI1 \in [-7.6, -3.8] \wedge XRCC4 \in [-5.2, -3.5] \wedge YAP1 \in [-3.6, -0.2]$"	Subgroup size: $47 \rightarrow 45$ Exceptionality: $12.0529 \rightarrow 11.9264$
"$ALDH1A1 \in [-7.0, -3.4] \wedge BSG \in [-8.9, -3.9] \wedge CA9 \in [-9.5, -4.5] \wedge CBX4 \in [-5.3, -3.7] \wedge CLDN4 \in [-4.9, -0.4] \wedge CXCL12 \in [-4.3, -1.5] \wedge ENO1 \in [-0.3, 1.2] \wedge ENO2 \in [-5.8, -2.4] \wedge EPHA1 \in [-7.2, -4.6] \wedge EPOR \in [-10.2, -6.6] \wedge ERCC5 \in [-5.2, -3.0] \wedge HIF1A \in [-5.7, -3.7] \wedge LOX \in [-2.9, -1.0] \wedge MMP9 \in [-2.8, 0.6] \wedge MYNN \in [-5.2, -2.8] \wedge PGK1 \in [-2.0, -0.2] \wedge PTEN \in [-6.5, -5.1] \wedge RAD23B \in [-1.0, 1.1] \wedge RAD50 \in [-3.6, -1.6] \wedge RB1 \in [-5.0, -2.4] \wedge SLC5A1 \in [-10.0, -3.3] \wedge TCF3 \in [-4.0, -2.1] \wedge XRCC1 \in [-5.2, -4.1] \wedge XRCC4 \in [-6.2, -4.1] \wedge XRCC5 \in [-2.2, -1.0]$"	"$ALDH1A1 \in [-7.0, -3.4] \wedge CA9 \in [-9.5, -4.5] \wedge CBX4 \in [-5.3, -3.7] \wedge ENO1 \in [-0.3, 1.2] \wedge EPHA1 \in [-7.2, -4.6] \wedge LOX \in [-2.9, -1.0] \wedge MMP9 \in [-2.8, 0.6] \wedge MYNN \in [-5.2, -2.8] \wedge SLC5A1 \in [-10.0, -3.3] \wedge TCF3 \in [-4.0, -2.1] \wedge XRCC1 \in [-5.2, -4.1]$"	Subgroup size: $12 \rightarrow 12$ Exceptionality: $27.4009 \rightarrow 26.0718$
"$ATM \in [-4.7, -1.6] \wedge BIRC5 \in [-4.2, -1.0] \wedge BNIP3L \in [-3.2, -1.2] \wedge BSG \in [-9.2, -3.9] \wedge CA9 \in [-10.7, -2.9] \wedge CBX4 \in [-5.3, -3.1] \wedge CD24 \in [-8.3, -4.5] \wedge CDKN3 \in [-6.1, -2.7] \wedge CXCL12 \in [-7.5, -2.0] \wedge DCBLD1 \in [-7.5, -4.5] \wedge ENO2 \in [-6.2, -2.4] \wedge ERCC5 \in [-5.7, -3.0] \wedge HK2 \in [-6.7, -2.2] \wedge HSPA4 \in [-4.0, -2.2] \wedge KDR \in [-6.0, -2.4] \wedge KRT17 \in [-3.0, 5.1] \wedge LOX \in [-7.1, -1.5] \wedge LOXL2 \in [-6.4, -1.8] \wedge MCM6 \in [-4.8, -2.5] \wedge MME \in [-9.6, -4.0] \wedge MRE11A \in [-6.1, -4.0] \wedge MTOR \in [-4.2, -2.3] \wedge MYC \in [-3.7, -0.5] \wedge MYNN \in [-5.0, -2.3] \wedge NBN \in [-3.9, -1.6] \wedge NOTCH1 \in [-6.3, -1.9] \wedge RAD23B \in [-1.4, 1.1] \wedge RAD50 \in [-4.0, -1.6] \wedge RASSF6 \in [-8.1, -1.6] \wedge RB1 \in [-5.0, -2.3] \wedge RELA \in [-3.5, -1.4] \wedge SDHA \in [-6.2, -2.9] \wedge SERPINB2 \in [-8.0, -0.6] \wedge TCF3 \in [-3.9, -2.1] \wedge TP53 \in [-5.0, -1.0] \wedge XPA \in [-5.3, -2.6] \wedge XPC \in [-5.7, -3.7] \wedge XRCC1 \in [-5.2, -2.5] \wedge XRCC4 \in [-5.8, -3.5] \wedge XRCC5 \in [-2.8, -0.8]$"	"$BIRC5 \in [-4.2, -1.0] \wedge BSG \in [-9.2, -3.9] \wedge CA9 \in [-10.7, -2.9] \wedge CD24 \in [-8.3, -4.5] \wedge DCBLD1 \in [-7.5, -4.5] \wedge LOX \in [-7.1, -1.5] \wedge LOXL2 \in [-6.4, -1.8] \wedge MCM6 \in [-4.8, -2.5] \wedge MME \in [-9.6, -4.0] \wedge MTOR \in [-4.2, -2.3] \wedge MYC \in [-3.7, -0.5] \wedge MYNN \in [-5.0, -2.3] \wedge RB1 \in [-5.0, -2.3] \wedge TCF3 \in [-3.9, -2.1] \wedge XPC \in [-5.7, -3.7] \wedge XRCC1 \in [-5.2, -2.5]$"	Subgroup size: $85 \rightarrow 87$ Exceptionality: $11.2782 \rightarrow 11.1260$