# Structured Treatment Modeling in Deep Survival Analysis via Hazard Factorization

**Natalia Hong** NATALIA.HONG@KEBLE.OX.AC.UK
*University of Oxford, United Kingdom*

**Krishnarajah Nirantharakumar** KRISHNARAJAH.NIRANTHARAKUMAR@KCL.AC.UK
*King's College London, United Kingdom*

**Christopher Yau** CHRISTOPHER.YAU@WRH.OX.AC.UK
*University of Oxford, United Kingdom; Health Data Research UK*

## Abstract

Deep learning models trained on electronic health records are increasingly used for clinical risk prediction, yet modeling heterogeneous treatment effects remains challenging. Most approaches treat treatment as an undifferentiated covariate (S-Learner), conflating treatment effects with baseline risk, while training separate models for treated and untreated patients (T-Learner) suffers from treatment imbalance and sparsity. We propose a structured hazard factorization that decomposes the hazard into a shared baseline component and a treatment-specific hazard ratio network, enabling direct estimation of time-varying, covariate-dependent hazard ratios without post-hoc computation. By sharing a baseline while isolating treatment effects, the framework acts as a hybrid between S- and T-Learners, improving efficiency and reducing majority-group dominance under imbalance. We further extend the model with differentiable subgroup assignment for regularized treatment effect estimation and inverse propensity weighting to adjust for confounding.

In simulations with known ground truth, our approach improves hazard ratio recovery while maintaining competitive survival prediction, and the subgroup extension recovers latent heterogeneity when assumptions hold. On two real-world clinical cohorts from the UK Clinical Practice Research Datalink, the framework produces time-varying hazard ratios and identifies subgroups characterized by established risk factors. Our results demonstrate that explicit hazard factorization provides useful inductive bias for incorporating treatment into deep survival models, bridging flexible neural architectures with hazard ratio estimation familiar to clinical practice.

**Keywords:** Deep Survival Models; Treatment Effect Heterogeneity; Hazard Ratio Estimation

**Data and Code Availability** All simulation experiments are reproducible via the accompanying code available on GitHub.[1] Analyses involving real-world clinical data were conducted using the UK Clinical Practice Research Datalink (CPRD), a primary care database subject to formal data governance requirements. Access to CPRD data is available via application through the CPRD website (www.cprd.com) to the Medicines and Healthcare products Regulatory Agency and requires relevant approvals.

**Institutional Review Board (IRB)** This study is based on synthetic simulation data and secondary analysis of de-identified primary care electronic health records from CPRD, which has been approved by the Independent Scientific Advisory Committee (ISAC) for Medicines and Healthcare products Regulatory Agency (MHRA) research with Study Reference ID 22_001903. Use of CPRD data for research purposes is covered by CPRD's ethics approval from the UK Health Research Authority.

## 1. Introduction

The widespread adoption of electronic health records (EHRs) has enabled development of clinical prediction models to support prognosis and treatment decisions (Goldstein et al., 2017). In most modeling approaches, treatment indicators, such as medication or

1. https://github.com/natlhong/HazardDeSurv

procedures, are included as static covariates, treated identically to demographics or laboratory measurements. This approach, analogous to the causal *S-Learner* (Künzel et al., 2019), employs a single model incorporating treatment as an input feature.

In clinical practice, treatment effects are rarely uniform. Patients with different characteristics respond differently to the same intervention, a phenomenon termed treatment effect heterogeneity (HTE) (Kravitz et al., 2004; Kent et al., 2018). When a single global model is trained, patterns from the majority group tend to dominate, while subgroup performance declines even if overall performance appears competitive. Training separate models for treated and untreated patients (*T-Learner*) can better capture HTE but becomes inefficient under treatment imbalance and sparse exposure. Other approaches such as the *X-Learner* and *R-Learner* (Künzel et al., 2019) aim to improve HTE estimation, with the X-Learner mitigating treatment imbalance by weighting pseudo-effect models toward the larger group, and the R-Learner leveraging residualization with propensity scores.

Beyond heterogeneity, treatment effects may vary *over time*. Classical Cox proportional hazards models (Cox, 1972) assume constant hazard ratios, an assumption frequently violated in practice (Stensrud and Hernán, 2025). Methods extending the Cox framework (e.g., DeepSurv (Katzman et al., 2018)) inherit this limitation, while most deep survival models do not natively produce hazard ratios, with the causal inference literature instead targeting alternative estimands such as differences in survival probabilities (Jeanselme et al., 2025; Curth et al., 2021; Hu et al., 2021) or restricted mean survival time (RMST) (Royston and Parmar, 2013; Nagpal et al., 2022).

We propose a structured hazard factorization framework that separates baseline risk from binary treatment effects in deep survival models. By parameterizing treatment through a dedicated hazard ratio network, the framework directly produces time-varying, covariate-dependent hazard ratios, avoiding unstable post-hoc contrasts. Hazard ratios are the standard estimand in classical clinical research, yet remain largely unavailable in existing deep survival approaches. Our framework therefore establishes a principled connection between flexible neural architectures and classical hazard ratio analysis. The multiplicative structure encodes the assumption that treatment modulates baseline risk rather than inducing separate dynamics, providing regularization under treatment imbalance while retaining parameter efficiency through a shared baseline hazard.

## 2. Methodology

### 2.1. Background

Our framework lies in the field of survival analysis, which studies time-to-event outcomes under right censoring, assuming non-informative censoring (independence of censoring and event time given covariates). Let $\mathbf{x} \in \mathbb{R}^p$ denote covariates and $a \in \{0, 1\}$ the binary treatment indicator. The observed outcome is $(T, \delta)$, where $T$ is the observed time and $\delta \in \{0, 1\}$ indicates whether an event occurred ($\delta = 1$) or the observation was censored ($\delta = 0$).

In prediction tasks, we are interested in the conditional event distribution given covariates,

$$\begin{aligned} F(t \mid \mathbf{x}, a) &= \Pr(T \leq t \mid \mathbf{x}, a) \\ &= 1 - \exp\left( - \int_0^t \lambda(s \mid \mathbf{x}, a)\, ds \right), \end{aligned}$$

where $\lambda(t \mid \mathbf{x}, a)$ denotes the hazard function, i.e., the instantaneous risk of an event at time $t$ given survival up to $t$ (derivation provided in Appendix A). This representation underpins our proposed methodology.

**Cox Proportional Hazards (PH).** The Cox model parameterizes the hazard function rather than the event time distribution directly, assuming a log-linear effect of covariates on the hazard:

$$\lambda(t \mid \mathbf{x}, a) = \lambda_0(t) \exp\big([\mathbf{x}, a]^\top \boldsymbol{\beta}\big),$$

where $\lambda_0(t)$ is an unspecified baseline hazard and $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{x}}, \beta_\alpha)$ denotes covariate coefficients. Under this model, the hazard ratio for treatment is $\exp(\beta_a)$, which is a constant independent of time.

**Cumulative Density Function (CDF)-Based Deep Survival Models.** Flexible neural approaches such as DeSurv (Danks and Yau, 2022), SuMo-Net (Rindt et al., 2022), and NeuralFineGray (Jeanselme et al., 2023) directly estimate the conditional cumulative distribution function:

$$F(t \mid \mathbf{x}, a) = \psi(u_\gamma(t, \mathbf{x}, a)),$$

where $u_\gamma$ is a neural network with monotonicity constraints in $t$, and $\psi$ is a link function mapping $[0, \infty)$ to $[0, 1]$. By modeling the CDF directly, these approaches provide flexible estimation without proportional hazards assumptions. However, they do not define a hazard function and therefore cannot produce hazard ratios directly.

### 2.2. Modified DeSurv Formulation

We build on the deep survival method, DeSurv, which models survival through a monotonic function:

$$u(t, \mathbf{x}, a) = \int_0^t g_\phi(s, \mathbf{x}, a)\, ds,$$

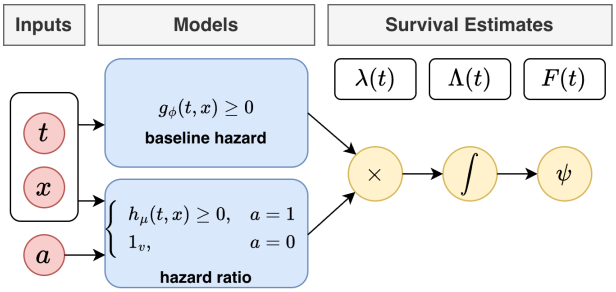
where $g_\phi > 0$ is enforced via a SoftPlus activation, and $u(t, \mathbf{x}, a)$ is integrated numerically. The original DeSurv uses the mapping $F(t \mid \mathbf{x}, a) = \tanh(u(t, \mathbf{x}, a))$, which has two issues: (1) unbounded loss when $g_\phi \to \infty$ for events (Appendix B), and (2) $u(t, \mathbf{x}, a)$ lacks standard survival interpretation.

We instead adopt the following mapping:

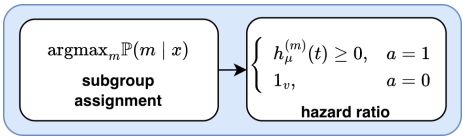
$$F(t \mid \mathbf{x}, a) = 1 - \exp\big(-u(t, \mathbf{x}, a)\big).$$

Under this mapping, $u(t, \mathbf{x}, a)$ corresponds to the cumulative hazard $\Lambda(t) = \int_0^t \lambda(s)\, ds$ (Kalbfleisch and Prentice, 2002). This connection to standard survival quantities enables our subsequent hazard factorization, while ensuring bounded loss during training.

### 2.3. Hazard Factorization




(a) General factorization framework.




(b) Subgroup extension with discrete assignment.

Figure 1: Proposed framework: the hazard $\lambda(t \mid \mathbf{x})$ factors into baseline $g_\phi(t, \mathbf{x})$ and treatment effect $h_\mu(t, \mathbf{x})$, whose product integrates to the cumulative hazard $\Lambda(t)$. The link $\psi(x) = 1 - \exp(-x)$ yields survival probabilities.

Consider the S-Learner approach: a single neural network $g_\phi(t, \mathbf{x}, a)$ models the hazard. While flexible, this has two drawbacks. First, $\text{HR}(t, \mathbf{x})$ must be computed post-hoc as a ratio of two function evaluations, amplifying uncertainty through error propagation. Second, under treatment imbalance, gradients predominantly flow from the majority class, potentially degrading minority performance.

We propose a **structured factorization** that explicitly separates baseline and treatment effects. For binary treatment $a \in \{0, 1\}$, we factorize:

$$\lambda(t \mid \mathbf{x}, a) = g_\phi(t, \mathbf{x}) \cdot h_\mu(t, \mathbf{x})^a = \begin{cases} g_\phi(t, \mathbf{x}) \cdot h_\mu(t, \mathbf{x}), & a = 1 \\ g_\phi(t, \mathbf{x}), & a = 0 \end{cases},$$

where $g_\phi$ is the baseline hazard network, and $h_\mu$ is the treatment effect network (Figure 1). Both are parameterized as multilayer perceptrons with SoftPlus output to ensure positivity. By construction, the treatment effect is given directly by the $h_\mu$ network:

$$\text{HR}(t, \mathbf{x}) = \frac{\lambda(t \mid \mathbf{x}, a{=}1)}{\lambda(t \mid \mathbf{x}, a{=}0)} = h_\mu(t, \mathbf{x}),$$

and the cumulative hazard under treatment becomes:

$$u(t, \mathbf{x}, a{=}1) = \int_0^t g_\phi(s, \mathbf{x}) \cdot h_\mu(s, \mathbf{x})\, ds.$$

This factorization leaves the original learning objective unchanged. Model parameters are estimated by minimizing the negative survival log-likelihood,

$$\mathcal{L} = -\sum_{i=1}^N \Big[1_{\delta_i=1} \log(F'(t_i)) + 1_{\delta_i=0} \log(1 - F(t_i))\Big]$$
$$= -\sum_{i=1}^N \Big[1_{\delta_i=1} \log(\lambda(t_i) S(t_i)) + 1_{\delta_i=0} \log(S(t_i))\Big]$$

where $S(t) = 1 - F(t)$ is the survival function and $F'(t) = f(t)$ is the corresponding density function.

**Relationship to Meta-Learners.** Our hazard factorization can be viewed as a hybrid between S- and T-Learners. Like S-Learners, the model shares a baseline hazard $g_\phi$ across treatment arms, allowing efficient use of all available data. Like T-Learners, it isolates treatment-specific components, preventing treatment effects from being absorbed into baseline risk estimation. The multiplicative structure imposes a regularizing constraint. Rather than learning arbitrary, arm-specific hazard functions, the model learns how treatment *modulates* a shared baseline. This inductive bias is particularly valuable under treatment imbalance, where estimating fully separate dynamics is statistically unstable.

### 2.4. Latent Subgroup Assignment

To further regularize treatment effect estimation and reduce variance under limited treatment data, we introduce a latent subgroup variable $Z \in \{1, \dots, M\}$ that induces parameter sharing in the hazard ratio network across individuals with similar treatment-response dynamics. Subgroup membership probabilities are modeled as

$$\pi_m(\mathbf{x}) := \mathbb{P}(Z = m \mid \mathbf{x}) = \frac{\exp(\zeta_m^\top \eta(\mathbf{x}))}{\sum_{j=1}^{M} \exp(\zeta_j^\top \eta(\mathbf{x}))},$$

where $\eta(\mathbf{x})$ is a learned feature representation of the covariates and $\{\zeta_m\}_{m=1}^{M}$ are subgroup-specific parameter vectors. Conditioned on subgroup assignment, the treatment effect is parameterized as a subgroup-specific, time-varying hazard ratio (HR)

$$h_\mu(t \mid Z{=}m) = h_\mu^{(m)}(t),$$

where $h_\mu^{(m)}$ depends only on time and is shared across individuals within subgroup $m$. This restriction reduces the number of treatment-effect parameters, lowering variance and improving stability. Covariates influence the treatment effect only through subgroup assignment, while the subgroup-specific functions $\{h_\mu^{(m)}\}$ define the temporal response profiles.

To enable end-to-end training through discrete subgroup assignments, we apply the Gumbel–Softmax relaxation (Jang et al., 2017). Let

$$G_m \sim \text{Gumbel}(0, 1),$$
$$\tilde{z}_m = \frac{\exp((\log \pi_m + G_m)/\tau)}{\sum_{j=1}^{M} \exp((\log \pi_j + G_j)/\tau)},$$

where $\tau > 0$ is a temperature parameter. As $\tau \to 0$, the relaxed assignment $\tilde{\mathbf{z}}$ approaches a one-hot vector, recovering discrete subgroup membership. In our experiments, we anneal $\tau$ during training (from 1.0 to 0.2), and at inference we use hard assignment $\hat{m} = \arg\max_m \pi_m(\mathbf{x})$.

### 2.5. Low-Rank Factorization for Individualized Effects

The subgroup-based formulation can be relaxed to allow individualized hazard ratios by replacing discrete subgroup assignments with a low-rank factorization of the hazard ratio. Specifically, we parameterize

$$h_\mu(t, \mathbf{x}) = \exp\Big(\sum_{k=1}^{K} \alpha_k(t)\, \beta_k(\mathbf{x})\Big),$$

where $\alpha : \mathbb{R}_+ \to \mathbb{R}^K$ captures time-dependent structure and $\beta : \mathbb{R}^p \to \mathbb{R}^K$ captures covariate-dependent structure. This formulation can be viewed as a continuous analogue of the subgroup model, replacing discrete assignment with a low-dimensional latent representation. The rank $K$ governs the expressiveness–stability trade-off. Small $K$ enforces strong regularization and shared structure, while larger $K$ permits more flexible individualized effects.

### 2.6. Propensity Score Weighting

In EHRs, treatment assignment is observational and confounded by covariates. We therefore consider a propensity-weighted variant of our subgroup model that incorporates *stabilized* inverse probability weighting (IPW) (Rosenbaum and Rubin, 1983; Robins et al., 2000). The procedure follows a two-step design in which propensity scores are estimated independently and then fixed during survival training, preventing downstream survival gradients from corrupting the propensity signal.

**Stage 1: Propensity Estimation.** We first estimate the propensity score $e(\mathbf{x}) = \mathbb{P}(A = 1 \mid \mathbf{x})$ using a separate multilayer perceptron trained with binary cross-entropy loss. Given the fitted propensities $\hat{e}(\mathbf{x})$, we construct stabilized inverse probability weights

$$w_i = \begin{cases} \dfrac{\bar{a}}{\hat{e}(\mathbf{x}_i)}, & a_i = 1, \\ \dfrac{1 - \bar{a}}{1 - \hat{e}(\mathbf{x}_i)}, & a_i = 0, \end{cases}$$

where $\bar{a} = N^{-1} \sum_i a_i$ is the empirical treatment prevalence. To improve numerical stability, propensity scores are clipped to $[0.01, 0.99]$ before weight computation (Austin and Stuart, 2015).

**Stage 2: Weighted Survival Learning.** The propensity model and weights are then held fixed, and the survival model is trained using a weighted negative log-likelihood,

$$\mathcal{L} = \sum_{i=1}^{N} w_i\, \ell_i(\theta),$$

where $\ell_i(\theta)$ denotes the individual negative survival log-likelihood contribution. This reweighting creates a pseudo-population in which treatment assignment is independent of measured confounders.

## 3. Experiments

### 3.1. Simulation Design

We conduct proof-of-concept simulations with known ground truth to evaluate each method's ability to recover time-varying hazard ratios under controlled violations of standard assumptions. Each survival dataset consists of $N = 20{,}000$ subjects with 10 observed covariates, generated under a Weibull proportional hazards model (Bender et al., 2005). Treatment assignment is *non-random* and depends on observed covariates through a non-linear logistic model, inducing confounding and treatment imbalance (approximately 40% treated). Censoring follows an independent exponential distribution with rate 0.05 and a maximum follow-up of 20 years, yielding approximately 40% observed events. An overview of the simulation settings is provided in Table 1, with full details given in Appendix C.

To study HTE recovery, we partition the population into two subgroups. Subgroup 1 (40%) has no treatment effect (HR = 1), while subgroup 0 (60%) experiences a beneficial treatment effect. Subgroup membership is determined by a latent marker. In **Scenario A**, this marker is directly observed and hazard ratios are constant over time, corresponding to a proportional hazards setting. In **Scenario B**, subgroup membership is unobserved and only indirectly reflected through three noisy proxy covariates $v_1, v_2, v_3$ (with moderate correlation 0.6–0.8 to the latent marker), while hazard ratios vary over time, violating the proportional hazards assumption.

**Scenario C** focuses on robustness rather than HTE recovery. In this setting, treated and untreated outcomes are generated from different distributional families (Weibull vs. Gompertz), such that no true multiplicative hazard ratio function exists. This scenario stress-tests the behavior of the proposed factorization when its structural assumptions are violated.

Table 1: Simulation scenarios.

| Scenarios | Hazard Ratios |
|---|---|
| A: Heterogeneous & Constant | Subgroup = 1: $\mathrm{HR}(t) = 1$<br>Subgroup = 0: $\mathrm{HR}(t) = \exp(-1)$ |
| B: Heterogeneous & Time-Varying | Subgroup = 1: $\mathrm{HR}(t) = 1$<br>Subgroup = 0: $\mathrm{HR}(t) = \exp(-0.075\,t)$ |
| C: Mis-specified (Non-Multiplicative) | Treated Hazard: Weibull<br>Untreated Hazard: Gompertz |

### 3.2. Real-World Primary Care Cohorts

We apply our method to two cohorts derived from the UK Clinical Practice Research Datalink (CPRD) Aurum database (Herrett et al., 2015), which contains anonymized longitudinal primary care records from general practices across England. These cohorts were selected to evaluate our framework in realistic observational settings where treatment assignment is confounded by indication and treatment effect heterogeneity is clinically plausible.

**RAAS-Inhibitor Cohort.** RAAS inhibitors (ACE and angiotensin receptor blockers) have established renoprotective effects in diabetic patients (Lewis et al., 2001), though benefits in broader hypertensive populations without pre-existing kidney disease remain less certain (Jafar et al., 2003). We study the association between RAAS inhibitor use and subsequent stage 3–5 CKD incidence among hypertensive patients aged 50–70 in London who remain CKD-free at 90 days following diagnosis ($N = 103{,}451$; 47.7% treated; 5.3% CKD incidence), with time zero defined at this landmark to allow for treatment initiation. Patients are followed for up to 10 years.

**Anticoagulant Cohort.** Anticoagulant therapy reduces stroke risk in patients with atrial fibrillation, with evidence suggesting broader mortality benefits (Hart et al., 2007). We study the association between anticoagulant use (warfarin and direct oral anticoagulants) and all-cause mortality among patients aged 60–80 diagnosed with atrial fibrillation in London ($N = 25{,}563$; 39.4% treated; 18.4% mortality), with time zero defined at 30 days post-diagnosis to allow for treatment initiation. Patients are followed for up to 5 years.

Both cohorts include covariates capturing demographics, comorbidities, and laboratory measurements (BMI, blood pressure, HbA1c, lipid profiles, and renal function markers).

### 3.3. Baselines and Implementation

We benchmark against several DeSurv-based baselines: (i) **Ignore Treatment**, which models the baseline hazard without including treatment as an input; (ii) **S-Learner**, which incorporates treatment as an additional covariate; and (iii) **T-Learner**, which trains separate models for treated and untreated individuals. We also compare to **DeepSurv** (Katzman et al., 2018), a neural-network extension of the Cox PH model. DeepSurv also functions as an S-Learner, incorporating treatment as a standard covariate.

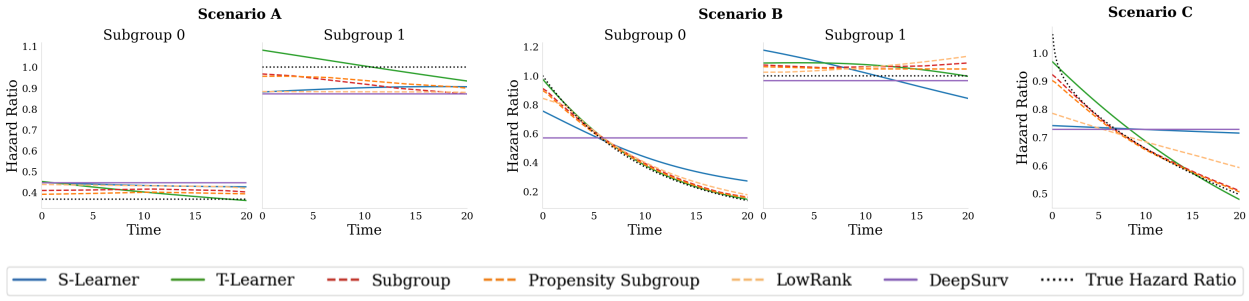



Figure 2: Average estimated hazard ratios over time. S- and T-Learner hazard ratios are obtained post-hoc from predictions under treatment $a = 0$ and $a = 1$. Solid lines correspond to benchmark methods, dashed lines to proposed variants, and the black dotted line denotes the ground-truth hazard ratio.

All models are trained with the Adam optimizer (Kingma and Ba, 2014) (learning rate $10^{-4}$) for up to 500 epochs with early stopping (patience 50). DeepSurv and DeSurv use hidden layers $[32, 32]$. In the factorized models, the subgroup network uses $[16, 8]$ with $M = 2$ subgroups; the low-rank model uses $[8, 8]$ (time) and $[8]$ (covariates) with rank $K = 2$; and the propensity network uses $[32, 16]$. Batch sizes are 256 for simulations and the Anticoagulant cohort, and 512 for the RAAS-Inhibitor cohort. Full details are provided in Appendix D.

### 3.4. Evaluation Metrics

**Survival Prediction.** We report standard survival prediction metrics, including time-dependent concordance ($\mathbf{C^{td}}$), integrated Brier score (**IBS**), integrated negative binomial log-likelihood (**INBLL**), and the negative survival log-likelihood (–**SLL**). Formal definitions are provided in Appendix E. For simulation studies with known ground truth, we additionally compute the **Integrated Squared Error (ISE)** between the predicted and true survival curves,

$$\text{ISE} = \frac{1}{N}\sum_{i=1}^{N}\int_0^{t_{\max}} \left(F(t \mid \mathbf{x}_i) - \hat{F}(t \mid \mathbf{x}_i)\right)^2, dt,$$

which directly measures deviation from the true survival distribution over time. In our analysis, we emphasize –SLL and ISE, as they constitute proper scoring rules (Rindt et al., 2022). DeepSurv does not define a full survival likelihood and cannot be evaluated using –SLL.

**Hazard Ratio Recovery.** In simulations where the data generating process is known, we evaluate recovery of the hazard ratio using the **Integrated Huber Error (HR-IHE)**:

$$\text{HR-IHE} = \frac{1}{N}\sum_{i=1}^{N}\int_0^{t_{\max}} L_\delta\Big(\text{HR}(t, \mathbf{x}_i) - \widehat{\text{HR}}(t, \mathbf{x}_i)\Big)\, dt,$$

where $L_\delta$ is the Huber loss:

$$L_\delta(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \delta \\ \delta\left(|x| - \frac{1}{2}\delta\right), & |x| > \delta \end{cases}.$$

We set $\delta = 0.2$. The Huber loss adds robustness to large deviations at isolated time points, which is important because hazard ratios are unbounded.

## 4. Results

### 4.1. Simulation Results

Figure 2 and Table 2 summarizes results across all simulation scenarios. In settings with HTE (Scenarios A and B), the proposed subgroup models consistently achieve a substantially lower hazard ratio error than baseline methods, with HR-IHE reductions of approximately 3–5× relative to S- and T-Learners. The gains are most pronounced in Scenario B, where there are time-varying treatment effects. Notably, the improvements in hazard ratio recovery do not compromise survival prediction, with the proposed models achieving competitive or superior survival metrics. In Scenario B, the Propensity Subgroup model attains the lowest ISE while simultaneously achieving the best HR-IHE, suggesting that the multiplicative inductive bias can improve both survival estimation and treatment effect recovery when the assumed structure approximately holds. The robustness of the approach is studied in Scenario C, where the multiplicative hazard assumption is deliberately violated

Table 2: Simulation results (mean ± std across 5 folds). Best results **bolded**.

| Scenario | | Model | $C^{td}$ ↑ | IBS ↓ | INBLL ↓ | -SLL ↓ | ISE ↓ | HR-IHE ↓ |
|---|---|---|---|---|---|---|---|---|
| A | | DeepSurv | **0.723 (0.007)** | **0.141 (0.002)** | **0.436 (0.005)** | – | **0.021 (0.001)** | 0.218 (0.014) |
| | DeSurv | Ignore Treatment | 0.704 (0.004) | 0.146 (0.001) | 0.449 (0.003) | 0.366 (0.006) | 0.118 (0.003) | 1.266 (0.023) |
| | DeSurv | S-Learner | 0.721 (0.006) | **0.141 (0.002)** | 0.437 (0.005) | 0.351 (0.004) | 0.028 (0.002) | 0.170 (0.032) |
| | DeSurv | T-Learner | 0.722 (0.007) | **0.141 (0.002)** | 0.437 (0.005) | **0.350 (0.004)** | 0.025 (0.002) | 0.245 (0.039) |
| | Proposed | Subgroup | 0.722 (0.006) | **0.141 (0.002)** | **0.436 (0.004)** | **0.350 (0.005)** | 0.023 (0.001) | **0.052 (0.025)** |
| | Proposed | Propensity Subgroup | 0.722 (0.006) | **0.141 (0.002)** | **0.436 (0.004)** | **0.350 (0.005)** | 0.022 (0.002) | **0.052 (0.024)** |
| | Proposed | LowRank | 0.722 (0.007) | **0.141 (0.002)** | 0.437 (0.005) | **0.350 (0.005)** | 0.026 (0.001) | 0.249 (0.027) |
| B | | DeepSurv | 0.708 (0.008) | 0.159 (0.003) | 0.481 (0.009) | – | 0.025 (0.003) | 0.482 (0.050) |
| | DeSurv | Ignore Treatment | 0.706 (0.006) | 0.161 (0.003) | 0.485 (0.008) | 0.321 (0.005) | 0.067 (0.005) | 1.121 (0.020) |
| | DeSurv | S-Learner | 0.708 (0.007) | 0.159 (0.003) | 0.482 (0.008) | 0.314 (0.006) | 0.031 (0.005) | 0.335 (0.075) |
| | DeSurv | T-Learner | 0.709 (0.006) | 0.159 (0.003) | 0.482 (0.007) | 0.312 (0.007) | 0.028 (0.002) | 0.312 (0.025) |
| | Proposed | Subgroup | **0.711 (0.006)** | 0.159 (0.003) | **0.480 (0.008)** | **0.309 (0.006)** | 0.020 (0.001) | **0.047 (0.009)** |
| | Proposed | Propensity Subgroup | **0.711 (0.006)** | **0.158 (0.003)** | **0.480 (0.008)** | **0.309 (0.006)** | **0.018 (0.001)** | **0.047 (0.012)** |
| | Proposed | LowRank | 0.710 (0.006) | 0.159 (0.003) | 0.481 (0.008) | 0.311 (0.007) | 0.026 (0.002) | 0.237 (0.030) |
| C | | DeepSurv | **0.712 (0.007)** | **0.150 (0.003)** | **0.459 (0.007)** | – | **0.012 (0.001)** | 0.220 (0.019) |
| | DeSurv | Ignore Treatment | 0.708 (0.006) | 0.151 (0.003) | 0.462 (0.007) | 0.356 (0.008) | 0.036 (0.001) | 0.898 (0.000) |
| | DeSurv | S-Learner | 0.711 (0.007) | **0.150 (0.003)** | 0.460 (0.007) | **0.352 (0.007)** | 0.015 (0.001) | 0.226 (0.020) |
| | DeSurv | T-Learner | 0.710 (0.006) | **0.150 (0.003)** | 0.460 (0.007) | **0.352 (0.007)** | 0.023 (0.002) | 0.260 (0.054) |
| | Proposed | Subgroup | 0.710 (0.006) | **0.150 (0.003)** | 0.460 (0.007) | **0.352 (0.007)** | 0.019 (0.002) | **0.040 (0.021)** |
| | Proposed | Propensity Subgroup | 0.710 (0.006) | **0.150 (0.003)** | 0.460 (0.007) | **0.352 (0.007)** | 0.020 (0.001) | 0.045 (0.021) |
| | Proposed | LowRank | 0.710 (0.006) | **0.150 (0.003)** | 0.460 (0.007) | **0.352 (0.007)** | 0.019 (0.001) | 0.198 (0.058) |

by generating treated and untreated outcomes from different distributional families. In this setting, the lowest predictive performance was achieved by DeepSurv, while our subgroup approach yields the lowest HR-IHE. The LowRank model shows intermediate performance. It improves over S- and T-Learners but does not match the Subgroup models for HR recovery. This suggests that when discrete subgroup structure is present, explicit subgroup assignment provides a more regularized and stable representation than continuous low-rank factorization.

Figure 2 provides a complementary qualitative view of hazard ratio estimation over time. In HTE scenarios, the S-Learner approaches (DeSurv: blue solid, DeepSurv: purple solid) produce hazard ratio trajectories that drift toward a population-level average, reflecting the behavior of a single global model optimized under heterogeneous effects. The T-Learner (green solid) mitigates this by fitting separate models for treated and untreated groups, which produces a subgroup-average hazard-ratio trajectory that more closely matches the ground truth. However, the substantially worse HR-IHE scores in Table 2 indicate that this comes at the cost of increased variance and large individual-level errors. DeepSurv (purple solid), consistent with its Cox-based formulation, yields time-constant hazard ratios and is therefore unable to represent time-varying treatment effects. In contrast, the proposed models (dashed) closely track the ground-truth hazard ratios (black dotted) across all scenarios. Even in Scenario C, where no true multiplicative hazard ratio exists, the proposed models produce correct estimates.

### 4.2. Real-World Primary Care Results

Table 3 reports the predictive performance on the two CPRD cohorts. In contrast to the simulation results, incorporating treatment information yields only marginal improvements over ignoring treatment at the population level. In the RAAS-Inhibitor cohort, the -SLL improves slightly from 0.138 (Ignore Treatment) to 0.137 (S-Learner and Subgroup models), while differences in the Anticoagulant cohort are negligible across all methods. These results suggest that, after adjustment for measured covariates, treatment effects are either modest at the population level, heavily confounded by unmeasured factors, or obscured by a high noise-to-signal ratio typical of observational EHR data.

Table 3: CPRD results (mean ± std across 5 folds). Best results **bolded**.

| Cohort | | Model | $C^{td}$ ↑ | IBS ↓ | INBLL ↓ | -SLL ↓ |
|---|---|---|---|---|---|---|
| RAAS-Inhibitor | | DeepSurv | 0.805 (0.009) | **0.039 (0.001)** | 0.150 (0.003) | – |
| | DeSurv | Ignore Treatment | 0.806 (0.011) | **0.039 (0.001)** | 0.148 (0.002) | 0.138 (0.003) |
| | DeSurv | S-Learner | **0.807 (0.010)** | **0.039 (0.000)** | 0.148 (0.002) | **0.137 (0.003)** |
| | DeSurv | T-Learner | 0.803 (0.008) | **0.039 (0.000)** | 0.149 (0.001) | 0.139 (0.002) |
| | Proposed | Subgroup | **0.807 (0.010)** | **0.039 (0.000)** | **0.147 (0.002)** | **0.137 (0.003)** |
| | Proposed | Propensity Subgroup | **0.807 (0.010)** | **0.039 (0.000)** | 0.148 (0.002) | 0.138 (0.003) |
| | Proposed | LowRank | 0.806 (0.011) | **0.039 (0.000)** | 0.148 (0.002) | 0.138 (0.002) |
| Anticoagulant | | DeepSurv | **0.720 (0.007)** | **0.095 (0.002)** | **0.315 (0.007)** | – |
| | DeSurv | Ignore Treatment | 0.717 (0.007) | 0.096 (0.002) | 0.316 (0.007) | **0.357 (0.004)** |
| | DeSurv | S-Learner | 0.718 (0.008) | **0.095 (0.002)** | 0.316 (0.007) | **0.357 (0.004)** |
| | DeSurv | T-Learner | 0.713 (0.007) | 0.096 (0.002) | 0.318 (0.007) | 0.359 (0.003) |
| | Proposed | Subgroup | 0.716 (0.007) | 0.096 (0.002) | 0.316 (0.007) | **0.357 (0.004)** |
| | Proposed | Propensity Subgroup | 0.716 (0.007) | 0.096 (0.002) | 0.316 (0.007) | **0.357 (0.003)** |
| | Proposed | LowRank | 0.716 (0.007) | 0.096 (0.002) | 0.317 (0.007) | 0.358 (0.003) |

Despite similar aggregate performance, meaningful differences emerge under subgroup analysis. Table 4 stratifies the RAAS-Inhibitor cohort by albumin-to-creatinine ratio (ACR), a clinically relevant marker of kidney function. Among patients with normal ACR ($\leq 3$), all models exhibit comparable performance. However, in the severely elevated ACR subgroup ($> 30$), the S-Learner shows degraded discrimination ($C^{td} = 0.756$) relative to both the Ignore Treatment baseline (0.774) and models that explicitly separate treatment effects. In contrast, the T-Learner and Propensity Subgroup models maintain better calibration, as reflected by lower IBS and -SLL. This divergence of minimal differences in aggregate metrics but pronounced subgroup-specific effects is consistent with the S-Learner's tendency to fit majority patterns, potentially masking heterogeneity. Additional subgroup analysis is provided in Appendix Table 9.

Figure 3 displays individualized hazard ratio trajectories for five randomly selected patients across models in the RAAS-Inhibitor cohort. DeepSurv produces constant hazard ratios over time, consistent with its Cox-based formulation. The DeSurv S-Learner yields non-smooth and erratic trajectories, while the T-Learner shows extreme sensitivity to individual patient characteristics, both reflecting numerical instability from post-hoc ratio computation. In contrast, the Subgroup model produces stable, smooth trajectories clustered around distinct values corresponding to learned subgroup assignments. The LowRank model yields individualized estimates with moderate time-variation and greater spread between patients, reflecting its continuous rather than discrete treatment effect parameterization.

### 4.3. Subgroup Analysis

To characterize patient profiles underlying the discovered subgroups, we perform a post-hoc analysis using logistic regression (Table 5). We extract hard subgroup assignments $\hat{z}_i = \arg\max_m \pi_m(\mathbf{x}_i)$ and fit a logistic regression model to predict subgroup membership from the original covariates. Accuracy reflects subgroup separability, and the top-3 features show the most predictive covariates, with frequencies indicating how often each feature is selected among the top predictors across the cross-validation folds.

In Scenario B, logistic regression achieves near-perfect accuracy (0.998), with proxy variables $v_1$, $v_2$, and $v_3$ (which have only moderate correlation with the true latent marker) consistently identified as top features. This confirms that the subgroup network successfully recovers latent structure from imperfect signals if they exist in the dataset. In the CPRD cohorts, accuracy remains high, though the defining characteristics are more heterogeneous. In the RAAS-Inhibitor cohort, prior hypertension medication use (HTN Drugs) is most discriminative, followed by creatinine and age, suggesting subgroup structure driven by baseline disease severity and renal function. In the Anticoagulant cohort, smoking status is most predictive, with platelet count and ethnicity also contributing, plausibly reflecting differences in cardiovascular risk profiles. These associations are

Table 4: RAAS-Inhibitor subgroup results (mean ± std across 5 folds) stratified by albumin-to-creatinine ratio (ACR). Best results **bolded**.

| | | ACR ≤ 3 (Normal) | | | | ACR > 30 (Severely Elevated) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Model** | **$C^{td}$ ↑** | **IBS ↓** | **INBLL ↓** | **-SLL ↓** | **$C^{td}$ ↑** | **IBS ↓** | **INBLL ↓** | **-SLL ↓** |
| | DeepSurv | **0.836 (0.023)** | **0.041 (0.005)** | 0.152 (0.016) | – | 0.764 (0.039) | 0.148 (0.019) | 0.477 (0.085) | – |
| DeSurv | Ignore Treatment | 0.835 (0.026) | **0.041 (0.004)** | **0.150 (0.017)** | **0.128 (0.017)** | **0.774 (0.076)** | 0.143 (0.020) | 0.443 (0.060) | **0.257 (0.064)** |
| DeSurv | S-Learner | 0.835 (0.027) | **0.041 (0.004)** | 0.151 (0.016) | **0.128 (0.017)** | 0.756 (0.072) | 0.143 (0.022) | 0.457 (0.074) | 0.265 (0.068) |
| DeSurv | T-Learner | 0.831 (0.022) | **0.041 (0.004)** | 0.152 (0.016) | 0.130 (0.016) | 0.760 (0.043) | **0.141 (0.017)** | **0.441 (0.048)** | 0.258 (0.044) |
| Proposed | Subgroup | **0.836 (0.025)** | **0.041 (0.005)** | **0.150 (0.017)** | **0.128 (0.017)** | 0.772 (0.069) | 0.143 (0.021) | 0.446 (0.061) | 0.259 (0.065) |
| Proposed | Propensity Subgroup | 0.835 (0.026) | **0.041 (0.004)** | 0.151 (0.016) | **0.128 (0.017)** | 0.772 (0.072) | **0.141 (0.022)** | 0.442 (0.063) | 0.258 (0.066) |
| Proposed | LowRank | 0.834 (0.026) | **0.041 (0.005)** | 0.151 (0.017) | **0.128 (0.018)** | 0.772 (0.068) | 0.143 (0.022) | 0.443 (0.066) | **0.257 (0.066)** |

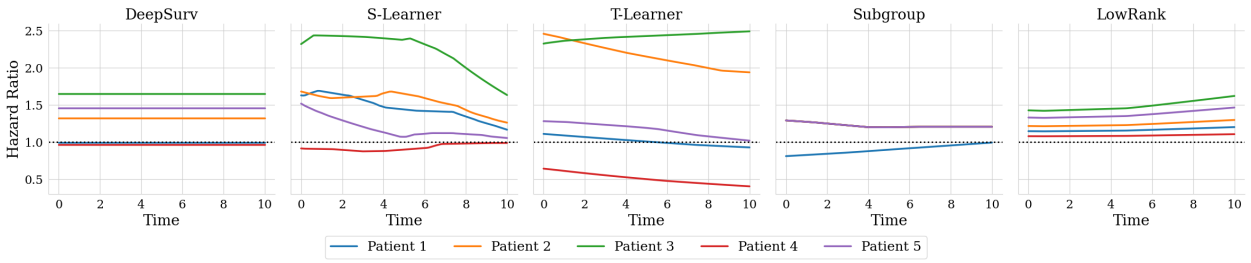



Figure 3: Individualized hazard-ratio estimates for five randomly selected patients across all models in the RAAS-Inhibitor cohort.

Table 5: Subgroup analysis via post-hoc logistic regression.

| **Dataset** | **Accuracy** | **Top 3 Features** |
|---|---|---|
| Simulation B | $0.998 \pm 0.001$ | $v_2$ (5/5), $v_1$ (5/5), $v_3$ (5/5) |
| RAAS-Inhibitor | $0.971 \pm 0.010$ | HTN Drugs (4/5), Creatinine (2/5), Age (2/5) |
| Anticoagulant | $0.957 \pm 0.017$ | Current Smoker (3/5), Platelet (2/5), Ethnicity (2/5) |

clinically interpretable but should be viewed as descriptive rather than causal.

Figure 4 visualizes estimated hazard ratios within each discovered subgroup. In the RAAS-Inhibitor cohort, our Subgroup model (red dashed) shows hazard ratios starting at 0.8 but increasing towards 1.0 by year 10 in Subgroup 0, while Subgroup 1 shows hazard ratios consistently above 1.2 across most methods. In the Anticoagulant cohort, Subgroup 0 shows hazard ratios below 1 (approximately 0.8-0.9), suggesting a protective treatment effect, while in Subgroup 1 hazard ratio estimates are closer to 1.0.

Kaplan-Meier survival curves stratified by discovered subgroup and treatment status (Figure 5) provide additional context. In the RAAS-Inhibitor cohort, Subgroup 0 exhibits high long-term survival (over 97% at 10 years) with minimal separation between treated and untreated patients. Subgroup 1 shows lower survival (approximately 87% at 10 years), with treated patients experiencing slightly worse outcomes than untreated patients. This is a pattern consistent with confounding by indication, where patients at higher baseline risk are more likely to receive RAAS-inhibitors. In the Anticoagulant cohort, Subgroup 0 shows slightly worse prognosis (approximately 78% survival at 5 years) compared to Subgroup 1 (approximately 80%), but within each subgroup, treated and untreated curves largely overlap. This suggests the treatment effect is too modest to produce visible Kaplan–Meier separation.

## 5. Discussion

This work shows that explicit hazard factorization provides a principled way to incorporate treatment into deep survival models. Across simulations, factorization substantially improves treatment effect estimation when effects are heterogeneous and/or time-varying and the multiplicative assumption approximately holds. Conceptually, the proposed framework represents a middle ground between classical meta-learners and this structural separation en-

Figure 4: Estimated hazard ratios over time for CPRD cohorts. Subgroups are inferred by the proposed subgroup model, and hazard ratios are averaged within each subgroup. Benchmark methods are evaluated on the same subgroup partitions.

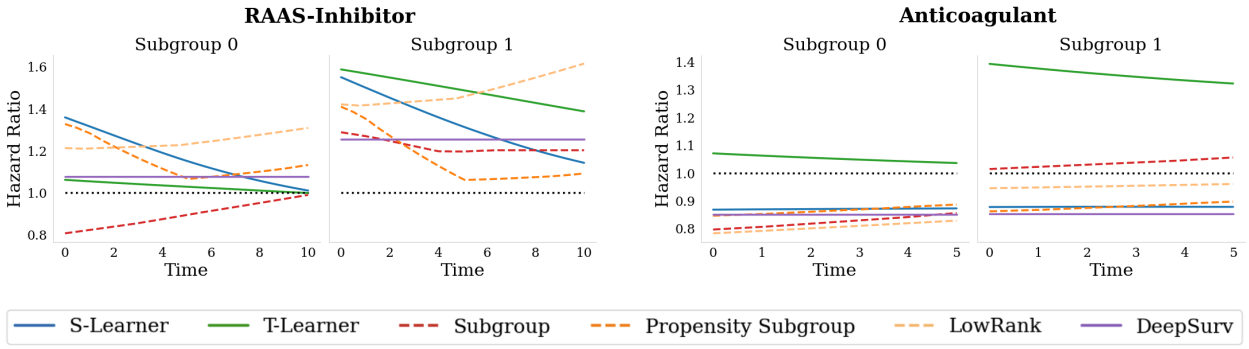
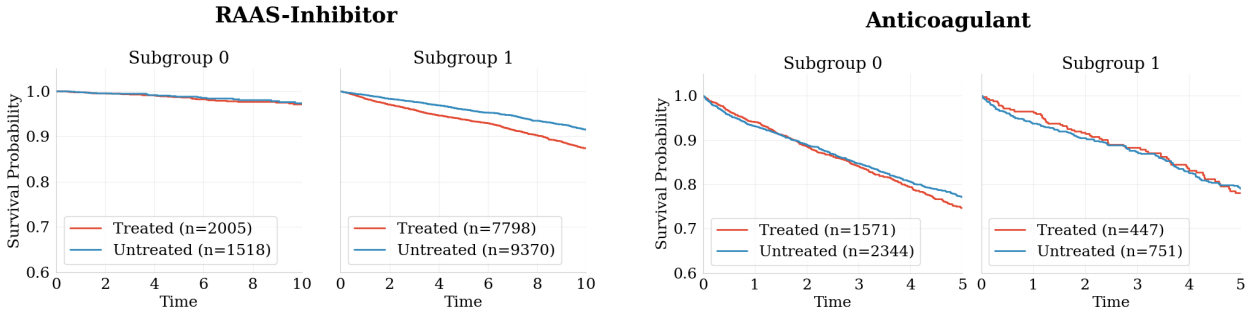



Figure 5: Kaplan-Meier survival curves stratified by discovered subgroup and treatment status.




ables direct estimation of time-varying hazard ratios, bridging flexible neural architectures with effect measures commonly used in clinical practice. However, individual-level time-varying hazard ratio estimates are inherently noisy, since each individual contributes information only at a single observed time point, motivating aggregation at the subgroup or population level for more stable estimates.

Our experiments focus on hazard ratio estimation rather than alternative summaries such as RMST. These quantities are complementary rather than competing, and RMST can be computed in our approach as well. RMST provides an absolute measure of survival benefit that can be preferable in some decision-making contexts, while hazard ratios express relative risk reduction and remain the dominant language of classical survival analysis.

The real-world experiments highlight several limitations. Population-level predictive gains from incorporating treatment are modest, reflecting challenges inherent to observational EHR data. Residual confounding from unmeasured factors, simplification of treatment as binary rather than accounting for dose, timing, or adherence, and noise in outcome ascertainment all limit recoverability of true effects.

While our experiments focus on binary treatment, the factorization naturally extends to categorical treatments via multiple treatment-specific components. We do not consider competing risks, which would require extending the framework to model cause-specific or subdistribution hazards, nor do we address continuous treatment regimes.

**Author Contribution** NH conceptualized the work, implemented the methods, executed the analyses, and drafted the manuscript. KN supplied clinical data and contributed clinical expertise. CY provided technical supervision, methodological guidance, and oversight of the writing process.

## Acknowledgments

NH acknowledges the receipt of studentship awards from the Health Data Research UK The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z) and the Alan Turing Enrichment Scheme. NH also received training support from the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). Christopher Yau is supported by an UKRI Turing AI Acceleration Fellowship (Ref: EP/V023233/1) and EPSRC grant (Ref: EP/Y018192/1). This study was undertaken as part of a National Institute for Health Research (NIHR) Intelligence for Multiple Long-Term Conditions (AIM) funded project: OPTIMising therapies, disease trajectories, and AI assisted clinical management for patients Living with complex multimorbidity (OPTIMAL study) Award ID: NIHR202632.[2] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. ISSN 1097-0258. doi: 10.1002/sim.2427.

Peter C. Austin and Elizabeth A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34:3661 – 3679, 2015.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24 (11):1713–1723, 2005.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1):1 – 3, 1950.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: learning heterogeneous treatment effects from time-to-event data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Dominic Danks and Christopher Yau. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7240–7256. PMLR, 2022.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.*, 24(1):198–208, January 2017.

Robert G Hart, Lesly A Pearce, and Maria I Aguilar. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann. Intern. Med.*, 146(12):857–867, June 2007.

Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd van Staa, and Liam Smeeth. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*, 44(3):827–836, 06 2015. ISSN 0300-5771. doi: 10.1093/ije/dyv098.

Liangyuan Hu, Jiayi Ji, and Fan Li. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat. Med.*, 40(21):4691–4713, September 2021.

Tazeen H Jafar, Paul C Stark, Christopher H Schmid, Marcia Landa, Giuseppe Maschio, Paul E de Jong, Dick de Zeeuw, Shahnaz Shahinfar, Robert Toto, Andrew S Levey, and AIPRD Study Group. Progression of chronic kidney disease: the role of blood pressure control, proteinuria, and angiotensin-converting enzyme inhibition: a patient-level meta-analysis. *Ann. Intern. Med.*, 139(4):244–252, August 2003.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

2. https://fundingawards.nihr.ac.uk/award/NIHR202632

Vincent Jeanselme, Chang Ho Yoon, Brian Tom, and Jessica Barrett. Neural fine-gray: Monotonic neural networks for competing risks, 2023.

Vincent Jeanselme, Chang Ho Yoon, Fabian Falck, Brian Tom, and Jessica Barrett. Identifying treatment response subgroups in observational time-to-event data, 2025.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition, 2002.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, Feb 2018. ISSN 1471-2288.

David M Kent, Ewout Steyerberg, and David van Klaveren. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*, 363, 2018. ISSN 0959-8138.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Richard L Kravitz, Naihua Duan, and Joel Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.*, 82(4):661–687, 2004.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *J. Mach. Learn. Res.*, 20:129:1–129:30, 2019.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019. ISSN 1091-6490.

E J Lewis, L G Hunsicker, W R Clarke, T Berl, M A Pohl, J B Lewis, E Ritz, R C Atkins, R Rohde, I Raz, and Collaborative Study Group. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N. Engl. J. Med.*, 345(12):851–860, September 2001.

Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. Counterfactual phenotyping with censored time-to-events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3634–3644, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850.

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

James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11 (5), 2000. ISSN 1044-3983.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ISSN 00063444, 14643510.

Patrick Royston and Mahesh K B Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.*, 13(1):152, December 2013.

Mats J Stensrud and Miguel A Hernán. Why use methods that require proportional hazards? *American Journal of Epidemiology*, 194(6):1504–1506, 01 2025. ISSN 0002-9262.

## Appendix A. Survival Relationships

The hazard function quantifies instantaneous risk at time $t$ given survival to $t$:

$$\lambda(t|\mathbf{x}) = \lim_{dt\to 0} \frac{\mathbb{P}(t \le T \le t + dt | T \ge t, \mathbf{x})}{dt} = \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})}$$

where $S(t) = \mathbb{P}(T > t) = 1 - F(t)$ is the survival function. The survival function relates to the cumulative hazard (Kalbfleisch and Prentice, 2002):

$$\begin{aligned}
\frac{d}{dt} S(t|\mathbf{x}) &= -f(t|\mathbf{x}) = -\lambda(t|\mathbf{x}) S(t|\mathbf{x}) \\
\frac{dS(t|\mathbf{x})}{S(t|\mathbf{x})} &= -\lambda(t|\mathbf{x})\, dt \\
\ln S(t|\mathbf{x}) &= -\int_0^t \lambda(u|\mathbf{x})\, du \\
S(t|\mathbf{x}) &= \exp\left(-\int_0^t \lambda(u|\mathbf{x})\, du\right) = \exp(-\Lambda(t|\mathbf{x}))
\end{aligned}$$

## Appendix B. Link Functions

We show that under the original DeSurv formulation, the objective can in certain cases, diverge to $-\infty$. This occurs because the log-likelihood permits arbitrarily large values of the gradient function $h(t)$ when initialized unfavorably. The right-censored survival log-likelihood is

$$\mathcal{L} = -\sum_{i=1}^{N} \Big[ 1_{\delta_i = 1} \log(F'(t_i) + \epsilon) + 1_{\delta_i = 0} \log(1 - F(t_i) + \epsilon) \Big],$$

where $\epsilon = 10^{-16}$ ensures numerical stability. The loss depends on the density term $F'(t_i)$ and the survival term $1 - F(t_i)$. DeSurv parameterizes

$$F(t) = \tanh\left(\int_0^t h(u)\, du\right),$$

so that event and censoring contributions to the loss function are:

$$\begin{aligned}
\text{Event, } \delta = 1 &: -\log\Big( (1 - \tanh^2 \int_0^t h(u)\, du)\, h(t) + \epsilon \Big), \\
\text{Censoring, } \delta = 0 &: -\log\Big( 1 - \tanh \int_0^t h(u)\, du + \epsilon \Big).
\end{aligned}$$

**Limit behavior.** As $h(u) \to 0$:

$$\mathcal{L} = \begin{cases} -\log(\epsilon) \approx 36.84, & \delta = 1 \\ -\log(1+\epsilon) \approx 0, & \delta = 0 \end{cases}.$$

As $h(u) \to \infty$:

$$\mathcal{L} = \begin{cases} -\log(\epsilon) - \log(h(t)) \to -\infty, & \delta = 1 \\ -\log(\epsilon) \approx 36.84, & \delta = 0 \end{cases}.$$

Thus, for event observations, the loss can decrease without bound as $h(t) \to \infty$, implying that the objective admits degenerate solutions where the model minimizes loss by predicting arbitrarily large hazards. This collapses the survival curve into a near-step function at time zero. Our mapping $F(t) = 1 - \exp(-u(t))$ avoids this as the loss is always lower bounded.

## Appendix C. Simulation

In this section, we denote the hazard function by $h(\cdot)$ to avoid notational conflict with $\lambda$, which is used for model parameters.

### C.1. Weibull Proportional Hazards Model

The Weibull proportional hazards model is a common choice for simulating survival data due to its closed-form expressions. The hazard function is defined as

$$h(t|\mathbf{x}) = \lambda p t^{p-1} \exp(\beta^T \mathbf{x}),$$

where $t$ represents survival time, $\lambda > 0$ is the baseline hazard rate or scale parameter, $p > 0$ is the shape parameter, and $\beta$ is the corresponding vector of regression coefficients. The survival function follows as

$$S(t|\mathbf{x}) = \exp\left(-\lambda t^p \exp(\beta^T \mathbf{x})\right).$$

Observed event times $t_E$ are generated using inverse transform sampling. Given $F(t|\mathbf{x})$ is a CDF and the survival function is defined as $S(t|\mathbf{x}) = 1 - F(t|\mathbf{x})$, setting

$$S(t_E|\mathbf{x}) = U, \quad U \sim \text{Uniform}(0, 1)$$

yields

$$\exp\left(-\lambda t_E^p \exp(\beta^T \mathbf{x})\right) = U.$$

Solving for $t_E$ results in

$$t_E = \left(\frac{-\log(U)}{\lambda \exp(\beta^T \mathbf{x})}\right)^{\frac{1}{p}}.$$

### C.2. Gompertz Proportional Hazards Model

The Gompertz proportional hazards model is another commonly used parametric model, particularly well-suited for capturing monotonic hazard trends over time. It is widely applied in aging and mortality studies. The hazard function is given by

$$h(t \mid \mathbf{x}) = \lambda \exp(\gamma t) \exp(\beta^\top \mathbf{x}),$$

where $\lambda > 0$ is the baseline hazard, $\gamma \in \mathbb{R}$ controls the shape of the hazard (increasing if $\gamma > 0$, decreasing if $\gamma < 0$), and $\beta$ are the regression coefficients. The corresponding survival function is given by

$$S(t \mid \mathbf{x}) = \exp\left(-\frac{\lambda}{\gamma}\left[\exp(\gamma t) - 1\right] \exp(\beta^\top \mathbf{x})\right).$$

Observed event times $t_E$ can again be generated using inverse transform sampling,

$$S(t_E \mid \mathbf{x}) = U, \quad U \sim \text{Uniform}(0, 1),$$

which implies

$$\exp\left(-\frac{\lambda}{\gamma}\left[\exp(\gamma t_E) - 1\right] \exp(\beta^\top \mathbf{x})\right) = U.$$

Solving for $t_E$ yields

$$t_E = \frac{1}{\gamma} \log\left(1 - \frac{\gamma \log(U)}{\lambda \exp(\beta^\top \mathbf{x})}\right).$$

### C.3. Time-Invariant and Time-Varying Hazard Ratios

To account for covariate effects that vary over time, the proportional hazards models can be extended to include time-dependent effects. The general forms for the hazard functions are:

$$\begin{aligned} \text{Weibull:} \quad & h(t \mid \mathbf{x}) = \lambda p t^{p-1} \exp(\eta(t \mid \mathbf{x})) \\ \text{Gompertz:} \quad & h(t \mid \mathbf{x}) = \lambda \exp(\gamma t) \exp(\eta(t \mid \mathbf{x})) \end{aligned}$$

A common specification for the time-varying component $\eta(t \mid \mathbf{x})$ is

$$\eta(t \mid \mathbf{x}) = \beta^\top \mathbf{x} + \gamma^\top \mathbf{x} \cdot g(t),$$

where $\beta$ captures time-invariant effects, $\gamma$ modulates the time-varying influence of covariates, and $g(t)$ is a user-defined function of time (e.g., $\log t$, $t$, or basis splines). When $\gamma^\top \mathbf{x} \neq \mathbf{0}$, the effect of covariates is no longer proportional over time.

If only the treatment indicator is allowed to interact with time, the proportional hazards assumption is relaxed for treatment, while being preserved for other covariates.

$$\eta(t \mid \mathbf{x}) = \beta^\top \mathbf{x} + \gamma_{\text{trt}} \cdot 1_{\text{treatment}} \cdot g(t).$$

Depending on the choice of $\eta(t \mid \mathbf{x})$, a closed-form expression for the survival function may not exist, requiring numerical integration and/or root-finding techniques to obtain simulated $t_E$.

### C.4. Simulation Details

We generate $N = 20{,}000$ observations following a multi-stage procedure designed to create realistic confounding and latent subgroup structure.

**Step 1: Latent Variables.** We first generate two underlying variables:

- $H \sim \mathcal{N}(0, 0.4^2)$: a continuous "health" variable representing underlying frailty (never observed)
- $G \sim \text{Bernoulli}(0.4)$: a binary "marker" determining treatment response subgroup

The marker $G$ is provided as an observed covariate in Scenario A, but is unobserved in Scenarios B and C.

**Step 2: Observed Covariates.** Ten observed covariates (age, sex, smoking status, prior condition, and continuous markers $v_1$–$v_5$) are generated as functions of the latent health variable $H$. In Scenario B, covariates $v_1$–$v_3$ additionally depend on the marker $G$, serving as noisy proxies (correlation approximately 0.6–0.8) for the unobserved subgroup membership.

**Step 3: Treatment Assignment.** Treatment is assigned via a logistic model depending on observed covariates, inducing confounding and yielding approximately 40% treatment prevalence.

**Step 4: Survival Times.** Survival times are generated from Weibull (Scenarios A, B) or mixed Weibull/Gompertz (Scenario C) distributions, with hazard depending on observed covariates. Treated subjects with $G = 0$ (responders) receive the treatment effect specified in Table 1; treated subjects with $G = 1$ (non-responders) and all untreated subjects follow the baseline hazard.

**Step 5: Censoring.** Censoring times $t_C$ are generated from an exponential distribution with rate 0.05, with administrative censoring at $t_{\max} = 20$ years, yielding approximately 40% observed events. The observed survival time and event indicator are then

$$\begin{aligned} T &= \min(t_E, t_C, t_{\max}), \\ \delta &= 1(t_E \leq \min(t_C, t_{\max})). \end{aligned}$$

## Appendix D. Experiments

We provide the full experimental configuration in Table 6. Data are split into 64% training, 16% validation, and 20% testing. No systematic hyperparameter tuning is performed for any method to ensure fair comparison. Instead, we conduct an ablation study (Appendix F) to assess sensitivity to key architectural choices. During training, subgroup assignments are sampled using the straight-through Gumbel–Softmax estimator, with temperature annealed as $\tau = \max(0.2, ; \tau_0(1 - \frac{e}{E})^2)$ where $\tau_0 = 1.0$, $e$ is the current epoch, and $E$ the total number of epochs, encouraging increasingly discrete assignments; at inference, assignments are obtained via arg max over the logits. For metrics requiring time integration (IBS, INBLL, ISE, HR-ISE), evaluation is performed on a discrete grid from $t = 0$ to $t_{\max}$ with step size 0.1.

Table 6: Experimental Configuration Across Datasets

| Experiments | Simulations | RAAS-Inhibitor | Anticoagulant |
|---|---|---|---|
| **Reproducibility** | | | |
| Random seeds | Seed = 42 for sklearn KFold, NumPy, PyTorch, CUDA | | |
| **Training Details** | | | |
| Batch size | 256 | 512 | 256 |
| Optimiser | Adam (learning rate = 1e−4, betas = (0.9, 0.999), eps = 1e−8) | | |
| Training epochs | 500 | | |
| Early stopping patience | 50 | | |
| **Survival Head (DeSurv) Architecture** | | | |
| Hidden layers | [32, 32] | | |
| **Subgroup Architecture** | | | |
| HR hidden layers ($h_\mu^{(m)}$) | [16, 8] | | |
| Subgroup hidden layers ($\pi_m$) | [16, 8] | | |
| **LowRank Architecture** | | | |
| Rank ($K$) | 2 | | |
| Time-dependent MLP ($\alpha$) | [8, 8] | | |
| Covariate-dependent MLP ($\beta$) | [8] | | |
| **Propensity Architecture** | | | |
| Rank ($K$) | [32, 16] | | |
| **Evaluation Settings** | | | |
| Max time | 20.0 | 10.0 | 5.0 |

## Appendix E. Evaluation Metrics

The classical survival metrics are implemented using the pycox library (Kvamme et al., 2019):

- **Time-Dependent Concordance ($\mathbf{C^{td}}$)** (Antolini et al., 2005) A measure of discrimination that quantifies how well the model orders predicted risks over time, estimated by calculating the proportion of pairs in concordance. Values closer to 1 reflect stronger concordance and thus better predictive ability.
- **Integrated Brier Score (IBS)** The Brier score (Brier, 1950) captures the squared difference between observed survival status and predicted survival probabilities, accounting for censoring. The IBS summarizes this across the follow-up period, with lower scores corresponding to more accurate models.
- **Integrated Negative Binomial Log-Likelihood (INBLL)** A likelihood-based criterion that evaluates binary survival predictions through the negative log-likelihood, adjusted for censoring and averaged over time. Smaller values indicate better performance.

# Appendix F. Additional Results

## F.1. Sample Size Ablation

Table 7 shows performance across sample sizes $N \in \{10k, 20k, 50k\}$. All models improve with more data. Our Subgroup models achieve consistent HR-IHE reduction across sample sizes. At $N = 50k$, Propensity Subgroup achieves HR-IHE $< 0.03$ in Scenarios A and C.

Table 7: Simulation sample size ablation study. Best results **bolded**.

| | | | $N = 10k$ | | $N = 20k$ | | $N = 50k$ | |
|---|---|---|---|---|---|---|---|---|
| | | **Model** | **ISE ↓** | **HR–IHE ↓** | **ISE ↓** | **HR–IHE ↓** | **ISE ↓** | **HR–IHE ↓** |
| **A** | | DeepSurv | 0.037 (0.005) | 0.382 (0.088) | 0.021 (0.001) | 0.218 (0.014) | 0.012 (0.002) | 0.100 (0.006) |
| | DeSurv | Ignore | 0.126 (0.005) | 1.287 (0.031) | 0.118 (0.003) | 1.266 (0.023) | 0.110 (0.003) | 1.283 (0.015) |
| | DeSurv | S-Learner | 0.046 (0.010) | 0.442 (0.206) | 0.028 (0.002) | 0.170 (0.032) | 0.012 (0.001) | 0.076 (0.006) |
| | DeSurv | T-Learner | 0.036 (0.003) | 0.396 (0.048) | 0.025 (0.002) | 0.245 (0.039) | 0.015 (0.001) | 0.174 (0.029) |
| | Proposed | Subgroup | 0.040 (0.011) | 0.351 (0.186) | *0.023 (0.001)* | **0.052 (0.025)** | *0.010 (0.000)* | *0.021 (0.007)* |
| | Proposed | Propensity Subgroup | *0.035 (0.004)* | **0.230 (0.033)** | **0.022 (0.002)** | **0.052 (0.024)** | **0.009 (0.000)** | **0.014 (0.004)** |
| | Proposed | LowRank | **0.034 (0.003)** | *0.288 (0.028)* | 0.026 (0.001) | 0.249 (0.027) | 0.013 (0.001) | 0.154 (0.018) |
| **B** | | DeepSurv | 0.035 (0.002) | 0.614 (0.041) | 0.025 (0.003) | 0.482 (0.050) | 0.018 (0.002) | 0.446 (0.006) |
| | DeSurv | Ignore | 0.072 (0.007) | 1.139 (0.027) | 0.067 (0.005) | 1.121 (0.020) | 0.061 (0.002) | 1.136 (0.013) |
| | DeSurv | S-Learner | 0.037 (0.003) | 0.546 (0.078) | 0.031 (0.005) | 0.335 (0.075) | 0.019 (0.001) | 0.211 (0.023) |
| | DeSurv | T-Learner | 0.041 (0.005) | 0.413 (0.053) | 0.028 (0.002) | 0.312 (0.025) | 0.018 (0.001) | 0.210 (0.020) |
| | Proposed | Subgroup | *0.030 (0.004)* | **0.046 (0.013)** | *0.020 (0.001)* | **0.047 (0.009)** | **0.013 (0.001)** | **0.047 (0.009)** |
| | Proposed | Propensity Subgroup | **0.028 (0.002)** | *0.076 (0.027)* | **0.018 (0.001)** | **0.047 (0.012)** | *0.014 (0.001)* | **0.047 (0.005)** |
| | Proposed | LowRank | 0.038 (0.003) | 0.302 (0.066) | 0.026 (0.002) | 0.237 (0.030) | 0.017 (0.001) | 0.165 (0.013) |
| **C** | | DeepSurv | **0.017 (0.002)** | 0.253 (0.034) | **0.012 (0.001)** | 0.220 (0.019) | **0.007 (0.000)** | 0.226 (0.009) |
| | DeSurv | Ignore | 0.043 (0.002) | 0.898 (0.000) | 0.036 (0.001) | 0.898 (0.000) | 0.029 (0.001) | 0.898 (0.000) |
| | DeSurv | S-Learner | *0.022 (0.002)* | 0.271 (0.039) | *0.015 (0.001)* | 0.226 (0.020) | 0.010 (0.001) | 0.190 (0.025) |
| | DeSurv | T-Learner | 0.035 (0.003) | 0.435 (0.101) | 0.023 (0.002) | 0.260 (0.054) | 0.012 (0.001) | 0.149 (0.020) |
| | Proposed | Subgroup | 0.025 (0.002) | *0.139 (0.075)* | 0.019 (0.002) | **0.040 (0.021)** | *0.009 (0.001)* | *0.034 (0.021)* |
| | Proposed | Propensity Subgroup | 0.026 (0.002) | **0.126 (0.067)** | 0.020 (0.001) | *0.045 (0.021)* | *0.009 (0.001)* | **0.027 (0.012)** |
| | Proposed | LowRank | 0.027 (0.004) | 0.300 (0.055) | 0.019 (0.001) | 0.198 (0.058) | 0.011 (0.001) | 0.125 (0.024) |

## F.2. Subgroups Number Ablation

Table 8 varies the number of subgroups $M \in \{2, 3, 5\}$. Proper scoring metrics (-SLL, ISE) are stable across $M$. HR-IHE shows modest sensitivity: $M = 2$ (matching ground truth) performs best in simulations, but $M = 3$ or $M = 5$ remain competitive.

## F.3. CPRD Subgroup Analysis

Table 9 extends the subgroup analysis of the RAAS-Inhibitor cohort to additional clinically relevant partitions such as age, prior hypertension drug use, and propensity score. Across most stratifications, model performance differences remain small, consistent with the aggregate results. The propensity score stratification reveals some variation in the high-propensity subgroup ($> 0.8$) where the S-Learner performs slightly worse on all metrics.

Table 8: Sensitivity to the number of subgroups.

| Experiment | M | -SLL ↓ | ISE ↓ | HR–IHE ↓ |
|---|---|---|---|---|
| **A** | 2 | 0.350 (0.005) | 0.023 (0.001) | 0.052 (0.025) |
| | 3 | 0.350 (0.005) | 0.023 (0.001) | 0.034 (0.023) |
| | 5 | 0.350 (0.005) | 0.025 (0.002) | 0.037 (0.021) |
| **B** | 2 | 0.309 (0.006) | 0.020 (0.001) | 0.047 (0.009) |
| | 3 | 0.309 (0.006) | 0.021 (0.001) | 0.062 (0.019) |
| | 5 | 0.309 (0.006) | 0.021 (0.002) | 0.051 (0.018) |
| **C** | 2 | 0.352 (0.007) | 0.019 (0.002) | 0.040 (0.021) |
| | 3 | 0.352 (0.007) | 0.018 (0.002) | 0.086 (0.040) |
| | 5 | 0.352 (0.007) | 0.018 (0.001) | 0.050 (0.026) |
| **RAAS-Inhibitor** | 2 | 0.137 (0.003) | - | - |
| | 3 | 0.138 (0.003) | - | - |
| | 5 | 0.138 (0.002) | - | - |
| **Anticoagulant** | 2 | 0.357 (0.004) | | |
| | 3 | 0.357 (0.003) | | |
| | 5 | 0.357 (0.003) | | |

Table 9: RAAS-Inhibitor subgroup (mean ± std across 5 folds). Best results **bolded**.

| | Model | $C^{td}$ ↑ | IBS ↓ | INBLL ↓ | -SLL ↓ | $C^{td}$ ↑ | IBS ↓ | INBLL ↓ | -SLL ↓ |
|---|---|---|---|---|---|---|---|---|---|
| | | Age ≤ 60 | | | | Age > 60 | | | |
| | DeepSurv | 0.809 (0.012) | **0.028 (0.001)** | 0.114 (0.004) | – | 0.780 (0.011) | 0.054 (0.002) | 0.198 (0.005) | – |
| DeSurv | Ignore Treatment | 0.809 (0.013) | **0.028 (0.001)** | 0.113 (0.003) | 0.106 (0.004) | 0.779 (0.013) | **0.053 (0.002)** | 0.194 (0.006) | 0.182 (0.004) |
| DeSurv | S-Learner | **0.813 (0.013)** | **0.028 (0.001)** | **0.112 (0.004)** | **0.104 (0.004)** | 0.779 (0.012) | **0.053 (0.001)** | 0.194 (0.005) | **0.181 (0.003)** |
| DeSurv | T-Learner | 0.808 (0.011) | **0.028 (0.001)** | 0.114 (0.003) | 0.106 (0.004) | 0.775 (0.011) | **0.053 (0.001)** | 0.196 (0.005) | 0.184 (0.002) |
| Proposed | Subgroup | 0.811 (0.013) | **0.028 (0.001)** | 0.113 (0.003) | 0.105 (0.004) | **0.781 (0.011)** | **0.053 (0.002)** | **0.193 (0.005)** | **0.181 (0.003)** |
| Proposed | Propensity Subgroup | 0.810 (0.013) | **0.028 (0.001)** | 0.113 (0.003) | 0.105 (0.004) | 0.780 (0.012) | **0.053 (0.002)** | 0.194 (0.006) | **0.181 (0.003)** |
| Proposed | LowRank | 0.810 (0.012) | **0.028 (0.001)** | 0.113 (0.003) | 0.105 (0.004) | 0.780 (0.013) | **0.053 (0.002)** | 0.194 (0.005) | 0.182 (0.003) |
| | | Hypertension Drugs = 0 | | | | Hypertension Drugs = 1 | | | |
| | DeepSurv | 0.797 (0.010) | 0.036 (0.001) | 0.141 (0.003) | – | **0.816 (0.012)** | 0.059 (0.002) | 0.211 (0.007) | – |
| DeSurv | Ignore Treatment | 0.799 (0.012) | 0.036 (0.001) | **0.138 (0.002)** | 0.131 (0.003) | 0.812 (0.013) | 0.059 (0.002) | 0.209 (0.005) | 0.176 (0.004) |
| DeSurv | S-Learner | 0.799 (0.011) | **0.035 (0.000)** | **0.138 (0.002)** | **0.130 (0.002)** | 0.815 (0.014) | **0.058 (0.001)** | **0.207 (0.005)** | **0.175 (0.006)** |
| DeSurv | T-Learner | 0.794 (0.009) | 0.036 (0.000) | 0.140 (0.002) | 0.133 (0.002) | 0.813 (0.010) | 0.059 (0.002) | 0.209 (0.005) | 0.178 (0.002) |
| Proposed | Subgroup | **0.800 (0.011)** | **0.035 (0.000)** | **0.138 (0.002)** | 0.131 (0.003) | 0.813 (0.011) | **0.058 (0.002)** | 0.208 (0.005) | **0.175 (0.004)** |
| Proposed | Propensity Subgroup | 0.799 (0.011) | 0.036 (0.001) | **0.138 (0.002)** | 0.131 (0.003) | 0.814 (0.012) | **0.058 (0.002)** | 0.208 (0.005) | 0.176 (0.003) |
| Proposed | LowRank | 0.799 (0.011) | **0.035 (0.000)** | **0.138 (0.002)** | 0.131 (0.002) | 0.812 (0.013) | **0.058 (0.002)** | 0.208 (0.006) | 0.176 (0.004) |
| | | Propensity* ≤ 0.2 | | | | Propensity* > 0.8 | | | |
| | DeepSurv | 0.786 (0.044) | 0.043 (0.004) | 0.162 (0.015) | – | 0.822 (0.006) | 0.049 (0.002) | 0.179 (0.006) | – |
| DeSurv | Ignore Treatment | 0.784 (0.043) | **0.042 (0.004)** | 0.161 (0.016) | 0.140 (0.018) | 0.823 (0.011) | **0.048 (0.002)** | **0.176 (0.008)** | **0.160 (0.005)** |
| DeSurv | S-Learner | **0.792 (0.043)** | **0.042 (0.004)** | **0.160 (0.015)** | **0.138 (0.016)** | 0.819 (0.011) | 0.049 (0.002) | 0.177 (0.007) | 0.162 (0.004) |
| DeSurv | T-Learner | 0.790 (0.042) | **0.042 (0.003)** | **0.160 (0.015)** | **0.138 (0.017)** | 0.817 (0.005) | 0.049 (0.002) | 0.178 (0.008) | 0.164 (0.005) |
| Proposed | Subgroup | 0.790 (0.043) | **0.042 (0.004)** | **0.160 (0.017)** | **0.138 (0.018)** | 0.823 (0.011) | **0.048 (0.002)** | **0.176 (0.008)** | **0.160 (0.005)** |
| Proposed | Propensity Subgroup | 0.790 (0.040) | **0.042 (0.004)** | **0.160 (0.016)** | 0.139 (0.017) | **0.824 (0.010)** | **0.048 (0.002)** | **0.176 (0.008)** | **0.160 (0.005)** |
| Proposed | LowRank | 0.790 (0.044) | **0.042 (0.004)** | **0.160 (0.016)** | **0.138 (0.018)** | 0.821 (0.010) | **0.048 (0.002)** | **0.176 (0.008)** | 0.161 (0.005) |

* Propensity partitions are based on estimates from a logistic regression model.