Simulation Design Report

Time-dependent Net Benefit Estimation under Informative Censoring

Leveraging IPCW for Robust Evaluation in Survival Data Analysis

Name:	Amir Farahmand
Report Title:	Simulation Design for tNB Estimation under Informative Censoring Using IPCW
Supervisors:	Harry Lee, Abdollah Safari

Introduction

Risk prediction models and biomarkers are central to modern clinical decision-making. A **biomarker** is a single measurable characteristic — such as a laboratory value, imaging finding, or genetic marker — that reflects biological processes or disease risk. A **risk prediction model**, in contrast, is a more general tool that can incorporate multiple biomarkers simultaneously to estimate an individual's probability of experiencing an event within a given time horizon.

In this work, we focus on **prediction models** for calculating the *time-dependent net benefit* (tNB). Prediction model research typically proceeds in two phases:

1. **Model development:** Building the model using available data, specifying predictors, estimating coefficients, and generating individual-level risk predictions.
2. **Model validation:** Assessing the model's performance in independent data to ensure generalisability beyond the development set.

Here, our simulation design addresses the **development phase** first, focusing on how informative censoring affects estimation of tNB.

Standard approaches for estimating time-dependent net benefit — such as the Kaplan–Meier (KM) method — assume **non-informative censoring**, meaning that the censoring process is independent of the event process. Under this assumption, the KM estimator yields unbiased estimates. However, under **informative censoring**, where the probability of being censored depends on biomarkers (and hence on prognosis), the KM estimator fails, producing biased tNB estimates. Therefore, we must adapt strategies that *explicitly account for the censoring mechanism* to recover unbiased estimates of tNB.

Several methods have been proposed to address biomarker-dependent censoring. Examples include:

- **Nearest-neighbour kernel–based Kaplan–Meier estimation**, which estimates survival functions locally in biomarker space.
- **Inverse probability of censoring weighting (IPCW)**, which reweights individuals according to the inverse of their probability of remaining uncensored.
- **Conditional IPCW**, a refinement of IPCW in which weights are estimated conditional on biomarkers or risk scores.

These methods share the goal of **capturing biomarker-dependent censoring** to reduce or eliminate bias in time-dependent net benefit estimation.

In this study, we focus on **conditional IPCW** as our primary adjustment method. Conditional IPCW can explicitly capture covariate (biomarker)-dependent censoring when the censoring survival function is correctly specified. By modelling the censoring distribution conditional on relevant covariates (biomarkers), we can adjust for bias introduced by informative censoring and recover unbiased estimates of tNB. This property directly addresses our central research question: to demonstrate the failure of the Kaplan–Meier estimator under informative censoring and to evaluate a method that can correct for it.

A key advantage of conditional IPCW is that it offers a **semi-parametric** framework. It is not fully non-parametric, which allows us to flexibly choose from a wide variety of modelling strategies to estimate the censoring distribution (e.g., Cox proportional hazards models, flexible parametric models, or machine learning methods) without committing to a fixed parametric form. This flexibility stands in contrast to nearest-neighbour kernel-based methods, which rely heavily on a smoothing parameter (bandwidth) and can be unstable when the sample size is moderate or the censoring rate is high.

Moreover, IPCW provides a **transparent framework** to incorporate covariates (biomarker) dependence directly into the weighting scheme. By conditioning on biomarkers in the censoring model, the method naturally accommodates the common and realistic scenario in prediction model development where censoring is related to prognostic biomarkers.

Finally, IPCW is **practical and widely used** in applied survival analysis, with readily available implementations in standard R packages such as `timeROC`. This practicality makes the method directly relevant for applied work and ensures that the insights from our simulation study will be accessible to practitioners.

The simulation design follows the **ADEMP** framework as introduced by **?**, and incorporates guidance from **?** to avoid pitfalls in simulation design and minimize the risk of questionable research practices (**QRPs**).

1 Aims – A

1.1 Purpose and Scope of the Simulation

This simulation study has a **dual purpose**:

1. **Proof-of-concept:** Demonstrate that under informative censoring, the Kaplan–Meier (KM) approach yields biased time-dependent net benefit (tNB) estimates, whereas inverse probability of censoring weighting (IPCW) can recover unbiased estimates when the censoring model is correctly specified.
2. **Robustness assessment:** Evaluate whether these conclusions hold across a range of realistic data-generating scenarios, varying hazard shape, censoring rate, and sample size.

The study focuses on censoring mechanisms where dropout probability depends on observed covariates, a common situation in clinical research.

1.2 Motivating Clinical Example

In longitudinal HIV cohort studies, CD4 counts are measured repeatedly as a key marker of immune function. Patients whose CD4 counts fall below 200 cells/mm^3 are at elevated risk of death and are also more likely to drop out of follow-up due to severe illness, stigma, or treatment changes. Such censoring is *informative* because it depends on prognosis-related covariates, violating the independent censoring assumption of standard survival estimators. Our simulation framework mirrors this feature to evaluate tNB estimation methods in settings with covariate-dependent informative censoring.

1.3 Central Research Question

Does IPCW provide unbiased and accurate estimates of time-dependent net benefit in the presence of informative censoring, compared with KM, which is valid only under non-informative censoring?

1.4 Theoretical Expectations

When censoring is non-informative, KM should yield unbiased estimates with slightly lower variance than IPCW due to greater efficiency. Under informative censoring, KM is expected to exhibit substantial bias, whereas IPCW should remain unbiased if the censoring model is correctly specified. In small samples, both estimators may show increased variability, but IPCW is anticipated to have much lower bias than KM in informative-censoring scenarios.

It is important to note that this unbiasedness property for KM under non-informative censoring and for IPCW under correctly specified informative censoring is an *asymptotic* result: it holds in large samples. In finite samples, both methods may exhibit some bias due to sampling variability and estimation error in the censoring model, with the potential for larger small-sample bias when the censoring proportion is high or the sample size is limited.

1.5 Scope and Assumptions

The primary focus is on covariate-dependent informative censoring modeled via a Cox proportional hazards model for the censoring distribution. Other dependency structures, such as those arising from unmeasured shared frailties or copula-based dependencies, are outside the scope of this study but could be addressed in future work. The validity of IPCW relies on correctly specifying the censoring model; misspecification can result in biased or inconsistent estimates.

2 Data-Generating Mechanisms – D

2.1 Covariate Generation

We generate two covariates for each simulated subject:

- One **continuous** covariate $Z_1 \sim \mathcal{N}(0, 1)$.
- One **binary** covariate $Z_2 \sim \text{Bernoulli}(0.5)$.

These covariates are included in both the *event-time* model and the *censoring-time* model. The motivation for using one continuous and one binary covariate is to mimic a simple but realistic clinical setting in which prognostic information comes from both types of variables. This design also facilitates introducing *informative censoring*, since the same covariates that affect survival can also be made to influence censoring times.

2.2 Survival Time Generation

We generate event times from a **proportional hazards (PH) model** of the form

$$h(t \mid Z_1, Z_2) = h_0(t) \exp\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2\right),$$

where:

- $h_0(t)$ is the baseline hazard function.
- β_0 is the *intercept* term, controlling the overall event rate and allowing calibration of the marginal survival probability at a given time point.
- β_1 and β_2 are the covariate effects, expressed on the log-hazard scale.

The proportional hazards framework is chosen because it is widely used in time-to-event analysis and allows us to:

1. Directly control the effect sizes of covariates via hazard ratios.
2. Adjust the intercept β_0 to achieve desired overall event rates across scenarios.
3. Flexibly model different shapes of the underlying hazard through the choice of $h_0(t)$.

For the continuous covariate effect β_1, it is common in simulation studies to choose a value such that a one-standard-deviation (1-SD) increase in Z_1 changes the hazard by approximately 50–100%. Similarly, for the binary covariate effect β_2, it is typical to choose a value such that the hazard for category 1 is 50–100% higher than for category 0. Setting $\beta_1 = 0.5$ and $\beta_2 = 0.5$ achieves hazard ratios of approximately $\exp(0.5) \approx 1.65$ for both covariates, which lies in this moderate, realistic effect-size range.

The intercept term β_0 controls the overall event rate and, together with the choice of baseline hazard $h_0(t)$, determines the marginal survival distribution. We calibrate β_0 so that the vast majority of events occur within a maximum follow-up time of approximately $t_{\max} \approx 6$ years. This ensures that in the simulated datasets, observation times are effectively bounded by six years, reflecting a realistic finite follow-up window and avoiding extremely long simulated survival times that would be implausible in a typical clinical cohort.

Operationally, β_0 is determined by solving for the value that yields a marginal survival probability close to zero at $t_{\max}$:

$$\mathbb{E}_{Z_1, Z_2}\left[S(t_{\max} \mid Z_1, Z_2; \beta_0)\right] \approx 0.001,$$

where $S(t \mid Z)$ is the survival function implied by the proportional hazards model with the chosen $h_0(t)$. This calibration is performed numerically for each baseline hazard specification so that the time-to-event distribution respects the $t_{\max}$ constraint while preserving the intended covariate effect sizes.

To capture a wide range of hazard shapes, we vary the specification of the baseline hazard:

- **Parametric baseline hazards:** We use the *generalized gamma* distribution for $h_0(t)$. The generalized gamma (GG) is a flexible three-parameter family with density function

$$f(t; \lambda, p, k) = \frac{|k|}{\lambda\, \Gamma(p)} \left(\frac{t}{\lambda}\right)^{kp-1} \exp\left[-\left(\frac{t}{\lambda}\right)^{k}\right], \qquad t > 0,$$

  where:

- $\lambda > 0$ is a scale parameter controlling the time scale of the distribution;
- $p > 0$ is a shape parameter influencing the tail heaviness;
- $k \neq 0$ is a shape parameter controlling skewness and the general form of the hazard function.

The GG family contains several well-known survival distributions as special cases:

- **Exponential:** $p = 1$, $k = 1$ (constant hazard).
- **Weibull:** $p = 1$, $k \neq 1$ (monotone hazard, increasing if $k > 1$, decreasing if $k < 1$).
- **Log-normal:** obtained as a limiting case $k \to 0$ with suitable reparameterisation.

By varying (p, k), the GG can produce constant, monotone, unimodal, or bathtub-shaped hazards, making it ideal for generating diverse baseline hazard scenarios.

Hazard shape scenarios:

1. *Constant hazard (Exponential case):* We set $p = 1$, $k = 1$, giving $h_0(t) = 1/\lambda$. The scale λ is chosen so that median survival is approximately 3.5 years, which for the exponential distribution implies:

$$\lambda \approx \frac{3.5}{\ln(2)} \approx 5.05 \quad \Rightarrow \quad h_0(t) \approx 0.198.$$

   This yields a constant hazard consistent with a moderate event rate and most events occurring before $t_{\max} \approx 6$ years.
2. *Weibull – increasing hazard:* We set $p = 1$, $k = 2.0$. The hazard increases linearly with time. The scale λ is tuned so that median survival is ≈ 3.5 years:

$$\lambda \approx \frac{3.5}{(\ln 2)^{1/2}} \approx 4.2.$$

3. *Weibull – decreasing hazard:* We set $p = 1$, $k = 0.5$. The hazard decreases over time, modelling situations with high early risk that declines. Scale λ tuned for median ≈ 3.5 years:

$$\lambda \approx \frac{3.5}{(\ln 2)^2} \approx 7.3.$$

4. *Log-normal – unimodal hazard:* We approximate the log-normal case via the $k \to 0$ limit of the GG, using the standard log-normal parameterisation for intuition. We choose parameters so that the hazard peaks at approximately 2.5 years and the median survival is ≈ 3.5 years:

$$\mu \approx \ln(2.5) \approx 0.92, \quad \sigma \approx 0.6.$$

   This shape models hazards that rise to a peak and then decline.
5. *Bathtub-shaped hazard:* We use the full GG flexibility with (p, k) set to produce high early hazard, a dip in mid-follow-up, and a late hazard increase. A reasonable choice is $p \approx 2.5$, $k \approx 0.5$, with λ tuned for median ≈ 3.5 years. This reflects situations such as post-surgical mortality patterns or device failure curves.

By fixing the median survival at ≈ 3.5 years for all shapes and calibrating λ accordingly, we ensure comparability of scenarios while preserving distinct hazard patterns. This allows us to evaluate estimator robustness across constant, monotone, unimodal, and bathtub-shaped hazards under controlled event-time distributions.

- **Non-parametric baseline hazard:** We additionally consider a smooth, non-monotone baseline hazard specified using cubic B-splines. The log-hazard is expressed as

$$\log h_0(t) = \gamma_0 + \sum_{j=1}^{J} \gamma_j B_j(t),$$

where $B_j(t)$ are spline basis functions with internal knots placed at 1.5, 3, and 5 years. We choose coefficients $(\gamma_1, \gamma_2, \gamma_3)$ to produce a hazard that increases early, decreases in mid-follow-up, and rises slightly towards the end, yielding a smooth non-monotone shape. The intercept γ_0 is calibrated so that the marginal median survival is approximately 3.5 years, ensuring comparability to the parametric baseline hazards. This design allows us to examine robustness of the estimators when the baseline hazard deviates from common parametric forms while remaining realistic and clinically plausible.

The combination of parametric and non-parametric baselines ensures that our simulation scenarios include both structured, theoretically grounded hazard shapes and more flexible, data-adaptive ones. This variety allows us to assess the robustness of time-dependent net benefit estimation methods across realistic hazard configurations.

2.3 Censoring Mechanism Generation

Censoring times are generated from a proportional hazards (PH) model of the form:

$$h_C(t \mid Z_1, Z_2) = h_{0C}(t) \exp\left(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2\right),$$

where $h_{0C}(t)$ denotes the baseline hazard of censoring. We specify an **exponential** baseline hazard,

$$h_{0C}(t) = \lambda_C,$$

which corresponds to a constant baseline hazard of censoring over time. The exponential baseline is sufficient for our purposes, as the aim of this component of the simulation is not to explore the effect of different censoring hazard shapes, but rather to study the impact of censoring that is *informative* versus *non-informative* on time-dependent net benefit estimation.

The parameters γ_0, γ_1, and γ_2 govern the overall level of the censoring hazard and its association with the covariates. In all scenarios, we fix γ_0 to a constant reference value that has no substantive effect on our comparisons, and vary only γ_1 and γ_2 to introduce or remove informativeness:

- **Non-informative censoring:** $\gamma_1 = \gamma_2 = 0$, so the censoring hazard does not depend on Z_1 or Z_2.
- **Informative censoring:** $\gamma_1 \neq 0$ and/or $\gamma_2 \neq 0$, so the censoring hazard depends on covariates that also affect event times.

The parameters $\gamma_0, \gamma_1, \gamma_2$ control the overall level of censoring and its dependence on the covariates. The intercept term γ_0 determines the baseline censoring hazard in the proportional hazards model. Since the baseline hazard rate λ_C is already tuned to achieve the desired marginal censoring proportion, we fix $\gamma_0 = 0$ for simplicity. This ensures that γ_1 and γ_2 can be interpreted purely as multiplicative effects relative to this baseline.

The parameter γ_1 governs the effect of the continuous covariate Z_1 on the censoring hazard. For interpretability, we express this as a hazard ratio for a one-standard-deviation (1-SD) increase in Z_1. Moderate covariate effects in survival analysis often correspond to hazard ratios in the range 1.5–2.0, which translate to γ_1 values of approximately 0.5 (HR $\approx$ 1.65) to 0.693 (HR $\approx$ 2.0). To maintain comparability with the event-time model, we set $\gamma_1 \approx 0.5$ in the primary scenarios.

The parameter γ_2 controls the effect of the binary covariate Z_2 on the censoring hazard, interpreted as the log hazard ratio comparing $Z_2 = 1$ to $Z_2 = 0$. Using the same reasoning, we set $\gamma_2 \approx 0.5$ in the primary scenarios, corresponding to a moderate hazard ratio of approximately 1.65.

These choices ensure that the magnitude of covariate effects on censoring is comparable to their effects on event risk, making the informative censoring scenario realistic and clinically plausible. To explore the impact of stronger or weaker dependence of censoring on the covariates, we also vary γ_1 and γ_2 in sensitivity analyses, increasing or decreasing their values while re-tuning λ_C to maintain the target overall censoring proportion. This allows us to assess the robustness of time-dependent net benefit estimation to different degrees of censoring informativeness.

When comparing scenarios with the same survival time distribution but different censoring mechanisms (informative vs non-informative), it is essential that the *overall censoring proportion* be the same in both scenarios. This ensures a fair comparison by isolating the effect of censoring informativeness rather than confounding it with differences in censoring rate. To achieve this, we adjust the baseline rate λ_C separately for each censoring mechanism so that the resulting proportion of censored individuals is matched to a pre-specified target (e.g., 20%, 30%, or 40%). This tuning is done numerically during the simulation setup so that across matched scenarios, the marginal censoring rate is identical despite differences in the covariate-dependence of the censoring process.

2.4 Risk Score Generation

For each simulated dataset, we generate a risk score for every individual based on a fitted Cox proportional hazards model using the simulated survival data. This procedure mirrors the way risk scores would be obtained in a real study from observed data.

Let t_0 denote the fixed evaluation time. The steps for generating the risk score are as follows:

1. **Fit a Cox model:** We fit a Cox proportional hazards regression model with the simulated covariates Z_1 and Z_2 as predictors:
$$h(t \mid Z_1, Z_2) = h_0(t) \exp\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2\right),$$
where β_0 is the intercept term and β_1, β_2 are the covariate effects. The model is fitted using the observed follow-up time $Z_i = \min(T_i, C_i)$ and event indicator δ_i for each subject.

2. **Estimate baseline survival:** From the fitted Cox model, we obtain the estimated baseline survival function $\widehat{S}_0(t)$ via the Breslow estimator. We then evaluate $\widehat{S}_0(t)$, the estimated baseline survival at the evaluation time t.

3. **Compute individual survival probability:** For each subject, we compute the linear predictor:
$$\widehat{\eta}_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2}.$$
The predicted survival probability at t for subject i is:
$$\widehat{S}_i(t) = \widehat{S}_0(t)^{\exp(\widehat{\eta}_i)}.$$

4. **Convert to predicted risk:** The predicted risk at t_0 for subject i is defined as:
$$\widehat{\pi}_{i,t} = 1 - \widehat{S}_i(t),$$
which represents the estimated probability of experiencing the event by t_0 given the subject's covariates.

The resulting $\widehat{\pi}_{i,t}$ values are used as the risk scores in the net benefit calculations. This approach ensures that risk estimation in the simulation mirrors common practice in applied survival analysis, where model-based predicted probabilities are derived from a fitted Cox model that includes an intercept term β_0.

2.5 Sample Size and Censoring Rate Scenarios

To comprehensively assess the performance of estimators under varying censoring conditions, we consider simulation scenarios with target marginal censoring rates of 10 % 20%, 30%, and 40%. This range spans from moderate to substantial censoring levels commonly encountered in practical survival analyses. Evaluating performance across these values allows us to investigate how increasing censoring impacts estimator bias.

We also vary the sample size across $n = 100, 500, 1000$, and 2000 subjects per simulated dataset. This range covers both small-sample settings, where estimation variability and finite-sample bias may be substantial, and large-sample settings, where asymptotic properties should dominate. Including both extremes enables us to examine the finite-sample behavior of the estimators and their convergence towards asymptotic performance.

2.6 Randomness and Reproducibility

A fixed random seed is used throughout all simulation scenarios to ensure reproducibility and to avoid replicating the same datasets across different settings.

3 Estimand – E

Our estimand is the *true* time-dependent net benefit at a fixed decision threshold z and evaluation time t_0, which we aim to estimate using different methods under varying simulation scenarios. The true net benefit serves as the benchmark in our simulation study, representing the value we would obtain if the event status of all individuals by time t_0 were known without censoring.

3.1 Definition of Net Benefit

For a binary outcome measured at a fixed time point, the net benefit (NB) of using a prediction model with decision threshold z can be expressed as a function of event prevalence, sensitivity, and specificity:

$$NB(z) = \text{prev} \cdot Se - (1 - \text{prev}) \cdot (1 - Sp) \cdot \frac{z}{1 - z},$$

where:

- prev is the proportion of individuals who experience the event,

- Se is the sensitivity, i.e., the probability of correctly classifying an event as high risk,
- Sp is the specificity, i.e., the probability of correctly classifying a non-event as low risk.

In the time-to-event setting, each component becomes time-dependent. The *time-dependent* net benefit at threshold z and time t is:

$$NB(z,t) = \text{prev(t)} \cdot Se(z,t) - (1 - \text{prev(t)}) \cdot (1 - Sp(z,t)) \cdot \frac{z}{1-z},$$

where prev_t, Se_t, and Sp_t denote the prevalence, sensitivity, and specificity evaluated at time t.

An equivalent and often more computation-friendly expression for the time-dependent net benefit is:

$$NB(z,t) = \frac{1}{n}\sum_{i=1}^{n} I\left(\pi_{i,t} > z\right)\left[\left(1 - S\left(t \mid \pi_{i,t} > z\right)\right) - S\left(t \mid \pi_{i,t} > z\right) \cdot \frac{z}{1-z}\right],$$

where $S(t \mid \pi_{i,t} > z)$ is the survival probability at time t among those with predicted risk exceeding z. This alternative formulation makes explicit the decomposition of net benefit into event and non-event components conditional on the decision rule $\pi_{i,t} > z$.

3.2 Time-Dependent Prevalence, Sensitivity, and Specificity

We adopt the **cumulative sensitivity** and **dynamic specificity** (C/D) definitions for survival data:

$$\text{prev(t)} = P(T_i \le t),$$

$$Se(z,t) = P(\pi_{i,t} > z \mid T_i \le t), \qquad Sp(z,t) = P(\pi_{i,t} \le z \mid T_i > t),$$

where:

- $\pi_{i,t}$ is the predicted risk score for individual i at time t,
- T_i is the event time for individual i,
- z is the decision threshold.

The cumulative sensitivity is the probability that an individual has a risk score exceeding z among those who experienced the event before or at time t. The dynamic specificity is the probability that an individual has a risk score less than or equal to z among those who remain event-free beyond time t. The prevalence prev_t is the proportion of individuals who experience the event by time t (cumulative incidence).

3.3 Population-Level Definition

The estimand is defined at the *population level* under the joint distribution of event times, censoring times, and marker values. In the simulation, this population quantity is approximated by the "true" value computed from the full simulated dataset before censoring is applied. We adopt the *opt-in* net benefit definition, treating "treat-none" as the reference strategy; thus, $NB = 0$ corresponds to no clinical benefit relative to treating no one. The decision threshold z is interpreted as the predicted risk level above which treatment would be initiated.

4 Estimation – M

For each simulated dataset, we fix the evaluation time to $t = 1$ and the decision threshold to $z = 0.25$. We then calculate the time-dependent net benefit $NB(z,t)$ using two alternative estimation approaches:

1. **Kaplan–Meier (KM) method** – suited for *non-informative* censoring.
2. **Inverse Probability of Censoring Weighting (IPCW) method** – suited for *informative* censoring.

Here, π_{i,t_0} denotes the predicted risk score for individual i at time t_0, $Z_i = \min(T_i, C_i)$ is the observed follow-up time, and $\delta_i = I(T_i \le C_i)$ is the event indicator.

4.1 KM method

The KM-based estimators of time-dependent cumulative sensitivity and dynamic specificity (Blanche et al., 2013) are:

$$\widehat{Se}_{KM}(z,t) = \frac{\left\{1 - \widehat{S}(t \mid \pi_{i,t} > z)\right\}\left(1 - \widehat{F}_{\pi_{i,t}}(z)\right)}{1 - \widehat{S}(t)},$$

$$\widehat{Sp}_{KM}(z,t) = \frac{\widehat{S}(t \mid \pi_{i,t} \leq z)\ \widehat{F}_{\pi_{i,t}}(z)}{\widehat{S}(t)},$$

where:

- $\widehat{S}(t \mid \pi_{i,t} > z)$ and $\widehat{S}(t \mid \pi_{i,t} \leq z)$ are the KM survival estimates at t in the respective subgroups,
- $\widehat{F}_{\pi}(z)$ is the empirical CDF of the predicted risk scores at z.

The KM estimate of the event prevalence at t is:

$$\widehat{\text{prev}_{KM}}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq t, \delta_i = 1),$$

i.e., the proportion of individuals who experience the event before or at t.

4.2 IPCW method

The IPCW-based estimators of time-dependent cumulative sensitivity and dynamic specificity (Blanche et al., 2013) are:

$$\widehat{Se}_{IPCW}(z,t) = \frac{\sum_{i=1}^{n} I(\pi_{i,t} > z, Z_i \leq t, \delta_i = 1)\ .\ w_i}{\sum_{i=1}^{n} I(Z_i \leq t, \delta_i = 1)\ .\ w_i},$$

$$\widehat{Sp}_{IPCW}(z,t) = \frac{\sum_{i=1}^{n} I(\pi_{i,t} \leq z, Z_i > t)\ .\ w_i'}{\sum_{i=1}^{n} I(Z_i > t)\ .\ w_i'},$$

where $\hat{S}_C(\cdot \mid \pi_{i,t})$ is the estimated censoring survival probability conditional on the predicted risk score.

The IPCW-adjusted event prevalence at t is:

$$\widehat{\text{prev}}^{\text{adj}}_{IPCW}(t) = \frac{1}{n}\sum_{i=1}^{n} w_i\ I(Z_i \leq t, \delta_i = 1),$$

with:

$$w_i = \hat{S}_C(Z_i \mid \pi_{i,t})^{-1}, w_i' = \hat{S}_C(t \mid \pi_{i,t})^{-1}$$

As stated earlier, we model the censoring distribution $\hat{S}_C$ using a Cox proportional hazards model, with the same predictors as in the event-time model, to capture the censoring survival function conditional on π_{i,t_0}. Accurate modeling of $\hat{S}_C$ is essential for IPCW to correctly adjust for informative censoring.

The IPCW weight $w_i = \hat{S}_C(Z_i \mid \pi_{i,t})^{-1}$ is the inverse of the estimated probability of remaining uncensored up to the observed event time Z_i for subject i. This is necessary because, to contribute to the prevalence estimate, a case must remain uncensored until its event time. If censoring depends on the predicted risk score $\pi_{i,t}$, higher-risk individuals may have different censoring probabilities; weighting by w_i rebalances the observed sample to represent the target population under informative censoring.

4.3 Proposed Algorithms and Bias Calculation

All could be found in the algorithms file.

5 Performance Measures – P

Our main objective in this simulation study is to evaluate the *bias* of time-dependent net benefit (NB) estimates under different censoring scenarios. The key scientific conclusion we aim to assess is:

> *Under informative censoring, the KM method fails, while IPCW, if correctly specified, produces appropriate estimates.*

For each simulated dataset, we have:

- The **true** net benefit, computed from a large uncensored dataset generated under the same scenario.
- The NB estimate from the **KM** method.
- The NB estimate from the **IPCW** method.

Bias definition

For each replication and each method, we compute the *absolute bias*:

$$\text{Absolute Bias} = \left|\overline{NB} - NB_{\text{true}}\right|,$$

which avoids the issue of sign cancellation that can occur when aggregating bias across replications.

Performance summaries

Across multiple replications for each simulation scenario, we report:

- **Mean absolute bias**: the average of the absolute bias values over all replications.
- **Standard error (SE) of bias**: the empirical standard deviation of bias values across replications.
- **Monte Carlo standard error (MCSE)** of the mean bias:

$$\text{MCSE}(\overline{\text{bias}}) = \frac{\widehat{\text{SD}}(\text{bias values})}{\sqrt{B}},$$

  where B is the number of simulation replications for the scenario.

Determining the number of simulation replications

We determine the required number of replications B using the general approach illustrated in the reference material (Appendix A.6 in the example). For each simulation scenario:

1. Perform a **pilot simulation** (e.g., 200 replications).
2. For each method, compute the variance of the bias values.
3. Identify the **worst-case variance** $\widehat{V}_{\max}$ across all methods and scenarios considered in the pilot.
4. Choose a target Monte Carlo standard error for the mean bias, e.g. MCSE ≤ 0.001.
5. Compute the required B as:

$$B_{\text{required}} = \frac{\widehat{V}_{\max}}{\text{MCSE}^2_{\text{target}}}.$$

6. Round B_{required} up to the nearest integer.

This ensures that the reported bias estimates are stable to the desired numerical precision across all scenarios.

6 Competing-Risk Extension (Cause-Specific Cox PH)

We extend the design to K competing causes (K=2 in most scenarios). For each subject, let (T, J) denote the event time and the cause label ($J \in \{1, \ldots, K\}$), with censoring time C. The observed data are

$$Z \;=\; \min(T, C), \qquad \delta \;=\; I(T \le C), \qquad \delta_k \;=\; I(\delta{=}1)\, I(J{=}k).$$

Throughout, we use conditional IPCW under the assumption

$$C \;\perp\!\!\!\perp\; (T, J) \mid X$$

(or equivalently conditional on a function of X, e.g., the predicted risk $\pi_{k,i,t}$).

6.1 Cause-Specific Data-Generating Mechanism

For cause $k \in \{1, \ldots, K\}$, we generate a latent event time T_k from a cause-specific proportional hazards (PH) model:

$$h_k(t \mid Z_1, Z_2) \;=\; h_{0k}(t)\, \exp\big(\beta_{0k} + \beta_{1k} Z_1 + \beta_{2k} Z_2\big),$$

with baseline hazard $h_{0k}(t)$ chosen from the same menu as in Section D (e.g., generalized gamma family to obtain constant, monotone, unimodal, or bathtub shapes). We set

$$T \;=\; \min_k T_k, \qquad J \;=\; \arg\min_k T_k.$$

Censoring times are generated from

$$h_C(t \mid Z_1, Z_2) \;=\; h_{0C}(t)\, \exp\big(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2\big), \qquad h_{0C}(t) = \lambda_C,$$

tuning λ_C to match target marginal censoring proportions (e.g., 10%, 20%, 30%, 40%). For comparability across scenarios, calibrate $h_{0k}(t)$ (and/or β_{0k}) so that target cause-specific cumulative incidences $F_k(t_{\max})$ are achieved at $t_{\max} \approx 6$ years (e.g., split total incidence across causes).

6.2 Cause-Specific Risk Score Generation (Development/Validation)

To develop and validate a risk score for a *cause of interest* k at time t, we fit a **cause-specific Cox PH** model for that cause, treating other causes as censoring:

$$h_k(t \mid X) \;=\; h_{0k}(t)\, \exp(\eta_k), \qquad \eta_k \;=\; \beta_{0k} + \beta_{1k} Z_1 + \beta_{2k} Z_2.$$

Let $\widehat{\Lambda}_{0k}(t)$ be the Breslow estimate and define

$$\widehat{\Lambda}_k(t \mid X_i) \;=\; \widehat{\Lambda}_{0k}(t)\, \exp(\widehat{\eta}_{k,i}), \qquad \widehat{S}(t \mid X_i) \;=\; \exp\Big(-\sum_{j=1}^{K} \widehat{\Lambda}_j(t \mid X_i)\Big).$$

The plug-in Aalen–Johansen (AJ) estimator of the **cause-k cumulative incidence** is

$$\widehat{F}_k(t \mid X_i) \;=\; \int_0^t \widehat{S}(u^- \mid X_i)\, d\widehat{\Lambda}_k(u \mid X_i),$$

computed numerically via Breslow increments. We set the cause-k **predicted risk** (score) as

$$\pi_{k,i,t} \;=\; \widehat{F}_k(t \mid X_i).$$

6.3 Cause-Specific Estimand for tNB

For cause k and threshold $z \in (0, 1)$,

$$\mathrm{prev}_k(t) \;=\; P(T \le t, J{=}k) \;=\; F_k(t), \qquad Se_k(z, t) \;=\; P(\pi_{k,i,t} > z \mid T \le t, J{=}k),$$

$$Sp_k(z, t) \;=\; P(\pi_{k,i,t} \le z \mid T > t),$$

and the **cause-k time-dependent net benefit** is

$$NB_k(z, t) \;=\; \mathrm{prev}_k(t)\, Se_k(z, t) \;-\; \big(1 - \mathrm{prev}_k(t)\big)\, \big(1 - Sp_k(z, t)\big)\, \frac{z}{1 - z}.$$

Here, controls are *dynamic*: individuals event-free of any cause at time t.

6.4 Estimation Methods for Competing Risks

(a) AJ-based plug-in (non-informative censoring). When censoring is non-informative,

$$\widehat{\text{prev}}_{k,AJ}(t) \;=\; \widehat{F}_k(t) \quad \text{(Aalen–Johansen)},$$

$$\widehat{Se}_{k,AJ}(z,t) \;=\; \frac{\widehat{F}_k(t \mid \pi_{k,i,t} > z)\,\big(1-\widehat{F}_{\pi_k}(z)\big)}{\widehat{F}_k(t)}, \qquad \widehat{Sp}_{k,AJ}(z,t) \;=\; \frac{\widehat{S}(t \mid \pi_{k,i,t} \le z)\,\widehat{F}_{\pi_k}(z)}{\widehat{S}(t)},$$

where $\widehat{S}(t) = 1 - \sum_{j=1}^{K} \widehat{F}_j(t)$ and $\widehat{F}_{\pi_k}(z)$ is the empirical CDF of $\pi_{k,i,t}$ at z.

(b) IPCW (covariate-dependent informative censoring). Let $\widehat{S}_C(u \mid \pi_{k,i,t})$ (or $\widehat{S}_C(u \mid X_i)$) be the estimated *censoring* survival. Define

$$w_i \;=\; \widehat{S}_C(Z_i \mid \pi_{k,i,t})^{-1}, \qquad w_i' \;=\; \widehat{S}_C(t \mid \pi_{k,i,t})^{-1}.$$

Then

$$\widehat{\text{prev}}_{k,IPCW}(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} w_i\, I(Z_i \le t,\ \delta_i{=}1,\ J_i{=}k),$$

$$\widehat{Se}_{k,IPCW}(z,t) \;=\; \frac{\sum_{i=1}^{n} I(\pi_{k,i,t} > z,\ Z_i \le t,\ \delta_i{=}1,\ J_i{=}k)\, w_i}{\sum_{i=1}^{n} I(Z_i \le t,\ \delta_i{=}1,\ J_i{=}k)\, w_i},$$

$$\widehat{Sp}_{k,IPCW}(z,t) \;=\; \frac{\sum_{i=1}^{n} I(\pi_{k,i,t} \le z,\ Z_i > t)\, w_i'}{\sum_{i=1}^{n} I(Z_i > t)\, w_i'}.$$

Finally,

$$\widehat{NB}_{k,IPCW}(z,t) \;=\; \widehat{\text{prev}}_{k,IPCW}(t)\,\widehat{Se}_{k,IPCW}(z,t) \;-\; \big(1-\widehat{\text{prev}}_{k,IPCW}(t)\big)\,\big(1-\widehat{Sp}_{k,IPCW}(z,t)\big)\,\frac{z}{1-z}.$$

Implementation notes. (1) Use the same censoring-model choices as in the single-event setting (e.g., Cox PH, flexible parametric, ML) but condition on X or on $\pi_{k,i,t}$ to capture covariate dependence. (2) For development, compute $\pi_{k,i,t}$ from the fitted cause-k Cox model; for validation, keep the model fixed and only compute scores. (3) If multiple causes are of interest, report $NB_k(z,t)$ for each k separately; $\widehat{S}(t)$ for specificity always refers to being event-free of *any* cause at t.

7 References

References