Simulation Study Design

Haochen Ding

1. Data Generation Plan

Each simulated dataset contains n individuals with covariates X, a binary sensitive attribute A, true event time T, censoring time C, and the observed outcome (Y, Δ).

Core variables:

- **Sensitive attribute (eg. sex):**

$$A \sim \text{Bernoulli}(p_A)$$

  p_A is proportion of group A.

- **Covariates:**

$$X = (X_1, X_2, ...) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

  μ and σ^2 depends on covariate itself's value range.

- **True event time:** Weibull proportional hazards model

$$\eta = \beta_A A + \beta_X^\top X, \qquad T = \Big(\frac{-\log U}{\lambda \exp(\eta)}\Big)^{1/k}, \qquad U \sim \text{Uniform}(0, 1),$$

  with λ (baseline rate) and k (shape parameter) controlling time scale and hazard monotonicity.

- **Censoring time:**

$$C \sim \text{Exp}\big(\mu \exp(\alpha\, \eta)\big),$$

- **Group-specific maximum follow-up:**

$$C_{\max}(A) = \begin{cases} C_{\max}(A_0), & A = 0, \\ C_{\max}(A_1), & A = 1. \end{cases}$$

  $C_{\max}(A_0) = C_{\max}(A_1)$ if there is no unbalanced follow-up.

- **Observed outcome:**

$$Y = \min(T, C, C_{\max}(A)), \quad \Delta = I\{T \le \min(C, C_{\max}(A))\}.$$

2. Model Assumptions

Scenario I: Under assumptions

- **Non-informative censoring:**

$$C \perp T \mid X, A$$

  Set $\alpha = 0$, so that $C \sim \text{Exp}(\mu)$, independent assumption holds.

- **No true group effect:**
  Set $\beta_A = 0$. Both groups have identical hazard functions and therefore identical survival distributions.

- **Balanced follow-up:** Set $C_{\max}(A_0) = C_{\max}(A_1)$.

Scenario II: Violating assumptions

- **Informative censoring:**
  Set $\alpha > 0$, so that $C \sim \text{Exp}\big(\mu \exp(\alpha\, \eta)\big)$. Since $\eta = \beta_X^\top X + \beta_A A$, censoring becomes correlated with event time and covariates, violating independent censoring assumption.

- **True group effect:**
  Set $\beta_A > 0$. Individuals with $A = 1$ experience systematically higher hazard (shorter expected survival time).

- **Unbalanced follow-up:** Set $C_{\max}(A_0) \neq C_{\max}(A_1)$. So two groups have different systematic maximum follow up time.

3. Design Choices

Weibull-distributed event times. Weibull distribution is used for generating event times because it offers a strong balance between realism and analytical simplicity. It flexibly models increasing or constant hazards through its shape parameter k. The scale parameter λ directly changes the baseline event rate. Also, Weibull model preserves the proportional hazards structure assumed by models like FCPH/FDCPH.

Exponential censoring mechanism. Censoring times are drawn from an exponential distribution for its constant hazard and direct interpretability. The rate parameter μ controls the average follow-up duration ($1/\mu$).

Normally distributed covariates. Covariates representing continuous clinical variables such as age or blood pressure. These data ofthen follows normal distribution when size is large. So normal distribution ensuring realistic variability across subjects while maintaining analytical simplicity.

Monte Carlo replications. Apply Monte Carlo design, generate datasets with different seeds / sample size under these conditions. Run models on these datasets and summarize scores based on all trails.

4. Parameters and Expected Effects

Parameter	Symbol	Example Values	Expected Effect
Sample size	n	5000	Larger n reduces variance.
Weibull shape	k	1.5	Controls hazard monotonicity.
Baseline rate	λ	0.004	Larger $\lambda \rightarrow$ earlier events.
Censoring rate	μ	0.1	Adjusts overall censoring proportion.
Group effect	β_A	0, 0.5	Introduces true hazard differences.
Covariate effect	β_X	Depends on X	Values reflects how covariate increase changes hazard
Risk–censoring link	α	0, 0.5	Higher $\alpha \rightarrow$ stronger violation of independence.
Follow-up cutoff	$C_{\max}(A)$	(10,10), (15,10)	Larger gap $\rightarrow$ stronger group disparity.
Group proportion	p_A	0.5, 0.7	Tests robustness to imbalance.
Fairness strength (in model)	γ, λ	0, 0.3, 0.5	Higher values increase fairness regularization.

5. Model Application

FAST (Do et al., 2023). FAST minimizes mutual information between the predicted survival representation $\hat{T}_\theta$ and sensitive attribute A:

$$\min_\theta \big[-\ell_{\text{surv}}(\theta) + \gamma\, I(A, \hat{T}_\theta) \big],$$

where ℓ_{surv} is the log-likelihood (eg. Cox). It enforces demographic parity by removing dependence between predictions and A. It uses (X, Y, Δ) for survival loss and $(A, \hat{T})$ for fairness penalty.

FCPH/FDCPH (Keya et al., 2023). These models regularize the risk or hazard outputs of Cox or DeepCox models:

$$\min_\theta \big[-L_X(\theta) + \lambda F_X(\theta) \big],$$

where F_X encodes group, individual, or intersectional fairness penalties $(F_g, F_i, F_\varepsilon)$. It uses (X, Y, Δ) for Cox likelihood and A in the fairness constraint. FDCPH replaces linear risk with a neural network for nonlinear relationships.

FIDP/FIPNAM ...