Multiple Imputation for Incomplete Survival Data with Missing Covariates

Toward Valid Causal Inference

Jooho Kim

Department of Statistics
Seoul National University

June 28, 2025

Outline

1. Multiple Imputation for Incomplete Survival Data
   - Weighted Analysis vs Imputation
   - Multiple Imputation
   - Multiple Imputation in Sub-sample Study

2. Causality of Hazard Ratio
   - Against Causal Interpretation of Hazard Ratio
   - For Causal Interpretation of Hazard Ratio

Motivation

- High-cost covariates (e.g., Blood test samples, Genomic data) are infeasible to collect for all members in the cohort.
- Commonly used sampling strategies are Nested case-control (NCC) and Case-cohort (CC) designs.

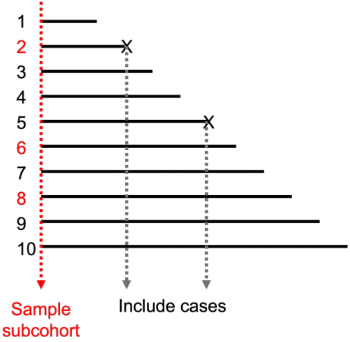
Units	Z	T	X_j
1			
$\vdots$			X_j^{obs}
n_j			
n_j+1			
$\vdots$			X_j^{mis}
N			

Table 1: Green = Observed, Red = Missing

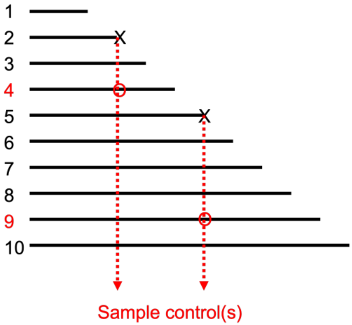
Two Sampling Designs

- In case-cohort design, a random **subcohort** is drawn at the beginning of the study and all cases outside the subcohort are included.

- In nested case-control design, controls are randomly sampled from those still at risk at each failure time.



(a) Case-cohort sampling (CC)



(b) Nested case-control sampling (NCC)

Notation

Symbol	Description
X_j^{obs}	Observed subvector of the expensive covariate $\in \mathrm{R}^{n_j}$
X_j^{mis}	Missing subvector of the expensive covariate $\in \mathrm{R}^{N-n_j}$
$X_j^{(m)}$	The mth imputed $X_j^{\mathrm{mis}} \in \mathrm{R}^{N-n_j}$
Z^i	Cheap covariates for unit $i \in \mathrm{R}^q$
T	Observed survival time
δ	Event type; failure$(=1)$ and censored$(=0)$
$\tilde{R}(t)$	Sampled risk set at time t

Weighted Partial Likelihood

- With the sub-sample data, we want to fit Cox proportional hazards model to quantify the hazard of the expensive covariate.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \prod_{i:\delta_i=1} \frac{\exp\left(\beta_{X_j} X_j^{\mathrm{obs},i} + \beta_Z Z^i\right)}{\sum_{k\in\tilde{R}(t_i)} w_k \exp\left(\beta_{X_j} X_j^{\mathrm{obs},k} + \beta_Z Z^k\right)}$$

where $w_i = \begin{cases} \delta_i + (1-\delta_i)\,\tilde{N}/\tilde{n} & \text{for CC} \\ \delta_i + (1-\delta_i)\left(1 - \prod_{k:t_k<t_i}\left(1 - \frac{m\delta_k}{n_{t_k}-1}\right)\right)^{-1} & \text{for NCC} \end{cases}$

- However, weighted partial likelihoods do not make use of cheap covariates.

Weighted Analysis vs Imputation

- Instead, we can **impute** the expensive covariate and use the full cohort.

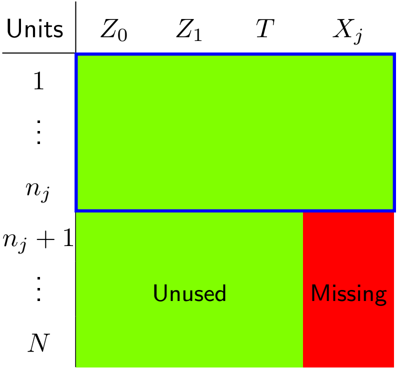


Table 2: Weighted Analysis

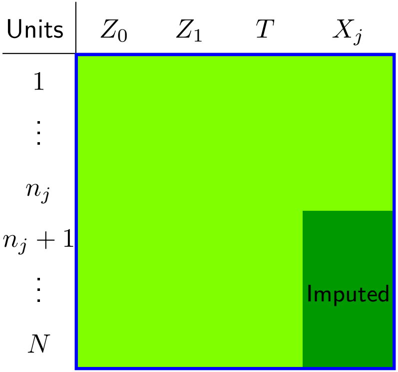


Table 3: Imputation

Multiple Imputation

1. Impute the missing value M times
2. Fit Cox PH model on each imputed data set,

$$\lambda^{(m)}(t) = \lambda_{\text{base}}(t) \exp(\hat{\beta}_{X_j}^{(m)} X_j + \hat{\beta}_{Z_1}^{(m)} Z_1)$$

3. Combine log hazard ratios, $\hat{\boldsymbol{\beta}}^{(m)}$, using Rubin's rule.

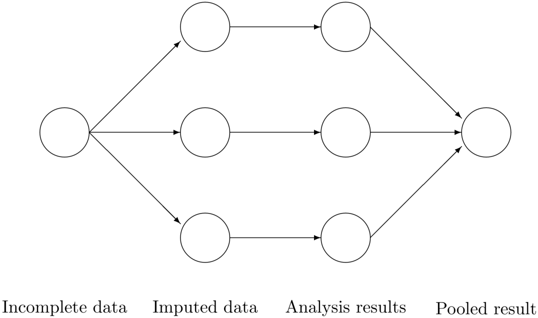


Figure 2: Main steps in multiple imputation

Multivariate Imputation by Chained Equation

- Let's look at how we obtain a single imputed data set in MICE algorithm.

Algorithm 1 MICE (Van Buuren, 2012)

Input: Incomplete cohort data with $\boldsymbol{X}^{\text{mis}}$
Output: Single imputed data set

```
for t = 1, ..., k do
    for j = 1, ..., p do
        Sample θ_j^(t) ~ π_j(θ_j | X_j^obs, X_{-j}^(t), Z, δ, T)
                       ∝ f_j(X_j^obs | X_{-j}^(t), Z, δ, T, θ_j) p_j(θ_j)
        Sample X_j^(t) ~ f_j(X_j^mis | X_{-j}^(t), Z, δ, T, θ_j^(t))
    end for
end for
```

for $t = 1, \ldots, k$ **do**
  for $j = 1, \ldots, p$ **do**
    Sample $\theta_j^{(t)} \sim \pi_j(\theta_j \mid X_j^{\text{obs}}, \boldsymbol{X}_{-j}^{(t)}, Z, \delta, T)$
    $\propto f_j(X_j^{\text{obs}} \mid \boldsymbol{X}_{-j}^{(t)}, Z, \delta, T, \theta_j)\, p_j(\theta_j)$
    Sample $X_j^{(t)} \sim f_j(X_j^{\text{mis}} \mid \boldsymbol{X}_{-j}^{(t)}, Z, \delta, T, \theta_j^{(t)})$
  end for
end for

where $\boldsymbol{X}_{-j}^{(t)} = (X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)}) \in \mathrm{R}^{(N-n_j)\times(p-1)}$

- MICE algorithm is different from Gibbs sampler. In Gibbs sampler,

$$\theta_j^{(t)} \sim \pi_j\big(\theta_j \mid X_j^{\text{obs}}, X_j^{(t-1)}, \boldsymbol{X}_{-j}^{(t)}, Z, \delta, T\big)$$

Rubin's Rule

Rubin's Combining Rule

The combined **mean** and **variance** estimates for β:

$$\hat{\beta} := \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}, \qquad \mathrm{var}(\hat{\beta}) := \bar{V} + \left(1 + \frac{1}{M}\right)B$$

where $\bar{V}$ and B are within and between imputation variances, respectively.

- Rubin's combining rule is based on Bayesian derivation.

$$p\big(\beta \mid X^{\mathrm{obs}}\big) = \int p\big(\beta \mid X^{\mathrm{mis}}, X^{\mathrm{obs}}\big)\, p\big(X^{\mathrm{mis}} \mid X^{\mathrm{obs}}\big)\, \mathrm{d}X^{\mathrm{mis}}$$

$$E\big(\beta \mid X^{\mathrm{obs}}\big) = E\Big[E\big(\beta \mid X^{\mathrm{mis}}, X^{\mathrm{obs}}\big) \mid X_{\mathrm{obs}}\Big]$$

$$\mathrm{Var}\big(\beta \mid X^{\mathrm{obs}}\big) = E\Big[\mathrm{Var}\big(\beta \mid X^{\mathrm{mis}}, X^{\mathrm{obs}}\big) \mid X^{\mathrm{obs}}\Big] + \mathrm{Var}\Big[E\big(\beta \mid X^{\mathrm{mis}}, X^{\mathrm{obs}}\big) \mid X^{\mathrm{obs}}\Big]$$

Rubin's Rule

- By $p\left(\beta \mid X^{\text{obs}}\right) \approx \frac{1}{M}\sum_{m=1}^{M} p\left(\beta \mid X^{(m)}, X^{\text{obs}}\right)$ where $X^{(m)} \sim p\left(X^{\text{mis}} \mid X^{\text{obs}}\right)$, we easily derive the Rubin's combining rule.

$$\begin{aligned} E\left(\beta \mid X^{\text{obs}}\right) &\approx \int \beta \frac{1}{M} \sum_{m=1}^{M} p\left(\beta \mid X^{(m)}, X^{\text{obs}}\right) \mathrm{d}\beta \\ &= \frac{1}{M} \sum_{m=1}^{M} E\left(\beta \mid X^{(m)}, X^{\text{obs}}\right) = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}^{(m)} \\ &=: \hat{\beta} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{Var}\left(\beta \mid X^{\text{obs}}\right) &\approx \frac{1}{M} \sum_{m=1}^{M} V^{(m)} + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2 \\ &= \bar{V} + \left(1 + \frac{1}{M}\right) B \\ &=: \text{var}(\hat{\beta}) \end{aligned} \tag{2}$$

Is Multiple Imputation Adequate in this Setting?

- For multiple imputation (MI) to be valid, we need several assumptions.
  1. **Missing at Random (MAR)** assumption
     $$\mathbb{P}(R = 1 \mid X, Z, \delta, T) = \mathbb{P}(R = 1 \mid Z, \delta, T)$$
     where R is missingness indicator and X is expensive covariate
  2. **Proper imputation** (Rubin, 1987)
  3. **Congeniality** (Meng, 1994) or **Compatibility** (Liu et al. 2014)
- Remember that the sampling designs rely on the failure indicator δ, and the failure time T.
- MAR assumption is met since the expensive covariate is missing by design!

What is Congeniality(Compatibility)?

Congeniality (Informal)

Analyst's model $P_A(\beta \mid X^{\text{com}})$ is **congenial** to imputer's model $P_I(X^{\text{mis}} \mid X^{\text{obs}})$ if:

1. $\mathrm{E}_A(\beta \mid X) = \hat{\beta}(X^{\text{com}}), \quad \mathrm{Var}_A(\beta \mid X) = \hat{V}(X^{\text{com}})$
2. $P_A(X^{\text{mis}} \mid X^{\text{obs}}) = P_I(X^{\text{mis}} \mid X^{\text{obs}})$

- A statistician can be the imputer and an epidemiologist can be the analyst.
- *Unconongeniality is generally "a rule not the exception".* (Xie & Meng, 2017)
- **Imputation model should be more general than the analysis model.**

MI Strategies for Sub-sample Study

- Keogh & White (2013) introduce two different imputation procedures.

  1. **Approximate Imputation Model**

  $$X = \theta_0 + \theta_Z^{\mathrm{T}} Z + \theta_\delta \delta + \theta_{\delta Z}^{\mathrm{T}} \delta Z + \theta_T \Lambda_{\text{base}}(T) + \theta_{ZT}^{\mathrm{T}} Z \Lambda_{\text{base}}(T) + \epsilon$$

  2. **Rejection Sampling**
     We add a rejection sampling step to the above imputation model.

- How should we construct the rejection rule?

  Target density

  $$f(X_j \mid \boldsymbol{X}_{-j}, Z, T, \delta) \propto f(T, \delta \mid X_j, \boldsymbol{X}_{-j}, Z, \beta) f(X_j \mid \boldsymbol{X}_{-j}, Z, \theta_j)$$

  Proposal density

  $$f(X_j \mid \boldsymbol{X}_{-j}, Z, \theta_j)$$

Rejection Sampling Method

- **Ratio of target density to proposal density**

$$\frac{f(T, \delta \mid X_j, \boldsymbol{X}_{-j}, Z, \beta) f(X_j \mid \boldsymbol{X}_{-j}, Z, \theta_j)}{f(X_j \mid \boldsymbol{X}_{-j}, Z, \theta)} = f(T, \delta \mid X_j, \boldsymbol{X}_{-j}, Z, \beta)$$
$$< c(T, \delta, \boldsymbol{X}_{-j}, Z, \beta)$$

- **Accept if,**

$$\begin{aligned} U &\leq \frac{f(T, \delta \mid X_j^*, \boldsymbol{X}_{-j}, Z, \beta)}{c(T, \delta, \boldsymbol{X}_{-j}, Z, \beta)} \\ &= \exp(-\Lambda_{\text{base}} e^{g(X_j^*, \boldsymbol{X}_{-j}, Z, \beta)}) \\ &= S^*(t) \end{aligned}$$

where $U \sim \text{Unif}(0, 1)$ and X_j^* is the imputed covariate.

Recent Work on MI for Sub-sample Study

- Borgan et al. (2023) proposed performing imputation only for the randomly selected **super-sample in order to reduce computational burden**.

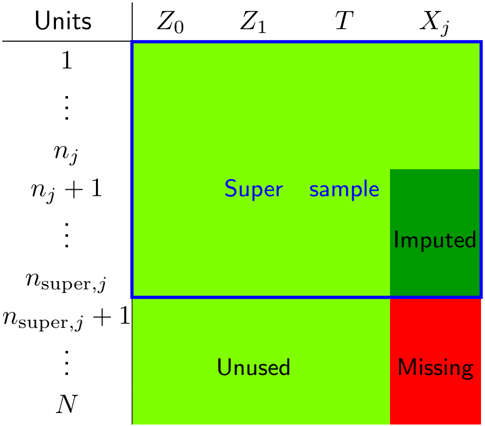


Table 4: Super-sample Weighted Analysis

Ongoing Work on MI for Sub-sample Study

- My current research extends the work of Borgan et al. (2023) by sampling units with **high influence** on both the estimator of interest($\hat{\beta}$) and the accuracy of imputation.

- Imputation Model Loss (Miao et al. 2021)

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(f_\theta(\boldsymbol{x}_i) - \boldsymbol{x}_i\right)^\top \operatorname{diag}(\boldsymbol{m}_i) \left(f_\theta(\boldsymbol{x}_i) - \boldsymbol{x}_i\right)}{2 \left\|\boldsymbol{m}_i\right\|_2^2}$$

- Influence Function for log hazard ratio (Reid & Crepeau, 1985)

$$\hat{\beta} - \beta = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IF}_i + o_p(N^{-1/2}), \quad \operatorname{var}(\hat{\beta}) \approx \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{IF}_i^2$$

- **Objective function:** $\operatorname{argmax}_{V_1, \ldots, V_{N-n}} (\sum_{i=1}^{N-n} V_i \, \mathrm{IF}_i^2 - \lambda \sum_{i=1}^{N-n} V_i)$ where V_i is a sampling indicator

Outline

Causal Interpretation of Hazard Ratio

- Why is hazard ratio (HR) difficult to causally interpret?
  1. HR may change over time.
  2. Period-specific HRs have **selection bias**.

$$\lambda_a(t) = \lim_{h \to 0} \frac{\Pr\left[t \leq T_i(a) < t + h \mid T_i(a) \geq t\right]}{h} \tag{3}$$

$$\mathrm{HR} = \frac{\lambda_1(t)}{\lambda_0(t)} = \lim_{h \to 0} \frac{\Pr\left[t \leq T_i(1) < t + h \mid {\color{red} T_i(1) \geq t}\right]}{\Pr\left[t \leq T_i(0) < t + h \mid {\color{blue} T_i(0) \geq t}\right]} \tag{4}$$

	$T(0) \geq t$	$T(0) < t$
$T(1) \geq t$	Always survivor	Protected
$T(1) < t$	Harmed	Never survivor

Table 5: Principal Strata at Time t

Marginal HR vs Conditional HR

- **Marginal HR**

$$\lambda(t) = \lambda_0(t)\exp(\beta_A A)$$

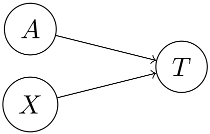
- **Conditional HR**

$$\lambda^\star(t) = \lambda_0^\star(t)\exp(\beta_A^\star A + \beta_X X)$$

- **Non-collapsibility**

$$\hat{\beta}_A \neq \hat{\beta}_A^\star$$

even if the following DAG is true:



where A is treatment and X is covariate, and T is survival time.

When does HR have causal interpretations?

Causal Proportional hazards assumption (Fay & Li, 2024)

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta) \quad \forall t$$

- Under this assumption, the following holds,

$$\frac{\log S_1(t)}{\log S_0(t)} = \exp(\beta) \quad \forall t$$

- Hazard ratio, $\exp(\beta)$, is a **population-level causal estimand** if the proportional hazards assumption holds.

- Note that this assumption is different from the usual PH assumption,

$$\frac{\lambda(t \mid x+1)}{\lambda(t \mid x)} = \exp(\beta_x)$$

Causal Hazard Ratio

Causal HR (Martinussen et al. 2020)

$$\mathrm{HR}_{\text{causal}} = \frac{\lambda_1^*(t)}{\lambda_0^*(t)} = \lim_{h\to 0} \frac{\Pr\left[t \le T_i(1) < t+h \mid T_i(1) \ge t, T_i(0) \ge t\right]}{\Pr\left[t \le T_i(0) < t+h \mid T_i(1) \ge t, T_i(0) \ge t\right]}$$

- Causal HR is the ratio of instantaneous risk at t for always survivors.
- However, it is not nonparametrically identifiable without strong assumptions.

	$T(0) \ge t$	$T(0) < t$
$T(1) \ge t$	Always survivor	Protected
$T(1) < t$	Harmed	Never survivor

Table 6: Principal Strata at Time t

Conclusion of Hazard Ratio

- In practice, researchers should be cautious in interpreting hazard ratios causally.
- **Alternative causal estimands**:
  - Survival probability causal effect(SPCE) at time t (Mao et al. 2018):

$$\Delta^{SPCE}(t) = S_1(t) - S_0(t)$$

  - Restricted average causal effect (RACE):

$$\Delta^{RACE}(\tau) = \int_0^\tau S_1(t)dt - \int_0^\tau S_0(t)dt$$

- These can offer more robust causal interpretations, especially under non-proportional hazards.

References I

- Self, S. G., & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case–cohort studies. *The Annals of Statistics*, 16(1), 64–81.
- Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case–control studies. *Biometrika*, 84(2), 379–394.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Van Buuren, S., & Van Buuren, S. (2012). Flexible imputation of missing data (Vol. 10, p. b1182). Boca Raton, FL: CRC Press.
- Murray, J. S. (2018). Multiple imputation: a review of practical and theoretical findings. (Technical report)
- Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.
- Xie, X., & Meng, X. L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, 27(4), 1485–1545.
- Keogh, R. H., & White, I. R. (2013). Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. *Statistics in Medicine*, 32(23), 4021–40413.

References II

- Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & Alzheimer's Disease Neuroimaging Initiative* (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4), 462–481.
- Borgan, Ø., Keogh, R. H., & Njøs, A. (2023). Use of multiple imputation in supersampled nested case–control and case–cohort studies. *Scandinavian Journal of Statistics*, 50(1), 13–37.
- Reid, N., & Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika*, 72(1), 1–9.
- Miao, X., Wu, Y., Chen, L., Gao, Y., Wang, J., & Yin, J. (2021). Efficient and effective data imputation with influence functions. *Proceedings of the VLDB Endowment*, 15(3), 624–632.
- Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1), 13–15.
- Fay, M. P., & Li, F. (2024). Causal interpretation of the hazard ratio in randomized clinical trials. *Clinical Trials*, 21(5), 623–635.
- Martinussen, T., Vansteelandt, S., & Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26, 833–855.
- Mao, H., Li, L., Yang, W., & Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine*, 37(26), 3745–3763.

Thank you for your attention!