Scalable and Efficient Multiple Imputation for Case-Cohort Studies via Influence-Based Supersampling

Jooho Kim

Department of Statistics
Seoul National University

The Korean Statistical Society, December 2025

Outline

1. Multiple Imputation for Case-Cohort Studies
2. Proposed Methodology
   - Influence Function-Based Supersampling
   - Weights for analysis
3. Simulation Results
4. Real Data Application

Some biomarkers are expensive to measure

Some biomarkers are expensive to measure

- **Cox proportional hazards model:** $\lambda(t) = \lambda_0(t)\exp\big(\boldsymbol{\beta}_{\boldsymbol{Z}}^\top \boldsymbol{Z} + \boldsymbol{\beta}_X^\top X\big)$ where X is expensive covariate and $\boldsymbol{Z}$ are low-cost covariates

Case-cohort sampling design

- A **case-cohort sample ($\mathcal{CC}$)** consists of a random subcohort ($\mathcal{SC}$) and all cases ($\mathcal{D}$) outside the subcohort.

- Covariates are **missing at random (MAR)** for individuals outside the case-cohort sample.

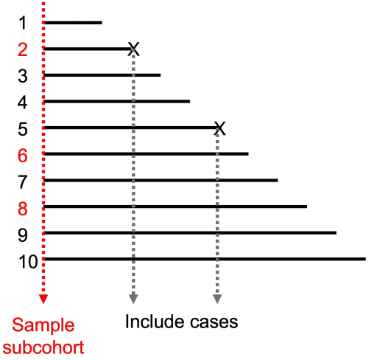


Figure 1: Case-cohort (CC) sampling

Analysis of case-cohort studies

Weighted partial likelihood

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \prod_{i=1}^{n_0} \prod_{t>0} \left\{ \frac{\exp\left(\boldsymbol{\beta}_{\boldsymbol{Z}}^\top \boldsymbol{Z}_i + \boldsymbol{\beta}_{\boldsymbol{X}}^\top \boldsymbol{X}_i\right)}{\sum_{j \in R(t)} w_j Y_j(t) \exp\left(\boldsymbol{\beta}_{\boldsymbol{Z}}^\top \boldsymbol{Z}_j + \boldsymbol{\beta}_{\boldsymbol{X}}^\top \boldsymbol{X}_j\right)} \right\}^{dN_i(t)},$$

where $w_j = \begin{cases} \dfrac{N-D}{n_{sc}-d} & \text{if } j \in \mathcal{SC} \setminus \mathcal{D} \\ 1 & \text{if } j \in \mathcal{D}, \end{cases}$ and $R(t) = \{i \in \mathcal{CC} \mid Y_i(t) = 1\}$

- **Sample size notation:**
  full cohort (Ω) : N
  subcohort ($\mathcal{SC}$) : n_{sc}
  cases in full cohort ($\mathcal{D}$) : D
  case-cohort sample ($\mathcal{CC}$) : $n_0 = n_{sc} + D$
  cases in subcohort: d

What does the data look like?

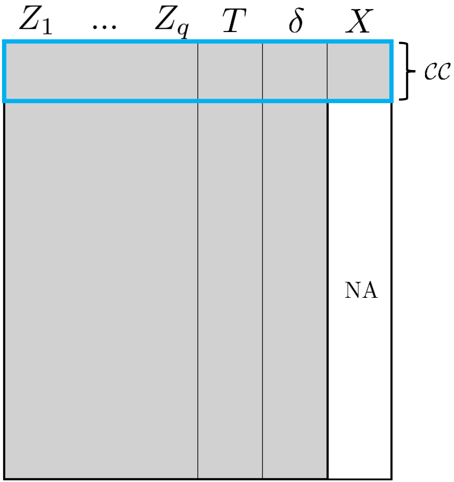


Figure 2: NA: missing, (T, δ): response variable, $\mathcal{CC}$: case-cohort sample

Using the full data through imputation

- **Multiple Imputation (MI)** is used to impute missingness.

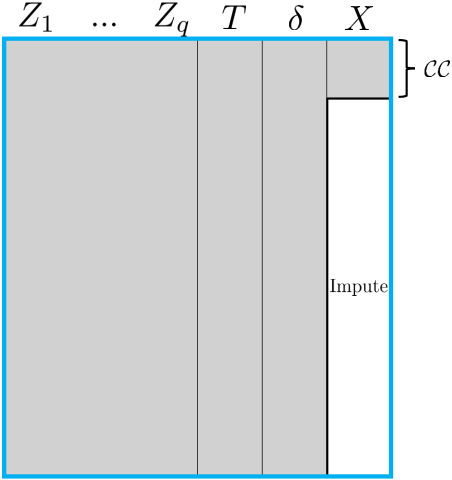
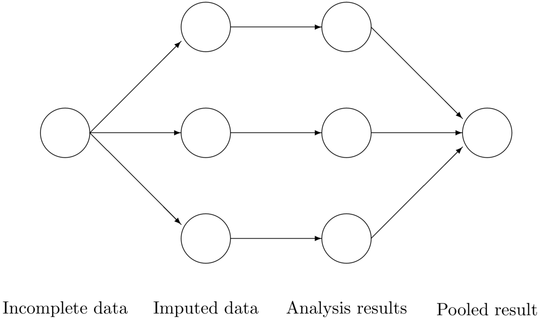


Figure 3: (T, δ): response variable, $\mathcal{CC}$: case-cohort sample

What is multiple imputation?



Incomplete data Imputed data Analysis results Pooled result

1. Impute the missing value M times (e.g., M=10)

2. Fit Cox model on each imputed data set, $\hat{\beta}^{(m)}, \; \forall m = 1, \ldots, M$

3. Combine estimators using Rubin's rule
$$\hat{\beta} \coloneqq \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}^{(m)}, \;\; \mathrm{var}(\hat{\beta}) \coloneqq \frac{1}{M}\sum_m V^{(m)} + (\frac{M+1}{M})\frac{1}{M-1}\sum_m (\hat{\beta}^{(m)} - \hat{\beta})^2$$

How do we obtain a single imputed dataset?

- **Multivariate Imputation by Chained Equation (MICE)**

Algorithm 1 MICE (Van Buuren, 2012)

Input: Incomplete dataset with $\boldsymbol{X}^{\mathrm{mis}}$

Output: Single imputed data set

```
for $j = 1, \ldots, p$ do
    for $\ell = 1, \ldots, L$ do
        Sample $\theta_j^{(\ell)} \sim \pi(\theta_j \mid \boldsymbol{X}_i^{(\ell-1)}, \boldsymbol{Z}_i, \delta_i, T_i; i \in \mathcal{SC})$
        Sample $X_{ij}^{(\ell)} \sim f(X_{ij} \mid \boldsymbol{X}_{i,-j}^{(\ell-1)}, \boldsymbol{Z}_i, \delta_i, T_i, \theta_j^{(\ell)}; i \in \Omega \setminus \mathcal{CC})$
    end for
end for
```

where $\boldsymbol{X}_{i,-j}^{(\ell)} = (X_{i1}^{(\ell)}, \ldots, X_{i,j-1}^{(\ell)}, X_{i,j+1}^{(\ell-1)}, \ldots, X_{ip}^{(\ell-1)})$

Nonlinear or interaction terms can induce bias in MICE

- **Compatibility** between imputation and analysis models

Nonlinear or interaction terms can induce bias in MICE

- **Compatibility** between imputation and analysis models
- **Substantive model compatible fully conditional specification (SMC-FCS)** by Bartlett et al. (*Stat Methods Med Res*, 2015)
- Accept imputed value $X_{ij}^{(\ell)}$ if

$$U \leq= \exp(-\Lambda_0(T)e^{g(X_{ij}^{(\ell)}, \boldsymbol{X}_{i,-j}, \boldsymbol{Z}_i, \beta)})$$

where $U \sim \mathrm{Unif}(0, 1)$.

Computational burden of multivariate missing data

- **High computational cost** of SMC-FCS

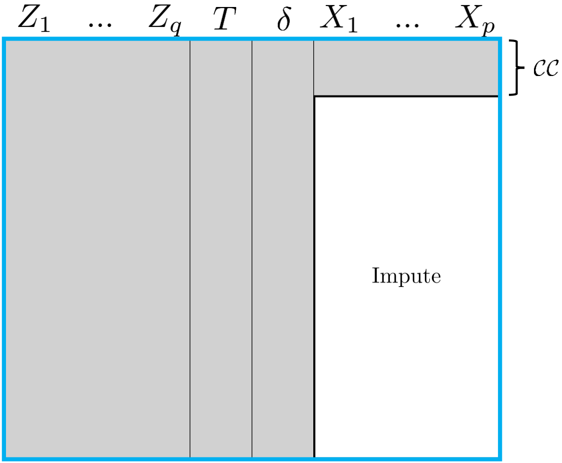


Figure 4: High-dimensional expensive covariates

Supersampling is helpful but...

- **Random supersampling** (RSS) of Borgan et al. (*Scand J Stat*, 2023)
- **Efficiency loss** of random supersampling

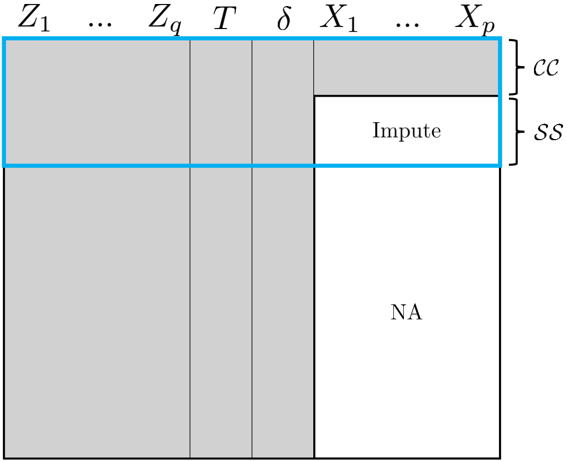


Figure 5: $\mathcal{CC}$: case-cohort sample, $\mathcal{SS}$: supersample

Influence function-based supersampling (ISS)

Influence function (IF)

The influence function ψ measures the first-order sensitivity of an estimator to an infinitesimal contamination at a point.

- We use IF to select observations **influential to the target parameter** (e.g., hazard ratio).
- For subsequent analysis, **probabilistic sampling** is required rather than deterministic selection.
- We want to find the **optimal inclusion probability** π_i^* for unit i.

Minimizing variance in the sampling stage

- Hazard ratio $\hat{\boldsymbol{\beta}}$ is an **asymptotically linear** estimator with influence function $\boldsymbol{\psi}_i$,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\psi}_i + o_p(1). \tag{1}$$

Minimizing variance in the sampling stage

- Hazard ratio $\hat{\boldsymbol{\beta}}$ is an **asymptotically linear** estimator with influence function $\boldsymbol{\psi}_i$,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\psi}_i + o_p(1). \tag{1}$$

- Using Horvitz–Thompson estimator for $\sum_{i=1}^{N} \boldsymbol{\psi}_i$ yields:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx \frac{1}{N^2} \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i^{\top} \tag{2}$$

with inclusion probability π_i

Minimizing variance in the sampling stage

- Hazard ratio $\hat{\boldsymbol{\beta}}$ is an **asymptotically linear** estimator with influence function $\boldsymbol{\psi}_i$,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\psi}_i + o_p(1). \tag{1}$$

- Using Horvitz–Thompson estimator for $\sum_{i=1}^{N} \boldsymbol{\psi}_i$ yields:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx \frac{1}{N^2} \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i^{\top} \tag{2}$$

  with inclusion probability π_i

- Minimizing the trace of the sampling variance leads to

$$\pi_i^* = \min\left\{\lambda \|\hat{\boldsymbol{\psi}}_i\|_2, \, 1\right\} \quad \text{subject to} \quad \sum_{i \in \Omega \setminus \mathcal{CC}} \pi_i = n_1 \tag{3}$$

  where n_1 is the supersample size.

Balanced sampling further improves efficiency

Balanced sampling (Deville and Tillé, 2004, *Biometrika*)

Find sampling indicator V_i subject to

$$\sum_{i \in \Omega \setminus \mathcal{CC}} \frac{V_i}{\pi_i^*} \boldsymbol{B}_i = \sum_{i \in \Omega \setminus \mathcal{CC}} \boldsymbol{B}_i, \tag{4}$$

for auxiliary variables $\boldsymbol{B}_i$,

- Using π_i^*, we **draw a supersample** that satisfies the balancing equations.
- We set auxiliary variables to the influence functions of low-cost covariates $\boldsymbol{B}_i = (\pi_i^*, \hat{\psi}_{i1}, \ldots, \hat{\psi}_{iq})$.

Calibrating the weights for unified analysis

Weight calibration (Deville and Särndal, 1992, *JASA*)

$$w_i^* = \underset{w_i}{\operatorname{argmin}} \sum_{i \in \Omega} \boldsymbol{V}_i \, d(w_i, w_i^0) \;\; \text{subject to} \; \sum_{i \in \Omega} \boldsymbol{V}_i w_i \boldsymbol{A}_i \;=\; \sum_{i \in \Omega} \boldsymbol{A}_i \qquad (5)$$

where $d(\cdot, \cdot)$: distance measure, $\boldsymbol{V}_i$: sampling indicator, $\boldsymbol{A}_i$: auxiliary variables.

- Weight calibration **enables unified analysis** while reducing variance.

$$\sum_{i \in \Omega \setminus \mathcal{D}} I\left(i \in \mathcal{SC} \setminus \mathcal{D}\right) w_i = (N - D) \frac{db_0}{db_0 + db_1}, \qquad (6)$$

$$\sum_{i \in \Omega \setminus \mathcal{D}} I(i \in \mathcal{SS}) w_i = (N - D) \frac{db_1}{db_0 + db_1}, \qquad (7)$$

$$\sum_{i \in \mathcal{D}} I(i \in \mathcal{D}) w_i = D, \qquad (8)$$

where db_0 and db_1 summarise influence in each subsample, $\mathcal{SC} \setminus \mathcal{D}$ and $\mathcal{SS}$.

Small Relative Bias of SMC-FCS

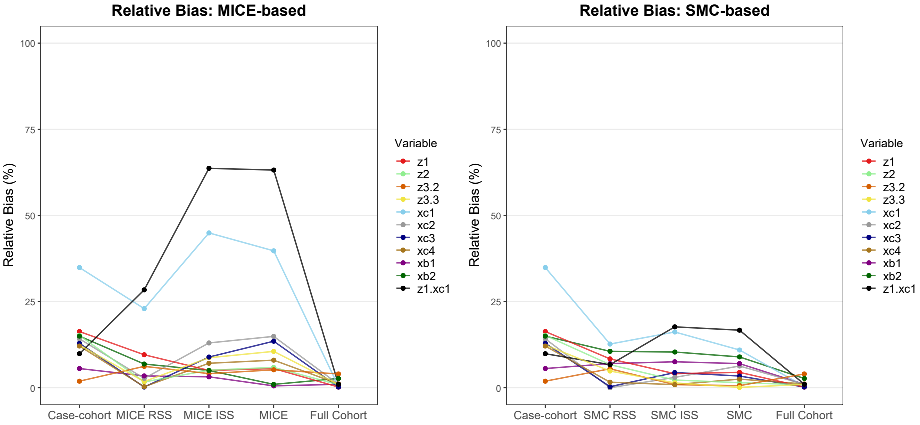
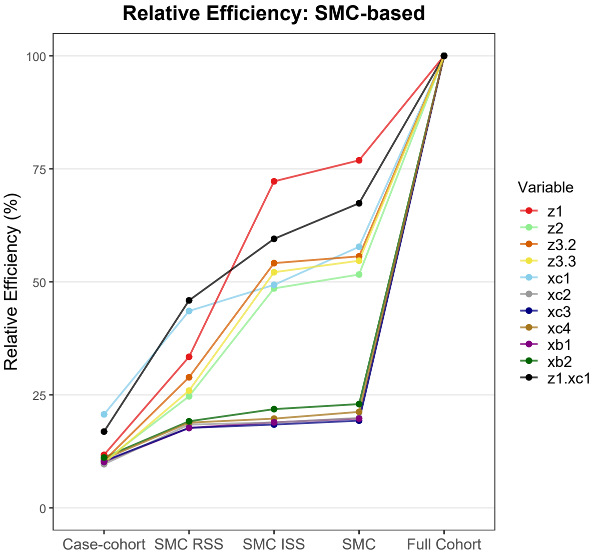


Figure 6: Interaction term in the analysis model

High relative efficiency of the proposed method



Real data analysis: NIH–AARP Diet and Health Study

Table 1: Runtime and bias of log hazard ratio estimates under SMC-FCS

	Runtime	**Sex**	**Race**			**Age group**		**Waist**	**Sex×Waist**
			Black	Hispanic	Asian/Other	60–64	65–71		
SMC	7.67 h	0.029	0.018	0.015	0.047	0.005	0.002	0.074	0.014
SMC RSS	7.84 min	0.058	0.252	0.106	0.004	0.014	0.023	0.071	0.001
SMC ISS	8.11 min	0.016	0.014	0.034	0.057	0.006	0.002	0.064	0.006

※ Smoking status, diabetes, and caloric intake are additionally adjusted for.
※ SMC RSS: Random supersampling in SMC-FCS
※ SMC ISS: Influence function-based supersampling in SMC-FCS

Discussion and future work

1. **Influence function-based sampling without case-cohort sampling**

Discussion and future work

1. **Influence function-based sampling without case-cohort sampling**
2. **Imputation model misspecification**

Discussion and future work

1. **Influence function-based sampling without case-cohort sampling**
2. **Imputation model misspecification**
3. **Beyond survival context, missing not at random (MNAR)**

Thank you for your attention!

How do we obtain a single imputed dataset?

- **Multivariate Imputation by Chained Equation (MICE)**

Algorithm 2 MICE (Van Buuren, 2012)

```
Input: Incomplete dataset with $\boldsymbol{X}^{\text{mis}}$
Output: Single imputed data set
for $j = 1, \ldots, p$ do
    for $\ell = 1, \ldots, L$ do
        Sample $\theta_j^{(\ell)} \sim \pi(\theta_j \mid \boldsymbol{X}_i^{(\ell-1)}, \boldsymbol{Z}_i, \delta_i, T_i;\, i \in \mathcal{SC})$
        Sample $X_{ij}^{(\ell)} \sim f(X_{ij} \mid \boldsymbol{X}_{i,-j}^{(\ell-1)}, \boldsymbol{Z}_i, \delta_i, T_i, \theta_j^{(\ell)};\, i \in \mathcal{SS})$
    end for
end for
```

where $\boldsymbol{X}_{i,-j}^{(\ell)} = (X_{i1}^{(\ell)}, \ldots, X_{i,j-1}^{(\ell)}, X_{i,j+1}^{(\ell-1)}, \ldots, X_{ip}^{(\ell-1)})$

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

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