Improving Efficiency and Stability in Nested Case-Control Studies for Rare Events Using Multiple Imputation

Sehee Jung[1]

[1]Department of Statistics and Data Science, Yonsei University, Seoul, South Korea.
[*]Address correspondence to: sh.jung9128@gmail.com

Abstract

Nested case-control (NCC) studies provide a cost-effective approach for investigating exposure-disease associations, particularly when collecting data from an entire cohort is infeasible. Traditional NCC analysis methods, such as the Thomas method and inverse probability weighting (IPW), offer theoretical consistency but face challenges in efficiency and stability, especially in the context of rare events. This study proposes multiple imputation (MI) as a robust alternative that utilizes all available cohort information, including surrogate variables and auxiliary covariates, to overcome these limitations. Simulation studies and a real-world application using ICU patient data demonstrate that MI reduces standard errors and improves precision compared to traditional methods. While simulation results revealed slight biases in MI estimates, empirical analyses highlighted its ability to produce less biased and more efficient estimates. These findings emphasize MI's potential as a reliable and effective method for NCC studies, striking a balance between robustness and precision while addressing the shortcomings of traditional NCC approaches.

Keywords: nested case-control study, multiple imputation, rare event

1 Introduction

In many large-scale prospective cohort studies, conducting expensive exposure measurements for all individuals is often impractical. Such exposures include biological measurements, genetic information, and precise dietary assessments in nutritional epidemiology. As a result, exposure-disease association studies are typically based on nested sub-studies within the cohort, which include all disease cases but only a subset of non-cases from the entire cohort. These sub-studies allow for the collection of complete exposure information.

Due to its cost-effectiveness in studying the temporal relationship between disease and exposure, nested case-control (NCC) sampling has emerged as a viable alternative to full cohort designs and case-control designs. In NCC studies, for each 'case' diagnosed with a specific disease, one or more controls are sampled from the risk set at the time of the case's event occurrence. The most widely used analytical approach for NCC data is Thomas's maximum partial likelihood estimation method, based on the Cox proportional hazards model [1, 2]. The consistency and asymptotic normality of Thomas's estimator have been rigorously established using counting process and martingale theory [3]. Further refinements to improve efficiency have also been proposed. For example, introduced the inverse probability weighting (IPW) method, where

individual log-likelihood contributions are weighted by the inverse probability of inclusion in the NCC study, resulting in greater estimation efficiency compared to Thomas's method [4].

Traditional nested case-control analyses, however, rely solely on data from the sampled individuals while ignoring valuable information available for the remaining individuals in the entire cohort. For instance, extensive data on low-cost covariates such as sex, height, weight, and smoking status may exist for the full cohort. Additionally, in some studies, high-cost exposure variables are partially missing for individuals outside the sub-study, but surrogate variables for these exposures may be available for the entire cohort. For example, while blood glucose levels, typically measured at high cost, may only be available for the sub-study participants, a surrogate measure such as mean arterial blood pressure, which is strongly correlated with blood glucose, could be collected across the full cohort. This limitation is particularly pronounced in studies involving rare events. In such scenarios, the small number of observed cases can lead to unstable estimates when applying traditional NCC analyses, worsening the challenges of achieving statistical precision and power.

To address these issues, this paper proposes a simple and effective approach that leverages multiple imputation (MI) to incorporate all available information from the entire cohort when analyzing nested case-control studies, especially in the presence of rare events. MI has gained increasing popularity in epidemiology and other research fields as a robust method for handling missing covariate data. In this study, we conceptualize the nested case-control study and the remaining cohort as a full cohort study with substantial missing data. Using this framework, we propose applying MI to integrate the full cohort information into the analysis of nested case-control and case-cohort studies. Missing data in the full cohort are imputed multiple times, and each imputed dataset is analyzed separately, with the resulting estimates combined to produce a final inference. The imputation models are constructed using the completely observed data from the nested case-control sample, allowing us to address the instability inherent in rare event studies while overcoming the limitations of existing methods.

The remainder of this paper is organized as follows. Section 2 summarizes the traditional analytical methods for nested case-control studies and explains how MI can be used to incorporate full cohort information into the analysis. Section 3 introduces the proposed method and presents simulation studies that evaluate our method against traditional NCC approaches. Section 4 provides an empirical application of the proposed method using health-related outcomes data in ICU-admitted patients. Finally, Section 5 discusses the results and presents our conclusions.

2 Method

2.1 Traditional analysis of nested case–control studies

We consider a single main exposure of interest Z_1 and a vector of adjustment variables $W = (W_1, \ldots, W_K)^\mathsf{T}$. Let $t_1 < t_2 < \cdots < t_I$ denote the ordered event times and i the identity of the case at time t_i. Throughout the paper, we assume a Cox proportional hazards model [5], that is, a hazard of the form

$$h(t; Z_1, W) = h_0(t) \exp\left(\beta_{Z_1} Z_1 + \beta_W^\mathsf{T} W\right), \tag{1}$$

where $h_0(t)$ is the baseline hazard and the log hazard ratios of interest are β_{Z_1} and $\beta_W = (\beta_{W_1}, \ldots, \beta_{W_K})^\mathsf{T}$. If Z_1 and W were available in the full cohort, the analysis would be using the partial likelihood [6]:

$$L_{\mathrm{FULL}} = \prod_i \frac{\exp\left(\beta_{Z_1} Z_{1,i} + \beta_W^\mathsf{T} W_i\right)}{\sum_{l \in R_i} \exp\left(\beta_{Z_1} Z_{1,l} + \beta_W^\mathsf{T} W_l\right)}, \tag{2}$$

where R_i denotes the risk set at t_i, that is, the set of all individuals with event or censoring time $\geq t_i$.

In nested case–control studies, m controls are sampled from $R_i \setminus \{i\}$ without replacement at each t_i, where $\delta_i = 1$. That is, for each case, m controls are randomly selected from the subjects still at risk at the time of the failure of the case. Note that the controls may include both failures and non-failures. Let S_i denote this set of m controls, and let $S = \{i : \delta_i = 1\} \cup \left(\bigcup_{i:\delta_i=1} S_i\right)$ denote all subjects included in the nested case–control study. Then, $\tilde{R}_i = R_i \cap S$ is the set of all subjects in the nested case–control study who are at risk at time t_i. Let n indicate the size of S. Thomas proposed maximizing the following partial likelihood to make inferences on β from nested case–control studies [1]:

$$L_{\mathrm{Thomas}}(\beta) = \prod_{i \in S} \left[\frac{\exp\left(\beta_{Z_1} Z_{1,i} + \beta_W^\mathsf{T} W_i\right)}{\sum_{j \in \{i\} \cup S_i} \exp\left(\beta_{Z_1} Z_{1,j} + \beta_W^\mathsf{T} W_j\right)} \right]^{\delta_i}. \tag{3}$$

The partial likelihood produces a model consistent estimator of log hazard ratios [7]. However, the denominators in the Thomas' likelihood use only $\{i\} \cup S_i$ and not all available subjects at risk, namely $\tilde{R}_i$. As we will see, such a 'partial risk set' approach is inherently inefficient because it uses only a subset of available subjects at risk. Samuelsen proposed maximizing the following partial likelihood using inverse probability weighting (IPW) [4]:

$$L_{\mathrm{IPW}}(\beta) = \prod_{i \in S} \left[\frac{\exp\left(\beta_{Z_1} Z_{1,i} + \beta_W^\mathsf{T} W_i\right)}{\sum_{j \in \tilde{R}_i} w_j \exp\left(\beta_{Z_1} Z_{1,j} + \beta_W^\mathsf{T} W_j\right)} \right]^{\delta_i}, \tag{4}$$

where $w_i = 1/p_i$, and p_i is the probability of subject i ever being included in the nested case–control study. Samuelsen proved the consistency of the resulting estimator: roughly, this is because the partial likelihood (4) is a design consistent estimator of Cox's partial likelihood (2), which in turn yields an estimating equation for a model consistent estimator for β. The normality of the estimator was left as a conjecture due to the complexity of the sampling scheme. To allow baseline hazard functions to differ across subgroups, I propose the use of the following partial likelihood where $S(i)$ is the index set of the subjects that share a common baseline hazard function with subject i:

$$L^S_{\mathrm{IPW}}(\beta) = \prod_{i \in S} \left[\frac{\exp\left(\beta_{Z_1} Z_{1,i} + \beta_W^\mathsf{T} W_i\right)}{\sum_{j \in \tilde{R}_i \cap S(i)} w_j \exp\left(\beta_{Z_1} Z_{1,j} + \beta_W^\mathsf{T} W_j\right)} \right]^{\delta_i}. \tag{5}$$

Kim computed the inclusion probabilities in a nested case–control study that ties in failure times. The probability for subject i was [8]:

$$p_i = \begin{cases} 1, & \text{if } \delta_i = 1, \\ 1 - \prod_{j:a_i < t_j < t_i} \left(1 - \min\left(1, \frac{m b_{ji}}{k_{ji} - b_{ji}}\right)\right), & \text{if } \delta_i = 0, \end{cases} \tag{6}$$

where k_{ji} is the size of $R_j \cap H_i$, where H_i is the set of subjects in the full cohort with the same values

of matching variables as subject i. In other words, k_{ji} is the number of subjects at risk at Y_j with the same values of the matching variables as subject i. Here, b_{ji} is the number of tied subjects in H_i that failed exactly at Y_j. In the absence of additional matching variables or ties in failure times, the inclusion probability simplifies to that of Samuelsen. The calculation of the minimum is for the late failure times when all subjects in $R_j \cap H_i$ are sampled because $k_{ji} - b_{ji} < mb_{ji}$.

2.2 Multiple imputation in the analysis of nested case–control studies

Multiple imputation was introduced by Rubin [9] and is now becoming widely used to handle missing exposure and covariate data in studies of different types, where the usual alternative approaches would be to restrict analysis to complete cases or, in the case of categorical exposures, to assign a 'missing' category. The key idea in using MI for missing exposure and covariate data is that the missing measurements are imputed by drawing a random value from the distribution of the missing variable conditional on all observed values. To account for the uncertainty in the imputed values, a number $D\,(D > 1)$ of imputed values are obtained for each missing data point, creating D complete imputed data sets. The resulting data sets are analysed separately but identically, and the resulting estimates are combined using 'Rubin's rules'. MI, as usually implemented, depends on the assumption that data are missing at random (MAR), that is, the missingness depends only on observed data. It has been shown that MI results in asymptotically unbiased estimates and correct standard errors provided that the imputation model is correctly specified. It is important that imputation models for missing exposure and covariate data include the outcome variable, for example, disease status [10]. A correctly specified imputation model is also compatible with the model for association between the explanatory variables and the outcome, which is referred to as congeniality [11].

We assume that there is a main univariate exposure of interest Z_1, which is observed in the nested case–control sample, but not in the remainder of the full cohort. All other covariates of interest for the analysis model, W, and any surrogates of the main exposure, Z_2, are assumed available in the full cohort. We emphasise, however, that the methods outlined do not technically depend either on there being any covariates, W, or on the existence of a surrogate exposure, Z_2. Any surrogate exposure is assumed to be independent, given Z_1, of the outcome.

We view the nested case–control sample, plus the remainder of the full cohort not in the sample, as a full-cohort study with a large amount of missing data. We propose imputing Z_1 using MI for individuals in the full cohort but not in the nested case–control sample. Note that data will be missing only in non-cases, as it is assumed that all cases observed in the full cohort are sampled into the nested case–control study. This results in several imputed data sets for the full cohort. A standard full-cohort analysis is then performed in each imputed data set, and the resulting hazard ratio estimates are combined. In the setting considered here, the missing data are missing by design, and hence, the MAR assumption required for MI is met trivially, provided the outcome variables are included in the imputation model.

We define $Z_{1,k}^{(d)}$ to be the value of the exposure Z_1 for the k-th individual in the d-th imputed data set. Where Z_1 is observed, therefore, $Z_{1,k}^{(d)}$ takes the same value for all $d\,(d = 1, \ldots, D)$. In a nested case–control study, Z_1 is observed for all cases and for the controls sampled at each event time. In the d-th imputed data set, the partial likelihood for the full cohort is

$$L_{\mathrm{FULL}(d)} = \prod_i \frac{\exp\left(\beta_{Z_1} Z_{1,i} + \beta_W^{\mathsf{T}} W_i\right)}{\sum_{l \in R_i} \exp\left(\beta_{Z_1} Z_{1,l}^{(d)} + \beta_W^{\mathsf{T}} W_l\right)}. \tag{7}$$

The resulting parameter estimates are denoted as $\hat{\beta}_{Z_1}^{(d)}$ and $\hat{\beta}_W^{(d)}$. Under Rubin's rules [9], pooled parameter estimates and standard errors are given by

$$\hat{\beta}_{Z_1} = \frac{1}{D}\sum_{d=1}^{D}\hat{\beta}_{Z_1}^{(d)}, \tag{8}$$

$$\text{var}(\hat{\beta}_{Z_1}) = A + \left(1 + \frac{1}{D}\right)B, \tag{9}$$

where A and B represent the within-imputation and between-imputation variance components, respectively. We have $A = \frac{1}{D}\sum_{d=1}^{D} A_d$, where A_d is the estimated variance of $\hat{\beta}_{Z_1}^{(d)}$, and $B = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\beta}_{Z_1}^{(d)} - \hat{\beta}_{Z_1}\right)^2$. Pooled estimates for $\hat{\beta}_W$ and corresponding standard errors are obtained in the same way.

Under MI, the imputed values for Z_1 in the full cohort are to be drawn from the conditional distribution of Z_1 given what is fully observed. Although it may seem unintuitive, it is crucial that the outcome, $\{\Delta, T\}$, in this case, is included in the imputation model [12]. Here, the outcome is $\{\Delta, T\}$, where $\Delta = 1$ for cases and $\Delta = 0$ for non-cases, and T denotes the event time for cases ($\Delta = 1$) and the censoring time for non-cases ($\Delta = 0$). The conditional distribution we wish to draw from is therefore the distribution of $Z_1 \mid Z_2, W, \Delta, T$. The distribution of $Z_1 \mid Z_2, W, \Delta, T$ is non-standard, and we discuss approximate imputation models which enables us to draw from the correct distribution.

A number of *ad hoc* methods for performing MI in proportional hazards regression models have been proposed, including using Δ and T or $\log(T)$ as variables in an imputation model. White and Royston [13] took a more principled approach and Keogh and White [10] extended this approach by including a surrogate exposure Z_2 in the imputation model. The fitting of an imputation model takes the form:

$$Z_1 = \theta_0 + \theta_{Z_2}Z_2 + \theta_W^{\mathsf{T}}W + \theta_\Delta\Delta + \theta_T H_0(T) + \theta_{WT}^{\mathsf{T}}WH_0(T) + \epsilon, \tag{10}$$

where $H_0(T)$ denotes the baseline cumulative hazard at observed time T, and ϵ are normally distributed residuals with mean 0 and variance σ_ϵ^2. For a given individual, an imputed value $Z_1^{(d)}$ is obtained using

$$Z_1^{(d)} = \hat{\theta}_0^{(d)} + \hat{\theta}_{Z_2}^{(d)}Z_2 + \hat{\theta}_W^{(d)T}W + \hat{\theta}_\Delta^{(d)}\Delta + \hat{\theta}_T^{(d)}H_0(T) + \hat{\theta}_{WT}^{(d)T}WH_0(T) + \epsilon^{(d)}, \tag{11}$$

where $\epsilon^{(d)}$ is a random draw from a normal distribution with mean 0 and variance $\sigma_\epsilon^{2(d)}$. Here, $\sigma_\epsilon^{2(d)}$, $\hat{\theta}_0^{(d)}$, $\hat{\theta}_{Z_2}^{(d)}$, $\hat{\theta}_W^{(d)}$, $\hat{\theta}_\Delta^{(d)}$, $\hat{\theta}_T^{(d)}$, and $\hat{\theta}_{WT}^{(d)}$ are draws from the posterior distributions of the parameter estimates from the imputation model. The imputation model in (2) was found to perform well in many circumstances, but to result in conservatively biased estimates when the covariates are strongly associated with the outcome. White and Royston found that replacing $H_0(T)$ by the Nelson–Aalen estimate of the cumulative hazard gave good results [13].

3 Simulation study

3.1 Description of the simulation

We assumed an underlying cohort of 30,000 individuals and considered the following simulation settings. Z_1 was the risk factor of interest, which was expensive to measure, and Z_2 served as its inexpensive surrogate. Z_1 and Z_2 were generated from a bivariate normal distribution with mean 0, unit variances, and correlation

$\mathrm{Corr}(Z_1, Z_2) = 0.75$. Additionally, two auxiliary covariates W_1 and W_2 were introduced. W_1 was drawn from a standard normal distribution, $W_1 \sim N(0,1)$, and W_2 was generated from a Bernoulli distribution with probability 0.5, $W_2 \sim \text{Bernoulli}(0.5)$. Event times were generated using a Cox proportional hazards model under a Weibull baseline hazard function. For event times T, the hazard function was specified as

$$\lambda(t|Z_1, W_1, W_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 W_1 + \beta_3 W_2), \tag{12}$$

where the baseline hazard $\lambda_0(t)$ followed a Weibull distribution with $\rho = 0.5$ and $\gamma = 1.5$, and the regression coefficients were given by $\beta_1 = \beta_2 = \log(2)$ and $\beta_3 = -1$. Random censoring times C were generated from an exponential distribution with rate parameter λ_c, which was chosen to achieve an approximate censoring rate of 99%. The observed time for each individual was defined as $X = \min(T, C)$, and the event indicator Δ was set to 1 if $T < C$, otherwise 0.

In the nested case–control (NCC) study, cases were defined as individuals experiencing an event ($\Delta = 1$), and controls were sampled from the risk set at the event time. Two control sizes were considered: 3 controls per case and 5 controls per case.

We also generated imputed datasets to account for the missing Z_1 values in the full cohort. Missing Z_1 values were imputed using a linear regression model with explanatory variables Z_2, W_1, W_2, Δ, and the Nelson–Aalen estimate of the baseline cumulative hazard $H_0(t)$. The imputation model took the following form:

$$Z_1 = \theta_0 + \theta_{Z_2} Z_2 + \theta_{W_1} W_1 + \theta_{W_2} W_2 + \theta_\Delta \Delta + \theta_T H_0(T) + \theta_{W_1 T}^{\mathsf{T}} W_1 H_0(T) + \theta_{W_2 T}^{\mathsf{T}} W_2 H_0(T) + \epsilon, \tag{13}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$. Posterior draws of the parameters θ were obtained to account for uncertainty in the imputation process, and multiple imputation (MI) was performed with $d = 10$ imputations. Further increasing the number of imputations did not result in any noticeable improvement in performance.

Four types of analyses were applied to the simulated data:

1. **Full-cohort analysis:** The full cohort data, where Z_1 is completely observed, is used to fit the Cox model.
2. **Thomas NCC analysis:** The NCC data are analyzed using Thomas' partial likelihood method.
3. **IPW-based NCC analysis:** The NCC data are analyzed using Cox regression with inverse probability weighting based on the inclusion probabilities.
4. **Multiple imputation analysis:** Missing Z_1 values in the NCC data are imputed using the proposed imputation model. The imputed datasets are analyzed, and the results are pooled using Rubin's rules.

The number of simulation replicates was set to 3000, and the resampling number was fixed at 2024. All simulations were performed using R.

3.2 Simulation study results

The simulation results are summarized in Table 1. The summary statistics are the average parameter estimates, the bias, the square root of the average squared standard errors (ASE), the empirical standard deviation of the 3000 estimates (ESD).

Table 1: Simultation study results

	Full Cohort				NCC-Thomas				NCC-IPW				MI			
	Mean	Bias	ASE	ESD	Mean	Bias	ASE	ESD	Mean	Bias	ASE	ESD	Mean	Bias	ASE	ESD
(1) Control Size: 3																
β_1	0.694	0.001	0.041	0.041	0.696	0.003	0.058	0.057	0.707	0.013	0.041	0.059	0.677	-0.016	0.043	0.040
β_2	0.693	0.000	0.041	0.041	0.696	0.003	0.058	0.059	0.707	0.014	0.041	0.059	0.687	-0.006	0.042	0.041
β_3	-1.001	-0.001	0.090	0.092	-1.004	-0.004	0.116	0.118	-1.016	-0.016	0.090	0.113	-0.994	0.006	0.091	0.091
(2) Control Size: 5																
β_1	0.694	0.001	0.041	0.041	0.696	0.003	0.053	0.053	0.708	0.015	0.041	0.053	0.681	-0.012	0.043	0.040
β_2	0.695	0.002	0.041	0.041	0.695	0.002	0.053	0.052	0.708	0.015	0.041	0.052	0.689	-0.004	0.042	0.041
β_3	-1.002	-0.002	0.090	0.090	-1.004	-0.004	0.106	0.106	-1.019	-0.019	0.090	0.106	-0.996	0.004	0.091	0.089

The results shows that NCC approaches suffer from a loss of efficiency compared to the Full Cohort method. This is expected because NCC relies on a subset of the data, which inherently reduces efficiency. This inefficiency is reflected in the higher ASE and ESD values observed for both NCC-Thomas and NCC-IPW when compared to the Full Cohort method.

Generally, NCC-IPW is known to achieve greater theoretical efficiency, though it tends to exhibit slightly higher bias [14]. For example, the theoretical variability (ASE) for NCC-IPW is smaller than that for NCC-Thomas, suggesting improved theoretical efficiency. However, the empirical variability (ESD) for NCC-IPW does not reflect this efficiency improvement, and in some cases, it appears inflated. This discrepancy between ASE and ESD for NCC-IPW is likely due to the instability caused by a high censoring rate and a low event occurrence (rare events). Under such conditions, the selection probabilities become small, leading to excessively large IPW weights and overestimation of ESD.

On the other hand, the MI method demonstrated significant improvements in efficiency compared to NCC-Thomas, although it exhibited slightly larger bias. Importantly, as the control size increased, the bias in MI reduced, owing to the increased amount of available data. This result highlights that MI can effectively improve efficiency while mitigating the limitations observed in NCC-IPW. Therefore, MI emerges as a promising alternative that is more efficient than NCC-Thomas and avoids the instability seen in NCC-IPW, particularly in the context of rare events.

4 Illustration: survival in ICU patients and blood glucose

Building on the findings from our simulation study, we extend our analysis to a real-world dataset to further evaluate the performance of the proposed methods. The analysis is conducted using the MIMIC-III dataset, which comprises de-identified health information from patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. The dataset includes survival time, defined as the duration between ICU admission and discharge, or the last recorded follow-up time when the exact discharge time is unavailable, leading to right-censored data. A total of 8,912 patients are included in this analysis.

The primary exposure of interest is blood glucose level (Glucose), which is only available for a subset of patients due to the practical limitations of blood sampling. To address this selective availability, mean blood pressure, a surrogate variable highly correlated with glucose, is used as an auxiliary measure. Additional covariates, including height, weight, temperature, heart rate, oxygen saturation, respiratory rate, and components of the Glasgow Coma Scale (eye-opening and verbal response), are included to adjust for

patient-specific characteristics.

Consistent with the setup in the simulation study, we employ a NCC design with a control size of 3 per case. Under this design, it is assumed that all observed cases from the full cohort are sampled, and missing data for the primary exposure (Glucose) occur exclusively among the non-cases. To handle the missing data and improve the robustness of our analysis, we apply multiple imputation (MI), leveraging both the surrogate variable and observed covariates to impute the unobserved glucose values.

Table 2: Results from a study of survival in ICU patients and blood glucose

	Full cohort		NCC-Thomas		NCC-IPW		MI	
	logHR	SE	logHR	SE	logHR	SE	logHR	SE
Glucose	-0.003	0.004	-0.001	0.005	-0.001	0.003	-0.005	0.004
Height	0.002	0.003	0.002	0.003	0.002	0.003	0.002	0.003
Weight	-0.002	0.001	-0.002	0.002	-0.001	0.001	-0.002	0.001
Temperature	-0.231	0.069	-0.217	0.082	-0.249	0.069	-0.227	0.069
Heart Rate	0.001	0.001	0.000	0.001	0.001	0.001	0.001	0.001
Oxygen saturation	-0.010	0.020	-0.012	0.022	-0.025	0.021	-0.009	0.020
Respiratory rate	-0.061	0.011	-0.055	0.012	-0.077	0.011	-0.061	0.011
Glasgow Coma Scale (eye.opening)	-0.081	0.041	-0.101	0.048	-0.083	0.042	-0.080	0.041
Glasgow Coma Scale (verbal response)	0.165	0.033	0.181	0.043	0.218	0.033	0.166	0.033

The results are shown in Table 2. Consistent with the findings from the simulation study, the use of MI) led to a reduction in standard errors (SE) compared to the NCC-Thomas approach. This effect is particularly evident for the primary exposure variable, Glucose, where the SE under MI is 0.004, compared to 0.005 under NCC-Thomas.

Interestingly, unlike the simulation results, the MI analysis demonstrates smaller deviations from the Full cohort estimates compared to those from both the NCC-Thomas and NCC-IPW methods. For instance, the logHR estimate for Glucose under MI is -0.005, which is closer to the Full cohort value of -0.003, whereas the NCC-Thomas and NCC-IPW methods yield logHR estimates of -0.001. This improved performance under MI can be attributed to the inclusion of covariates that are highly correlated with the primary exposure variable (Glucose). The inclusion of these covariates in the imputation model likely contributed to the process by providing additional information for the missing values.

In conclusion, the MI analysis not only reduces standard errors but also yields less biased estimates, particularly for the primary exposure of interest. These findings highlight the robustness and efficiency of MI in addressing challenges associated with the NCC design, aligning well with the theoretical advantages observed in the simulation study.

5 Discussion

This study demonstrates the potential of MI as an effective analytical approach for NCC studies, particularly in the context of rare events. By leveraging all available cohort information, including surrogate variables and auxiliary covariates, MI addresses the inherent inefficiencies of traditional NCC methods while maintaining model consistency. The findings from both simulation studies and the real-world application using ICU patient data highlight several key advantages and considerations for MI method.

The widely used Thomas and IPW methods in NCC studies highlight the context in which MI offers

significant advantages. The Thomas method relies on a partial likelihood based on partial risk sets, yielding consistent estimates but suffering from inefficiency due to its inability to fully utilize the available risk set information. The IPW method, on the other hand, achieves better theoretical efficiency by weighting contributions based on the inverse probability of inclusion. However, in the context of rare events, the IPW method often fails to perform reliably. This is because the high variability in inclusion probabilities can lead to unstable weights, resulting in inflated standard errors and decreased precision.

MI overcomes these issues by leveraging all available cohort information, effectively transforming the problem of missing exposure data into a manageable imputation task. The inclusion of surrogate variables and auxiliary covariates in the imputation model enhances the stability of estimates, particularly for rare events. In the empirical analysis using ICU patient data, MI not only reduced standard errors compared to both Thomas and IPW methods but also produced estimates closer to those derived from the full cohort analysis. This result highlights MI's robustness in real-world applications where missing data patterns are complex and event rates are low.

While simulation studies showed slight biases in MI estimates under certain settings, the practical advantages of MI in empirical studies, such as improved efficiency and reduced variability, underscore its value as a reliable analytical approach for NCC studies. By effectively addressing the instability of IPW and the inefficiency of Thomas, MI demonstrates its capability to balance robustness and precision, making it a promising method for analyzing NCC designs in epidemiological research.

Despite its advantages, MI is not without challenges. The validity of MI depends on correctly specifying the imputation model and the assumption that data are missing at random (MAR). While the MAR assumption is met in the context of NCC studies by design, careful consideration is required to ensure model congeniality and inclusion of all relevant variables. Furthermore, the computational cost of performing multiple imputations and pooling results can be non-trivial, particularly for large datasets. Future work should continue to refine and expand the use of MI to address the evolving needs of observational and clinical research.

The code used for the simulation studies in this paper is available at `https://github.com/sh-jung9128/NCC-MI`.

References

1. Liddell F, McDonald J, and Thomas D. Methods of cohort analysis: appraisal by application to asbestos mining. Journal of the Royal Statistical Society: Series A (General) 1977;140:469–83.
2. Oakes D. Survival times: aspects of partial likelihood. International Statistical Review/Revue Internationale de Statistique 1981:235–52.
3. Goldstein L and Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. The Annals of Statistics 1992:1903–28.
4. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. Biometrika 1997;84:379–94.
5. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 1972;34:187–202.
6. Cox DR. Partial likelihood. Biometrika 1975;62:269–76.

7. Borgan Ø and Langholz B. Nonparametric estimation of relative mortality from nested case-control studies. Biometrics 1993:593–602.

8. Kim RS and Kaplan RC. Analysis of secondary outcomes in nested case-control study designs. Statistics in medicine 2014;33:4215–26.

9. Rubin DB. Multiple imputation for nonresponse in surveys. Vol. 81. John Wiley & Sons, 2004.

10. Keogh RH and White IR. Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. Statistics in Medicine 2013;32:4021–43.

11. Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical science 1994:538–58.

12. Moons KG, Donders RA, Stijnen T, and Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology 2006;59:1092–101.

13. White IR and Royston P. Imputing missing covariate values for the Cox model. Statistics in medicine 2009;28:1982–98.

14. Kim RS. Analysis of nested case-control study designs: revisiting the inverse probability weighting method. Communications for statistical applications and methods 2013;20:455.