Published in final edited form as:

Stat Med. 2011 December 10; 30(28): 3328–3340. doi:10.1002/sim.4380.

Reweighting estimators for Cox regression with missing covariate data: Analysis of insulin resistance and risk of stroke in the Northern Manhattan Study

Qiang Xu^{a,*}, Myunghee Cho Paik^b, Tatjana Rundek^c, Mitchell S. V. Elkind^d, and Ralph L. Sacco^c

^aCenter for Drug Evaluation and Research, U.S. Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993, U.S.A.

^bDepartment of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A.

^cDepartment of Neurology, Miller School of Medicine, University of Miami, Miami, FL 33136, U.S.A.

^dDepartment of Neurology, Columbia University, New York, NY 10032, U.S.A.

Abstract

Incomplete covariates often obscure analysis results from a Cox regression. In an analysis of the Northern Manhattan Study (NOMAS) to determine the influence of insulin resistance on the incidence of stroke in non-diabetic individuals, insulin level is unknown for 34.1% of the subjects. The available data suggest that the missingness mechanism depends on outcome variables, which may generate biases in estimating the parameters of interest if only using the complete observations. This article aimed to introduce practical strategies to analyze the NOMAS data and present sensitivity analyses by using the reweighting method in standard statistical packages. When the data set structure is in counting process style, the reweighting estimates can be obtained by built-in procedures with variance estimated by the jackknife method. Simulation results indicate that the jackknife variance estimate provides reasonable coverage probability in moderate sample sizes. We subsequently conducted sensitivity analyses for the NOMAS data, showing that the risk estimates are robust to a variety of missingness mechanisms. At the end of this article, we present the core SAS and R programs used in the analysis.

Keywords

missing covariate; proportional hazards model; weighted estimating equation

1. Introduction

Survival times with covariates are frequently modeled using the Cox proportional hazard regression model [1]. A problem emerges when one or more risk factors are missing. Complete-case (CC) analysis, which excludes subjects with missing data and performs Cox regression based on the completely observed subjects, is a convenient way of handling the missingness issues. However, CC analysis not only loses efficiency from discarding the incomplete data but, more seriously, may also generate biased estimators when the missingness mechanism depends on the outcome variables.

We are confronted with such a problem in the analysis of the Northern Manhattan Study (NOMAS) data set. The NOMAS is an ongoing prospective, population-based cohort of 3298 subjects, and it aims to identify risk factors for stroke and global cardiovascular diseases [2]. The cohort was recruited between 1993 and 2001. Subjects were eligible if they were at least 40 years of age and lived in a predefined geographic area of northern Manhattan for at least 3 months in a household with a telephone. All participants underwent a thorough baseline examination.

In this paper, we are particularly interested in the effect of insulin resistance on the incidence of stroke among 2533 nondiabetic subjects in the NOMAS cohort. Insulin resistance is measured by the Homeostatic Model Assessment of Insulin Sensitivity (HOMA) index ($\text{HOMA index} = \text{fasting insulin } [\mu\text{U/mL}] \times \text{fasting glucose } [\text{mmol/L}] / 22.5$): a HOMA value over 3 is defined as insulin resistance with a sensitivity of 86% and specificity of 94% [3]. The main analysis consists of a Cox regression for the outcome of strokes on insulin resistance, adjusted for other known modifiable risk factors listed in detail in Section 4.

We noticed that out of the 2533 nondiabetic subjects, 864 (34.1%) did not have their HOMA index recorded, creating a potential hurdle to the application of usual Cox models. After a close look at the data set, we found that the subjects who enrolled earlier appeared to have less blood samples available for fasting insulin and glucose analysis and thus were missing HOMA index, a common phenomenon caused by regulatory issues in the early stage of cohort studies. As a result, the subjects having HOMA values tended to have shorter follow-up time and less chance to develop strokes simply because they were enrolled more recently. Under such circumstance, the coefficient estimates from Cox regressions implemented in statistical software, such as SAS and R, which automatically exclude observations with missing values, may yield a distorted impression of the influence of insulin resistance on strokes.

There are many approaches for dealing with incomplete covariates in Cox regression, mainly including the likelihood-based methods [4–6] and the inverse probability weighting (IPW) methods [7–10]. A more recent approach is the reweighting method proposed by Xu, Paik, Luo, and Tsai [11]. Implementing these methods requires specific programming, thus creating a gap between academic research and routine application. The primary purpose of this paper was to find a practical strategy that requires little project-oriented programming and that can be used to perform a bias-correcting analysis on a daily basis. We here apply

the reweighting method to the NOMAS analysis for three reasons. Compared with the likelihood methods, the implementation of this method can be facilitated by built-in procedures in standard statistical softwares with the variance of the reweighting estimator (RWE) estimated using the jackknife method [12]. Moreover, the RWE may potentially gain efficiency over the inverse probability weighting estimator (IPWE) [11]. Finally, the reweighting method allows display of sensitivity analyses for the missingness mechanism assumption.

The basic idea of the reweighting method is to first weight each complete subject by the inverse of the observation probability to correct the selection bias introduced by missingness, and then to improve estimation efficiency, impose the marginal observation probability given a risk set at time t on the pseudo-unbiased samples generated from the first step. Like the aforementioned likelihood methods and IPW methods, the reweighting method relies on the missing at random (MAR) assumption as well; that is, the observation probability depends on the completely observed variables, including the follow-up time, the failure indicator, and/or the complete covariates, but not on the missing covariate itself. In the NOMAS case, the MAR assumption might be plausible. However, the validity of this assumption cannot be verified based on the observed data. To test whether the reweighting estimates are robust against the MAR assumption, we designed a sensitivity analysis in which the reweighting estimates were examined under various scenarios of missingness mechanism assuming that the HOMA index was not missing at random (NMAR).

This paper focuses on illustrating a number of aspects in the analysis of the NOMAS with the reweighting method. In Section 2, we mathematically introduce the reweighted estimating equation. In Section 3, we discuss the computational challenges in the analysis and describe the strategy to these challenges so that the reweighting method can be performed by built-in procedures in standard softwares, particularly in SAS. As an important component of method implementation, in Section 3, we also give general rules for building models of the missingness mechanism. We elaborate step by step the NOMAS analysis and present, in Section 4.1, the resulting RWEs along with the IPWEs, compared with the CC results, with sensitivity analyses for the MAR assumption presented in Section 4.2. Section 5 evaluates the performance of the jackknife variance estimators of the RWEs and IPWEs in studies with modest sample sizes through simulations. Finally, we generalize the proposed strategy to multivariate cases in Section 6 whereby the implementation in R are also supplemented.

2. Reweighted estimating equation

For the convenience of subsequent presentation, we first introduce the following notation: $T = \min(T^\circ, C)$ and $\delta = I(T^\circ \leq C)$, where T° is the time to strokes and C is the censoring time. Assume that C is independent of T° given $\mathbf{Z}$, a covariate vector that in this example can be partitioned into $\mathbf{Z}^m$, the HOMA index subject to missingness, and $\mathbf{Z}^c$, the completely observed confounders. Let R denote the observation indicator; that is, $R = 1$ if the HOMA index is observed and $R = 0$ otherwise. Suppose that $(T_i, \delta_i, \mathbf{Z}_i, R_i), i = 1, \dots, n$, are independent and identically distributed random quintuplets. The missingness mechanism is determined by the distribution of R , a Bernoulli with probability $\pi(\mathbf{W})$, where $\mathbf{W} = (\mathbf{Z}^c, T, \delta)$

under the MAR assumption. Let $N(t)$ be the counting process, that is, $N(t) = \mathcal{I}(T \leq t)$, and $Y(t)$ be the at-risk indicator, that is, $Y(t) = I(T > t)$.

Our interest is in estimating β , the coefficients of Cox regression, to study the association between the predictors and stroke. If there were no missing values in HOMA, we would obtain the maximum partial likelihood estimator by solving the following estimating equation:

$$\mathbf{U}(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ \mathbf{Z}_i - \frac{\mathbf{S}^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i(t) = 0, \quad (1)$$

where

$$\mathbf{S}^{(k)}(\beta, t) = \frac{1}{n} \sum_{j=1}^n Y_j(t) \mathbf{Z}_j^{\otimes k} \exp(\beta^T \mathbf{Z}_j), \quad k=0, 1, 2$$

with $\mathbf{Z}^{\otimes 0} = 1$, $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$, $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}^T$ and $\tau = \sup\{t : P(Y(t) > 0)\}$. The maximum partial likelihood estimators can be easily produced in standard statistical softwares.

To apply the reweighting method, we assume that the HOMA index is missing at random, and thus, the missing mechanism depends on $\mathbf{W}$. The reweighted estimating equation is as follows:

$$\mathbf{U}_{RW}(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau R_i \frac{\hat{\pi}^*(t)}{\pi(\mathbf{W}_i)} \left\{ \mathbf{Z}_i - \frac{\mathbf{S}_{RW}^{(1)}(\beta, t)}{S_{RW}^{(0)}(\beta, t)} \right\} dN_i(t), \quad (2)$$

where

$$\mathbf{S}_{RW}^{(k)}(\beta, t) = \frac{1}{n} \sum_{j=1}^n R_j \frac{\hat{\pi}^*(t)}{\pi(\mathbf{W}_j)} Y_j(t) \mathbf{Z}_j^{\otimes k} \exp(\beta^T \mathbf{Z}_j), \quad k=0, 1, 2.$$

The imposed time-variant selection probability $\hat{\pi}^*(t)$ is the empirical estimator of the marginal observation probability given a risk set at time t ; that is,

$$\hat{\pi}^*(t) = \frac{\sum_{i=1}^n R_i Y_i(t)}{\sum_{i=1}^n Y_i(t)}. \quad (3)$$

The RWE of β can be obtained by solving $\mathbf{U}_{RW}(\beta) = 0$. The reweighted estimating equation is similar to that of the IPW method except for imposing $\hat{\pi}^*(t)$ on observed subjects who have been first weighted by the inverse of the observation probability.

3. Implementation of the reweighting method

Implementation of the reweighting method in statistical packages, such as in SAS, has two computational challenges: the imposed $\hat{\pi}^*(t)$ is a time-variant weight, which cannot be

directly handled in built-in procedures; and estimating the theoretical variance of the RWE needs specific programming. As a result, the reweighting method cannot be performed on a daily basis. Besides the computing difficulties, correct identification of the missingness mechanism is challenging and also essential for validly accomplishing the reweighting method. In this section, we illustrate the strategies to solve these issues.

If the weight in a weighted estimating equation is time-constant, the Newton–Raphson algorithm is already built in standard packages. For example, in SAS, the weighted coefficient estimators for Cox regression can be produced using PHREG procedure with WEIGHT statement, which so far does not allow directly involving time-variant weights. Nevertheless, if we first transform the data set structure to the counting process style of input, in which every subject is represented by multiple observations, each identifying a semiclosed time interval $t_1; t_2$ and a status indicator that tells if an event occurs at time t_2 , then a time-variant weight can also be incorporated using PHREG procedure to avoid intensive programming. We developed a SAS macro SURVSPLIT, which is presented in Appendix A, to facilitate the data set transformation. For the reweighting analysis, after transformation, the imposed selection probability $\hat{\pi}^*$ at t_2 is then calculated using Equation (3) for each time interval, and the RWEs can be obtained by the counting process accommodating PHREG procedure as follows:

```
proc phreg data= countdata; /* counting style data generated from macro
" SURVSPLIT" */
model (t1,t2)*status(0)= Z; /* status=0 if the event has not occur at t2 */
weight reweight;
run;
```

The reweight equals $\hat{\pi}^*(t_2) / \pi(\mathbf{W}_i)$. As demonstrated, with counting process format, the reweighting estimation requires nothing more than an ordinary weighted procedure.

To draw inference about the coefficients of interest needs consistent estimators of the variance of the RWE. However, the immediate variance estimators from PHREG are no longer valid. We here suggest using the nonparametric jackknife method to estimate variances based upon the observation that the reweighted estimating equation is asymptotically equivalent to the sum of independent and identically distributed random variables

$$\mathbf{u}_{RW}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau R_i \frac{\pi^*(t)}{\pi(\mathbf{W}_i)} \{ \mathbf{Z}_i - \mathbf{e}(\boldsymbol{\beta}, t) \} dM_i(t),$$

where $\mathbf{e}(\boldsymbol{\beta}, t) = E \left\{ Y(t) \pi^*(t) \mathbf{Z} \exp(\boldsymbol{\beta}^T \mathbf{Z}) \right\} / E \left\{ Y(t) \pi^*(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}) \right\}$ and

$M(t) = N(t) - \int_0^t Y(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}) \lambda_o(t) dt$ [11]. One form of the jackknife estimator of variance, v_J , is

$$\frac{n-p}{n} \sum_{i=1}^n (\hat{\beta}_{-i} - \hat{\beta}) (\hat{\beta}_{-i} - \hat{\beta})', \quad (4)$$

where $\hat{\beta}_{-i}$ is the estimate of β obtained on the basis of the sample with the i th observation deleted and p is the dimension of covariates. Let $I(\beta)$ be the information matrix, then, v_J is asymptotically equivalent to

$$\frac{n-p}{n} \sum_{i=1}^n \left\{ \left[I_{-i}(\hat{\beta}) \right]^{-1} \mathbf{u}_{RW_{-i}}(\hat{\beta}) \mathbf{u}_{RW_{-i}}(\hat{\beta})' \left[I_{-i}(\hat{\beta}) \right]^{-1} \right\}, \quad (5)$$

which is also asymptotically equivalent to the theoretical variance of the RWEs. A detailed description of the theoretical variance can be found in [11]. Note that if $\hat{\beta}_{-i}$ is obtained by performing one step of the Newton–Raphson algorithm, with β_0 being the starting value, then (4) and (5) are exactly the same.

The jackknife variance estimator v_J has several advantages in terms of computational feasibility. The estimation procedure only requires a simple loop, which can be easily programmed in SAS. Also, the associated codes can be generalized to other data sets with minimum project-specific programming. Moreover, v_J does not require estimating the baseline hazard, a not-so-easy task, whereas the theoretical sandwich estimator does.

The observation probability $(\pi \mathbf{W}_i)$ is usually unknown. The RWEs of β are asymptotically unbiased as long as $(\pi \mathbf{W}_i)$ is consistently estimated. Therefore, it is crucial to correctly specify the model for observation mechanism $(\pi \mathbf{W}_i)$. For the NOMAS analysis, we first assumed MAR and used a logistic regression model

$\pi(\mathbf{W}; \alpha) = \exp(\alpha' \mathbf{W}) / [1 + \exp(\alpha' \mathbf{W})]$ to estimate the unknown observation probability, where α was a vector of nuisance parameters for the logistic model with R being the response variable, and then solved the reweighting estimating equation (2) with the true replaced by its estimate $\hat{\pi}(\mathbf{W}; \hat{\alpha})$. The computing process largely followed the strategies elaborated previously.

A logistic model may seem a rather simplistic choice. However, note that the exact functional form is not of particular interest here, but the plausible estimators of observation probability a model may provide. Selection of the predictors for the observation model is a key to obtain accurate estimators. Generally, a rich collection of predictors from complete variables $\mathbf{W}$ not only prevents model under-specification but may also help reduce the standard errors (SEs) of the estimators for the parameters of interest.

4. Analysis of the Northern Manhattan Study

Previous sections outline the reweighting method and computing techniques involved in its implementation. In Section 4.1, we elaborate the analysis of the NOMAS data set in a step-by-step fashion to help analysts follow our recommended strategies. In addition, we illustrate how to investigate the robustness of the NOMAS results against the MAR assumption through sensitivity analyses in Section 4.2.

4.1. Application of the reweighting method

As introduced earlier, whether insulin resistance predicts stroke in nondiabetic subjects is the core question we are trying to answer. Insulin resistance is often found in people with obesity, hypertension, low HDL, and hypertriglyceridemia, which are known risk factors for the incidence of strokes. As a result, our main analysis consisted of a Cox regression for the outcome of strokes on insulin resistance, adjusted for the aforesaid predictors and other potential confounders including smoking status, alcohol consumption, activity intensity, and demographics. To be specific, the variables adjusted in the Cox regression were age, sex (female as reference), race/ethnicity dummy variables (white as reference), socioeconomic status estimated by completion of high school education, current smoking status, moderate alcohol use or not, presence of moderate to heavy physical activity, systolic blood pressure, diastolic blood pressure, obesity measured by waist circumference, fasting HDL level, and LDL level.

After excluding individuals previously diagnosed with diabetes and those of a race/ethnicity other than black, white, or Hispanic, 2533 subjects were included in this analysis. In Table I, we present outcome variables and risk factors stratified by insulin resistance status. Because of regulatory issues, subjects with missing HOMA index were enrolled earlier and thus tended to have longer follow-up. The average times to strokes among 864 incompletely observed subjects and 1669 complete subjects were 8.3 and 7.2 years, respectively, with 102 (65.4%) out of a total of 156 strokes contributed by incomplete subjects. In addition, compared with complete subjects, those with missing HOMA index were more likely to have higher HDL level (46.6 vs. 48.5 mg/dL) and lower LDL level (131.2 vs. 128.5 mg/dL) and to be non-Hispanic (43.3 vs. 59.0%), older (68.0 vs. 70.9 years), and educated (46.4 vs. 52.5%).

The first step of applying the reweighting method, as described in Section 3, was to build a logistic regression to estimate π , the probability of observing the HOMA index, for each NOMAS subject given his or her related complete variables $\mathbf{W}$. For a rich collection of the predictors for π , we considered the follow-up time, the stroke status, and all complete main effects in the Cox regression as well as their interactions that were significant at the 0.2 level. Table II shows the fitted logistic regression, which is consistent with what we have observed in Table I. The low p -values of the follow-up time T and the stroke indicator δ , along with the magnitudes of the coefficients, indicate a strong relationship between two outcome variables and the missingness in the HOMA index, a scenario where the CC analysis may generate biased estimate of the effect of insulin resistance on the incidence of strokes.

After merging the predicted observation probability $\hat{\pi}$ from the logistic model built in the first step to the original NOMAS data set, we transformed the data set structure to the counting process style of input by calling the SAS macro

```
SURVSPLIT (data=NOMAS, time=stroke_time, event=stroke)
```


where stroke_time and stroke were the outcome variables of time to stroke and stroke status, respectively. Four new variables were automatically generated by the macro after transformation: an index variable `_i_` that ranked the episodes of split time intervals, variables `start` and `stop` that recorded t_1 and t_2 for each interval $(t_1; t_2]$, respectively, and `istat`, an indicator of stroke status at t_2 . We provide a detailed example of applying this macro in Appendix A.

On the basis of the transformed data set, calculating time-variant $\hat{\pi}^*(t)$ was rather straightforward. Combined with the statement `BY stop`, a `FREQ` procedure on the observation indicator `R` provided us the proportion of observed subjects at time `stop`, that is, the empirical estimate of the marginal observation probability as formulated in Equation (3). After merging $\hat{\pi}^*(t)$ to the transformed data set by `stop`, we were then ready to calculate `reweight`, the weight in the reweighting method, which simply was $\hat{\pi}^*(t) / \hat{\pi}$.

Up to this point, data preparation for implementing the reweighting method was finished. To obtain the reweighting estimate of the effect of insulin resistance through the introduced Cox model with known risk factors adjusted, we took advantage of the built-in `PHREG` procedure in the form to accommodate counting process data structure as follows:

```
proc phreg data= countNOMAS; /* Transformed NOMAS data set */
model (start,stop)*istat(0)= HOMA black hispanic age ...;
weight reweight;
run;
```

We present the resulting RWEs of the Cox model coefficients and their jackknife SE estimates in Table III, along with those from the IPW method and CC method. In the IPW analysis, we used the same model listed in Table II.

All three methods show that insulin resistance is significantly associated with a more than twofold increased risk of stroke at α level of 0.05 after adjusting for other confounders, suggesting a potential need of primary preventive therapies targeted at insulin resistance. It is interesting to note that compared with the CC analysis, the reweighting method and the IPW method reduce the benefit of moderate to heavy alcohol consumption by over 38%, while increasing the risk of stroke among smokers by over 80%. Most SEs of the reweighting estimates are no bigger than those of the IPW estimates.

To further visualize the impact of insulin resistance on stroke, Figure 1 displays its marginal effect through Kaplan–Meier (KM) curves by using the reweighting method and the CC method, both methods demonstrating a harmful effect of insulin resistance. The KM curves also show that within each stratum of insulin resistance status, the reweighting probabilities of having no strokes are lower over time compared with the CC estimated probabilities. The IPW KM estimates largely overlap with the reweighting curves and therefore are not included here for a clearer view.

4.2. Sensitivity analysis

Thus far the analysis is based on the MAR assumption, which is not verifiable based on the observed data. Even after conditioning on a rich collection of possible predictors, the missingness mechanism might still depend on the missing HOMA index itself. Sensitivity analyses thus are needed to support the previous findings.

The reweighting method can be easily adapted to conduct sensitivity analyses for the MAR assumption. In the NOMAS data set, if the HOMA index was not missing at random, that is, observation depended on HOMA values, the π model should include the HOMA index as an independent predictor. Realizing the fact that with $\hat{\pi}$ the reweighting estimating equation involved only complete observations, we re-estimated observation probability for each complete subject by modifying the previously fitted logistic model as follows:

$$\text{logit}(\pi) = \alpha'_W \mathbf{W} + \alpha_H \text{HOMA} + \alpha'_{HW} \text{HOMA} * \mathbf{W},$$

where α_W and $\mathbf{W}$ remained the same as in Table II. What made the current model different from the previous one was the newly added main effect of the HOMA index and interactions between HOMA and $\mathbf{W}$ with coefficients α_H and α_{HW} , respectively. The magnitudes of α_H and α_{HW} measured the deviation of the missingness mechanism from the MAR assumption; in other words, the larger the magnitudes of α_H and α_{HW} were, the more severe the missingness was away from the MAR assumption.

To be specific, we treated the effect of the HOMA index on the missingness mechanism as fixed. After assigning values from -4 to 4 to α_H and α_{HW} , the parameters of interest were re-estimated using the reweighting method. The same computing and programming techniques as those illustrated in detail in Section 4.1 for the main analysis of the NOMAS under MAR assumption were adopted for the sensitivity analyses. The only difference in the whole procedure was in the first step, where observation probability $\hat{\pi}(\mathbf{W}, \text{HOMA})$ replaced $\hat{\pi}(\mathbf{W})$ to incorporate contributions from HOMA values. The fluctuation in the resulting estimates of the coefficients of the Cox model demonstrated the behavior of the reweighting estimates when the MAR assumption did not hold.

To present in this paper the robustness of the estimate of the insulin resistance effect, we focus on the changes in α_{HT} and α_{HS} , the interactions between the HOMA index and outcome variables time to stroke and stroke indicator, respectively, while setting other elements in α_{HW} , that is, interactions between HOMA index and observed covariates as zero, because interactions with outcome variables have greater influence on the robustness of results. Figure 2 presents the reweighting estimates for insulin resistance and their 95% CIs with α_{HT} being zero and α_{HS} ranging from -4 to 4 under three different values of α_H : $\alpha_H = 2$ in plot (a); $\alpha_H = 0$ in plot (b); and $\alpha_H = -2$ in plot (c). We notice that insulin resistance's harmful effect decreases as α_{HS} increases; however, from -2 to 2 , a practical range of α_{HS} in most situations, all three plots steadily support the previous conclusion that the insulin resistance increases the risk of stroke. Analyses by the IPW method is also displayed in Figure 2. The fluctuation pattern of the reweighting estimates is similar, but of less variation, when α_{HT} varies in the same range with α_{HS} being zero; as a result, the same

conclusion regarding the effect of insulin resistance is reached. The related figures hence are not presented here for the interest of space.

The sensitivity analyses for the effect of the observed covariates are relatively easy. We only considered α_H , that is, set $\alpha_{HW} = 0$, and let α_H vary from -4 to 4. The re-estimated effects of these variables, as anticipated, fluctuate less noticeably compared with those of the insulin resistance. In Figure 3, we present the graphs of sensitivity analyses for age, sex, and smoker, as they were significant in the main analysis. Figure 3 suggests that although the effect estimates may lose significance depending on the severity of non-ignorability in missingness, the direction of impact does not flip significantly over wide range of α_H , and thus, the conclusions drawn from the previous analysis are robust against the MAR assumption. Sensitivity analyses for other observed covariates demonstrate the same robustness.

It is interesting to point out that the RWEs and IPWEs in Figure 2(b) when α_{HS} is zero and in Figure 3 when α_H is zero are exactly the estimates that we previously obtained in Table III under the MAR assumption.

5. Simulation study

The NOMAS study has large sample size, making the asymptotic results reliable; however, it is not always the case in clinical research, especially for studies in rare diseases. We conducted simulations to evaluate the performance of the RWEs and their jackknife variance estimators under moderate sample sizes. The CC analyses and the IPW estimators are also presented.

There were two simulation settings. Both considered two covariates, one missing covariate Z^m and one observed covariate Z^c . In the first setting, Z^m and Z^c were independent, and both followed Bernoulli(0.5); in the second setting, $Z^m|Z^c$ followed Normal(Z^c , 1) where Z^c followed Bernoulli(0.5). The hazard function was specified by $\lambda(t|Z) = \lambda_o(t) \exp(\beta^T Z)$ equal to 1. The censoring times followed exponential distributions with mean $0.5Z^c + 1$ and mean 1 in the first and second settings, respectively, generating 55% and 50% cases. Cohort sizes of 250 and 500 were chosen for both simulation settings.

The selection mechanism followed Bernoulli distribution wherein $\pi = \{1 + \exp[-(-1 - 2Z^c + 3.5T - 0.5\delta)]\}^{-1}$, leading to 68% of cases and 50% of censors missing in Z_m in the first setting and 67% of cases and 55% of censors missing in the second setting. The simulation results of the two settings are presented in Tables IV and V, respectively.

One thousand replicates in each simulation setting were generated for $\beta = (-\ln(2), \ln(2))$. In the tables, the excrusus (α) indicates that the estimates were obtained under the true π model; ($\hat{\alpha}$) indicates that the estimates were obtained on the basis of the estimated nuisance parameters. Bias is the average difference between a parameter and its estimates. Sample SE equals the square root of the sample variance of the 1000 parameter estimates. Mean theoretical SE, SE_T , is the average of the 1000 theoretical SE estimates. Mean jackknife SE,

SE_J , is the average of the jackknife SE estimates. The $95\%CP_T$ is the 95% confidence coverage probability based on the theoretical SE estimates, and the $95\%CP_J$ is the 95% confidence coverage probability based on the jackknife SE estimates.

Tables IV and V show that when sample size reaches 500, biases of the RWEs and IPWEs are negligible. The jackknife SE estimates are close to the sample SEs, especially with the sample size of 500 where the differences between them in both tables are in the neighborhood of $\pm 5\%$ of the sample SEs. In contrast, the theoretical SE estimates tend to be less than the sample SEs: with sample size of 250, the theoretical SE estimates underestimate the sample SEs by up to 25% in both tables, and with sample size of 500, the underestimation is still up to 14% in Table IV and 21% in Table V. In addition, the 95% coverage probabilities based on jackknife SE estimates in both tables are closer to the nominal value compared with the coverage probabilities based on theoretical SE estimates irrespective of sample sizes.

In summary, jackknife SE estimates appear to be appropriate substitutes for theoretical ones in the sense that they are closer to sample SEs and have better 95%CP for both RWEs and IPWEs. We therefore recommend that jackknife SE estimates be routinely used especially for studies with relatively small sample sizes that may lead to incorrect inferences if not using the jackknife estimates.

6. Discussion

In this particular analysis of the NOMAS data set, the missingness occurred primarily in the HOMA index alone. However, oftentimes, there might be multiple missing covariates when building a Cox regression. It is not difficult to generalize the previous computational strategies to data sets with arbitrary nonmonotone missingness.

Let R° be the overall missingness indicator; that is, $R^\circ = 1$ if all elements in $\mathbf{Z}^m$ are observed.

The reweighting method weights each complete subject by $\hat{\pi}_1^*(t) / \pi_1$, where

$\hat{\pi}_1^*(t) = \sum_{i=1}^n R_i^\circ Y_i(t) / \sum_{i=1}^n Y_i(t)$ and π_1 is the probability of being completely observed.

When π_1 is unknown, under the MAR assumption, generalized logistic models in which

$\pi_i = \exp(\boldsymbol{\alpha}_i^T \mathbf{W}) / \sum_{i=1}^M \exp(\boldsymbol{\alpha}_i^T \mathbf{W})$ is postulated can be used to estimate it. The M denotes the number of different missingness patterns. For example, if there are two missing covariates, then the maximum number of missingness patterns is 4: missing in the first covariate, in the second one, in both, or in neither. We are interested in the probability of the last scenario. For model identification, we set that $\boldsymbol{\alpha}_M = \mathbf{0}$. In SAS, the estimation of π_1 can be facilitated using CATMOD procedure. Once we obtain the reweighting weights, the rest of the analysis procedure follows the strategies illustrated in Section 3.

Building a generalized logistic model with multiple missing covariates can be difficult when there are very few observations in a particular missingness pattern. In such circumstances, we need to carefully investigate the possible reasons for each covariate's missingness. If legitimate, for example, some covariates are missing for similar reasons and missingness overlaps largely, then we can combine those patterns to reduce the complexity.

Although covariates $\mathbf{Z}$ in the reweighted estimating equation demonstrated in Section 2 are time-constant, the reweighting method actually can deal with time-variant covariates as well with the same computational technique. With the counting process style of input, the values of time-variant covariates at time t_2 may be programmed to fit into the transformed data set, and the subsequent weighted PHREG procedure handles time-variant weights and time-variant covariates at once.

The reweighting analysis may also be conducted in R. The built-in function `survSplit` in the survival package plays the same role as the SAS macro that we developed. We used the same name for the convenience of analysts. The weighted Cox regression based on the data in counting process format then can be produced using function `coxph`. We present the detailed codes in Appendix A.

This paper aimed at introducing an applicable and efficient strategy to account for missing covariates in Cox regression. The IPW method is easy to implement without transformation of data structure because there are no time-variant weights involved. However, the reweighting method offers potential gain in the estimation efficiency, as demonstrated in simulations, especially when the predicted observation probability is small for some observations. Intuitively, the reweighting estimators may control the overweighting in the IPW method of the subjects with small probabilities by the re-selection step. Both approaches need the MAR assumption and allow sensitivity analyses by modifying π model accordingly.

We can further improve the estimation efficiency by using the augmented reweighting method, which is thoroughly studied by Xu *et al.* [11]. This procedure can no longer be conducted in standard packages and hence is not of interest in this paper.

APPENDIX A

The SAS macro SURVSPLIT transforms data sets with time-to-event variables into counting style of input and adds COUNT as the prefix to the original data set names for the generated data sets. To invoke SURVSPLIT, the original data set, the event indicator with 1 for developed cases and 0 for censors, and the failure time need to be specified for parameters data, event, and time, respectively. The macro is as follows.

id	failtime	status	age	group
1	10	0	70	1
2	3	1	62	1
3	2	0	58	0
4	12	1	75	0

For example, the data set EXAMPLE is

id	failtime	status	age	group	start	stop	istat	_i_
1	10	0	70	1	0	3	0	1
1	10	0	70	1	3	10	0	2
2	3	1	62	1	0	3	1	1
3	2	0	58	0	0	2	0	1
4	12	1	75	0	0	3	0	1
4	12	1	75	0	3	12	1	2

After applying %SURVSPLIT(data=EXAMPLE, event=status,time=failtime), the transformed data set COUNTEXAMPLE has the following structure:

```
%macro SURVSPLIT(data= , event= , time= );
  %local i j;
  data case(keep=&time) ; set &data; if &event=1; run;
  proc sort data=case nodup out=cut; by &time; run;
  proc transpose data=cut out=tcut(drop=_name_);run;
  data _null_;
    call symput('ncase',trim(left(put(_nobs,12)))));
    if 0 then set cut nobs=_nobs; stop;
  run;
  data _null_; set tcut;
    %do i=1 %to &ncase; call symput("t&i", col&i); %end;
  run;
  data count&data;
    array tt{1:&ncase} _temporary_ ;
    %do j=1 %to &ncase; tt[&j]=SYMGET("t&j"); %end;
    set &data; start = 0; stop = 0; istat = 0;
    do _i_=1 to &ncase;
      if ( &time=tt[_i_] ) then do;
        stop= &time; istat = &event; output;
      end;
      else if ( &time >tt[_i_] ) then do;
        stop= tt[_i_] ; istat= 0; output;
      end;
      start = stop;
    end;
    if ( &time >tt[&ncase] ) then do;
      stop = &time; istat = &event; output;
    end;
  run;
  proc delete data=case cut tcut;run;
```

```
%mend SURVSPLIT;
```

where `istat` indicates event status at time stop.

In R, the built-in function `survSplit` conducts similar transformation of the data structure:

```
COUNTEXAMPLE <- survSplit(data=EXAMPLE,cut=unique(failtime[status= 1]),
end="failtime",start="start",event="status", episode="i")
```

Note that R replaces the original `failtime` and `status` by newly created stop times and associated status indicators, different from the output of the SAS macro that keeps original variables and creates variables `stop` and `istat` for each time interval. The Cox regression based on `COUNTEXAMPLE` can be implemented in R with `coxph(Surv(start, failtime, status)~age+group, data=COUNTEXAMPLE)`.

References

1. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
2. Elkind MSV, Sciacca RR, Boden-Albala B, Rundek T, Paik MC, Sacco RL. Moderate alcohol consumption reduces risk of ischemic stroke: the Northern Manhattan Study. *Stroke*. 2006; 37:13–19. [PubMed: 16306464]
3. Rundek T, Gardener H, Xu Q, Goldberg RB, Wright CB, Boden-Albala B, Disla N, Paik MC, Elkind MSV. Insulin resistance and risk of ischemic stroke among non-diabetic individuals from the Northern Manhattan Study. *Archives of Neurology*. 2010; 67(10):1195–1200. [PubMed: 20937946]
4. Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 1999; 94:896–908.
5. Chen HY. Double-semiparametric method for missing covariates in Cox regression models. *Journal of the American Statistical Association*. 2002; 97:565–576.
6. Herring AH, Ibrahim JG. Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*. 2001; 96:292–302.
7. Pugh, M.; Robins, J.; Lipsitz, S.; Harrington, D. Technical Report. Harvard School of Public Health, Dept. of Biostatistics; 1994. Inference in the Cox Proportional Hazards Model With Missing Covariate Data..
8. Qi L, Wang CY, Prentice R. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100:1250–1263.
9. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
10. Wang CY, Chen HY. Augmented inverse probability weighted estimator for Cox regression missing covariate regression. *Biometrics*. 2001; 57:414–419. [PubMed: 11414564]
11. Xu Q, Paik MC, Luo XD, Tsai WY. Reweighting estimators for Cox regression with missing covariates. *Journal of the American Statistical Association*. 2009; 104:1155–1167.
12. Lipsitz SR, Parzen M. A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics*. 1996; 52:291–298. [PubMed: 8934598]

Figure 1. Kaplan–Meier estimates by insulin resistance (IR) status from complex-case (CC) and reweighting (RW) analyses.

Figure 2.
Sensitivity analyses for the Homeostatic Model Assessment of Insulin Sensitivity index.
IPW, inverse probability weighting; RW, reweighting.

Figure 3. Sensitivity analyses for the observed covariates. IPW, inverse probability weighting; RW, rereighting.

Table I

Summary of the Northern Manhattan Study data set by insulin resistance status.

Variable	HOMA 3 (<i>n</i> = 416)	HOMA <3(<i>n</i> = 1253)	No HOMA (<i>n</i> = 864)
Time to stroke (mean (SD), years)	6.7 (2.2)	7.4 (2.4)	8.3 (3.3)
Stroke (%)	20 (4.8)	34 (2.7)	102 (11.8)
Black (%)	70 (16.8)	260 (20.8)	260 (30.1)
Hispanic (%)	274 (65.9)	673 (53.7)	354 (41.0)
White (%)	72 (17.3)	320 (25.5)	250 (28.9)
Age (mean (SD), years)	67.3 (9.8)	68.3 (10.6)	70.9 (10.8)
Male (%)	140 (33.7)	475 (37.9)	326 (37.7)
Female (%)	276 (66.3)	778 (62.1)	538 (62.3)
High school education (%)	161 (38.7)	614 (49.0)	454 (52.5)
Moderate to heavy activity (%)	19 (4.6)	133 (10.6)	97 (11.2)
Current smoker (%)	62 (14.9)	223 (17.8)	149 (17.3)
Moderate alcohol consumption (%)	124 (29.8)	480 (38.3)	288 (33.3)
Waist (mean (SD), inches)	38.9 (5.0)	35.5 (4.5)	36.5 (5.2)
Diastolic (mean (SD), mmHg)	85.2 (11.0)	82.8 (11.1)	82.2 (11.3)
Systolic (mean (SD), mmHg)	145.4 (20.3)	142.1 (21.8)	143.1 (20.9)
HDL (mean (SD), mg/dL)	42.6 (12.4)	47.9 (14.4)	48.5 (15.3)
LDL (mean (SD), mg/dL)	130.4 (36.6)	131.4 (35.2)	128.5 (35.7)

HOMA, Homeostatic Model Assessment of Insulin Sensitivity; SD, standard deviation.

Table II

Summary of the logistic regression that was used to estimate the observation probability of the Homeostatic Model Assessment of Insulin Sensitivity index.

Parameter	Estimate	Standard error	<i>p</i> -value
Intercept	2.429	2.690	0.367
Time to stroke	−0.138	0.061	0.024
Stroke	−2.120	0.200	0.000
Black	−0.168	0.129	0.191
Hispanic	0.364	0.134	0.006
Age	−0.030	0.005	0.000
Sex	−0.184	0.103	0.073
High school education	−0.003	0.111	0.981
Moderate to heavy activity	−0.477	0.257	0.063
Current smoker	−0.304	0.694	0.662
Moderate alcohol consumption	0.130	0.099	0.189
Waist	0.018	0.071	0.798
Diastolic	0.065	0.036	0.071
Systolic	−0.027	0.019	0.152
HDL	−0.005	0.004	0.189
LDL	0.011	0.004	0.008
Time to stroke*smoker	0.061	0.041	0.139
Time to stroke*LDL	−0.007	0.003	0.041
Smoker*LDL	−0.007	0.003	0.041
Smoker*HDL	0.012	0.009	0.176
Systolic*waist	0.001	0.001	0.125
Diastolic*waist	−0.002	0.001	0.081
Education*activity	0.445	0.314	0.156

Table III

The proportional hazard models of the Northern Manhattan Study using various approaches for the missingness in the HOMA index.

Effects	Complete case		IPW		RW	
	β	SE	β	SE	β	SE
HOMA	* 0.829	0.306	* 0.778	0.327	* 0.915	0.316
Black	-0.521	0.428	-0.789	0.392	-0.607	0.386
Hispanic	-0.502	0.413	-0.680	0.380	-0.546	0.388
Age	* 0.055	0.016	* 0.065	0.015	* 0.064	0.015
Sex	* 0.876	0.313	* 0.833	0.376	* 0.850	0.366
Heduc	-0.454	0.339	-0.344	0.313	-0.377	0.312
Actmodheavy	-0.805	0.729	-0.501	0.556	-0.385	0.515
Smoker	0.447	0.343	* 0.904	0.344	* 0.801	0.338
Etmod	-0.176	0.303	-0.190	0.301	-0.183	0.298
Waist	-0.031	0.032	-0.032	0.031	-0.045	0.031
Diastolic	* 0.031	0.015	0.020	0.015	0.026	0.015
Systolic	0.011	0.008	0.008	0.007	0.009	0.007
HDL	-0.002	0.011	-0.001	0.012	-0.002	0.011
LDL	0.000	0.004	-0.002	0.005	-0.002	0.005

HOMA, Homeostatic Model Assessment of Insulin Sensitivity; IPW, inverse probability weighting; RW, rereighting; SE, standard error.

* Indicates significant relationship under α level of 0.05.

Table IV

Simulation setting 1 with jackknife variance estimator studying the performance of jackknife variance estimators for RWEs and IPWEs.

Appro.	Bias		Sample SE		Mean SE _J		Mean SE _T		95%CP _J		95%CP _T	
	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c
Cohort size 250												
CC	-0.087	-0.464	0.355	0.325	0.352	0.347	0.328	0.334	0.953	0.743	0.932	0.711
RWE- α	-0.010	-0.077	0.385	0.338	0.394	0.353	0.326	0.299	0.937	0.960	0.887	0.918
RWE - $\hat{\alpha}$	-0.013	-0.071	0.395	0.329	0.409	0.346	0.324	0.298	0.929	0.958	0.885	0.935
IPWE- α	-0.018	-0.106	0.458	0.408	0.467	0.423	0.355	0.321	0.919	0.927	0.852	0.862
IPWE - $\hat{\alpha}$	-0.022	-0.102	0.474	0.405	0.488	0.421	0.352	0.321	0.909	0.928	0.836	0.864
Cohort size 500												
CC	-0.091	-0.465	0.220	0.225	0.233	0.230	0.225	0.229	0.957	0.480	0.945	0.474
RWE- α	-0.003	-0.046	0.263	0.224	0.269	0.234	0.243	0.215	0.943	0.953	0.925	0.932
RWE - $\hat{\alpha}$	-0.006	-0.037	0.265	0.211	0.276	0.225	0.245	0.216	0.943	0.964	0.920	0.952
IPWE- α	-0.002	-0.060	0.324	0.270	0.325	0.282	0.280	0.240	0.930	0.941	0.894	0.908
IPWE - $\hat{\alpha}$	-0.006	-0.050	0.329	0.260	0.336	0.275	0.281	0.241	0.925	0.946	0.884	0.922

CC, complex case; IPWE, inverse probability weighting estimator; RWE, reweighting estimator; SE, standard error.

Table V

Simulation setting 2 with jackknife variance estimator studying the performance of jackknife variance estimators for RWEs and IPWEs.

Appro.	Bias		Sample SE		Mean SE _J		Mean SE _T		95%CP _J		95%CP _T	
	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c	β_m	β_c
Cohort size 250												
CC	-0.127	-0.403	0.205	0.441	0.221	0.463	0.200	0.430	0.928	0.871	0.915	0.844
RWE- α	-0.015	-0.062	0.219	0.450	0.241	0.475	0.193	0.377	0.966	0.951	0.914	0.902
RWE - $\hat{\alpha}$	-0.019	-0.057	0.223	0.442	0.248	0.477	0.193	0.375	0.958	0.952	0.896	0.910
IPWE- α	-0.020	-0.095	0.254	0.543	0.273	0.566	0.200	0.406	0.964	0.923	0.884	0.852
IPWE - $\hat{\alpha}$	-0.023	-0.092	0.260	0.541	0.282	0.576	0.200	0.405	0.956	0.930	0.872	0.848
Cohort size 500												
CC	-0.114	-0.406	0.145	0.294	0.147	0.305	0.137	0.295	0.889	0.747	0.860	0.724
RWE- α	-0.012	-0.041	0.166	0.309	0.161	0.312	0.142	0.275	0.935	0.951	0.898	0.922
RWE - $\hat{\alpha}$	-0.013	-0.034	0.167	0.307	0.164	0.309	0.142	0.275	0.935	0.943	0.895	0.920
IPWE- α	-0.016	-0.062	0.195	0.379	0.186	0.376	0.154	0.309	0.939	0.927	0.876	0.883
IPWE - $\hat{\alpha}$	-0.016	-0.057	0.196	0.380	0.190	0.376	0.155	0.309	0.937	0.936	0.872	0.888

CC, complex case; IPWE, inverse probability weighting estimator; RWE, reweighting estimator; SE, standard error.