Use of multiple imputation in supersampled nested case-control and case-cohort studies

Ørnulf Borgan[1], Ruth H. Keogh[2], and Aleksander Njøs[1]

[1]Department of Mathematics, University of Oslo

[2]London School of Hygiene and Tropical Medicine

Running headline: MI in supersampled case-control studies

Abstract

Nested case-control and case-cohort studies are useful for studying associations between covariates and time-to-event when some covariates are expensive to measure. Full covariate information is collected in the nested case-control or case-cohort sample only, while cheaply measured covariates are often observed for the full cohort. Standard analysis of such case-control samples ignores any full cohort data. Previous work has shown how data for the full cohort can be used efficiently by multiple imputation of the expensive covariate(s), followed by a full-cohort analysis. For large cohorts this is computationally expensive or even infeasible. An alternative is to supplement the case-control samples with additional controls on which cheaply measured covariates are observed. We show how multiple imputation can be used for analysis of such supersampled data. Simulations show that this brings efficiency gains relative to a traditional analysis and that the efficiency loss relative to using the full cohort data is not substantial.

Keywords: case-cohort, Cox regression, expensive covariates, large cohorts, missing covariate information, multiple imputation, nested case-control

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/sjos.12624

1 Introduction

Large cohort studies are widely used in epidemiology to investigate the effect of risk factors and other covariates on the time to an event of interest, like disease diagnosis or death. Cox regression is commonly used to analyse data from such studies. Estimation for Cox's model requires information on risk factors and other covariates for all individuals in the cohort, also for rare diseases where most individuals will not experience the event of interest. To avoid the collection of expensive covariate information, such as biological measurements, for all individuals in the cohort, it may be advantageous to adopt a cohort sampling design. For these designs, information on risk factors and other covariates are recorded for all individuals who experience the event of interest ("cases"), but only for a sample of the individuals who do not experience the event ("controls").

There are two main types of cohort sampling designs: *nested case-control studies* and *case-cohort studies*. The two types of cohort sampling designs differ in the way controls are selected. For nested case-control sampling (Thomas, 1977), one for each case selects at random a small number of controls from those at risk at the case's event time, and a new sample of controls is selected for each case. For case-cohort sampling (Prentice, 1986), a subcohort is selected at random from the full cohort at the outset of the study, and the individuals in the subcohort are used as controls at all event times when they are at risk.

As described in Section 2, the traditional analyses of nested case-control and case-cohort data only use information for individuals in the case-control sample, i.e., for the cases and controls/subcohort. However, while expensive covariate measurements may be available only for the case-control sample, there may be other cheaply measured covariates that are available for all individuals in the full cohort. The traditional analyses of sampled cohort data ignore this information. The sampled cohort plus the additional data available on the remainder of the full cohort may be viewed as a full cohort study with a large missing data problem, in which the expensive covariate measurements are missing by design for individuals outside the case-control sample. One approach for handling this missing data problem, is maximum likelihood estimation for the full cohort; see, e.g., Scheike & Martinussen (2004), Scheike & Juul (2004), Saarela *et al.* (2008), and Zeng & Lin (2014, 2018). The paper by Kulathinal & Arjas (2006) is also worth mentioning. They consider a Bayesian analysis of case-cohort data using the full cohort likelihood

and Bayesian data augmentation. Another popular approach for handling missing data problems, is multiple imputation (MI), which essentially is an approximation to a full Bayesian analysis. MI has good frequentist properties, and it is easier to implement than a full Bayesian analysis. In the context of Cox regression for sampled cohort data, MI has been studied by Keogh & White (2013) and Keogh *et al.* (2018); see also the review by Keogh (2018).

In studies with very large cohorts, maximum likelihood estimation or the use of MI for the entire cohort may be computationally very demanding or even infeasible. Focusing on MI, we in this paper investigate a middle way between a traditional sampled cohort study and a sampled cohort study with MI for the full cohort. The idea is to select a supersample of the sampled cohort by adding more controls to a nested case-control study and enlarging the subcohort for a case-cohort study. The expensive covariate measurements are imputed for the individuals in the supersample who are not in the original case-control sample, but not for the other individuals in the full cohort. The data for the imputed supersample may then be analysed using the traditional methods for analysing sampled cohort data.

The outline of the paper is as follows. In Section 2 we briefly describe Cox's regression model for the full cohort, and review the traditional ways of analysing sampled cohort data using this model. MI in Cox regression with cohort data and sampled cohort data is reviewed in Section 3, and in Section 4 we describe how MI may be modified when imputation is only performed for individuals in the supersample. In Section 5 we present a simulation study that illustrates how MI for the supersample compares with MI for the full cohort and the traditional methods for analysing sampled cohort data. Finally, in Section 6, we discuss our results and point out some directions for further research. Some simulation results that supplement those of Section 5 are given as online supporting material on the journal's web page.

2 Traditional analysis of sampled cohort data

We start out with a review of the traditional ways of analysing sampled cohort data. For ease of presentation, we assume that there is a single covariate X that is expensive to measure, while $\boldsymbol{Z}$ is a vector of cheaply measured covariates. The situation with more than one expensive covariate is briefly discussed in Section 6. The hazard rate for an

individual with covariate values $X = x$ and $\boldsymbol{Z} = \boldsymbol{z}$ is denoted $h(t \,|\, x, \boldsymbol{z})$. The time variable t may be age, time since the onset of a disease, or some other time scale relevant to the problem at hand. We assume that the covariates are related to the hazard rate by Cox's proportional hazards model

$$h(t \,|\, x, \boldsymbol{z}) = h_0(t) \exp\left(x\beta + \boldsymbol{z}^{\mathsf{T}}\boldsymbol{\gamma}\right). \tag{1}$$

Here β and $\boldsymbol{\gamma}$ are regression coefficients that describe the effects of the covariates on the hazard, while the baseline hazard $h_0(t)$ corresponds to the hazard of an individual with all covariates equal to zero.

We consider a cohort $\mathcal{C} = \{1, 2, \ldots, N\}$ of N independent individuals, and suppose for a moment that both X and $\boldsymbol{Z}$ are observed for the full cohort. The values of $(X, \boldsymbol{Z})$ for individual i are denoted $(x_i, \boldsymbol{z}_i)$. We do not observe the event of interest for all individuals. Due to censoring, we for each individual $i \in \mathcal{C}$ only observe (T_i, D_i), where T_i is the minimum of an event time and a censoring time, and $D_i = 1$ if the event is observed, and $D_i = 0$ otherwise. The values of T_i and D_i are denoted t_i and d_i. We assume throughout that the censoring and event times for individual i are independent given $X_i, \boldsymbol{Z}_i$, and that the censoring time and X_i are independent given $\boldsymbol{Z}_i$. The latter assumption is not required for the methods in this section, but it is needed when using MI in Sections 3 and 4.

Then for the full cohort, the regression coefficients β and $\boldsymbol{\gamma}$ in (1) are estimated by the values $\widehat{\beta}$ and $\widehat{\boldsymbol{\gamma}}$ that maximize Cox's partial likelihood

$$L_{\mathrm{coh}}(\beta, \boldsymbol{\gamma}) = \prod_{i \in \mathcal{E}} \frac{\exp\left(x_i\beta + \boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{\gamma}\right)}{\sum_{j \in \mathcal{R}(t_i)} \exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}. \tag{2}$$

Here $\mathcal{R}(t) = \{j \,|\, t_j \geq t\}$ is the risk set at time t, and $\mathcal{E} = \{j \,|\, d_j = 1\}$ is the set of all cases. Further the Breslow estimator of the cumulative baseline hazard $H_0(t) = \int_0^t h_0(u)du$ may be given as $\widehat{H}_{0,\mathrm{coh}}(t; \widehat{\beta}, \widehat{\boldsymbol{\gamma}})$, where

$$\widehat{H}_{0,\mathrm{coh}}(t; \beta, \boldsymbol{\gamma}) = \sum_{t_i \leq t} \frac{d_i}{\sum_{j \in \mathcal{R}(t_i)} \exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}. \tag{3}$$

It is well known that the maximum partial likelihood estimators for the full cohort have

similar properties as ordinary maximum likelihood estimators (Andersen & Gill, 1982). In particular, standard errors of the parameter estimates may be obtained from the observed information matrix.

As described below, nested case-control and case-cohort data are traditionally analysed using modified versions of Cox's partial likelihood and the Breslow estimator.

2.1 Nested case-control studies

The controls in a nested case-control study are selected as follows. If the event of interest is observed for an individual i at time t_i, one selects at random a small number m of controls from $\mathcal{R}(t_i)\setminus\{i\}$, i.e., the risk set with the case excluded. The set of size $m+1$ that consists of the case and the sampled controls is denoted a sampled risk set and denoted $\widetilde{\mathcal{R}}(t_i)$. The expensive covariate is measured for the individuals in the sampled risk sets, but it is not needed for the other individuals in the full cohort.

Traditionally, estimation of the regression coefficients in a nested case-control study is based on the partial likelihood

$$L_{\mathrm{ncc}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i\in\mathcal{E}} \frac{\exp\left(x_i\beta + \boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{\gamma}\right)}{\sum_{j\in\widetilde{\mathcal{R}}(t_i)} \exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}. \tag{4}$$

Note that (4) is similar to the full cohort partial likelihood (2), except that the sum in the denominator is only over individuals in a sampled risk set. For computing one may use standard software for Cox regression, like coxph in the survival library in R, formally treating the label of the sampled risk sets as a stratification variable in the Cox regression.

The cumulative baseline hazard may be estimated by a Breslow-type estimator given as $\widehat{H}_{0,\mathrm{ncc}}(t;\widehat{\beta},\widehat{\boldsymbol{\gamma}})$, where

$$\widehat{H}_{0,\mathrm{ncc}}(t;\beta,\boldsymbol{\gamma}) = \sum_{t_i\le t} \frac{d_i}{\sum_{j\in\widetilde{\mathcal{R}}(t_i)} w(t_i)\exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}, \tag{5}$$

and $\widehat{\beta}$ and $\widehat{\boldsymbol{\gamma}}$ are obtained by maximizing (4). In (5), the weights are $w(t_i) = N(t_i)/(m+1)$, where $N(t) = |\mathcal{R}(t)|$ is the number of individuals at risk at time t.

The maximum partial likelihood estimators $\widehat{\beta}$ and $\widehat{\boldsymbol{\gamma}}$ obtained by maximizing (4) enjoy similar large sample properties as ordinary maximum likelihood estimators (Goldstein &

Langholz, 1992; Borgan *et al.*, 1995).

2.2 Case-cohort studies

For the case-cohort design, a subcohort $\widetilde{\mathcal{C}}$ of size n is selected at random from the full cohort at the outset of the study. The individuals in the subcohort are used as controls at all event times when they are at risk. The expensive covariate is measured for all individuals in the subcohort as well as for cases occurring outside the subcohort, but it is not needed for the other individuals in the full cohort.

Different methods have been suggested for estimating the regression coefficients for case-cohort data. The methods are based on weighted pseudo-likelihoods of the form

$$L_{\mathrm{cch}}(\beta, \boldsymbol{\gamma}) = \prod_{i \in \mathcal{E}} \frac{\exp\left(x_i\beta + \boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{\gamma}\right)}{\sum_{j \in \widetilde{\mathcal{S}}(t_i)} w_j \exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}, \tag{6}$$

but the methods differ in the choice of weights w_j and sampled risk sets $\widetilde{\mathcal{S}}(t_i)$.

For Prentice's original pseudo-likelihood, all weights are one and $\widetilde{\mathcal{S}}(t_i) = \mathcal{S}(t_i) \cup \{i\}$, where $\mathcal{S}(t_i) = \{j \,|\, t_j \geq t_i, j \in \widetilde{\mathcal{C}}\}$ is the set of all subcohort individuals at risk at time t_i. Self & Prentice (1988) considered the modification where $\widetilde{\mathcal{S}}(t_i) = \mathcal{S}(t_i)$ also when the case is not in the subcohort.

In Prentice's pseudo-likelihood, a case outside the subcohort only makes a contribution at its event time. In order to make use of the information from the cases at all times when they are at risk, we may adopt an inverse probability weighted (IPW) pseudo-likelihood (Kalbfleisch & Lawless, 1988). Then we let $\widetilde{\mathcal{S}}(t_i) = \{j \,|\, t_j \geq t_i, j \in \widetilde{\mathcal{C}} \cup \mathcal{E}\}$, where $\mathcal{E}$ is the set of all cases. So now $\widetilde{\mathcal{S}}(t_i)$ consists of all subcohort individuals at risk at time t_i together with all cases who are at risk at that time. The weights are given as $w_j = 1/p_j$, where p_j is the probability that individual j is included in the case-control sample. The cases are included with probability one, so the weights are $w_j = 1$ for all cases (whether they are in the subcohort or not). We let $N^{(0)}$ and $n^{(0)}$ be the number of individuals in the cohort and subcohort, respectively, who do not experience the event of interest. Then an individual who is not a case, is included in the subcohort with probability $p_j = n^{(0)}/N^{(0)}$, and hence may be given the weight $w_j = N^{(0)}/n^{(0)}$.

Pseudo-likelihoods of the form (6) are not partial likelihoods, so the maximum pseudo-

likelihood estimators do not enjoy similar large sample properties as ordinary maximum likelihood estimators. In particular, standard errors cannot be computed directly from the observed pseudo-information matrix. But it has been shown that the maximum pseudo-likelihood estimators are approximately multivariate normally distributed, and estimators for their variance-covariance matrices have been worked out; see, e.g., Self & Prentice (1988); Therneau & Li (1999); Borgan *et al.* (2000), and Samuelsen *et al.* (2007). The estimators based on (6) may be computed using the cch-command in the survival package in R. Here method="Prentice", method="SelfPrentice", and method ="LinYing" give the three estimators mentioned above.

For case-cohort data we may estimate the cumulative baseline hazard by a Breslow-type estimator given as $\widehat{H}_{0,\mathrm{cch}}(t;\widehat{\beta},\widehat{\boldsymbol{\gamma}})$, where

$$\widehat{H}_{0,\mathrm{cch}}(t;\beta,\boldsymbol{\gamma}) = \sum_{t_i \le t} \frac{d_i}{\sum_{j\in\widetilde{\mathcal{S}}(t_i)} \widetilde{w}_j \exp\left(x_j\beta + \boldsymbol{z}_j^{\mathsf{T}}\boldsymbol{\gamma}\right)}, \tag{7}$$

and $\widehat{\beta}$ and $\widehat{\boldsymbol{\gamma}}$ are obtained by maximizing (6). For the IPW estimator, the sampled risk sets $\widetilde{\mathcal{S}}(t_i)$ and weights in (7) are as given above, so $\widetilde{w}_j = 1$ for cases and $\widetilde{w}_j = N^{(0)}/n^{(0)}$ for non-cases. For the Prentice estimator, we may use $\widetilde{w}_j = N/n$ for all individuals in the subcohort and let $\widetilde{\mathcal{S}}(t_i) = \mathcal{S}(t_i)$ (Prentice, 1986, p. 6).

3 Multiple imputation (MI) in Cox regression

In this section we outline methods for MI of missing covariate data in Cox regression in analysis of cohort data, and their extension to sampled cohort data in which an expensive covariate is missing for everyone outside the case-control sample. We assume data are missing at random (MAR), which can be expressed as the assumption that the probability of missingness in a given variable is independent of missing data conditional on observed data (Rubin, 1987; Carpenter & Kenward, 2013, Chapter 1).

3.1 MI in full cohort studies

We refer to the book of Carpenter & Kenward (2013), for example, for an overview of the theory of MI, and here focus on its use when the analysis model is a Cox regression. In brief, the MI procedure involves imputing missing values in covariates by taking a random

draw from the estimated posterior distribution of the missing data given the observed data (including the outcome). This is repeated K times to give K 'complete' imputed data sets in which missing values in covariates are filled in. The analysis model, in our case a Cox regression, is then fitted to each of the K imputed data sets to give K estimates of the regression coefficients. Pooled estimates and a corresponding variance-covariance matrix are then obtained using 'Rubin's rules' (e.g., Carpenter & Kenward, 2013, p. 45-46).

First consider a full cohort study in which a single covariate X has missing data, and $\boldsymbol{Z}$ denotes a vector of fully observed covariates. It is assumed that outcome information (T, D) is observed for all individuals in the cohort. The aim in MI is therefore to impute missing values of X from the conditional distribution of X given $\boldsymbol{Z} = \boldsymbol{z}$, $T = t$ and $D = d$. It has been shown that when the analysis model is a Cox regression, as in (1), this conditional distribution is not of any standard form, such as a normal distribution, which presents a challenge for obtaining imputations. To address this, two main approaches have been described for imputation of missing data on covariates in Cox regression.

The first is that of White & Royston (2009) who derived an approximate form for the imputation model, which is a regression of X on $\boldsymbol{z}$, d, and $\widehat{H}(t)$, where $\widehat{H}(t)$ denotes the Nelson-Aalen estimate of the cumulative hazard at the individual's event or censoring time t. We refer to this approach as MI-Approx.

The second approach, described by Bartlett *et al.* (2015), uses rejection sampling and an iterative procedure to impute missing values of X from the conditional distribution that is compatible with the outcome (or 'substantive') model, which here is a Cox proportional hazards model. The basis of the method of Bartlett *et al.* (2015) is that we, for each iteration, draw a potential value for X from a 'proposal distribution', and then use a 'rejection rule' to decide whether to accept the potential value as a draw from the distribution of interest, i.e., the conditional distribution of X given $\boldsymbol{Z} = \boldsymbol{z}$, $T = t$ and $D = d$. In a given iteration, the rejection rule is to accept a potential value x^* as an imputed value for a missing value of X if

$$\begin{aligned} U &\leq \exp\{-H_0^*(t)e^{x^*\beta^* + \boldsymbol{z}^\mathsf{T}\boldsymbol{\gamma}^*}\} && \text{if } d = 0, \\ U &\leq H_0^*(t)\exp\{1 + x^*\beta^* + \boldsymbol{\gamma}^{(k)\mathsf{T}}\boldsymbol{z} - H_0^*(t)e^{x^*\beta^* + \boldsymbol{z}^\mathsf{T}\boldsymbol{\gamma}^*}\} && \text{if } d = 1. \end{aligned} \tag{8}$$

Here U denotes a random draw from a standard uniform distribution, $\beta^*, \boldsymbol{\gamma}^*$ denote pos-

terior draws of the model parameters, obtained after fitting the Cox regression analysis model to the current imputed data, and $H_0^*(t) = \widehat{H}_{0,\text{coh}}(t; \beta^*, \boldsymbol{\gamma}^*)$ denotes the cumulative baseline hazard obtained using these parameter draws; cf. formula (3). This approach is referred to as the 'substantive model compatible full conditional specification' (SMC-FCS) method. Following earlier work, we refer to it as MI-SMC.

MI-Approx involves specification of the model for each covariate with missingness given the other covariates and the outcome, which in our context is $f(X|\boldsymbol{Z}, T, D)$. For continuous X a conditional normal distribution is often assumed as an approximation. This approximation is derived from assuming X is normally distributed conditional on $\boldsymbol{Z}$, in the case of continuous X. MI-SMC requires specification of the model $f(X|\boldsymbol{Z})$, which typically takes a more standard form than $f(X|\boldsymbol{Z}, T, D)$. For a continuous X it is common to assumes that $f(X|\boldsymbol{Z})$ is a normal distribution. Because MI-SMC accommodates non-linear terms in the Cox regression model, it is possible under this approach to instead specify a distribution for a transformation $g(X)$ (e.g. $g(X) = \log X$) given $\boldsymbol{Z}$, with the imputations of $g(X)$ then being back-transformed for use in the Cox regression including X and $\boldsymbol{Z}$. We refer to Bartlett *et al.* (2015) for a detailed discussion of compatibility of models used in MI and imputation model mis-specification, including extensions to the situation in which there is more than one covariate with missingness.

For overviews of the MI-Approx and MI-SMC methods in the content of Cox regression, see Carpenter & Kenward (2013, Chapter 8), Keogh *et al.* (2018), and Keogh (2018). The MI-Approx method has been found to perform well in a range of circumstances, in particular for rare events. But it does not apply when there are non-linear terms involving the missing covariate X in the proportional hazards model, including interactions between X and $\boldsymbol{Z}$. MI-SMC accommodates non-linear terms (including interactions) because it obtains draws of X in such as way that they are drawn from a distribution that is compatible with the proportional hazards model. MI-approx can be implemented using the mice package in R and MI-SMC using the smcfcs package.

3.2 MI in sampled cohort studies

Nested case-control and case-cohort studies can be viewed as full cohort studies with data missing for the expensive covariate X. The methods for MI in a full cohort study

outlined in the preceding Section 3.1 can be applied directly to impute missing values of X for individuals in the full cohort who were not sampled to the nested case-control or case-cohort study. The use of MI in this way for both nested case-control and case-cohort studies was described by Keogh & White (2013), who assessed the two imputation approaches using simulation studies.

There are two main ways in which the use of MI in this setting differs from its use in a more standard missing data setting. Firstly, for sampled cohort data the expensive covariate is typically unmeasured for a very large proportion of the full cohort. This approach therefore involves imputing a large proportion of values for the expensive covariate, and one might expect that the consequences of mis-specifying the imputation model could be severe in this situation. Secondly, in the standard MI setting, the assumption that data are missing at random (MAR) is a crucial assumption that we cannot be certain is met. However, the MAR assumption for sampled cohort data follows by the design of the studies, and in this sense the use of MI in this situation is 'safer' than in the standard setting.

4 Supersampling of sampled cohorts

We now describe more in detail how we obtain a supersample for nested case-control and case-cohort data, and how we may use MI to impute missing covariate values for the supersample. In general, a supersampled nested case-control study is a standard nested case-control study augmented with additional controls for each case, and a supersampled case-cohort study is a standard case-cohort study augmented with an additional random subcohort. In both types of study, the expensive covariate X is observed only for individuals in the original case-control sample, whereas $\boldsymbol{Z}$, T, and D are observed for all individuals in the supersample. The analysis makes use of the individuals in supersample, but it does not use the remaining individuals in the full cohort. We start out by considering a supersampled case-cohort study since imputation is more straightforward for this design than for the nested case-control design.

4.1 Supersampling for case-cohort data

For a case-cohort study, we have a subcohort $\widetilde{\mathcal{C}}$ of size n selected by simple random sampling from the full cohort $\mathcal{C}$. The expensive covariate is measured for individuals in the case-control sample $\widetilde{\mathcal{C}} \cup \{j \,|\, d_j = 1\}$, but not for the remaining individuals in the cohort. We now select another random sample $\mathcal{C}^*$ of size n^* from $\mathcal{C} \setminus \widetilde{\mathcal{C}}$, i.e., from the individuals in the cohort who are not in the original subcohort. In this way we obtain a supersampled case-cohort study with subcohort $\widetilde{\mathcal{C}} \cup \mathcal{C}^*$ of size $n + n^*$. Note that for the supersampled case-cohort data, the expensive covariate X is not measured for individuals in $\mathcal{C}^* \cap \{j \,|\, d_j = 0\}$, i.e., for individuals in the supersample who are not in the original case-cohort sample. The inexpensive covariate, $\boldsymbol{Z}$, is observed for all individuals in the supersample. Because X is missing for individuals in $\mathcal{C}^* \cap \{j \,|\, d_j = 0\}$ we use MI to impute it.

In Section 3.2 we outlined how a case-cohort study within a full cohort may be viewed as a full cohort study with missing data, and imputation methods for full cohort data can therefore be used. The supersampled case-cohort data is not a full cohort study, and therefore a modified approach is needed to perform the imputation. Keogh *et al.* (2018) have discussed how the MI-Approx and MI-SMC methods can be used for imputing missing data on covariates within a case-cohort sample, assuming missingness is at random. The supersampled case-cohort study represents a similar situation, except that the data on X are missing by design and therefore the missingness in X is at random. Hence we may use the imputation methods described by Keogh *et al.* (2018) as summarized in the following paragraphs.

In MI-Approx the imputation model is a regression of X on $\boldsymbol{z}$, d, and $\widehat{H}(t)$. Here $\widehat{H}(t)$ is the Nelson-Aalen estimate of the cumulative hazard at the individual's event or censoring time t. Since the Nelson-Aalen estimate only depends on the T_i's and D_i's, it may be computed from the available data for the full cohort, and it is not computationally intensive to obtain for large cohorts. The MI-Approx method can then be applied directly in a supersampled case-cohort study, by fitting the imputation model using all individuals in the supersampled data.

The MI-SMC approach requires an estimate $H_0^*(t)$ of the cumulative baseline hazard at each person's event or censoring time; cf. formula (8). As outlined in Section 2.2, the

cumulative baseline hazard can be estimated from a case-cohort study using $\widehat{H}_{0,\mathrm{cch}}(t;\beta,\boldsymbol{\gamma})$ as given in equation (7). This estimator can also be applied to the supersampled case-cohort data. For the MI-SMC approach we then use $H_0^*(t) = \widehat{H}_{0,\mathrm{cch}}(t;\beta^*,\boldsymbol{\gamma}^*)$ computed at each iteration of the MI-SMC procedure using posterior draws β^*, $\boldsymbol{\gamma}^*$ of the model parameters and the most recent set of imputed values of X. The MI-SMC approach also involves estimating the distribution of $X|\boldsymbol{Z}$, as draws of X from this distribution are used as potential imputed values in the rejection sampling. The parameters of the distribution of $X|\boldsymbol{Z}$ can be estimated from the supersampled subcohort, which is a random sample from the full cohort.

After obtaining imputed values of X for individuals in the supersample using MI-Approx or MI-SMC, a standard case-cohort analysis is applied to each imputed data set and the parameter estimates and their standard errors combined using Rubin's rules.

4.2 Supersampling for nested case-control data

In a nested case-control study, we for each $i \in \mathcal{E}$ select at random m controls from $\mathcal{R}(t_i) \setminus \{i\}$. The sampled risk set $\widetilde{\mathcal{R}}(t_i)$ consists of the case i and its sampled controls. The expensive covariate is measured for individuals in the case-control sample $\cup_{i\in\mathcal{E}}\widetilde{\mathcal{R}}(t_i)$, but not for the remaining individuals in the cohort. For each $i \in \mathcal{E}$ we now add more controls by selecting at random m^* individuals from $\mathcal{R}(t_i) \setminus \widetilde{\mathcal{R}}(t_i)$. The set of these additional controls is denoted $\mathcal{R}^*(t_i)$. In this way we obtain a supersampled nested case-control study with sampled risk sets $\widetilde{\mathcal{R}}(t_i) \cup \mathcal{R}^*(t_i)$ of size $m + m^*$. Note that for the supersampled nested case-control data, the expensive covariate is not measured for individuals in $(\cup_{i\in\mathcal{E}}\mathcal{R}^*(t_i)) \cap \{j \mid d_j = 0\}$, i.e., for individuals in the supersample who are not in the original case-control sample.

As described in Section 3.1, the aim of MI for cohort data is to impute a missing value of X from the conditional distribution of X given the observed data $\boldsymbol{Z} = \boldsymbol{z}$, $T = t$ and $D = d$, and this may be achieved using the exact or an approximate conditional distribution. As the extended subcohort $\widetilde{\mathcal{C}} \cup \mathcal{C}^*$ for supersampled case-cohort data is a random sample from the full cohort, this also applies (with $d = 0$) for supersampled case-cohort data (Section 4.1). However, the controls in a supersampled nested case-control study are not a random sample from the full cohort. Rather, if an event occurs at time s,

all individuals with $T > s$ are potential controls, including individuals who later become a case. Therefore, in a supersampled nested case-control study, the missing value of X for a control at time s should be imputed from the conditional distribution of X given $\boldsymbol{Z} = \boldsymbol{z}$ and $T > s$.

To see how this imputation model relates to the one for the full cohort and supersampled case-cohort data, we may for an individual with $T > s$ introduce $T^{(s)} = s$ and $D^{(s)} = 0$, corresponding to the censored event time and event indicator we would have if the individual had been censored at time s. Then the imputation model for a control at time s may be given as the conditional distribution of X given $\boldsymbol{Z} = \boldsymbol{z}$, $T^{(s)} = s$ and $D^{(s)} = 0$. This is similar to the imputation model for the full cohort and supersampled case-cohort data, but we condition on $T^{(s)} = s$ and $D^{(s)} = 0$ rather than $T = t$ and $D = d$. Thus, for MI-Approx, we should use the Nelson-Aalen estimate evaluated at the time of the case to which a control is matched, and not the control's own censoring time.

To apply MI-SMC in a supersampled nested case-control study, we can use the cumulative baseline hazard estimator (5). To obtain imputations for control individuals it is appropriate to use the estimate $H_0^*(s) = \widehat{H}_{0,\mathrm{ncc}}(s; \beta^*, \boldsymbol{\gamma}^*)$ obtained at the event time s of the case to which the control is matched. We also require estimates of the parameters of the distribution of $X|\boldsymbol{Z}$, which can be obtained from a regression of X on $\boldsymbol{Z}$ in the controls.

Some individuals can feature as controls for more than one case, and in both MI-approx and MI-SMC we then obtain a new imputed value for each duplicate. After obtaining imputed values of X for individuals in the supersample, a standard nested case-control analysis is applied to each imputed data set and the parameter estimates and their standard errors combined using Rubin's rules.

5 Simulation study

5.1 Simulation aims and implementation

We use a simulation study to illustrate the methods described above and to assess their performance. The simulation study is structured following the guidance by Morris *et al.* (2019).

Aims

The aims of the simulation study are to illustrate the proposed multiple imputation methods for analysis of supersampled nested case-control and case-cohort studies. Specifically we wish to check that the methods give approximately unbiased estimates of the parameters of interest where expected, and to investigate the efficiency gains from the use of supersampled data compared with the original study data. We also wish to check that the standard errors obtained by Rubin's rules are approximately unbiased, and that the estimates have good coverage.

Data-generating mechanisms

Data are generated for a full cohort of N individuals. Three covariates, (X, Z_1, Z_2), are generated for each individual. X is the expensive covariate, which we will later assume is observed only in the nested case-control or case-cohort study, and (Z_1, Z_2) are cheaply measured covariates. We generated Z_1 from a standard normal distribution, and independently Z_2 from a Bernoulli distribution with probability 0.5. The expensive covariate X was generated by

$$X = 0.25Z_1 + 0.25Z_2 + \varepsilon \tag{9}$$

where ε, independent of Z_1 and Z_2, is a random variable with mean 0 and variance 1. The distribution of ε was assumed to follow a normal distribution or a log-normal distribution (shifted to have mean 0). The latter was generated as $\varepsilon = e^Y - e^{\sigma^2/2}$, where Y is $N(0, \sigma^2)$-distributed with $\sigma = 0.694$. Event times T_E were generated from the Weibull hazard model

$$h(t|x, z_1, z_2) = \lambda_E \kappa t^{\kappa-1} e^{\beta_X x + \beta_{Z1} z_1 + \beta_{Z2} z_2 + \beta_{XZ1} x z_1}.$$

We used shape parameter $\kappa = 4$ and values of the scale parameter λ_E as described below, and considered a scenario with no interaction between X and Z_1 using $\beta_X = 1$, $\beta_{Z1} = 0.5$, $\beta_{Z2} = 1$, $\beta_{XZ1} = 0$, and a scenario with an interaction between X and Z_1 using $\beta_X = 1$, $\beta_{Z1} = 0.5$, $\beta_{Z2} = 1$, $\beta_{XZ1} = 0.5$. The scenario with interaction was only considered when ε in (9) was normally distributed. Censoring times T_C were generated from a Weibull hazard model using shape and scale parameter values $\kappa = 4$ and λ_C, and independent of

the covariates. Administrative censoring was imposed at time 15 years. The observed time for each individual is therefore $T = \min(T_E, T_C, 15)$ and D denotes the event indicator.

We consider two sample sizes for the cohort of $N = 5000$ and $N = 25000$. For all six scenarios (two sample sizes, two distributions of X without interaction with Z_1, one distribution of X with interaction with Z_1), the values of λ_E and λ_C were chosen to give approximately 250 cases in the cohort (5% of individuals having the event when $N = 5000$ and 1% having the event when $N = 25000$) and approximately 35% of the individuals administratively censored at time 15 years. Values used for λ_E and λ_C are shown in Supplementary Table S1.

A nested case-control sample was obtained within the full cohort using all cases and one control per case. A case-cohort sample was obtained by including all cases and a random subcohort of 5% of the individuals when $N = 5000$ and 1% when $N = 25000$, resulting in the subcohort being approximately the same size as the number of controls in the nested case-control sample.

For both study designs we obtained a small superset sample and a large superset sample. For the nested case-control study we obtained a small superset sample by selecting 3 additional controls for each case, giving 4 controls per case in total. For the large superset sample we selected an additional 11 controls for each case, giving 12 controls per case in total. For the case-cohort study we obtained a small superset sample by adding an additional random subcohort of 15% of the individuals when $N = 5000$ and an additional random subcohort of 3% of the individuals when $N = 25000$, both of which correspond to a four-fold increase in the size of the subcohort. The large superset sample was obtained by increasing the size of the subcohort 12 fold, corresponding to an additional 55% of individuals when $N = 5000$ and an additional 11% of individuals when $N = 25000$.

In each simulated cohort, the covariate X was set to be missing for individuals not in the original nested case-control or case-cohort data sets, except for when a full-cohort analysis is being performed.

Target of analysis

The estimands of interest for the simulation study are the log hazard ratios (log HRs) $\beta_X, \beta_{Z1}, \beta_{Z2}$, and β_{XZ1} (in scenarios with the interaction), and their standard errors.

Methods

For each simulated full cohort we performed the following analyses:

(i) a full cohort analysis assuming X, Z_1, Z_2 are fully observed on all individuals;

(ii) a traditional analysis of the nested case-control study and of the case-cohort study, excluding individuals outside the case-control sample;

(iii) an analysis of the nested case-control or case-cohort sample in addition to the rest of the cohort, assuming X is observed only in the nested case-control or case-cohort sample and imputing X for individuals in the remainder of the full cohort using the methods described in Section 3.2;

(iv) an analysis of the supersampled nested case-control or case-cohort study using a small superset, assuming X is missing for the individuals in the superset who are not in the original nested case-control or case-cohort study, and using the imputation methods described in Section 4;

(v) as in (iv) but using the large superset.

In (iii) (Z_1, Z_2) are fully observed for individuals in the full cohort who are not in the nested case-control or case-cohort study. In (iv) and (v) (Z_1, Z_2) are fully observed in the supersampled nested case-control or case-cohort study. In all MI analyses we use 10 imputed data sets. In the MI-SMC analyses we used 500 iterations, and found that using a lower number (200 iterations for example) could lead to bias; in particular for situation (iii) with cohort size $N = 25000$. In the case-cohort study analyses we obtained estimates using both the Prentice and IPW estimators. When the Prentice estimator was used in the analysis, we also used the Prentice method to obtain the cumulative baseline hazard estimates in MI-SMC, and similarly for the IPW estimator; see the final paragraph of Section 2.2.

Performance measures

We generated $n_{\text{sim}} = 1000$ data sets for each of the six full cohort scenarios. In each analysis we obtain estimates of the log HRs $\beta_X, \beta_{Z1}, \beta_{Z2}$, and β_{XZ1} (in scenarios including

the interaction term), and their model-based standard errors and 95% confidence intervals (CIs). For each log HR we obtain an estimate of the bias $n_{\text{sim}}^{-1}\sum_{j=1}^{n_{\text{sim}}}(\hat{\beta}_j-\beta)$ and the empirical standard deviation ('Emp SE') $\sqrt{(n_{\text{sim}}-1)^{-1}\sum_{j=1}^{n_{\text{sim}}}(\hat{\beta}_j-\bar{\hat{\beta}})^2}$, where $\hat{\beta}_j$ denotes the estimate from the jth simulated data set, β denotes the true value, and $\bar{\hat{\beta}}$ denotes the average of the n_{sim} estimates. We also obtained the average of the model-based estimates of the standard errors ('Model SE') of the log HR estimates, and the 95% coverage, meaning the percentage of the n_{sim} 95% CIs that contain the true value. A well-performing method would have bias close to zero, Emp SE similar to Model SE, and coverage close to the nominal level of 95%. To ease the comparison of the analysis methods, we for each method also report the root mean squared error ('RMSE') and the relative efficiency ('Rel Eff') of the log HR estimates relative to a full cohort analysis. RMSE is given as the square root of the sum of the squared bias and Emp SE^2, and it provides information about the bias-variance trade-off for a method. Rel Eff is defined as the empirical variance from a full cohort analysis (Emp SE^2) divided by the empirical variance from a given analysis method. For methods that are approximately unbiased, Rel Eff informs us how a method compares to a full cohort analysis. For each performance measure we estimated the Monte Carlo standard errors (Morris *et al.*, 2019).

Software

The simulation was conducted in R and code enabling the simulation to be replicated are provided at `https://github.com/ruthkeogh/supersampling`. We used the survival package, including the cch function for analysis of case-cohort studies. The multiple imputation for the MI-Approx method was implemented using mice. For the MI-SMC imputation approach we used a modified version of the smcfcs function. For use with a case-cohort study (standard or supersampled) modifications were made to smcfcs to allow an analysis using the Prentice or IPW estimators. For use with a nested case-control study (standard or supersampled) modifications were made to smcfcs so that potential imputations of X considered in the rejection sampling are obtained from a regression of X on (Z_1, Z_2) using data from the controls (which includes some cases) instead of non-cases. A further modification is that the cumulative baseline hazard estimate used in the imputation procedure is obtained at the event time of the case in each individual's matched set, rather than at

their own censoring time. The modified version of smcfcs also restricts the function to those components required for the analyses performed in this simulation, which can help others to more easily follow the code to gain a better understanding of the analysis.

5.2 Simulation results

Tables 1 and 2 show the results for the full cohort of size 25000 for the nested case-control study for the situations without the $X \times Z_1$ interaction and with the $X \times Z_1$ interaction when ε in (9) is normally distributed. The results when there is no interaction and ε follows a lognormal distribution (shifted to have mean zero) are given in Table 3. Note that for the situations of Tables 1 and 2, the imputation model is correctly specified, while it is mis-specified for the situation of Table 3. Tables 4–6 show the corresponding results for the case-cohort setting using a full cohort of size 25000. In the main text we show results from case-cohort analyses obtained using the IPW estimator. Results using the Prentice estimator were also obtained and are shown in Supplementary Tables S2–S4. Results for the full cohort of size 5000 are shown in Supplementary Tables S5–S7 for nested case-control and S8–S10 for case-cohort.

As we expect, the full cohort analysis gives unbiased log HR estimates, correct standard errors (comparing Emp SE with Model SE) and coverage at the nominal level.

`[Insert Table 1 about here]`

5.2.1 Nested case-control results

We first consider the situation without interaction (Table 1) when the imputation model is correctly specified. Here a traditional nested case-control analysis gives a minor bias in the estimates for all parameters. But the root mean square errors are very close to the empirical SEs, so the bias is of little importance compared to the variability of the estimates. The averages of the model-based SEs ('Model SE') are close to the empirical standard deviations of the estimates ('Emp SE'), and the method has coverage around 95%. The relative efficiency of the log HR estimates relative to the full cohort estimates is somewhat below 20% for $\hat{\beta}_X$ and $\hat{\beta}_{Z1}$ and somewhat below 30% for $\hat{\beta}_{Z2}$.

Method (iii), which uses the full cohort with X imputed for the individuals who are

not in the nested case-control sample, gives unbiased log HR estimates, correct SEs and fairly good coverage. The results are very similar using MI-Approx and MI-SMC. Method (iii) results in substantial gains in efficiency relative to the traditional analysis; the relative efficiencies are about 30% for $\hat{\beta}_X$ and about 70% for $\hat{\beta}_{Z1}$ and $\hat{\beta}_{Z2}$. Thus, as known from earlier work, the gain in efficiency is largest for the coefficients of Z_1 and Z_2 which are assumed observed in the full cohort.

The methods using a supersampled nested case-control study perform well using both MI-Approx and MI-SMC, giving approximately unbiased log HR estimates both for a small and a large superset. The SEs are mainly correct and the coverage is close to 95%. An exception is the MI-SMC estimate of β_X, where the empirical SE is somewhat larger than the model-based SE and the coverage is 90%.

As may be expected, the efficiencies for the supersampled nested case-control studies relative to a full cohort analysis are between the relative efficiencies for the traditional nested case-control analysis and method (iii) with imputation for the full cohort. More specifically, we find that MI-Approx with a small/large superset was 79/82% efficient for estimation of β_X relative to method (iii). For β_{Z1} the corresponding relative efficiencies were 57/76%, and for β_{Z2} they were 80/92%. This shows that a large part of the information contained in the full cohort may be extracted by using a supersample. Not surprisingly, using a large superset typically results in smaller SEs and higher relative efficiencies, especially for the coefficients of Z_1 and Z_2, though the gains in efficiency are not substantial. Again an exception is the MI-SMC estimate of β_X, where the empirical SE is slightly larger for the large superset than the small superset.

[Insert Table 2 about here]

We then consider the situation with interaction (Table 2). Here a traditional nested case-control analysis gives somewhat biased estimates for all parameters, except for the interaction parameter β_{XZ1}. But the root mean square errors are fairly close to the empirical SEs, so the biases are clearly of less importance than the variability of the estimates. The model SEs tend to be slightly too small, but coverage is close to 95%. The relative efficiency of the log HR estimates relative to the full cohort estimates are all between 8% and 17%.

As noted in the last paragraph of Section 3.1, MI-approx does not apply for the interaction model. This is the reason why this imputation method gives substantial bias in parameter estimates (for the full cohort and the supersamples) for all parameters except β_{Z2}. But MI-SMC performs well for the interaction model. Method (iii) with MI-SMC imputation for the full cohort gives unbiased estimates, correct SEs and good coverage. Further, there is a huge gain in efficiency relative to the traditional analysis; the relative efficiencies are around 60% for $\hat{\beta}_X$ and $\hat{\beta}_{Z2}$, and around 90% for $\hat{\beta}_{Z1}$ and $\hat{\beta}_{XZ1}$.

The supersampled nested case-control studies perform fairly well for the interaction model when using MI-SMC. There is a small bias in the estimates of β_X, β_{Z1} and β_{XZ1}, and the model SEs tend to be slightly too large for the small superset. The coverage is fairly close to 95% for all parameters, but a bit higher for the small than the large superset. The relative efficiencies for the small superset are about 30% for $\hat{\beta}_X$ and $\hat{\beta}_{Z2}$, 40% for $\hat{\beta}_{Z1}$, and 20% for $\hat{\beta}_{XZ1}$. For the large superset the corresponding values are about 40%, 50%, and 25%. These relative efficiencies are clearly lower than those of method (iii). But the relative efficiencies for the small superset are about 2.5 times higher than those of the traditional nested case-control analysis, while they are at least 3 times higher for the large superset. So there is a lot to gain by using a supersampled nested case-control study compared to a traditional analysis.

[Insert Table 3 about here]

Finally, we consider the situation where there is no interaction, but the imputation model is mis-specified (Table 3). Here a full cohort analysis gives similar results as for the correctly specified imputation model, except that the estimate of β_X has clearly lower SE than in Table 1. Also a traditional nested case-control analysis gives similar results as for the correctly specified imputation model, but the estimates are a bit more biased and the SEs are somewhat larger than in Table 1. Imputation using MI-Approx gives substantial bias for all parameter estimates both when X is imputed for the full cohort [method (iii)] and when imputation is restricted to a supersample [methods (iv) and (v)]. Method (iii) with MI-SMC imputation for the full cohort also gives a substantial bias. But method (iv), which uses MI-SMC imputation with a small superset, gives almost unbiased estimates, correct SEs, and coverage close to 95%. Further, the relative efficiencies of method (iv) are

more than twice those of the traditional nested case-control analysis. MI-SMC imputation with the large superset [method (v)] gives somewhat larger biases, and the model SE for the estimate of β_X is too low, which results in poor coverage for this parameter.

For the situation without interaction, the results for cohort size 5000 (Supplementary Table S5) are quite similar to those obtained with cohort size 25000. For the situation with interaction, the broad picture for cohort size 5000 (Supplementary Table S6) is similar to that for cohort size 25000. But the traditional method has slightly less bias for cohort size 5000 and most SEs are a bit smaller. The relative efficiencies of the supersampling methods using MI-SMC are similar or larger for cohort size 5000 compared to cohort size 25000. Finally, when the imputation model is mis-specified, the results for cohort size 5000 (Supplementary Table S7) are broadly speaking similar to those obtained with cohort size 25000. But using MI-SMC with a large superset results in a larger bias for estimation of β_X and a smaller bias for estimation of β_{Z2}.

5.2.2 Case-cohort results

We first consider the situation without interaction (Table 4). Here the traditional IPW estimator for case-cohort data gives a fairly large bias in the estimate of β_X and some bias in the estimate of β_{Z1}. The model SEs tend to be too small, and coverage is too low, in particular for β_X and β_{Z1}.

[Insert Table 4 about here]

Method (iii), which uses the full cohort with X imputed for the individuals who are not in the case-cohort sample, gives result which are close to those obtained with method (iii) for nested case-control data (Table 1). Thus, both for MI-Approx and MI-SMC, we obtain unbiased log HR estimates, correct SEs and good coverage. Also the relative efficiencies are close to those obtained with method (iii) for nested case-control data.

The methods using supersampled case-cohort data give approximately unbiased estimates for all parameters. For estimation of β_{Z1} and β_{Z2}, the empirical SEs are fairly close to those obtained for the supersampled nested case-control studies. But the model SEs are slightly larger, and as a consequence of this the coverages are a bit higher than 95%. For estimation of β_X, the empirical SEs are higher for MI-SMC than MI-Approx, but the

model SEs are quite similar.

[Insert Table 5 about here]

For the situation with interaction (Table 5), the traditional IPW estimator gives a fairly large bias in the estimates of β_X and β_{Z1}, and the model SEs are too small for all estimates. As for nested case-control studies, MI-Approx gives substantial bias for the interaction model. But MI-SMC performs well, and for method (iii) with MI-SMC imputation for the full cohort we obtain results that are close to those obtained with nested case-control data. For supersampled case-cohort data using MI-SMC, the model SEs tend to be somewhat too large and coverage too high. The empirical SEs are about the same as those obtained with supersampled nested case-control data, except for estimation of the interaction parameter β_{XZ1} where the empirical SEs for supersampled nested case-control data are about 50% larger than for supersampled case-cohort data.

[Insert Table 6 about here]

The results when there is no interaction, but the imputation model is mis-specified (Table 6), are fairly similar to those for nested case-control sampling. But the classical method for case-cohort has lower SEs for estimation of β_X than nested case-control, and the supersampled case-cohort data give a substantial bias in the estimate of β_{Z2} both for a small and a large superset.

For cohort size 5000 and no interaction (Supplementary Table S8), the traditional IPW estimator typically gives less biased estimates with smaller SEs compared to the results for cohort size 25000. However, the estimates for the supersampling methods tend to be slightly more biased when $N = 5000$ than for $N = 25000$. But also for these estimates, the SEs tend to be somewhat smaller for $N = 5000$ than for $N = 25000$. For the situation with interaction (Supplementary Table S9), both the traditional estimates and the supersampling estimates tend to be somewhat less biased and have smaller SEs for cohort size 5000 than for cohorts size 25000. The relative efficiencies of the supersampling methods using MI-SMC are somewhat larger for cohort size 5000 compared to cohort size 25000.

Supplementary Tables S2-S4 show results from case-cohort analyses using the Prentice estimator. These show that typically the traditional Prentice estimator gives less bias, but larger SEs than the IPW estimator. For supersampled data there is no systematic difference in the biases of the two methods, but the IPW estimator gives clearly lower SEs than Prentice's estimator.

6 Discussion

In this paper we have proposed a supersampling approach for nested case-control and case-cohort studies, in which a nested case-control or case-cohort sample is supplemented with additional controls. An expensive covariate is assumed to be measured only on the original nested case-control or case-cohort sample, but cheaply measured covariates are measured for all individuals in the supersample.

We outlined a method for analysis of supersampled data, which involves imputing missing data on the expensive covariate for the individuals in the supersample who are not in the original nested case-control or case-cohort sample. Our focus was on two MI methods, one using an approximate imputation model (MI-Approx) and one that uses rejection sampling to obtain draws of the missing covariate from a posterior distribution that is compatible with the Cox proportional hazards model used for the main analysis (MI-SMC). Both imputation approaches were originally proposed to enable imputation of missing covariate data in Cox regression analyses based on full cohorts under a missing at random assumption (White & Royston, 2009; Bartlett *et al.*, 2015), before being modified for use in nested case-control and case-cohort samples with missing data (Keogh *et al.*, 2018). The supersampled nested case-control or case-cohort study can be considered as a nested case-control or case-cohort study with a larger number of controls and with missing data on the expensive covariate. Hence the MI-Approx and MI-SMC methods described by Keogh *et al.* (2018) can be applied. However, when imputing in a supersampled nested case-control study, one should use the time of the case to which a control is matched and not the control's own censoring time.

We used a simulation study to assess the proposed imputation methods for analysis of supersampled nested case-control and case-cohort studies, focusing on a large underlying cohort with a low event rate, which is realistic for situations in which these study designs

are likely to be used. Small and large supersets were considered, in which the number of controls per case (nested case-control study) or the size of the subcohort (case-cohort study) was increased 4 fold or 12 fold compared with the original sample. We compared analyses of supersampled data using MI-Approx and MI-SMC with a full cohort analysis, with a traditional analysis of the original nested case-control or case-cohort data, and with an MI analysis of the nested case-control or case-cohort data where the expensive covariate is imputed for all individuals in the cohort who are not in the original nested case-control or case-cohort data sample. The supersampled data sets with small and large supersets can be viewed as lying in between the extremes of using the original sample with no information from the rest of the cohort, and using all available information from the rest of the cohort.

In the absence of an interaction term in the analysis model the MI-Approx and MI-SMC analyses of the supersampled studies both perform well when the imputation model is correctly specified, giving approximately unbiased estimates of association and approximately correct standard errors and coverage. This was true for both nested case-control and case-cohort studies. The good performance of MI-Approx in these situations could be expected, since we consider a situation with a low event rate. We showed substantial gains in efficiency in the supersample analyses compared with the traditional analyses of nested case-control and case-cohort data. These gains were greater for the fully observed covariates compared to the covariate missing by design and imputed for the individuals in the supersample who were not in the original nested case-control or case-cohort sample.

Our simulation results for a correctly specified imputation model further showed that when the analysis model includes an interaction between the covariate with missingness and another covariate, MI-Approx gives biased estimates while MI-SMC gives approximately unbiased estimates. This was expected and has been shown in other studies using the two methods. Theory shows the MI-Approx method does not work well in combination with non-linear terms in the analysis model, whereas MI-SMC handles this situation because the imputations are drawn from a distribution that is compatible with the analysis model. MI-Approx also gives biased estimates in the absence of an interaction term when the imputation model is mis-specified. For this situation also MI-SMC imputation for the full cohort gives a substantial bias. This illustrates that the consequences of mis-

specifying the imputation model may be severe when imputing the X-values for a large cohort. But MI-SMC imputation for supersampled data gives quite good results, in particular for nested case-control for the small superset. Thus one advantage of supersampling compared with using information from the full cohort, is that the impact of a mis-specified imputation model will be less severe as a lower proportion of the data is imputed.

When the distribution of X given $\boldsymbol{Z}$ is skewed, as is the case for our mis-specified imputation model scenario, in additional investigations we explored whether the performance of MI-SMC imputation may be improved by assuming $\log X\,|\boldsymbol{Z}$ to be normally distributed when performing the imputations, which approximately matches the data generating distribution for X. Then, as shown in Supplementary Tables S11 and S12, results from MI-SMC imputation are improved for full cohort imputation method for both nested case-control and case-cohort studies. However, the estimates obtained for the supersampled data when using this model in the MI-SMC imputation tend to have more bias. We note that the true distribution of $X|\boldsymbol{Z}$ differs in the full cohort and in the supersampled data. These findings demonstrate the importance of checking that the assumed imputation model is approximately correctly specified.

The MI-SMC approach required a large number of iterations, and it is considerably more computationally intensive than MI-Approx; see Supplementary Table S13 for an example of computing times for the various methods. Thus computational efficiency is an advantage of the MI-Approx method when there are no interactions (or other non-linear terms). We used 500 iterations for the MI-SMC method, which is considerably larger than the standard 10 iterations that are typically used in MI in a general context. Nevertheless, small biases from the MI-SMC approach may be attributed to minor lack of convergence. It would be advantageous if the software could be extended to incorporate straightforward ways of assessing convergence in MI-SMC. When using MI-SMC we recommend using imputed values from MI-Approx as starting values for the iterative procedure.

In our description of the methods and in the simulation study we focused on a setting in which there is a single expensive covariate, X, that is observed only in the original nested case-control or case-cohort sample. The imputation methods used therefore only needed to impute a single covariate. In general there may be more than one covariate that can only be observed on the original sample. Both imputation methods used (MI-Approx

and MI-SMC) extend easily to allow several covariates to be imputed, and both do this using the chained equations approach (White *et al.*, 2011; Carpenter & Kenward, 2013).

There are a number of further investigations that would be of interest to consider in future work. We did not consider auxiliary variables that are correlated with the expensive missing covariate, but which are not of independent interest as covariates in the analysis model. Inclusion of such auxiliary variables in the imputation model has previously been found to improve efficiency of estimates of the coefficient for the expensive covariate, and would be of interest to consider in further work in the supersampling context. Further, we have only considered the classical versions of the nested case-control and case-cohort designs, where the controls/subcohort are selected by simple random sampling (Thomas, 1977; Prentice, 1986). Langholz & Borgan (1995) and Borgan *et al.* (2000) have developed stratified versions of the cohort sampling designs that make it possible to incorporate information on auxillary variabes into the sampling process in order to obtain more informative samples of controls/subcohort. One then classifies the cohort individuals into a number of strata according to their values of the auxilary variables, and selects the controls/subcohort by stratified random sampling. In this way, one may increase the variation of the expensive covariate in the case-control sample, and thereby achieve a more efficient estimation of the effect of this covariate. It would be of interest to study the use of MI for supersampled cohort data, where the original sample and/or the supersample is selected by stratified random sampling.

To enable efficient use of data in supersampled nested case-control and case-cohort studies we focused on the use of MI. Alternative methods of analysis for supersampled studies would be of interest to consider. The supersampling approach could be viewed as a two-stage design in which a nested case-control or case-cohort sample is obtained (Phase I sample) and then certain information is only obtained in a Phase II sample, and one possible alternative analysis approach would be to use inverse probability weighting.

In summary, use of supersampling in nested case-control and case-cohort studies, combined with an MI analysis, can enable gains in efficiency over traditional analyses of these sampled cohort studies, which ignore information available on individuals in the cohort but not in the case-control sample. In many situations the cohort underlying a nested case-control or case-cohort sample is very large. Making use of data for the full cohort

on cheaply measured covariates using MI may be extremely computationally expensive or even infeasible in this case. It may also result in bias if the imputation model is mis-specified. Supersampling provides a practical alternative which combines computational feasibility and robustness for mis-specified imputation models with good efficiency relative to using the full cohort information.

Supporting information

Additional information for this article is available online.

References

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100–1120.

Bartlett, J. W., Seaman, S. R., White, I. R. & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat. Methods Med. Res.* **24**, 462–487.

Borgan, Ø., Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23**, 1749–1778.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6**, 39–58.

Carpenter, J. R. & Kenward, M. G. (2013). *Multiple imputation and its application.* Wiley, New York.

Goldstein, L. & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20**, 1903–1928.

Kalbfleisch, J. D. & Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Stat. Med.* **7**, 149–160.

Keogh, R. H. (2018). Multiple imputation for sampled cohort data. In *Handbook of statistical methods for case-control studies* (eds Ø. Borgan, N. E. Breslow, C. Chatterjee, M. H. Gail, A. Scott & C. J. Wild), 373–390. CRC Press, Boca Raton.

Keogh, R. H., Seaman, S., Bartlett, J. & Wood, A. (2018). Multiple imputation of missing data in nested case-control and case-cohort studies. *Biometrics* **74**, 1438–1449.

Keogh, R. H. & White, I. R. (2013). Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. *Stat. Med.* **32**, 4021–4043.

Kulathinal, S. & Arjas, E. (2006). Bayesian inference from case-cohort data with multiple end-points. *Scand. J. Stat.* **33**, 25–36.

Langholz, B. & Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

Morris, T. P., White, I. R. & Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38**, 2074–2102.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Saarela, O., Kulathinal, S., Arjas, E. & Läärä, E. (2008). Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. *Stat. Med.* **27**, 5991–6008.

Samuelsen, S. O., Ånestad, H. & Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scand. J. Stat.* **34**, 103–119.

Scheike, T. H. & Juul, A. (2004). Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics* **5**, 193–206.

Scheike, T. H. & Martinussen, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scand. J. Stat.* **31**, 283–293.

Self, S. G. & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64–81.

Therneau, T. M. & Li, H. (1999). Computing the Cox model for case-cohort designs. *Lifetime Data Anal.* **5**, 99–112.

Thomas, D. C. (1977). Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining," by F. D. K. Liddell, J. C. McDonald & D. C. Thomas. *J. Roy. Statist. Soc. Ser. A* **140**, 469–491.

White, I. R. & Royston, P. (2009). Imputing missing covariate values for the Cox model. *Stat. Med.* **28**, 1982–1998.

White, I. R., Royston, P. & Wood, A. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med.* **30**, 377–399.

Zeng, D. & Lin, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Amer. Statist. Assoc.* **109**, 371–383.

Zeng, D. & Lin, D. Y. (2018). Maximum likelihood estimation for case-cohort and nested case-control studies. In *Handbook of Statistical Methods for Case-Control Studies* (eds Ø. Borgan, N. E. Breslow, C. Chatterjee, M. H. Gail, A. Scott & C. J. Wild), 391–404. CRC Press, Boca Raton.

Corresponding author:
Ørnulf Borgan
Department of Mathematics, University of Oslo
P.O.Box 1053, Blindern
0316 Oslo, Norway
email: borgan@math.uio.no

Table 1: Nested case-control: Results for scenario with full cohort size $N = 25000$ and no interaction between X and Z_1. SS: supersampled. MC errors were: ≤ 0.009 for Bias, ≤ 0.006 for EmpSE, ≤ 0.005 for Model SE, ≤ 0.009 for Coverage, ≤ 0.046 for Rel Eff.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.002	0.060	0.059	0.943	1.000	0.060
Traditional NCC	-	0.029	0.152	0.148	0.945	0.155	0.154
NCC within full cohort	MI approx	-0.012	0.105	0.105	0.931	0.323	0.105
NCC within full cohort	MI SMC	0.004	0.108	0.104	0.926	0.304	0.108
SS NCC, small superset	MI approx	0.017	0.119	0.122	0.954	0.251	0.120
SS NCC, small superset	MI SMC	0.010	0.125	0.125	0.950	0.229	0.125
SS NCC, large superset	MI approx	0.010	0.116	0.112	0.931	0.265	0.116
SS NCC, large superset	MI SMC	-0.002	0.132	0.117	0.901	0.205	0.132
Results for β_{Z1}							
Full cohort	-	-0.000	0.059	0.060	0.946	1.000	0.059
Traditional NCC	-	0.010	0.139	0.135	0.944	0.180	0.140
NCC within full cohort	MI approx	-0.009	0.069	0.073	0.957	0.726	0.070
NCC within full cohort	MI SMC	0.002	0.071	0.073	0.956	0.704	0.071
SS NCC, small superset	MI approx	0.004	0.091	0.094	0.950	0.420	0.091
SS NCC, small superset	MI SMC	0.004	0.092	0.096	0.949	0.415	0.092
SS NCC, large superset	MI approx	0.001	0.079	0.081	0.960	0.557	0.079
SS NCC, large superset	MI SMC	0.003	0.084	0.084	0.949	0.493	0.084
Results for β_{Z2}							
Full cohort	-	0.005	0.153	0.151	0.942	1.000	0.153
Traditional NCC	-	0.015	0.282	0.280	0.955	0.293	0.282
NCC within full cohort	MI approx	-0.011	0.182	0.177	0.938	0.706	0.182
NCC within full cohort	MI SMC	0.000	0.183	0.177	0.939	0.693	0.183
SS NCC, small superset	MI approx	-0.002	0.204	0.206	0.953	0.561	0.204
SS NCC, small superset	MI SMC	0.002	0.209	0.209	0.953	0.535	0.209
SS NCC, large superset	MI approx	-0.007	0.190	0.185	0.940	0.648	0.190
SS NCC, large superset	MI SMC	-0.002	0.197	0.193	0.944	0.604	0.197

Table 2: Nested case-control: Results for scenario with full cohort size $N = 25000$ and interaction between X and Z_1. SS: supersampled. MC errors were: ≤ 0.014 for Bias, ≤ 0.010 for EmpSE, ≤ 0.007 for Model SE, ≤ 0.015 for Coverage, ≤ 0.076 for Rel Eff.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.009	0.093	0.091	0.946	1.000	0.094
Traditional NCC	-	0.081	0.262	0.240	0.953	0.127	0.274
NCC within full cohort	MI approx	0.254	0.146	0.152	0.623	0.409	0.292
NCC within full cohort	MI SMC	0.011	0.115	0.116	0.951	0.652	0.116
SS NCC, small superset	MI approx	0.096	0.174	0.182	0.955	0.286	0.199
SS NCC, small superset	MI SMC	0.037	0.166	0.179	0.964	0.313	0.170
SS NCC, large superset	MI approx	0.133	0.164	0.156	0.872	0.321	0.212
SS NCC, large superset	MI SMC	0.028	0.151	0.152	0.945	0.381	0.153
Results for β_{Z1}							
Full cohort	-	0.014	0.119	0.121	0.957	1.000	0.119
Traditional NCC	-	0.070	0.291	0.278	0.947	0.166	0.300
NCC within full cohort	MI approx	0.453	0.125	0.191	0.267	0.896	0.470
NCC within full cohort	MI SMC	0.015	0.124	0.126	0.963	0.922	0.125
SS NCC, small superset	MI approx	0.097	0.186	0.208	0.962	0.407	0.210
SS NCC, small superset	MI SMC	0.058	0.186	0.199	0.966	0.406	0.195
SS NCC, large superset	MI approx	0.153	0.159	0.184	0.922	0.554	0.221
SS NCC, large superset	MI SMC	0.073	0.167	0.171	0.945	0.506	0.182
Results for β_{Z2}							
Full cohort	-	0.006	0.147	0.146	0.953	1.000	0.147
Traditional NCC	-	0.057	0.431	0.412	0.954	0.116	0.434
NCC within full cohort	MI approx	-0.051	0.180	0.192	0.943	0.664	0.187
NCC within full cohort	MI SMC	0.000	0.195	0.190	0.945	0.565	0.195
SS NCC, small superset	MI approx	0.025	0.260	0.277	0.971	0.319	0.261
SS NCC, small superset	MI SMC	0.010	0.267	0.284	0.970	0.303	0.267
SS NCC, large superset	MI approx	0.012	0.221	0.226	0.952	0.440	0.222
SS NCC, large superset	MI SMC	-0.010	0.229	0.239	0.955	0.410	0.229
Results for β_{XZ1}							
Full cohort	-	-0.005	0.062	0.062	0.954	1.000	0.063
Traditional NCC	-	-0.000	0.219	0.212	0.951	0.081	0.219
NCC within full cohort	MI approx	-0.422	0.057	0.115	0.006	1.196	0.426
NCC within full cohort	MI SMC	-0.006	0.068	0.069	0.956	0.843	0.068
SS NCC, small superset	MI approx	-0.116	0.118	0.159	0.943	0.278	0.166
SS NCC, small superset	MI SMC	-0.032	0.139	0.155	0.962	0.201	0.143
SS NCC, large superset	MI approx	-0.184	0.094	0.129	0.753	0.440	0.207
SS NCC, large superset	MI SMC	-0.054	0.123	0.126	0.928	0.256	0.134

Table 3: Nested case-control: Results for scenario with $N = 25000$ and no interaction between X and Z_1, when $X|Z_1, Z_2$ has a non-normal distribution. MI-Approx assumes a normal distribution for $X|Z_1, Z_2, \widehat{H}_0(s)$ and MI-SMC assumes a normal distribution for $X|Z_1, Z_2$. SS: supersampled.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.002	0.027	0.026	0.945	1.000	0.027
Traditional NCC	-	0.052	0.172	0.160	0.950	0.025	0.179
NCC within full cohort	MI approx	-0.311	0.064	0.054	0.000	0.180	0.317
NCC within full cohort	MI SMC	-0.260	0.058	0.047	0.000	0.220	0.266
SS NCC, small superset	MI approx	-0.332	0.061	0.073	0.008	0.197	0.337
SS NCC, small superset	MI SMC	0.020	0.110	0.102	0.937	0.061	0.112
SS NCC, large superset	MI approx	-0.365	0.059	0.055	0.001	0.213	0.369
SS NCC, large superset	MI SMC	-0.038	0.101	0.072	0.784	0.072	0.108
Results for β_{Z1}							
Full cohort	-	-0.001	0.060	0.060	0.952	1.000	0.060
Traditional NCC	-	0.028	0.184	0.175	0.955	0.108	0.186
NCC within full cohort	MI approx	0.120	0.083	0.088	0.733	0.527	0.146
NCC within full cohort	MI SMC	0.171	0.088	0.089	0.526	0.469	0.193
SS NCC, small superset	MI approx	0.154	0.104	0.112	0.762	0.339	0.186
SS NCC, small superset	MI SMC	-0.003	0.113	0.117	0.956	0.287	0.113
SS NCC, large superset	MI approx	0.165	0.092	0.093	0.592	0.434	0.189
SS NCC, large superset	MI SMC	-0.024	0.098	0.099	0.937	0.379	0.101
Results for β_{Z2}							
Full cohort	-	0.006	0.138	0.140	0.960	1.000	0.138
Traditional NCC	-	0.037	0.377	0.364	0.955	0.134	0.378
NCC within full cohort	MI approx	0.197	0.189	0.196	0.843	0.531	0.273
NCC within full cohort	MI SMC	0.290	0.205	0.200	0.698	0.449	0.355
SS NCC, small superset	MI approx	0.294	0.239	0.242	0.809	0.331	0.379
SS NCC, small superset	MI SMC	-0.028	0.252	0.255	0.945	0.299	0.254
SS NCC, large superset	MI approx	0.297	0.220	0.206	0.708	0.393	0.370
SS NCC, large superset	MI SMC	-0.088	0.221	0.218	0.932	0.387	0.238

Table 4: Case-cohort with IPW estimator: Results for scenario with full cohort size $N = 25000$ and no interaction between X and Z_1. SS: supersampled. MC errors were: ≤ 0.011 for Bias, ≤ 0.008 for EmpSE, ≤ 0.005 for Model SE, ≤ 0.012 for Coverage, ≤ 0.040 for Rel Eff.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.002	0.060	0.059	0.943	1.000	0.060
Traditional case-cohort	-	0.081	0.209	0.159	0.821	0.082	0.224
Case-cohort within full cohort	MI approx	-0.012	0.107	0.111	0.931	0.309	0.108
Case-cohort within full cohort	MI SMC	0.004	0.109	0.110	0.931	0.301	0.109
SS case-cohort, small superset	MI approx	0.015	0.121	0.160	0.988	0.242	0.122
SS case-cohort, small superset	MI SMC	0.031	0.156	0.163	0.934	0.147	0.159
SS case-cohort, large superset	MI approx	-0.004	0.111	0.133	0.968	0.286	0.111
SS case-cohort, large superset	MI SMC	0.005	0.153	0.142	0.901	0.153	0.153
Results for β_{Z1}							
Full cohort	-	-0.000	0.059	0.060	0.946	1.000	0.059
Traditional case-cohort	-	0.036	0.191	0.158	0.877	0.096	0.194
Case-cohort within full cohort	MI approx	-0.009	0.074	0.074	0.940	0.635	0.075
Case-cohort within full cohort	MI SMC	0.002	0.075	0.074	0.948	0.623	0.075
SS case-cohort, small superset	MI approx	0.007	0.105	0.121	0.978	0.321	0.105
SS case-cohort, small superset	MI SMC	0.014	0.113	0.123	0.957	0.274	0.114
SS case-cohort, large superset	MI approx	-0.005	0.083	0.094	0.970	0.509	0.083
SS case-cohort, large superset	MI SMC	0.005	0.092	0.099	0.961	0.413	0.092
Results for β_{Z2}							
Full cohort	-	0.005	0.153	0.151	0.942	1.000	0.153
Traditional case-cohort	-	0.021	0.340	0.306	0.926	0.202	0.340
Case-cohort within full cohort	MI approx	-0.006	0.185	0.180	0.935	0.681	0.185
Case-cohort within full cohort	MI SMC	0.007	0.183	0.181	0.943	0.698	0.183
SS case-cohort, small superset	MI approx	0.003	0.211	0.236	0.961	0.526	0.211
SS case-cohort, small superset	MI SMC	0.015	0.227	0.245	0.969	0.451	0.228
SS case-cohort, large superset	MI approx	-0.005	0.191	0.200	0.956	0.639	0.191
SS case-cohort, large superset	MI SMC	0.011	0.209	0.213	0.950	0.535	0.209

Table 5: Case-cohort with IPW estimator: Results for scenario with full cohort size $N = 25000$ and interaction between X and Z_1. SS: supersampled. MC errors were: ≤ 0.015 for Bias, ≤ 0.011 for EmpSE, ≤ 0.007 for Model SE, ≤ 0.016 for Coverage, ≤ 0.053 for Rel Eff.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.009	0.093	0.091	0.946	1.000	0.094
Traditional case-cohort	-	0.106	0.276	0.210	0.822	0.114	0.296
Case-cohort within full cohort	MI approx	0.270	0.153	0.156	0.584	0.369	0.310
Case-cohort within full cohort	MI SMC	0.009	0.121	0.120	0.951	0.596	0.121
SS case-cohort, small superset	MI approx	0.160	0.186	0.265	0.976	0.250	0.245
SS case-cohort, small superset	MI SMC	0.035	0.182	0.211	0.970	0.263	0.185
SS case-cohort, large superset	MI approx	0.212	0.163	0.227	0.926	0.327	0.267
SS case-cohort, large superset	MI SMC	-0.008	0.160	0.176	0.963	0.340	0.160
Results for β_{Z1}							
Full cohort	-	0.014	0.119	0.122	0.957	1.000	0.120
Traditional case-cohort	-	0.112	0.308	0.256	0.890	0.150	0.328
Case-cohort within full cohort	MI approx	0.452	0.130	0.192	0.271	0.845	0.470
Case-cohort within full cohort	MI SMC	0.016	0.125	0.127	0.957	0.914	0.126
SS case-cohort, small superset	MI approx	0.233	0.176	0.358	0.984	0.460	0.292
SS case-cohort, small superset	MI SMC	0.066	0.177	0.217	0.971	0.452	0.189
SS case-cohort, large superset	MI approx	0.334	0.148	0.313	0.961	0.649	0.365
SS case-cohort, large superset	MI SMC	0.052	0.145	0.179	0.984	0.673	0.154
Results for β_{Z2}							
Full cohort	-	0.006	0.147	0.147	0.953	1.000	0.147
Traditional case-cohort	-	0.002	0.490	0.384	0.854	0.091	0.489
Case-cohort within full cohort	MI approx	-0.037	0.190	0.198	0.952	0.604	0.193
Case-cohort within full cohort	MI SMC	0.007	0.200	0.194	0.943	0.544	0.200
SS case-cohort, small superset	MI approx	-0.005	0.269	0.399	0.994	0.299	0.269
SS case-cohort, small superset	MI SMC	0.008	0.281	0.372	0.990	0.275	0.281
SS case-cohort, large superset	MI approx	-0.025	0.221	0.298	0.988	0.447	0.222
SS case-cohort, large superset	MI SMC	0.005	0.237	0.294	0.985	0.387	0.237
Results for β_{XZ1}							
Full cohort	-	-0.005	0.063	0.062	0.954	1.000	0.063
Traditional case-cohort	-	-0.021	0.148	0.122	0.894	0.180	0.149
Case-cohort within full cohort	MI approx	-0.426	0.058	0.115	0.005	1.149	0.430
Case-cohort within full cohort	MI SMC	-0.005	0.070	0.070	0.950	0.794	0.070
SS case-cohort, small superset	MI approx	-0.217	0.104	0.250	0.994	0.364	0.241
SS case-cohort, small superset	MI SMC	-0.025	0.091	0.128	0.992	0.475	0.094
SS case-cohort, large superset	MI approx	-0.329	0.087	0.215	0.821	0.522	0.340
SS case-cohort, large superset	MI SMC	-0.032	0.084	0.110	0.989	0.553	0.090

Table 6: Case-cohort: Results for scenario with $N = 25000$ and no interaction between X and Z_1, when $X|Z_1, Z_2$ has a non-normal distribution. MI-Approx assumes a normal distribution for $X|Z_1, Z_2, \widehat{H}(t)$ and MI-SMC assumes a normal distribution for $X|Z_1, Z_2$. SS: supersampled.

Method	MI method	Bias	Emp SE	Model SE	Coverage	Rel eff	RMSE
Results for β_X							
Full cohort	-	0.002	0.027	0.026	0.944	1.000	0.027
Traditional case-cohort	-	0.040	0.069	0.046	0.642	0.157	0.079
Case-cohort within full cohort	MI approx	-0.340	0.065	0.055	0.000	0.175	0.346
Case-cohort within full cohort	MI SMC	-0.276	0.060	0.049	0.000	0.206	0.283
SS case-cohort, small superset	MI approx	-0.260	0.071	0.103	0.229	0.147	0.269
SS case-cohort, small superset	MI SMC	0.052	0.053	0.043	0.637	0.264	0.074
SS case-cohort, large superset	MI approx	-0.314	0.068	0.078	0.015	0.161	0.321
SS case-cohort, large superset	MI SMC	-0.025	0.079	0.050	0.824	0.118	0.083
Results for β_{Z1}							
Full cohort	-	-0.001	0.061	0.060	0.951	1.000	0.061
Traditional case-cohort	-	0.020	0.192	0.147	0.858	0.099	0.193
Case-cohort within full cohort	MI approx	0.131	0.087	0.090	0.705	0.489	0.157
Case-cohort within full cohort	MI SMC	0.187	0.092	0.091	0.455	0.433	0.209
SS case-cohort, small superset	MI approx	0.141	0.109	0.175	0.963	0.309	0.178
SS case-cohort, small superset	MI SMC	-0.050	0.113	0.126	0.945	0.289	0.123
SS case-cohort, large superset	MI approx	0.134	0.095	0.128	0.887	0.406	0.165
SS case-cohort, large superset	MI SMC	-0.049	0.104	0.118	0.939	0.339	0.115
Results for β_{Z2}							
Full cohort	-	0.006	0.138	0.140	0.960	1.000	0.138
Traditional case-cohort	-	0.006	0.446	0.330	0.852	0.096	0.446
Case-cohort within full cohort	MI approx	0.225	0.195	0.197	0.802	0.503	0.297
Case-cohort within full cohort	MI SMC	0.328	0.215	0.205	0.635	0.413	0.393
SS case-cohort, small superset	MI approx	0.220	0.228	0.359	0.986	0.367	0.317
SS case-cohort, small superset	MI SMC	-0.126	0.239	0.249	0.922	0.334	0.270
SS case-cohort, large superset	MI approx	0.225	0.205	0.260	0.923	0.456	0.304
SS case-cohort, large superset	MI SMC	-0.144	0.219	0.239	0.902	0.397	0.262