

Nested case–control studies: should one break the matching?

Ørnulf Borgan · Ruth Keogh

Received: 1 August 2014 / Accepted: 6 January 2015 / Published online: 23 January 2015
© Springer Science+Business Media New York 2015

Abstract In a nested case–control study, controls are selected for each case from the individuals who are at risk at the time at which the case occurs. We say that the controls are matched on study time. To adjust for possible confounding, it is common to match on other variables as well. The standard analysis of nested case–control data is based on a partial likelihood which compares the covariates of each case to those of its matched controls. It has been suggested that one may break the matching of nested case–control data and analyse them as case–cohort data using an inverse probability weighted (IPW) pseudo likelihood. Further, when some covariates are available for all individuals in the cohort, multiple imputation (MI) makes it possible to use all available data in the cohort. In the paper we review the standard method and the IPW and MI approaches, and compare their performance using simulations that cover a range of scenarios, including one and two endpoints.

Keywords Case–cohort · Competing risks · Cox regression · Inverse probability weighting · Matching · Multiple imputation · Nested case–control

Electronic supplementary material The online version of this article (doi:[10.1007/s10985-015-9319-y](https://doi.org/10.1007/s10985-015-9319-y)) contains supplementary material, which is available to authorized users.

Ø. Borgan (✉)

Department of Mathematics, University of Oslo, P.O.Box 1053, Blindern, 0316 Oslo, Norway
e-mail: borgan@math.uio.no

R. Keogh

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street,
London WC1E 7HT, UK
e-mail: Ruth.Keogh@lshtm.ac.uk

1 Introduction

Cox regression is commonly used to assess the influence of risk factors and other covariates on mortality or morbidity. Estimation in Cox's model is based on a partial likelihood, which at each observed death or disease occurrence compares the covariate values of the individual who experienced the event of interest to those of all individuals at risk. Thus Cox regression requires collection of covariate information for all individuals in the cohort, including when only a small fraction of the individuals experience the event of interest. This may be very expensive in large cohorts. Further, when covariate measurements are based on biological material stored in biobanks, it will imply a waste of valuable material that one may want to save for future studies.

Cohort sampling designs, where covariate information is collected 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") then offer useful alternatives that may save valuable biological material and drastically reduce the workload of data collection and error checking. Further, as most of the statistical information is contained in the cases, such studies may still be sufficient to give reliable answers to the questions of interest.

There are two main types of cohort sampling designs: *nested case-control studies* and *case-cohort studies*; see e.g. Keogh and Cox (2014, Chaps. 7, 8) for a review. The two types of cohort sampling designs differ in the way controls are selected. In a nested case-control study, controls are selected for each case from the individuals at risk at the time at which the case occurs. In the parlance of classical case-control studies (e.g. Breslow 1996), one says that the controls are *matched* on study time. To adjust for possible confounding, it is common to match on other variables as well. This is achieved by selecting controls with the same values of the confounding variables as the case. Nested case-control data are traditionally analysed using a partial likelihood similar to the one for the full cohort. In a case-cohort study one *does not match* the controls to the cases. Instead a subcohort is selected from the full cohort, and the individuals in the subcohort are used as controls at all event times when they are at risk.

If one wants to apply a cohort sampling design, a choice between a nested case-control and a case-cohort study has to be made. The choice between the two designs depends on a number of issues, and it has to be made on a case by case basis; see e.g. Borgan and Samuelsen (2013, Sect. 17.5) for a discussion of the considerations one should make to arrive at a useful design for a given study. In particular, if one wants to use an assembled cohort to study more than one endpoint (e.g. more than one disease) the case-cohort design may be preferable, since then one may use the individuals in the subcohort as controls for all endpoints. But in situations where a careful matching on confounders is needed to avoid bias, the nested case-control design may be preferable.

It has been suggested that one may break the matching in a nested case-control study and treat the nested case-control data as if they were case-cohort data with a non-standard sampling scheme for the subcohort (e.g. Samuelsen 1997; Chen 2001). The data may then be analysed using an inverse probability weighted (IPW) pseudo likelihood. By doing so, one may overcome the limitations of a traditional analysis of nested case-control data, in particular that controls cannot be reused across studies of different endpoints. However, one also runs the risk of introducing bias in the analysis.

For a traditional analysis of nested case-control data or when using an IPW pseudo likelihood, only data for the cases and the sampled controls are used in the analysis. In many situations some covariates will be known for all cohort members, but this information is disregarded in the partial likelihood and the IPW pseudo likelihood. More recently, methods have been suggested that make use of all the information in the cohort. One then considers estimation for nested case-control data as a missing data problem, where the covariates only known for the cases and the controls are missing by design for the remaining individuals in the cohort. Estimation may then be performed using the EM algorithm for the full cohort likelihood (Scheike and Juul 2004) or, which is computationally less demanding, by using multiple imputation (MI) (Keogh and White 2013). It should be noted that the full likelihood approach and MI break the matching between the cases and the controls.

It is the purpose of this paper to compare the performance of the IPW pseudo likelihood and MI with the traditional partial likelihood analysis of nested case-control data, considering studies of both one and two endpoints. By investigating a number of different simulation scenarios, we will clarify when it may be beneficial to break the matching in nested case-control data and when problems may occur if the matching is broken. Further we will investigate the gain one may obtain by using the MI approach for the full cohort, and point out the pitfalls that have to be avoided when this approach is used.

The paper is organised as follows. In Sect. 2 we focus on a single endpoint and outline the traditional analysis of nested case-control data using a partial likelihood. The alternative analysis which breaks the matching and uses an IPW pseudo likelihood is discussed in Sect. 3, and in Sect. 4 we consider analyses which make use of all available data for the full cohort. We extend the methods to the situation with two endpoints in Sect. 5. In Sect. 6 the different methods of analysis for nested case-control data are compared using simulation studies covering a range of scenarios, including one and two endpoints. The paper finishes with a discussion in Sect. 7.

2 Nested case-control studies with one endpoint

We will first consider the situation where there is only one endpoint of interest. This may correspond to the onset of a specific disease or death from a given cause. The situation with more than one endpoint is considered in Sect. 5. We begin by outlining the analysis of full cohort data and then extend to nested case-control data.

2.1 Cohort data

We consider a cohort $\mathcal{C} = \{1, \dots, n\}$ of n independent individuals. For each individual $i \in \mathcal{C}$ we have two vectors of covariates. The vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ contains the covariates of main interest and other variables that we will adjust for in the analysis, while $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})'$ is a vector of additional confounding variables that we will match on when selecting the controls (cf. Sect. 2.2). We assume that the hazard rate $h_i(t)$ for the time of the event of interest for the i th individual depends on both \mathbf{x}_i and \mathbf{z}_i and that it takes the form

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\gamma}' \mathbf{z}_i). \quad (1)$$

Due to censoring, we do not observe the event times. For each individual $i \in \mathcal{C}$ we only observe (T_i, D_i) , where T_i is the minimum of the event time and a censoring time, and $D_i = 1$ if T_i equals the event time and $D_i = 0$ otherwise. We assume that censoring is independent (e.g. Kalbfleisch and Prentice 2002, Sects. 1.3, 6.2), which in particular implies that censoring may depend on the covariates \mathbf{x}_i and \mathbf{z}_i .

The risk set $\mathcal{R}(t) = \{i \mid T_i \geq t\}$ is the collection of all individuals who are under observation just before time t . For ease of notation we write \mathcal{R}_i for the risk set at time T_i , i.e. $\mathcal{R}_i = \mathcal{R}(T_i)$, and introduce $\mathcal{E} = \{i \mid D_i = 1\}$ for the set of all cases. The vectors of regression coefficients in (1) are estimated by the values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that maximize Cox's partial likelihood

$$L_{\text{co}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i \in \mathcal{E}} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\gamma}' \mathbf{z}_i)}{\sum_{j \in \mathcal{R}_i} \exp(\boldsymbol{\beta}' \mathbf{x}_j + \boldsymbol{\gamma}' \mathbf{z}_j)}. \quad (2)$$

It is well known that the maximum partial likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are approximately multivariate normally distributed around the true values of the parameter vectors with a covariance matrix that may be estimated by the inverse information (Andersen and Gill 1982).

2.2 Nested case–control data

We assume that the vectors \mathbf{z}_i of confounding variables in (1) are observed for the full cohort. Further, in Sects. 2–6, we assume that \mathbf{z}_i can only take a finite number of different values $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(k)}$. This implies that numeric confounders have to be categorized; in particular when matching on age in the simulations (Sect. 6) we will match on age in whole years. In the final Sect. 7 we discuss briefly how one may match directly on a numeric confounder.

Now the cohort may be divided into k strata according to the values of the \mathbf{z}_i . Further, if an individual i in stratum s experiences the event of interest at time T_i , one selects at random m controls by simple random sampling from the remaining individuals at risk in stratum s . The set consisting of the case i and the m controls is called a *sampled risk set* and is denoted $\tilde{\mathcal{R}}_i$, and the nested case–control sample consists of the collection of all sampled risk sets. The covariates \mathbf{x}_i are ascertained for the individuals in the case–control sample, but are not needed for the remaining individuals in the cohort. Note that the selection of controls is done independently at the different event times. Thus subjects may serve as controls for multiple cases, and a case may serve as control for other cases that experienced the event when the case was at risk.

For nested case–control data one cannot estimate the effects of the variables \mathbf{z}_i used in the matching of controls to cases. However, the vector of regression coefficients $\boldsymbol{\beta}$ for the covariates \mathbf{x}_i may be estimated by $\hat{\boldsymbol{\beta}}$, the value of $\boldsymbol{\beta}$ maximizing the partial likelihood

$$L_{\text{ncc}}(\boldsymbol{\beta}) = \prod_{i \in \mathcal{O}} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i)}{\sum_{j \in \tilde{\mathcal{O}}_i} \exp(\boldsymbol{\beta}' \mathbf{x}_j)}, \tag{3}$$

cf. Oakes (1981) and Borgan et al. (1995). The estimator $\hat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed around the true value of $\boldsymbol{\beta}$ with a covariance matrix that may be estimated by the inverse information obtained from (3).

Note that the partial likelihood (3) remains unchanged if the cohort model (1) is replaced by a stratified Cox model, where the hazard for an individual i in stratum s takes the form

$$h_i(t) = h_{0s}(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i). \tag{4}$$

This shows that by matching, we obtain valid inference for $\boldsymbol{\beta}$ under quite weak assumptions on the effects of the confounding variables \mathbf{z}_i .

The traditional analysis of nested case–control data outlined in this section can be implemented easily using standard software, for example using the `coxph` function in R with stratification by the sampled risk set identifiers.

3 Breaking the matching and inverse probability weighting

In the partial likelihood (3), a case and its controls are included only at the event time of the case. If we break the matching the covariate information for the cases and the controls may be used at all times when they are at risk. One may then estimate the regression coefficients in (1) by maximizing an inverse probability weighted (IPW) pseudo likelihood of the form

$$L_{\text{ipw}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i \in \mathcal{O}} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\gamma}' \mathbf{z}_i) w_i}{\sum_{j \in \mathcal{S}_i} \exp(\boldsymbol{\beta}' \mathbf{x}_j + \boldsymbol{\gamma}' \mathbf{z}_j) w_j}. \tag{5}$$

Here \mathcal{S}_i is the set consisting of *all* cases and controls in the nested case–control sample who are at risk at time T_i . The *weights* are $w_j = 1/p_j$, where the p_j are appropriate inclusion probabilities. The weight for a given individual is the same at all time points at which the individual is at risk.

The inclusion probabilities are $p_j = 1$ for cases, while for controls (who do not later become a case) they may be estimated in different ways. Following Samuelsen (1997) and Støer and Samuelsen (2013), we may for control individual j use the Kaplan–Meier type estimate

$$p_j = 1 - \prod \left(1 - \frac{m}{n_i - 1} \right), \tag{6}$$

where the product is over all event times T_i when individual j is a possible matched control, and n_i is the number at risk at time T_i who satisfy the matching criteria. Thus, if individual i is in stratum s , then n_i is the number at risk in this stratum at time T_i .

Another possibility is to estimate the inclusion probabilities by logistic regression (Saarela et al. 2008; Støer and Samuelsen 2013). One then restricts attention to the non-cases and assumes a logistic regression model for the sampling indicators O_j ; $j \in \mathcal{C} \setminus \mathcal{E}$. The sampling indicators are 1 for sampled controls (who do not later become a case) and 0 for non-sampled individuals. The follow-up times and the matching variables are used as covariates in the logistic regression model, i.e.

$$p_j = P(O_j = 1 | T_j, D_j = 0, \mathbf{z}_j) = \frac{\exp(\alpha_0 + \alpha_1 T_j + \boldsymbol{\alpha}'_2 \mathbf{z}_j)}{1 + \exp(\alpha_0 + \alpha_1 T_j + \boldsymbol{\alpha}'_2 \mathbf{z}_j)}. \quad (7)$$

As a modification, we may use a generalized additive model where the linear term $\xi_1 T_j$ in (7) is replaced by a smooth function $f(T_j)$ of the follow-up times (Samuelsen et al. 2007; Støer and Samuelsen 2013).

The pseudo likelihood (5) does not possess likelihood properties, so we cannot estimate the covariance matrix by the inverse information. However, the robust sandwich estimator has shown to be adequate in most situations (Samuelsen et al. 2007; Saarela et al. 2008; Støer and Samuelsen 2012) and will be used in our simulations (Sect. 6).

The inverse probability weighting method described in this section can be implemented using the R package `multipleNCC` (Støer and Samuelsen 2014).

4 Using the full cohort

In the previous sections we have described how data from a nested case–control study may be analysed under the standard approach using the partial likelihood (3) or by breaking the matching and using a weighted pseudo likelihood analysis (5). For both methods, only data for the cases and the sampled controls are used in the analysis, though the confounding variables \mathbf{z}_i are used in the sampling of controls and to obtain the weights p_j under the logistic regression approach. However, in many situations some of the covariates of main interest will be known for all cohort members, but this information is not used by the methods described in Sects. 2.2 and 3.

In this section we describe two approaches to make use of all data that is available in the full cohort. To this end we assume that the covariate vectors \mathbf{x}_i may be partitioned as $\mathbf{x}_i = (\mathbf{x}_i^{(cc)'} , \mathbf{x}_i^{(all)'})'$, where $\mathbf{x}_i^{(all)}$ is observed for all individuals in the cohort while $\mathbf{x}_i^{(cc)}$ is only observed for the individuals in the nested case–control sample. Thus, if we denote by \mathcal{O} the set of cases and controls, the available data are $(T_i, D_i, \mathbf{x}_i^{(all)}, \mathbf{z}_i)$ for $i \in \mathcal{C}$, while we additionally observe $\mathbf{x}_i^{(cc)}$ for $i \in \mathcal{O}$.

4.1 A full likelihood approach

In a nested case–control study, information on $\mathbf{x}_i^{(cc)}$ is missing by design for $i \in \mathcal{C} \setminus \mathcal{O}$. Provided that $(T_i, D_i, \mathbf{x}_i, \mathbf{z}_i); i \in \mathcal{C}$; are independent, the full cohort likelihood (conditional on $\mathbf{x}_i^{(all)}$ and \mathbf{z}_i) is the product of contributions from the nested case–control sample \mathcal{O} and from the remainder of the cohort $\mathcal{C} \setminus \mathcal{O}$ and takes the form

$$\begin{aligned}
 L_{\text{full}} \propto & \prod_{i \in \mathcal{O}} p(t_i, d_i | \mathbf{x}_i, \mathbf{z}_i) dG(\mathbf{x}_i^{(\text{cc})} | \mathbf{x}_i^{(\text{all})}, \mathbf{z}_i) \\
 & \times \prod_{i \in \mathcal{C} \setminus \mathcal{O}} \int_{\mathcal{X}_{\text{cc}}} p(t_i, d_i | \mathbf{x}_i, \mathbf{z}_i) dG(\mathbf{x}_i^{(\text{cc})} | \mathbf{x}_i^{(\text{all})}, \mathbf{z}_i)
 \end{aligned} \tag{8}$$

(Scheike and Juul 2004; Saarela et al. 2008). Here $p(t_i, d_i | \mathbf{x}_i, \mathbf{z}_i)$ is the conditional density of (T_i, D_i) given $\mathbf{x}_i, \mathbf{z}_i$, and $G(\mathbf{x}_i^{(\text{cc})} | \mathbf{x}_i^{(\text{all})}, \mathbf{z}_i)$ is the conditional distribution of $\mathbf{x}_i^{(\text{cc})}$ given $\mathbf{x}_i^{(\text{all})}$ and \mathbf{z}_i . Further the integral in (8) is over the space \mathcal{X}_{cc} of all possible values of the covariate vectors $\mathbf{x}_i^{(\text{cc})}$.

To achieve a full maximum likelihood solution, we need to specify the conditional distributions in (8). The conditional distribution of (T_i, D_i) given $\mathbf{x}_i, \mathbf{z}_i$ will depend both on the distribution of the event times and the distribution of the censoring times. However, if the censoring distribution does *not* depend on $\mathbf{x}_i^{(\text{cc})}$, it may be disregarded when considering the likelihood. If we then assume the Cox model (1) for the event times and partition the vector of regression coefficients for $\mathbf{x}_i = (\mathbf{x}_i^{(\text{cc})'}, \mathbf{x}_i^{(\text{all})}')'$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{\text{cc}}, \boldsymbol{\beta}'_{\text{all}})'$, the conditional density of (T_i, D_i) given $\mathbf{x}_i, \mathbf{z}_i$ may be given as

$$\begin{aligned}
 p(t_i, d_i | \mathbf{x}_i, \mathbf{z}_i) = & \left[h_0(t_i) \exp\left(\boldsymbol{\beta}'_{\text{cc}} \mathbf{x}_i^{(\text{cc})} + \boldsymbol{\beta}'_{\text{all}} \mathbf{x}_i^{(\text{all})} + \boldsymbol{\gamma}' \mathbf{z}_i\right) \right]^{d_i} \\
 & \times \exp\left\{-\exp\left(\boldsymbol{\beta}'_{\text{cc}} \mathbf{x}_i^{(\text{cc})} + \boldsymbol{\beta}'_{\text{all}} \mathbf{x}_i^{(\text{all})} + \boldsymbol{\gamma}' \mathbf{z}_i\right) \int_0^{t_i} h_0(u) du\right\}. \tag{9}
 \end{aligned}$$

We also need to specify the conditional distribution of $\mathbf{x}_i^{(\text{cc})}$ given $\mathbf{x}_i^{(\text{all})}$ and \mathbf{z}_i (cf. Sect. 4.2), or we may adopt a non-parametric approach and assume that the conditional distribution has point masses at the observed covariate values (Scheike and Juul 2004). When maximizing the likelihood, it is assumed that the cumulative baseline hazard $H_0(t)$ is a step function with jumps at the observed event times. So in (9), $h_0(t_i)$ is replaced by $\Delta H_0(t_i)$ and $\int_0^{t_i} h_0(u) du$ is replaced by $\sum \Delta H_0(t_j)$, where the sum is over all event times t_j not larger than t_i .

There does not appear to exist any ready-made software for implementing the full likelihood approach, and we will not consider the full likelihood approach in our simulations (Sect. 6).

4.2 Multiple imputation

As noted in the preceding section, in a nested case–control study, information on $\mathbf{x}_i^{(\text{cc})}$ is missing by design for $i \in \mathcal{C} \setminus \mathcal{O}$. Another approach to analysis which makes use of the data available in the full cohort is to treat this as a missing data problem and to use MI (Rubin 1987) to impute the values of $\mathbf{x}_i^{(\text{cc})}$ which are missing for individuals outside the nested case–control sample. MI is now a widely used approach to handling missing data and is familiar to many researchers; see e.g. Carpenter and Kenward (2013) for a review. Its application to the use of full cohort data in nested case–control studies was described by Keogh and White (2013).

The key idea in using MI for missing data is that the missing values are imputed by drawing random values from the joint distribution of the partially observed variables conditional on all fully observed variables, including the outcome. To account for the uncertainty in the imputed values a number $M > 1$ of imputed values are obtained for each missing data point, creating M complete imputed data sets. The resulting data sets are analysed separately but identically and the resulting estimates are combined using Rubin's Rules (Rubin 1987). MI results in asymptotically unbiased estimates and correct standard errors provided the imputation model is correctly specified. An alternative to specifying a joint model is to instead specify a separate univariate model for each partially observed variable conditional on all other variables; this is called the full conditional specification (Van Buuren 2007). This last approach is simpler than using a joint model, especially when there are several partially missing variables of different types, e.g. binary and numeric.

To describe how MI is implemented for nested case-control data, we begin by considering the situation in which there is only one partially observed covariate, $x_i^{(cc)}$, assumed to be numeric. White and Royston (2009, Appendix A2) show that if the conditional distribution of $x_i^{(cc)}$ given $\mathbf{x}_i^{(all)}$ and \mathbf{z}_i is normal with a mean that is linear in $\mathbf{x}_i^{(all)}$ and \mathbf{z}_i , and if censoring does not depend on $x_i^{(cc)}$, then the conditional distribution of $x_i^{(cc)}$ given $\mathbf{x}_i^{(all)}$, \mathbf{z}_i and the outcome (d_i, t_i) , may be approximated by a linear regression model for $x_i^{(cc)}$ with $\mathbf{x}_i^{(all)}$, \mathbf{z}_i , d_i , and the cumulative baseline hazard $H_0(t_i)$ as covariates. Moreover, for computational purposes, one may replace the cumulative baseline hazard by the Nelson-Aalen estimate $\widehat{H}(t_i)$. This motivates the imputation model

$$x_i^{(cc)} = \theta_0 + \boldsymbol{\theta}'_1 \mathbf{x}_i^{(all)} + \boldsymbol{\theta}'_2 \mathbf{z}_i + \theta_3 d_i + \theta_4 \widehat{H}(t_i) + \epsilon_i, \quad (10)$$

where the ϵ_i are normally distributed with mean 0 and constant variance σ_ϵ^2 . After fitting (10) for the individuals in the nested case-control sample, the next step is to take a draw of each of the model parameters from their posterior (estimated) joint distribution. To explain this procedure further, we let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_0, \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \widehat{\theta}_3, \widehat{\theta}_4)$ denote the estimated regression coefficients from (10) with variance-covariance matrix \mathbf{V} , and let $\widehat{\sigma}_\epsilon^2$ denote the estimated residual variance. Draws σ_ϵ^* are obtained using $\sigma_\epsilon^* = \widehat{\sigma}_\epsilon \sqrt{(n - J)/g}$, where J is the length of the vector $\widehat{\boldsymbol{\theta}}$ and g is a draw from a chi-squared distribution with $n - J$ degrees of freedom. Values $\boldsymbol{\theta}^*$ are then obtained using $\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}} + \sigma_\epsilon^* \widehat{\sigma}_\epsilon^{-1} \mathbf{u} \mathbf{V}^{1/2}$, where \mathbf{u} is a vector of random draws from a standard normal distribution and $\mathbf{V}^{1/2}$ denotes the Cholesky decomposition of \mathbf{V} (White et al. 2011). We let $\boldsymbol{\theta}^{*(m)}$ and $\sigma_\epsilon^{2*(m)}$ denote the m th set of parameter draws obtained using this procedure. The imputed values for the m th imputed data set are obtained using

$$x_i^{(cc)(m)} = \theta_0^{*(m)} + \boldsymbol{\theta}_1^{*(m)'} \mathbf{x}_i^{(all)} + \boldsymbol{\theta}_2^{*(m)'} \mathbf{z}_i + \theta_3^{*(m)} d_i + \theta_4^{*(m)} \widehat{H}(t_i) + \epsilon_i^*, \quad (11)$$

where ϵ_i^* is a random draw from a normal distribution with mean 0 and variance $\sigma_\epsilon^{2*(m)}$. This is repeated M times to obtain M imputed data sets. For a missing binary covariate, the linear regression imputation model in (10) is replaced by a logistic regression model

and the procedure for obtaining parameter draws is adjusted accordingly (White et al. 2011). The use of parameter values drawn from their posterior distribution, as opposed to using the imputation model parameter estimates themselves in each imputation, is necessary to create so-called ‘proper’ imputations (Rubin 1987). If this is not done then the estimated variances of the combined estimates of the parameters of interest (see below) will be an underestimate (e.g. Carpenter and Kenward 2013, pp. 63–64).

In the m th imputed data set ($m = 1, \dots, M$) a full cohort analysis is performed using the partial likelihood (2). The resulting parameter estimates are denoted $\widehat{\beta}^{(m)}$ and $\widehat{\gamma}^{(m)}$. According to Rubin’s Rules the combined estimates $\widehat{\beta}$ and $\widehat{\gamma}$ are given by the average of $\widehat{\beta}^{(m)}$ and $\widehat{\gamma}^{(m)}$ over the M imputed data sets and their variance is estimated by $\mathbf{W} + (1 + 1/M)\mathbf{B}$, where \mathbf{W} and \mathbf{B} are the within- and between-imputation components of variance.

The MI procedure outlined above can be extended to account for a vector of partially missing covariates $\mathbf{x}_i^{(cc)}$. We let $\mathbf{x}_{i(-j)}^{(cc)}$ denote $\mathbf{x}_i^{(cc)}$ with the j th element removed. Then, for a numeric covariate, the imputation model for element j of $\mathbf{x}_i^{(cc)}$, corresponding to (10), is

$$x_{ij}^{(cc)} = \theta_{0j} + \theta'_{1j}\mathbf{x}_i^{(all)} + \theta'_{2j}\mathbf{x}_{i(-j)}^{(cc)} + \theta'_{3j}\mathbf{z}_i + \theta_{4j}d_i + \theta_{5j}\widehat{H}(t_i) + \epsilon_{ij}. \tag{12}$$

In the multivariable situation, an iterative procedure is used to obtain imputed values for $\mathbf{x}_i^{(cc)} = (x_{i1}^{(cc)}, \dots, x_{ip}^{(cc)})'$; see White et al. (2011) for an overview. Starting with $x_{i1}^{(cc)}$, the imputation model (12) is fitted in those with complete data and draws of the parameters in model (12) are taken from their posterior distribution. Missing values of $x_{i1}^{(cc)}$ are then imputed using the model. This procedure is repeated for each partial missing variable $x_{i2}^{(cc)}, \dots, x_{ip}^{(cc)}$ in turn, using the imputed values of $x_{i1}^{(cc)}, \dots, x_{i,j-1}^{(cc)}$ when fitting the model for $x_{ij}^{(cc)}$. The whole procedure is then repeated a number of times until convergence of the parameter estimates, following which a final draw of the parameters is taken from their posterior and a final set of imputed values obtained for each $x_{ij}^{(cc)}$. This forms the first imputed data set. The iterative procedure is repeated to obtain M imputed data sets.

In the derivation of the imputation model (10), it is assumed that censoring does not depend on the partially missing covariate $x_i^{(cc)}$. When this is the case, one may consider censoring as a competing risk and adopt the approach for two end-points described in Sect. 5.4.

The MI approach described in this section can be implemented for example using the `mice` package in R (Van Buuren and Groothuis-Oudshoorn 2011) or the `mi` package in Stata.

5 Nested case–control studies with more than one endpoint

In some situations one would like to use an assembled cohort to study two or more end-points (e.g. more than one disease). In a classical nested case–control study (Sect. 2.2), the controls are matched to their cases, so new controls have to be selected for each

endpoint. However, by breaking the matching one may use the controls selected for one endpoint as controls also for another endpoint.

For simplicity, we will consider the situation with two endpoints and assume that the occurrence of one endpoint precludes the occurrence of the other. The situation may then be described by a competing risks model with two causes of failure corresponding to the two endpoints of interest. We will here outline how the results in Sects. 2, 3, and 4 may be modified to cover the situation with two endpoints.

5.1 Cohort and nested case–control data

For the i th individual in the cohort we assume that the hazard for the e th endpoint is given by a Cox model of the form

$$h_{ei}(t) = h_{e0}(t) \exp(\boldsymbol{\beta}'_e \mathbf{x}_i + \boldsymbol{\gamma}'_e \mathbf{z}_i), \quad (13)$$

where $\boldsymbol{\beta}_e = (\beta_{e1}, \dots, \beta_{ep})'$ and $\boldsymbol{\gamma}_e = (\gamma_{e1}, \dots, \gamma_{eq})'$ are vectors of regression coefficients for endpoint e ; $e = 1, 2$. For individual i we now observe (T_i, D_{1i}, D_{2i}) , where T_i is an event time for one of the two endpoints or a censoring time. Further $D_{ei} = 1$ if T_i equals the event time for the e th endpoint and $D_{ei} = 0$ otherwise; $e = 1, 2$. We write $\mathcal{E}_e = \{i \mid D_{ei} = 1\}$ for the set of cases for endpoint e .

For cohort data we may then estimate the regression coefficients $\boldsymbol{\beta}_e$ and $\boldsymbol{\gamma}_e$ for the e th endpoint by maximizing a partial likelihood of the form (2), but with the product restricted to $i \in \mathcal{E}_e$. For nested case–control data, the controls are selected as described in Sect. 2.2. However, since the controls are matched to the cases, there will be a separate set of controls for each of the two endpoints. Here we may estimate $\boldsymbol{\beta}_e$ by maximizing a partial likelihood of the form (3), where again the product is for $i \in \mathcal{E}_e$.

5.2 Inverse probability weighting

If we break the matching, all controls may be used for both endpoints. The regression coefficients $\boldsymbol{\beta}_e$ and $\boldsymbol{\gamma}_e$ for the e th endpoint may then be estimated by maximizing a pseudo likelihood of the form (5), where the product is over $i \in \mathcal{E}_e$ and \mathcal{S}_i is the set consisting of the cases and controls for both endpoints who are at risk at time T_i . Further, we now estimate the inclusion probabilities p_j using the controls for both endpoints. In particular, in (6) the product is over the event times for both endpoints when individual j is a possible matched control, while in (7) the sampling indicators O_j are 1 for sampled controls for both endpoints (who do not later become a case for any of the two endpoints).

5.3 Maximum likelihood

When there are two endpoints the likelihood (8) should be modified by replacing $p(t_i, d_i \mid \mathbf{x}_i, \mathbf{z}_i)$ by $p(t_i, d_{1i}, d_{2i} \mid \mathbf{x}_i, \mathbf{z}_i)$, the conditional density of (T_i, D_{1i}, D_{2i})

given $\mathbf{x}_i, \mathbf{z}_i$. Further, assuming the Cox model (13) for the hazards of the two endpoints, the conditional density of (T_i, D_{1i}, D_{2i}) given $\mathbf{x}_i, \mathbf{z}_i$ may be given as

$$p(t_i, d_{1i}, d_{2i} | \mathbf{x}_i, \mathbf{z}_i) = \prod_{e=1}^2 \left\{ \left[h_{e0}(t_i) \exp \left(\boldsymbol{\beta}'_{e,cc} \mathbf{x}_i^{(cc)} + \boldsymbol{\beta}'_{e,all} \mathbf{x}_i^{(all)} + \boldsymbol{\gamma}'_e \mathbf{z}_i \right) \right]^{d_{ei}} \times \exp \left(- \exp \left(\boldsymbol{\beta}'_{e,cc} \mathbf{x}_i^{(cc)} + \boldsymbol{\beta}'_{e,all} \mathbf{x}_i^{(all)} + \boldsymbol{\gamma}'_e \mathbf{z}_i \right) \int_0^{t_i} h_{e0}(u) du \right) \right\}. \tag{14}$$

As for one endpoint (Sect. 4.1), we need to specify the conditional distribution of $\mathbf{x}_i^{(cc)}$ given $\mathbf{x}_i^{(all)}$ and \mathbf{z}_i or adopt a non-parametric approach where the conditional distribution has point masses at the observed covariate values. Also when maximizing the likelihood, we assume that the cumulative baseline hazards $H_{e0}(t)$ of the two endpoints are step functions with jumps at the observed event times. So in (14), $h_{e0}(t_i)$ is replaced by $\Delta H_{e0}(t_i)$ and $\int_0^{t_i} h_{e0}(u) du$ is replaced by $\sum \Delta H_{e0}(t_j)$, where the sum is over all event times for endpoint e with t_j not larger than t_i .

5.4 Multiple imputation

By following the workings of White and Royston (2009, Appendix A) it is straightforward to extend the imputation model in (10) to two (or more) endpoints. For a univariate partially observed numeric covariate $x_i^{(cc)}$, the imputation model in (10) is modified as follows

$$x_i^{(cc)} = \theta_0 + \boldsymbol{\theta}'_1 \mathbf{x}_i^{(all)} + \boldsymbol{\theta}'_2 \mathbf{z}_i + \boldsymbol{\theta}'_3 \mathbf{d}_i + \boldsymbol{\theta}'_4 \widehat{\mathbf{H}}(t_i) + \epsilon_i. \tag{15}$$

Here $\mathbf{d}_i = (d_{1i}, d_{2i})'$ is a vector of event type indicators for the i th individual and $\widehat{\mathbf{H}}(t_i) = (\widehat{H}_1(t_i), \widehat{H}_2(t_i))'$ is a vector of Nelson–Aalen estimates of the cumulative hazards for the two endpoints. The generalisation to the situation of a vector of partially observed covariates, $\mathbf{x}_i^{(cc)}$, is by using a similar extension to imputation model (12).

6 Simulation study

In a standard analysis of nested case–control data using the partial likelihood (3), the controls are matched to the cases. As noted at the end of Sect. 2.2 the matching implies that the standard analysis gives valid inference for $\boldsymbol{\beta}$ under quite weak assumptions on the effects of the confounding variables. In an IPW and MI analysis, the controls are no longer matched to the cases, i.e. the matching is broken. In this section we will use simulations to investigate (i) problems that may occur if one breaks the matching, and (ii) benefits one may obtain by breaking the matching.

We focus on a single numeric partially observed covariate $x^{(cc)}$ and begin by considering a ‘basic’ situation with one endpoint of interest (Sect. 6.1). We then extend this to incorporate a number of special issues which may affect the results, including

the presence of batch effects in the measurements of the partially observed covariate (Sect. 6.2), interaction terms between the partially observed covariate and other covariates in the model for the hazard (Sect. 6.3), mis-specification of the form of the covariates in the hazard model, and issues which may affect the MI approach (Sect. 6.4). Finally, we consider the situation with two endpoints (Sect. 6.5).

For all situations (described in detail below) we assume an underlying cohort of 10,000 individuals recruited at three equal sized centres. The individuals are between 50 and 70 years at recruitment and they are followed until an event of interest occurs, until closure of the study after 15 years, or until censoring before that time. For all the scenarios, we assume that 2% of the individuals experience the event of interest, 10% drop-out before the end of the study, and 20% are censored by a competing event. (With two endpoints, 2% of the individuals experience the first endpoint and 1% the second endpoint.)

From a simulated cohort with 10,000 individuals, we select a nested case–control sample with 1 or 3 controls per case, additionally matched on age and centre. For each simulated cohort we then estimate the regression coefficients using the methods of Sects. 2.1, 2.2, 3, and 4.2 (or the corresponding methods of Sect. 5 for two endpoints) and report summaries based on 1,000 simulated cohorts. For the IPW method we compute estimates using the inclusion probabilities (6) and (7), as well as the modification of (7) using a generalized additive model. However, as there were only minor differences between the three IPW methods, we in the tables below only give results for the logistic inclusion probabilities (7). The results for MI are based on $M = 10$ imputed data sets.

6.1 Basic simulation

Description of the simulation. We assume that there are two covariates of main interest: one “expensive” covariate $x^{(\text{cc})}$ that is only observed for the cases and the controls, and one “cheap” covariate $x^{(\text{all})}$ that is observed for all individuals in the cohort. We also consider two additional confounding variables: the age of an individual and the centre where the individual is recruited. Throughout we assume that age at recruitment is uniformly distributed on the integers 50, 51, . . . , 69, and that individuals are uniformly distributed over the three centres. The information on age and centre is given by the covariates $\mathbf{z} = (z_1, z_2, z_3)'$, where $z_1 = \text{“age} - 60\text{”}$ and z_2 and z_3 are indicators for centres 2 and 3, respectively.

Conditional on age and centre, the distribution of $\mathbf{x} = (x^{(\text{cc})}, x^{(\text{all})})'$ is bivariate normal with mean $(\xi_1 z_1 + \xi_2 z_2 + \xi_3 z_3, 0)'$, standard deviations 1, and correlation 0.70. For the parameters of the conditional mean we chose the values $\xi_1 = 0.05$, $\xi_2 = 0.50$, and $\xi_3 = -0.50$ corresponding to a variation in $x^{(\text{cc})}$ of one standard deviation due to age and one standard deviation due to centre.

Given the covariates, an event time T_e , measured from the time of recruitment, is generated from a proportional hazards model

$$h(t | \mathbf{x}, \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x} + \boldsymbol{\gamma}'\mathbf{z}), \quad (16)$$

where $\beta = (\beta_{cc}, \beta_{all})'$, $\gamma = (\gamma_1, \gamma_2, \gamma_3)'$, and the baseline hazard $h_0(t)$ takes a Weibull form with shape parameter a . The event time may be censored due to drop-out, censoring by a competing risk, or closure of the study. More specifically the censoring time T_c is given as $T_c = \min(T_{c1}, T_{c2}, 15)$, where the (potential) time of drop-out T_{c1} is exponentially distributed with rate λ_c and the (potential) time T_{c2} of censoring by a competing risk is generated from a proportional hazards model with $x^{(cc)}$ and $z_1 = \text{“age} - 60\text{”}$ as covariates:

$$h_c(t | \mathbf{x}, \mathbf{z}) = h_{c0}(t) \exp \left(\delta_1 x^{(cc)} + \delta_2 z_1 \right). \tag{17}$$

The baseline hazard $h_{c0}(t)$ is of the Weibull form with shape parameter a_c . The observed event time is then given by $T = \min(T_e, T_c)$ while the event indicator is specified as $D = I\{T_e < T_c\}$.

Events in the cohort are generated from the Cox model (16) with Weibull shape $a = 5$. The regression coefficients for the covariates of main interest are $\beta_1 = \beta_2 = 0.50$. For the confounding variables, age has the regression coefficient $\gamma_1 = 0.10$, while the regression coefficients for the centres are $\gamma_2 = 0.30$ for centre 2 and $\gamma_3 = -0.30$ for centre 3 (centre 1 is the reference). For censoring by the competing risk, we assume Weibull shape $a_c = 5$ and regression coefficients $\delta_1 = 0$ for $x^{(cc)}$ and $\delta_2 = 0.10$ for age. Thus for the basic simulation model, censoring by the competing risk does not depend on $x^{(cc)}$.

For all situations involving the basic simulation model and its extensions in Sects. 6.2–6.4, the scale parameters of the baseline Weibull hazards in (16) and (17), and the rate of drop out λ_c are adjusted to give 2% events in the cohort, 10% drop-outs, and 20% censoring by the competing event.

Results. The results from the basic simulation for a nested case–control sample with 1 control per case are summarised in Table 1. The corresponding results when there are 3 controls per case are shown in Supplementary Table 1.

As we expect, the standard analysis of the nested case–control data results in almost unbiased estimates of the parameters of interest, though the standard errors are almost 80% larger than under a full cohort analysis. The loss of efficiency of the standard nested case–control analysis is reduced when the number of controls per case increases. With 3 controls per case the standard errors for the standard analysis are about 30% larger than for the full cohort.

When there is 1 control per case, breaking the matching in the nested case–control data and performing an IPW analysis results in some upwards bias in the estimates of both β_{cc} and β_{all} and the mean squared differences from the cohort estimates are larger than for the standard analysis. Further the standard errors are somewhat underestimated, but the coverage is close to the nominal level. The bias in the estimate of β_{cc} is smaller, though still present, when the number of controls per case increases to 3, though the bias in β_{all} disappears and the standard errors are unbiased. The standard nested case–control analysis and an IPW analysis give estimates with about the same standard errors, so the IPW analysis results in no gain in efficiency compared with the standard analysis, despite enabling use of more controls at each event time.

Table 1 Results from the basic simulation

	Estimate	Model SE	Emp SD	MSE	Cov
Results for β_{cc}					
Full cohort	0.502	0.099	0.100	–	0.946
Nested case–control: standard analysis	0.513	0.176	0.176	0.021	0.953
Nested case–control: IPW	0.534	0.177	0.187	0.025	0.940
Nested case–control: MI	0.499	0.157	0.153	0.014	0.941
Results for β_{all}					
Full cohort	0.495	0.099	0.098	–	0.955
Nested case–control: standard analysis	0.506	0.176	0.175	0.022	0.952
Nested case–control: IPW	0.532	0.177	0.187	0.026	0.935
Nested case–control: MI	0.498	0.127	0.121	0.006	0.963

The nested case–control sample has one control per case. All results are based on 1,000 simulated data sets where the true log hazard ratios are $\beta_{cc} = \beta_{all} = 0.50$. ‘Estimate’ is the mean estimate of β_{cc} or β_{all} . ‘Model SE’ is the square-root of the mean estimated variance of the parameter estimates. ‘Emp SD’ is the empirical standard deviation of the estimates. ‘MSE’ is the mean of the squared differences between the estimates for each method and those from the full cohort. ‘Cov’ is the obtained coverage of a 95 % nominal confidence interval

The MI analysis of the nested case–control data gives unbiased estimates of β_{cc} and β_{all} and correct model-based standard errors. This approach also results in smaller standard errors and mean squared differences from the cohort estimates than the standard nested case–control analysis. As we expected, the gain in efficiency is greater for the parameter associated with the covariate which is observed in the full cohort and when there is only 1 control per case.

6.2 Batch effects

Description of the simulation. Next, we assume that the covariate $x^{(cc)}$ is a biomarker that is determined by a biochemical analysis. Then it is quite common to analyse a case and its control in the same batch to control for possible batch effects (Rundle et al. 2005). In a standard analysis using the partial likelihood (3) a batch effect is taken care of by the matching on batch (in addition to matching on the other confounding variables \mathbf{z}). But when the matching is broken in an IPW or MI analysis, bias may be introduced. To investigate this, we generate cohort data as described for the basic simulation model (Sect. 6.1), but in the data used for estimation we add a common measurement error to $x^{(cc)}$ for a case and its controls. The measurement errors are assumed to be normal with mean zero and standard deviation 0.50 or 0.25, which is a half or a quarter of the standard deviation of the random variation in $x^{(cc)}$.

Results. For a nested case–control study with 1 control per case and with measurement error standard deviation 0.50, the results from analyses when there are batch effects in the measurements are shown in Table 2. Corresponding results when there are 3

Table 2 Results from the basic simulation extended to include a batch effect with standard deviation 0.50 in the measurement of $x^{(cc)}$

	Estimate	Model SE	Emp SD	Cov
Results for β_{cc}				
Nested case-control: standard analysis	0.513	0.176	0.176	0.953
Nested case-control: IPW	0.370	0.138	0.146	0.806
Nested case-control: MI	0.498	0.158	0.152	0.953
Results for β_{all}				
Nested case-control: standard analysis	0.506	0.176	0.175	0.952
Nested case-control: IPW	0.645	0.160	0.171	0.858
Nested case-control: MI	0.499	0.128	0.121	0.960

The nested case-control sample has one control per case. All results are based on 1,000 simulated data sets where the true log hazard ratios are $\beta_{cc} = \beta_{all} = 0.50$. 'Estimate' is the mean estimate of β_{cc} or β_{all} . 'Model SE' is the square-root of the mean estimated variance of the parameter estimates. 'Emp SD' is the empirical standard deviation of the estimates. 'Cov' is the obtained coverage of a 95 % nominal confidence interval

controls per case are shown in Supplementary Table 2. Note that in this scenario it is not relevant to consider a full cohort analysis.

In the standard nested case-control analysis the batch effects are eliminated in the partial likelihood and the standard analysis therefore results in unbiased estimates of β_{cc} and β_{all} and correct standard errors. In fact, for the standard analysis the results of Tables 1 and 2 are identical.

The IPW analysis gives biased estimates of β_{cc} with the bias being towards the null. The estimate of β_{all} is also biased, but in the direction away from the null. Some of the bias in the estimate of β_{all} is alleviated when the number of controls per case is increased to 3, though there remains a substantial bias. The standard error estimates are also biased in a similar pattern as described for the basic simulation. The reason for biased estimates under the IPW analysis is that the batch effects do not cancel each other out in the pseudo likelihood (5), since the denominator in the pseudo likelihood at a given time now includes individuals with a range of batch effects. This results in a measurement error in the values of $x^{(cc)}$ used in the pseudo likelihood, which is known to result in biased estimates of the parameters associated with $x^{(cc)}$ and with the adjustment variables.

In Table 2 we assumed a reasonably large batch effect. The biases in the estimates from the IPW analysis are clearly smaller when the batch effects are smaller; cf. Supplementary Table 3 which show results from the situation with one control per case and measurement error standard deviation 0.25.

The MI approach appears to work well even in the presence of batch effects, giving apparently unbiased estimates of both β_{cc} and β_{all} , plus correct standard errors. As in the basic simulation, there are gains in efficiency to be made by using the MI approach in place of the standard nested case-control analysis, especially in the estimation of β_{all} and especially when there is only 1 control per case in the nested case-control sample. When we perform MI we fit an imputation model which has the partially missing covariate as the outcome, as in (10). Random measurement error in an outcome

variable used in a regression does not give rise to bias in the estimated regression coefficients, hence the regression coefficients in the imputation model are consistently estimated. There remain batch effects in the measurements for individuals for whom $x^{(cc)}$ is observed. This has negligible impact here, however, because the proportion of individuals with $x^{(cc)}$ observed is small. If the nested case–control sample made up a more substantial part of the cohort then some bias in the MI estimates may be anticipated, though this is not likely to be a common scenario.

6.3 Interaction

Description of the simulation. It is known that the MI approach described in Sect. 4.2 may give biased estimates for interactions between $x^{(cc)}$ and the other covariates (Keogh and White 2013). To investigate this further, we simulate event times from a Cox model with interaction between $x^{(cc)}$ and $x^{(all)}$. Specifically, the cohort is simulated as described in Sect. 6.1, except that in the Cox model (16) we now have $\mathbf{x} = (x^{(cc)}, x^{(all)}, x^{(cc)}x^{(all)})'$ and $\boldsymbol{\beta} = (\beta_{cc}, \beta_{all}, \beta_{int})'$ with $\beta_{cc} = \beta_{all} = \beta_{int} = 0.50$. Using MI, we handled the interaction between the partially observed covariate $x^{(cc)}$ and the fully observed covariate $x^{(all)}$ by using the ‘passive’ approach in which $x^{(cc)}$ is imputed in the usual way [i.e. using (10) and (11)] and the interaction is obtained by multiplying the imputed value of $x^{(cc)}$ by $x^{(all)}$.

Results. The results are shown in Table 3 for a nested case–control study with 1 control per case. Corresponding results for a study with 3 controls per case are shown in Supplementary Table 4.

The MI approach results, as we expect, in fairly substantial bias in the estimates of the main effect terms, β_{cc} and β_{all} , and the interaction term, β_{int} , and substantial loss of coverage both for 1 and 3 controls per case. The bias arises due to a lack of compatibility, or ‘incongeniality’, between the imputation model for the missing covariate and the hazard model for the outcome of interest (Meng 1994). When there is 1 control per case, the standard method gives large standard deviations and large mean squared differences from the cohort estimates. But with 3 controls per case, the standard method performs much better. The IPW results show some minor upwards bias in the main effect estimates, though the bias is much less than that found using MI and the interaction term is approximately unbiased. The IPW analysis gives considerable gain in efficiency in estimation of the interaction term compared with the standard nested case–control analysis, in particular when there is 1 control per case. There is also a gain in efficiency in the main effect terms, though the standard errors are under-estimated and the coverage is too low.

6.4 Further extensions to the basic simulation

To assess the sensitivity of the methods to certain assumptions, we performed a number of additional simulations, based on extensions to the basic simulation model of Sect. 6.1:

Table 3 Results from the basic simulation extended to include an interaction between $x^{(cc)}$ and $x^{(all)}$

	Estimate	Model SE	Emp SD	MSE	Cov
Results for β_{cc}					
Full cohort	0.508	0.113	0.115	-	0.945
Nested case-control: standard analysis	0.530	0.268	0.280	0.064	0.943
Nested case-control: IPW	0.546	0.204	0.236	0.041	0.907
Nested case-control: MI	0.421	0.186	0.226	0.034	0.861
Results for β_{all}					
Full cohort	0.510	0.119	0.121	-	0.944
Nested case-control: standard analysis	0.526	0.278	0.289	0.067	0.952
Nested case-control: IPW	0.554	0.215	0.256	0.052	0.892
Nested case-control: MI	0.741	0.194	0.163	0.068	0.792
Results for β_{int}					
Full cohort	0.498	0.048	0.050	-	0.941
Nested case-control: standard analysis	0.526	0.155	0.167	0.025	0.952
Nested case-control: IPW	0.510	0.075	0.082	0.004	0.910
Nested case-control: MI	0.327	0.083	0.078	0.032	0.428

The nested case-control sample has one control per case. All results are based on 1,000 simulated data sets where the true log hazard ratios are $\beta_{cc} = \beta_{all} = \beta_{int} = 0.50$. ‘Estimate’ is the mean estimate of β_{cc} , β_{all} or β_{int} . ‘Model SE’ is the square-root of the mean estimated variance of the parameter estimates. ‘Emp SD’ is the empirical standard deviation of the estimates. ‘MSE’ is the mean of the squared differences between the estimates for each method and those from the full cohort. ‘Cov’ is the obtained coverage of a 95% nominal confidence interval

- (1) *Mis-specified hazard model.* We consider two types of mis-specification of the hazard model (16). Firstly, we consider mis-specification of the way the confounders \mathbf{z} influence the hazard. Here we assume that the basic simulation is altered to include a squared term in the covariate $z_1 = \text{“age} - 60\text{”}$. The log hazard ratio associated with this non-linear term was chosen to be -0.005 , representing a realistic scenario, or -0.05 , representing a rather extreme scenario. Secondly, following Scott and Wild (1986, 2002), we assume that the effect of the covariate $x^{(cc)}$ itself is mis-specified, and that the true hazard also includes a squared term in $x^{(cc)}$ with coefficient equal to 0.15 or -0.17 (but we fit a Cox model without the squared term). The values of the coefficients of the squared term were chosen so that the likelihood ratio test based on the partial likelihood (3) has about 50% power (at the 5% level) to detect the curvature. For both types of mis-specification of the hazard model, the other log hazard ratio parameters are as given in Sect. 6.1.
- (2) *Censoring by a competing risk depending on $x^{(cc)}$.* The MI approach for a single endpoint (Sect. 4.2) assumes that censoring does not depend on $x^{(cc)}$. To investigate how sensitive MI is to this assumption, we generate data as described for the basic simulation, but with $\delta_1 = 1$ in the competing risk hazard model (17).
- (3) *Mis-specification of the conditional distribution of $x^{(cc)}$.* The imputation model (10) is derived by assuming that the conditional distribution of $x^{(cc)}$ given $x^{(all)}$ and \mathbf{z} is normal with a mean that is linear in $x^{(all)}$ and \mathbf{z} (White and Royston 2009). We

Table 4 Results from the basic simulation extended to include a squared term in $x^{(cc)}$ in the hazard model with coefficient -0.17

	Estimate	Model SE	Emp SD	MSE
Results for β_{cc}				
Full cohort	0.222	0.099	0.092	–
Nested case–control: standard analysis	0.300	0.167	0.168	0.025
Nested case–control: IPW	0.229	0.159	0.163	0.017
Nested case–control: MI	0.293	0.156	0.157	0.020
Results for β_{all}				
Full cohort	0.497	0.099	0.098	–
Nested case–control: standard analysis	0.508	0.164	0.168	0.018
Nested case–control: IPW	0.528	0.166	0.177	0.020
Nested case–control: MI	0.464	0.123	0.124	0.007

The nested case–control sample has one control per case. All results are based on 1,000 simulated data sets where the true log hazard ratios are $\beta_{cc} = \beta_{all} = 0.50$. ‘Estimate’ is the mean estimate of β_{cc} or β_{all} . ‘Model SE’ is the square-root of the mean estimated variance of the parameter estimates. ‘Emp SD’ is the empirical standard deviation of the estimates. ‘MSE’ is the mean sum of the squared differences between the estimates from each method and those from the full cohort

investigate how sensitive MI is to this assumption using two scenarios: (i) either the assumption of normality in $x^{(cc)}$ is violated by generating $x^{(cc)}$ to be log-normal with standard deviation 1, or (ii) the assumption of a linear mean is violated by altering the generation of $x^{(cc)}$ to have mean $\xi_1 z_1 + \xi_2 z_2 + \xi_3 z_3 + \xi_4 z_1^2$, where $z_1 = \text{“age} - 60\text{”}$. In (ii) we use parameter values $\xi_4 = -0.005$ and $\xi_4 = -0.05$, and the other parameters remain unchanged.

The results from a mis-specified hazard model are shown in Table 4, Supplementary Tables 5, 6, 7; those from allowing the censoring by a competing risk to depend on $x^{(cc)}$ in Supplementary Tables 8 and 9; and those from mis-specification of the conditional distribution of $x^{(cc)}$ in Supplementary Tables 10, 11, 12.

When the hazard model is mis-specified due to ignoring a modest quadratic effect of a confounding variable (log hazard ratio -0.005 , Supplementary Table 5), there is a small upwards bias in the estimates for the IPW approach, while there is no bias for the other methods. For a strong quadratic effect (log hazard ratio -0.05 , Supplementary Table 6) the bias for the IPW approach becomes larger, but the other methods still give essentially unbiased estimates.

When the mis-specification of the hazard model is due to a quadratic effect of the covariate of main interest $x^{(cc)}$, it is of little interest to compare the estimates with the coefficient $\beta_{cc} = 0.50$ for the linear term. What is of interest here is to see how the estimates for the various methods compare with those from the full cohort. Both for a negative curvature (Table 4) and a positive curvature (Supplementary Table 7), the IPW estimates are the ones closest to the full cohort estimates, while the standard method and MI estimation gives estimates that deviate more from those obtained for the full cohort.

When the censoring by a competing risk depends on $x^{(cc)}$ the results from the MI analysis appear unaffected, with the estimates remaining approximately unbiased (Supplementary Table 8). However, if the censoring by the competing risk is increased to 50 %, MI gives an estimate of β_{cc} that is a bit too low (Supplementary Table 9). We allowed a rather large effect of $x^{(cc)}$ on the censoring, so our results suggest that the MI approach is quite robust to departures from the assumption that the censoring distribution does not depend on the partially missing covariates in this setting. The IPW analysis shows some upwards bias, but this is no greater than that found under the basic simulation.

When $x^{(cc)}$ is log-normal (Supplementary Table 10) the full cohort and standard nested case-control analysis are unaffected. The point estimates from the IPW analysis also appear to be unaffected, though there remains some upwards bias as in other scenarios, while the standard errors are too low, resulting in reduced coverage. The MI analysis gives bias in both β_{cc} (towards the null) and β_{all} (away from the null) due to mis-specification of the imputation model. Here the model is badly mis-specified. A modest, and what we may consider realistic, non-linear term in age in the model for the mean of $x^{(cc)}$ results in a minor downward bias in the MI estimates of β_{cc} and a minor upwards bias in the IPW estimates of $x^{(all)}$ (Supplementary Table 11). However, a larger non-linear effect in age in the model for the mean of $x^{(cc)}$ results in substantial bias in both the IPW and MI estimates, with the bias being slightly more severe using MI (Supplementary Table 12). The larger non-linear effect for age is likely to be unrealistic, but we show the results for illustration of what could happen in an extreme scenario.

6.5 Two endpoints

Description of the simulation. Finally, we use simulations to investigate a scenario with two endpoints, applying the methods described in Sect. 5. The simulation for two endpoints follows the basic simulation for one endpoint, except that (potential) event times for the two endpoints are generated using cause specific hazard functions of the form (16). The observed time T for a given individual is the time of whichever occurs first of endpoint 1, endpoint 2, censoring by a competing risk [using the hazard model in (17)], random drop-out and the end of follow-up. The log hazard ratio parameters in (16) are assumed to be the same for both endpoints with values as given in Sect. 6.1. The scale parameters of the Weibull baseline hazards in (16) and (17), and the drop-out rate λ_c , are chosen so that 2 % of individuals in the cohort are observed to have endpoint 1, 1 % to have endpoint 2, 20 % are censored by a competing risk, and 10 % drop out.

Nested case-control samples are taken for each endpoint with 1 control per case for both endpoints or 3 controls per case for both endpoints. Both $x^{(cc)}$ and $x^{(all)}$ are observed for individuals in the nested case-control samples, but not in the remainder of the cohort, where only $x^{(all)}$ and \mathbf{z} are observed. In the nested case-control analyses for endpoint 1 using IPW and MI, individuals from the nested case-control sample based on endpoint 2 contribute to the risk set at each event time, and vice versa. In the MI analysis individuals who are not in the

nested case–control sample also contribute to the risk sets, with imputed values for $x^{(cc)}$.

Results. The results for nested case–control samples for two endpoints with 1 control per case are summarised in Table 5. The corresponding results when there are three controls per case are shown in Supplementary Table 13.

The IPW analysis appears to give a slight upwards bias in the estimates, as in the basic simulations, which disappears once there are three controls per case. There remains some slight bias in the estimated standard errors using the IPW analysis, as seen under the basic simulation. The IPW analysis results in a gain in efficiency relative to the standard nested case–control analysis. Focusing on the scenario with 1 control per case, for endpoint 1 the relative efficiencies of the IPW estimates (computed as the ratio of the empirical variance for the standard analysis to the empirical variance for IPW) are 1.17 for $x^{(cc)}$ and 1.09 for $x^{(all)}$. For endpoint 2 the relative efficiencies of the IPW estimates are 1.90 for $x^{(cc)}$ and 1.77 for $x^{(all)}$. The gain in efficiency is much greater for endpoint 2, which is the rarer endpoint, because the relative increase in the number of controls is larger than for endpoint 1.

The MI analysis gives unbiased estimates of the log hazard ratio parameters for both endpoints, and the increase in efficiency relative to the standard analysis is greater than for the IPW analysis. Again focusing on the scenario with 1 control per case, for endpoint 1 the relative efficiencies of the MI estimates are 1.53 for $x^{(cc)}$ and 2.14 for $x^{(all)}$. For endpoint 2 the relative efficiencies of the IPW estimates are 2.39 for $x^{(cc)}$ and 2.82 for $x^{(all)}$.

For nested case–control studies with three controls per case, the gains in efficiency are smaller, but still not insubstantial, in particular for endpoint 2 and in particular using the MI analysis.

6.6 Summary of simulation results

The standard analysis of nested case–control studies provided almost unbiased estimates and achieved coverage close to the nominal 95% for all situations considered. But the loss in efficiency relative to a full cohort analysis was substantial, especially when there was only one control per case.

Our simulations show that there are ways in which we can gain efficiency in the analysis of nested case–control data by breaking the matching between cases and controls (IPW analysis) and by making use of all available data in the cohort (MI analysis). In the situation of a single endpoint and with linear covariate effects in the hazard and no interactions involving the covariates of main interest (the basic simulation), MI gave gains in efficiency relative to the standard nested case–control analysis, in particular for the covariate $x^{(all)}$ known for the full cohort and in particular when there was one control per case. However, using an IPW analysis in this scenario gave no benefits in terms of efficiency, and resulted in possible small bias in the parameter estimates and their estimated standard errors.

For the situation with two endpoints, we obtained gains in efficiency by breaking the matching and reusing the controls for one endpoint as controls for the other endpoint,

Table 5 Results for the situation with two endpoints

	Estimate	Model SE	Emp SD	MSE	Cov
Endpoint 1: Results for $\beta_{1,cc}$					
Full cohort	0.504	0.100	0.099	–	0.946
Nested case-control: standard analysis	0.512	0.176	0.173	0.020	0.959
Nested case-control: IPW	0.528	0.154	0.160	0.015	0.942
Nested case-control: MI	0.499	0.139	0.140	0.010	0.948
Endpoint 1: Results for $\beta_{1,all}$					
Full cohort	0.495	0.100	0.097	–	0.952
Nested case-control: standard analysis	0.499	0.175	0.171	0.019	0.963
Nested case-control: IPW	0.510	0.153	0.164	0.016	0.932
Nested case-control: MI	0.497	0.119	0.117	0.004	0.950
Endpoint 2: Results for $\beta_{2,cc}$					
Full cohort	0.503	0.141	0.141	–	0.954
Nested case-control: standard analysis	0.538	0.256	0.263	0.049	0.957
Nested case-control: IPW	0.526	0.188	0.191	0.017	0.952
Nested case-control: MI	0.497	0.171	0.170	0.010	0.947
Endpoint 2: Results for $\beta_{2,all}$					
Full cohort	0.502	0.141	0.139	–	0.956
Nested case-control: standard analysis	0.512	0.254	0.257	0.046	0.958
Nested case-control: IPW	0.519	0.188	0.193	0.017	0.949
Nested case-control: MI	0.505	0.155	0.153	0.004	0.957

The nested case-control samples for each endpoint have one control per case. All results are based on 1,000 simulated data sets where the true log hazard ratios are $\beta_{e,cc} = \beta_{e,all} = 0.50$ ($e = 1, 2$). ‘Estimate’ is the mean estimate of the log hazard ratio β_{cc} or β_{all} . ‘Model SE’ is the square-root of the mean estimated variance of the parameter estimates. ‘Emp SD’ is the empirical standard deviation of the estimates. ‘MSE’ is the mean of the squared differences between the estimates for each method and those from the full cohort. ‘Cov’ is the obtained coverage of a 95% nominal confidence interval

and vice versa. The gains in efficiency were found using both IPW and MI analyses and were particularly substantial for the rare endpoint when we have one control per case.

The impact of laboratory batch effects in covariate measurements on hazard ratio estimates is eliminated in a nested case-control study by processing samples from cases and controls within a matched set within the same batch and using the standard analysis. But when the measurement error is fairly large, breaking the matching and using an IPW analysis gave severely biased estimates.

The MI approach gave unbiased estimates and substantial gains in efficiency in the presence of a batch effect and in most other situations we have considered. But there exist scenarios in which the imputation procedure described in Sect. 4.2 can result in bias. In particular, for the situation with interaction between the covariates $x^{(cc)}$ and $x^{(all)}$ the MI analysis gave severely biased estimates. However, the IPW approach worked well for estimation of the interaction term, and gave large gains in efficiency in estimation of the interaction parameter relative to the standard analysis.

The standard analysis and the IPW approach are valid for censoring depending on the partially observed covariate, and the simulations indicate that also MI is quite robust to censoring depending on $x^{(cc)}$. Mis-specification of the model for how the hazard depends on confounders results in some bias in the IPW analysis due both to a mis-specified Cox model and mis-specification of the weights model. But when the mis-specification of the hazard model is due to a quadratic effect of the covariate of main interest $x^{(cc)}$, the IPW method gives estimates closer to the full cohort estimates than the standard method and MI estimation. Further, strong non-linear associations between the covariate $x^{(cc)}$ and confounders result in bias for IPW and MI due to mis-specifications in the imputation and weight models, though non-linear associations of realistic magnitude result in small bias.

We have presented IPW results from using weights estimated by logistic regression as in (7). Estimating the weights using generalised additive models gave very similar results to the standard logistic approach. Using the Kaplan–Meier weights (6) tended to give greater bias than the regression-based weights.

In the paper of [White and Royston \(2009\)](#) on the use of imputation models such as those in (10) and (12) it was suggested that some efficiency could be gained by additionally including interactions between the covariates and the Nelson–Aalen estimates, though simulation studies showed very similar results from the two models. We investigated the impact of extending the imputation models in this way in our simulation studies but also found the results from the extended models to be almost identical to those from the simpler models used.

7 Discussion

Nested case–control studies are commonly analysed using the partial likelihood (3). This is a “safe approach” that provides unbiased estimates and a straightforward statistical analysis under quite weak assumptions on the effects of the confounding variables.

An alternative to the standard analysis is to break the matching and use an inverse probability weighted (IPW) pseudo likelihood. As shown in our simulations, this may result in substantial efficiency gains when an assembled cohort is used to study more than one endpoint. But when close matching is needed to control for a confounder (like a batch effect), bias may be introduced by breaking the matching. Another analysis option is MI, which makes use of all information available for the full cohort. This yields improved estimates for both one and two endpoints, in particular for the covariates available for all the individuals in the cohort.

In an IPW or a MI analysis, the controls are no longer matched to the cases. One therefore has to control for the confounders by including them in the Cox regression. Thus the IPW and MI analyses require more careful modelling than the standard analysis, and one runs the risk of bias due to model mis-specification. However, our simulations indicate that the effects of the covariates of main interest are estimated without much bias unless the model for the effects of the confounders is badly mis-specified.

In our simulations there was a tendency to bias in the IPW estimates, also for correctly specified models. It is possible that the weights used in IPW could be improved

to reduce this bias, but since the weights models are in a sense ad hoc it is not clear how this would best be done.

As illustrated in our simulations, the MI approach of Sect. 4.2 may give biased estimates when there is interaction between the partially observed covariate and other covariates. The bias arises due to a lack of compatibility, or ‘incongeniality’, between the imputation model for the missing covariate and the hazard model for the outcome of interest (Meng 1994). Bartlett et al. (2014) describe an MI approach which accommodates the outcome model using rejection sampling. A Stata package is forthcoming. In the context of nested case-control data, Keogh and White (2013) implemented this approach and found it to perform well.

Our simulations also showed that MI may give biased estimates if the partially observed covariate has a strongly non-normal distribution. The imputation may then be performed on a transformed scale and back-transformed before using the imputed values in the partial likelihood analysis. However, the imputation approach described in Sect. 4.2 is not suitable in this case because the non-linearity of the transformed partially observed covariate in the hazard model gives rise to an inconsistency between the imputation model and the outcome model. The problem is the same as arises for a model with interactions and the more complex imputation procedures of Bartlett et al. (2014) could be used to obtain unbiased estimates using MI.

We have shown how MI may improve the estimates by using all the available data in the cohort. An alternative to MI is a full maximum likelihood (ML) approach as described in Sects. 4.1 and 5.3. However, as no standard software is available for the full ML approach, we have not included it in our simulations. But it would be of interest to study the performance of the ML approach for our simulation scenarios.

For ease of presentation, we have focused on right-censored event times. But all the methods we have considered may also be used when event times are subject to left-truncation and right-censoring. That this is the case for cohort data and the standard analysis of nested case-control data is well known (e.g. Aalen et al. 2008, Sects. 4.1 and 4.3). Further the IPW pseudo likelihood allows for left-truncated data by an appropriate modification of the inclusion probabilities (Støer and Samuelsen 2012), while for a full maximum likelihood approach the conditional density of (T_i, D_i) given $\mathbf{x}_i, \mathbf{z}_i$ in (8) should be replaced by the conditional density of (T_i, D_i) given $\mathbf{x}_i, \mathbf{z}_i, T_i > v_i$, where v_i is the left-truncation time for the i th individual (Saarela et al. 2008). For the MI approach we may follow the arguments of White and Royston (2009, Appendix A) to see that left-truncation is handled by replacing the Nelson-Aalen estimate $\hat{H}(t_i)$ in the imputation model (10) by $\hat{H}(t_i) - \hat{H}(v_i)$.

When selecting the matched controls, we have used information on the at risk status of the individuals and their values of the confounders \mathbf{z}_i (Sect. 2.2). However, sometimes one in addition for all cohort members knows the value of a surrogate variable of an exposure of main interest. One may then select a more efficient set of controls by means of stratified (or counter-matched) sampling within strata defined by the surrogate (Langholz and Borgan 1995). The surrogate variable may also be used to obtain improved weights for the IPW method by means of calibration (Støer and Samuelsen 2012) and to improve the imputation model (10) when using MI (Keogh and White 2013).

We have assumed that the variables used for matching can only take a finite number of different values. Then it is possible to select controls with exactly the same values of the matching variables as a case. However, when matching on a numeric confounder, it is quite common to use nearest available neighbour matching or caliper matching. Then the matching will not be perfect and (3) will no longer be a partial likelihood. But estimation for nested case–control data may still be performed using (3), and if the matching is close the results will not differ much from those presented in Sect. 6.

In some situations with nested case–control data, one may want to study the effect of a time-varying covariate, like the cumulative dose of a potential carcinogen. The standard analysis of nested case–control data is valid for time-dependent covariates, and so is the IPW analysis if one may obtain the full trajectories of time-dependent covariates for cases and controls. However, multiple imputation and the full maximum likelihood approach have not been extended to accommodate time-varying covariates in this context.

Acknowledgments Most of this research was done when Ørnulf Borgan was visiting the Department of Medical Statistics at London School of Hygiene and Tropical Medicine the spring of 2014. The department is acknowledged for its hospitality and for providing the best working facilities. We also want to thank Nathalie Støer for letting us use her new R package `multipleNCC` before it was made publicly available.

References

- Aalen OO, Borgan Ø, Gjessing HK (2008) *Survival and event history analysis: a process point of view*. Springer, New York
- Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. *Ann Stat* 10:1100–1120
- Bartlett JW, Seaman SR, White IR, Carpenter JR (2014) Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*. doi:[10.1177/0962280214521348](https://doi.org/10.1177/0962280214521348)
- Borgan Ø, Samuelsen SO (2013) Nested case–control and case–cohort studies. In: Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH (eds) *Handbook of survival analysis*. Chapman and Hall/CRC Press, Boca Raton, Florida, pp 343–367
- Borgan Ø, Goldstein L, Langholz B (1995) Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann Stat* 23:1749–1778
- Breslow NE (1996) *Statistics in epidemiology: the case–control study*. *J American Stat Assoc* 91:14–28
- Carpenter JR, Kenward MG (2013) *Multiple imputation and its application*. Wiley, New York
- Chen K (2001) Generalized case–cohort estimation. *J R Stat Soc Ser B* 63:791–809
- Kalbfleisch JD, Prentice RL (2002) *The statistical analysis of failure time data*, 2nd edn. Wiley, Hoboken
- Keogh RH, Cox DR (2014) *Case–control studies*. Cambridge University Press, Cambridge
- Keogh RH, White IR (2013) Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. *Stat Med* 32:4021–4043
- Langholz B, Borgan Ø (1995) Counter-matching: a stratified nested case–control sampling method. *Biometrika* 82:69–79
- Meng X (1994) Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 9:538–558
- Oakes D (1981) Survival times: aspects of partial likelihood (with discussion). *Int Stat Rev* 49:235–264
- Rubin DB (1987) *Multiple imputation for nonresponse in surveys*. Wiley, New York
- Rundle AG, Vineis P, Ahsan H (2005) Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiol Biomark Prev* 14:1899–1907
- 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 SO (1997) A pseudolikelihood approach to analysis of nested case–control studies. *Biometrika* 84:379–394

- Samuelsen SO, Ånestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. *Scand J Stat* 34:103–119
- Scheike TH, Juul A (2004) Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics* 5:193–206
- Scott AJ, Wild CJ (1986) Logistic models under case-control or choice based sampling. *J R Stat Soc Ser B* 48:170–182
- Scott AJ, Wild CJ (2002) Logistic models under case-control or choice based sampling. *J R Stat Soc Ser B* 64:207–219
- Støer NC, Samuelsen SO (2012) Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Anal* 18:261–283
- Støer NC, Samuelsen SO (2013) Inverse probability weighting in nested case-control studies with additional matching—a simulation study. *Stat Med* 32:5328–5339
- Støer NC, Samuelsen SO (2014) multipleNCC: weighted Cox-regression for nested case-control data. <http://CRAN.R-project.org/package=multipleNCC>, R package version 1.0
- Van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 16:219–242
- Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. *J Stat Softw* 45:1–67
- White IR, Royston P (2009) Imputing missing covariate values for the Cox model. *Stat Med* 28:1982–1998
- White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 30:377–399