Statistics
in Medicine

Research Article

Received 17 June 2015, Accepted 16 May 2016 Published online 16 June 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7012

Martingale residual-based method to control for confounders measured only in a validation sample in time-to-event analysis

Rebecca M. Burne and Michal Abrahamowicz*†

Unmeasured confounding remains an important problem in observational studies, including pharmacoepidemiological studies of large administrative databases. Several recently developed methods utilize smaller validation samples, with information on additional confounders, to control for confounders unmeasured in the main, larger database. However, up-to-date applications of these methods to survival analyses seem to be limited to propensity score calibration, which relies on a strong surrogacy assumption. We propose a new method, specifically designed for time-to-event analyses, which uses martingale residuals, in addition to measured covariates, to enhance imputation of the unmeasured confounders in the main database. The method is applicable for analyses with both time-invariant data and time-varying exposure/confounders. In simulations, our method consistently eliminated bias because of unmeasured confounding, regardless of surrogacy violation and other relevant design parameters, and almost always yielded lower mean squared errors than other methods applicable for survival analyses, outperforming propensity score calibration in several scenarios. We apply the method to a real-life pharmacoepidemiological database study of the association between glucocorticoid therapy and risk of type II diabetes mellitus in patients with rheumatoid arthritis, with additional potential confounders available in an external validation sample. Compared with conventional analyses, which adjust only for confounders measured in the main database, our estimates suggest a considerably weaker association.

Keywords: unmeasured confounding bias; imputation; Cox proportional hazards model; time-varying covariates; observational studies

1. Introduction

In observational studies, and particularly in post-marketing studies of drug safety, often the best source of sufficiently large data is found in administrative databases [1, 2]. These large databases, although they have sufficient power to detect clinically important but rare adverse events, often have limited or no information on some important confounders, for example, lifestyle characteristics such as body mass index (BMI) and smoking, measures of disease severity, and laboratory tests [3]. Lack of information on such confounders can lead to important biases, and may suggest associations where there are none [4]. Thus, development of new statistical methods, which can eliminate or reduce biases because of unmeasured confounding that is imperative [5].

One possible solution to this problem may involve the use of smaller datasets with more detailed confounder information, which we will refer to as validation samples. Such a dataset could be a subset of the large database selected for further measurements, or could be external to the large database, arising from a different source population with similar clinical characteristics. In the context of pharmacoepidemiological studies, the validation sample may represent a relatively small subset of subjects enrolled in a clinical research study and individually evaluated by study investigators, or a clinical dataset or registry

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 1A1, Canada
**Correspondence to: Michal Abrahamowicz, Department of Epidemiology, Biostatistics and Occupational Health, McGill University Health Centre, 687 Pine Avenue West, V building, Montreal, Quebec H3A 1A1, Canada.*
†E-mail: Michal.Abrahamowicz@mcgill.ca

collected externally in order to monitor a particular disease or drug [6–8]. Recently developed methods that use the validation sample to adjust for confounders unmeasured in large main databases include propensity score calibration [7], BayesPS [9], and two-stage calibration [10].

Propensity score calibration (PSC) accounts for unmeasured confounding by adapting regression calibration to propensity score analysis [7]. The method involves calculation of (i) the 'gold standard' propensity score, which adjusts for all confounders available within the validation sample and (ii) an 'error-prone' propensity score, which adjusts for only those confounders measured in the main database. It uses the relationship between these two scores to impute a corrected propensity score value in the main database that is then used as an adjustment or matching variable in the final analysis of the drug-outcome association. The performance of PSC, however, depends strongly on an assumption of surrogacy, which requires that the error-prone propensity score is conditionally independent of the outcome given in the gold standard propensity score [7]. It has been shown that PSC eliminates unmeasured confounding bias if surrogacy is met, but can perform quite poorly if this assumption is violated [11].

BayesPS addresses the same problem using a Bayesian paradigm, together with propensity scores [9]. In contrast to PSC, BayesPS does not depend on the assumption of surrogacy and performs much better than PSC in simulations where surrogacy is violated [9]. However, it was developed for a binary outcome and has not been extended to time-to-event analysis [9]. Thus, this method cannot be directly applied in many pharmacoepidemiological studies, which typically rely on a prospective or retrospective cohort design for which time-to-event analysis is most appropriate.

Two-stage calibration is a method developed recently by Lin and Chen (2014) [10]. It uses propensity scores as confounder summaries in both validation and main datasets, and then uses the approach of Chen and Chen (2000) to correct the regression estimate for the exposure obtained from the validation sample [12].

Use of validation samples may help in addressing the problem of unmeasured confounding in many observational database studies of drug safety or effectiveness. However, such applications require the development and validation of methods that allow for an accurate use of validation samples in time-to-event analyses of cohort studies. To address this challenge, we propose and validate a new method of adjusting for unmeasured confounders in time-to-event analysis, using additional confounder information available in a validation sample. To the best of our knowledge, no such methods, which attempt to adjust for unmeasured confounding using validation samples, have previously been suggested specifically for survival analysis. It should be noticed that while the problem of unmeasured confounders with a validation sample appears similar to the problem of missing data, in administrative databases the confounders of interest are not recorded for any subjects, so that missingness depends only on the mechanism which determines who is included in the validation sample. Thus, in our context, the pertinent question is how the characteristics of the subjects included in the validation sample relate to the full cohort of the eligible study participants included in the large main database. Similar to existing methods for using validation sample data, summarized earlier, we initially develop our method assuming that the validation sample is a random sample of the main database. In the terminology used in the missing data literature, this would correspond to the missing completely at random (MCAR) assumption [13]. Then, in sensitivity analyses, we consider cases where inclusion in the validation sample does depend on values of (i) the measured confounders (missing at random (MAR)) or (ii) the unmeasured confounders (missing not at random (MNAR)). We also discuss the case where the validation sample is external that arises from a different population, possibly with a different event rate.

Section 2 describes the motivation for the method and its implementation. In Section 3, we present a simulation study to assess the performance of the proposed method in several scenarios and compare it with alternative methods applicable in survival analysis. Section 4 illustrates an application of our method to account for unmeasured confounding of the relationship between glucocorticoid therapy and type II diabetes mellitus (DM) in a large cohort of rheumatoid arthritis (RA) patients. The paper is concluded with a discussion.

2. Methods: martingale residual-based imputation

2.1. Motivation

We propose a method for imputing confounders unmeasured in the main database in time-to-event analyses using a validation sample with complete confounder information. Our method involves the use of martingale residuals in imputation models for the unmeasured confounders. The martingale residual was

originally proposed in time-to-event analysis for the purpose of model assumption checking [14], and can be thought of as a measure of the 'excess' events observed over what is predicted by the fitted model. Following Barlow and Prentice (1988) [14] and Therneau *et al.* (1990) [15], the martingale residual for individual i at time t is defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\left\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(s)\right\} d\hat{\Lambda}_0(s), \tag{1}$$

where $Y_i(s)$ is an indicator function denoting whether individual i is at risk at time $s \leqslant t$, $N_i(t)$ is a counting process for the observed events by individual i at time t ($N_i(t) = 0$ until individual i has an event, after which time $N_i(t) = 1$), $\hat{\Lambda}_0(s)$ is the Breslow estimate of the cumulative hazard [16], and $\mathbf{Z}_i(s)$ is the vector of possibly time-dependent covariates for individual i at time s.

The rationale behind using the martingale residual to impute unmeasured confounders has two components. Firstly, including information on the outcome in the imputation model, in addition to data on measured covariates or predictors, yields less biased final regression coefficients [13, 17]. However, in the setting of survival analysis, the event times are not directly observed for some subjects, and thus, outcomes are comprised of both (i) observed time and (ii) censoring status. The martingale residual in (1) combines information on both components of the outcome, and, in addition, accounts for the effects of all measured covariates including the exposure of primary interest. Thus, inclusion of the martingale residual in the imputation model accounts for the individual subjects' outcomes and may improve the imputation. Secondly, the martingale residual may partly explain those outcomes that appear unlikely given the observed covariates, and thus may indicate presence of unmeasured confounding. Consider, for example, an important unmeasured continuous risk factor U, with hazard increasing with increasing U. Subjects who had an early event may be expected to have, on average, a high value of U. The martingale residual quantifies the 'excess' events experienced by an individual at time t, and is bounded by $-\infty$ and 1. A value close to 1 indicates that an individual experienced an event when their predicted risk from the model that accounted only for measured covariates was low. Thus, this individual has 'excess risk' that is not well explained by the measured covariates and could possibly be explained by an unobserved high value of U. In contrast, a highly negative value of martingale residual would indicate that the model predicts a high risk, while the individual has not yet had an event. Because it would appear that the model predicts a higher risk than the individual's true risk, it is probable that the value for U is low. In the Appendix, we provide a more formal rationale for the use of the martingale residuals in the imputation model.

2.2. *Implementation*

2.2.1. Non-time-varying data and internal validation sample. In this section, we present the data set-up and implementation in the case of an internal validation sample and analyses restricted to time-invariant variables. Denote the full data (main database which includes the internal validation sample) as $\{t_i, \delta_i, X_i, \mathbf{C}_i, \mathbf{U}_i\}$, for $i = 1, \ldots, N$. Let (t_i, δ_i) denote the time and censoring indicator for each individual, X_i their (binary) exposure, $\mathbf{C}_i = (C_{1i}, \ldots, C_{pi})$ the vector of fully observed confounders, and $\mathbf{U}_i = (U_{1i}, \ldots, U_{ki})$ the confounders measured only in the validation sample (which we will refer to as 'unmeasured'). Suppose that for those individuals in the main database with subscripts $i = 1, \ldots, n$, the confounders $U_{1i}, \ldots, U_{ki}$ are missing. For the m individuals in the validation sample (with subscripts $i = n+1, \ldots, N = n+m$) these $U_{1i}, \ldots, U_{ki}$ are measured, as well as $\mathbf{C}_i$.

The implementation of the proposed method includes the following four steps:

Step 1: Fit a Cox proportional hazards (PH) model to the full data, using data on only exposure X and measured confounders $\mathbf{C}$:

$$\lambda(t) = \lambda_0(t) \exp\left\{\gamma_1 X + \boldsymbol{\gamma}_2' \mathbf{C}\right\}. \tag{2}$$

From this fitted model, obtain the martingale residual $\hat{M}_i$ for $i = 1, \ldots, N$.

Step 2: The martingale residual is then used in the imputation models for the unmeasured confounders. In the validation sample, where information on the unmeasured confounders $U_{1i}, \ldots, U_{ki}$ is available, fit the separate 'training' models for the imputation step for each $U_1, \ldots, U_k$.

2.1: If U_j is continuous, in the validation sample fit a multivariable linear regression model dependent on exposure, the martingale residual, and all measured confounders

$$E\left[g\left(U_j\right)\right] = \alpha_0 + \alpha_1 X + \alpha_2 \hat{M} + \boldsymbol{\alpha}_3' \mathbf{C}, \tag{3}$$

where $g(.)$ is the identity function if U_j has an approximately Normal distribution (determined by, for example, $p \geqslant 0.05$ for a Wilks–Shapiro test), and some reasonable transformation such as the log function otherwise.

2.2: If U_j is binary, in the validation sample fit a multivariable logistic regression model:

$$\text{logit}\left[P\left(U_j = 1\right)\right] = \alpha_0 + \alpha_1 X + \alpha_2 \hat{M} + \boldsymbol{\alpha}_3' \mathbf{C}. \tag{4}$$

Step 3: Then, based on the models estimated in Step 2, impute the missing confounder in the main database:

3.1: If U_j is continuous, impute $g(U_{ij})$ for each $i = 1, \ldots, n$ from a normal distribution with mean $E\left[g\left(U_{ij}\right)\right] = \left(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\boldsymbol{\alpha}}_3\right) \mathbf{Z}_i'$ obtained from the fitted model in (3), and variance $V_i = \hat{\sigma}^2 \mathbf{z_i} (Z'Z)^{-1} \mathbf{z_i'}$, the prediction variance, where $\mathbf{z}_i = \left(1, X_i, \mathbf{C}_i, \hat{M}_i\right)$ is individual i's vector of covariates, and $\mathbf{Z} = \left(\mathbf{z}_{n+1}', \ldots \mathbf{z}_N'\right)'$ is the design matrix of covariates from individuals in the validation sample. In this way, uncertainty in the estimates from model (3), which depends on the size of the validation sample, is taken into account. Then, if $g(U_{ij}) \neq U_{ij}$, back-transform using $g^{-1}(\cdot)$ to obtain imputed U_j on the original scale.

3.2: For binary U_j, impute U_{ij} for $i = 1, \ldots, n$ from a Bernoulli distribution with $p_i = P(U_{ij} = 1) = \text{expit}\left\{\hat{\alpha}_0 + \hat{\alpha}_1 X_i + \hat{\alpha}_2 \hat{M}_i + \hat{\boldsymbol{\alpha}}_3' \mathbf{C}_i\right\}$ from (4).

It is possible to multiply impute each U_j from fitted model (3) or (4) in order to incorporate uncertainty in the imputation. We examine the influence of multiple imputation on coverage rates of confidence intervals (CIs) in simulations (Section 3.1).

Step 4: Once these confounders have been imputed for $i = 1, \ldots, n$ in the main database, fit the final Cox PH model to all subjects $i = 1, \ldots, N$

$$\lambda(t) = \lambda_0(t) \exp\left\{\beta X + \boldsymbol{\gamma}_1' \mathbf{C} + \boldsymbol{\gamma}_2' \mathbf{U}\right\}, \tag{5}$$

to estimate the exposure effect β, adjusted for both measured confounders and imputed values of unmeasured confounders.

2.2.2. Time-varying data and external validation samples. Implementation of our method, described earlier for time-invariant data, extends easily to analyses where both confounders and exposure are time-varying, but such an extension requires some key assumptions about the data generation process. Here, we assume that time-varying confounders are not affected by the past exposure, or unmeasured determinant(s) of exposure, and thus, do not act as time-varying mediators of the exposure effect on the outcome [18]. Under this assumption, the treatment effect could be accurately estimated, if all confounders were fully measured, by fitting the 'true' model that simply adjusts for all relevant confounders

$$\lambda(t) = \lambda_0(t) \exp\left\{\beta X(t) + \boldsymbol{\gamma}_1' \mathbf{C}(t) + \boldsymbol{\gamma}_2' \mathbf{U}(t)\right\}. \tag{6}$$

We also assume that only current values of the confounders, observed at time t, affect (i) the current hazard at t, i.e. that the event of interest is acute in nature, and (ii) current exposure (treatment assignment). The aforementioned assumptions about the data-generating processed are summarized in the directed acyclic graph in Figure 1.

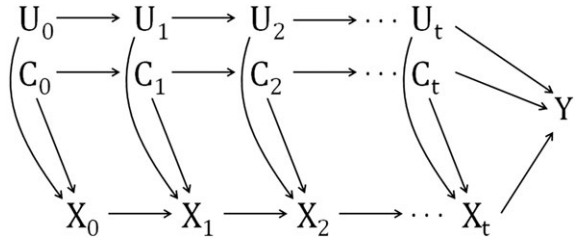

Figure 1. Directed acyclic graph (DAG) showing the assumed data-generating process for time-varying data.

The implementation in the time-varying case proceeds along similar steps to the time-invariant case. First, the martingale residual $M_i(t)$ is estimated for each follow-up time t for each individual from the Cox PH model:

$$\lambda(t) = \lambda_0(t) \exp\left\{\gamma_1 X(t) + \boldsymbol{\gamma}_2' \mathbf{C}(t)\right\}. \tag{7}$$

Then, each time-dependent confounder $U_j(t)$, $j = 1, \ldots, k$, is imputed using time-varying analogs to Equation (3) or (4) for, respectively, continuous and binary confounders: $E\left[g\left(U_j(t)\right)\right] = \alpha_0 + \alpha_1 X(t) + \alpha_2 \hat{M}(t) + \boldsymbol{\alpha}_3' \mathbf{C}(t)$, or logit $\left[P\left(U_j(t) = 1\right)\right] = \alpha_0 + \alpha_1 X(t) + \alpha_2 \hat{M}(t) + \boldsymbol{\alpha}_3' \mathbf{C}(t)$. Note that we are implicitly making the assumption that only current values, observed at time t, of the exposure, measured confounders, and martingale residual are informative about the concurrent value of the unmeasured confounder $U_j(t)$.

While above we assume that the validation sample is considered a random sample of the main database, we also consider cases where the validation sample is external to the main database. One slight modification is necessary for applications where the validation sample is external, that is, obtained from a separate database rather than being a random sample from the main database. In this case, in Step 1, the martingale residuals are obtained separately for the validation sample and the main database, with identical models (Equation (2) or (7)) estimated, independently of each other, in each dataset. The training and imputation steps (Steps 2 and 3) proceed as described in the previous section, but the final model (Step 4) is run in the main database only (excluding the validation sample).

3. Simulations

To evaluate our method's performance, we simulated a hypothetical prospective study designed to assess the risk of an adverse event due to exposure to a drug, assuming some confounders unmeasured in the main database are available in a validation sample. We first assessed the method, and compared it with existing methods applicable for time-to-event analyses, in the case where both the exposure and covariates were time-fixed. In this simulation study, we varied several parameters: the true effect of drug exposure; the strength and direction of unmeasured confounding, including variation of whether and to what extent the surrogacy assumption was violated; the censoring mechanism; the size of the validation sample; the mechanism of inclusion into the validation sample; and the event rate in the validation sample (in the case of an external validation sample). Then, we further assessed the method assuming both exposure and confounders were time-varying, and an external validation sample, while altering the strength and direction of unmeasured confounding across the simulated scenarios. All simulations were run using R software [19]. Simulation code is available on GitHub (http://github.com/RMBurne/MR-based-imputation).

3.1. Time-fixed covariates

3.1.1. Data generation. In simulations, we assumed two continuous measured confounders $\mathbf{C_i} = \left(C_{1i}, C_{2i}\right)$ and two unmeasured confounders $\mathbf{U_i} = \left(U_{1i}, U_{2i}\right)$, where U_{1i} was continuous and U_{2i} binary. For most simulations (except select scenarios described later) the four confounders were assumed to be independent of each other, and the validation sample and main database were simulated identically, thus assuming that the validation sample is a random subset of the main database.

For $i = 1, \ldots, N = 10,000$, we generated $\{C_{1i}, C_{2i}, U_{1i}, U_{2i}\}$ where each of $C_{1i}, C_{2i}, U_{1i} \sim \mathcal{N}(0, 1)$, and $U_{2i} \sim \text{Bern}(p)$. Binary exposure X_i was generated for each individual from a Bernoulli distribution with probability conditional on both observed and unobserved confounders, assuming a logistic model

$$P(X_i = 1) = \text{expit}\left\{\xi_0 + \boldsymbol{\xi}_1' \mathbf{C}_i + \boldsymbol{\xi}_2' \mathbf{U}_i\right\}. \tag{8}$$

Using the method of inversion [20], the true event time for each individual, t_i, was generated conditional on simulated exposure and confounders from the *a priori* specified Cox PH model (Equation (5)), assuming an underlying exponential event rate with a constant hazard.

Because the value of the martingale residual depends in part on the censoring mechanism, both administrative and random censoring were presumed in different simulation scenarios. In the basic scenario, administrative censoring was implemented such that 10% of subjects were observed to have the (uncensored) event. For these scenarios with only administrative censoring at the end of the study,

all subjects whose event times fell after the 10th percentile were censored at that time. For scenarios where a combination of random and administrative censoring was assumed, first censoring times were generated from an exponential distribution, and then an administrative cut-off was taken as the 10th percentile of the observed event times. Further scenarios were assumed with only random censoring. In the first, both the underlying hazard and censoring mechanism were assumed to be exponentially distributed; and in the second, they were assumed to be Weibull, in such a way that censoring was around 40%.

The main simulations, described earlier, rely on the assumption that the validation sample is a random sample of the main database, which corresponds to the assumption of MCAR in the missing data literature [13]. Whereas similar assumptions underlie PSC and other methods relevant for our setting, insofar as the validation sample is required to be comparable with [7] or 'exchangeable' with [9] the main sample, in some real-life applications this may not be the case. Therefore, in sensitivity analyses, we assessed the performance of the methods in two more complex situations where inclusion into the validation sample may depend on the values of, respectively, (i) covariates measured in the main database (**C**) and (ii) values of unmeasured confounders (**U**). For case (i), we assumed that probability of inclusion into the validation sample increased with increasing values of the continuous variable C_1, so that the distribution of this confounder was shifted toward higher values in the validation sample compared with the main database. We also assumed that there was some correlation between C_1 and one unmeasured confounder, U_2, as we believe that this is a plausible scenario, and it may make it more difficult to obtain unbiased estimates. For case (ii), we assumed that the probability of inclusion into the validation sample increased with increasing values of the continuous unmeasured confounder U_1. Finally, we have added a further simulation scenario, in which the event rate in the validation sample is twice that of the main database, which may occur if subjects for whom additional confounders are measured were included in a hospital-based clinical study and, thus, represent a high-risk subpopulation.

Once this full data was generated, a random sample of size m was taken for the validation sample, and then $\{U_{1i}, U_{2i}\}$ were deleted for the remaining $i = 1, \ldots, n = N - m$. With this data set-up, the imputation and analysis proceeded as described in Section 2.2.

Details and values of all parameters used in different scenarios can be found in Appendix A of the Supporting Information.

3.1.2. Analysis of simulated datasets. For each simulation scenario, we generated and analyzed 1000 independent datasets, and compared our method with three alternative methods. Firstly, we fit the conventional Cox PH model, which we will refer to as 'standard' adjustment, which adjusted only for the confounders measured in the main database. As the second method, we implemented PSC as described by Stürmer *et al.* (2005) for Cox regression analysis [7]. The aforementioned methods were also compared with a method which incorporated outcome information in the imputation model through inclusion of two additional terms: the log survival time and the event indicator. A similar method has previously been used in studies comparing multiple imputation methods for survival data [21, 22]. Comparison with this method enables us to examine the potential advantages of representing the information on the individual subjects' outcomes by a single value given by the martingale residual rather than by two separate variables for (i) the censoring indicator and (ii) the follow-up time. The results of our martingale residual-based imputation and the three alternative methods were compared with respect to bias, standard deviation (SD), and root mean square error (RMSE) of the estimated effect (adjusted log hazard ratio) of the exposure.

3.1.3. Coverage rate of the martingale residual-based imputation estimate. Because the SD of the estimate in our model (5) does not take into account the variation arising from the estimation of the martingale residual and imputation of the unmeasured confounders, the naïve CI based on the covariance matrix of model (5) is likely to underestimate the true sampling variance. We therefore investigated if adequate coverage can be achieved by either of two alternative approaches: (i) multiple martingale residual-based imputation or (ii) bootstrap.

First, to assess whether multiple imputation of the confounders using martingale residual-based imputation would improve the accuracy of the variance estimation and coverage, we repeated the imputation and further steps 10 times for each simulated sample. Standard errors and CIs were obtained using standard multiple imputation rules [23].

The second approach involved bootstrap resampling. Because of the computational burden of combining bootstrap with simulations, we performed bootstrapping for only six scenarios and limited it to 500 simulated datasets for each scenario. All four steps of martingale residual-based imputation were

replicated in each resampled dataset. Three hundred bootstrap samples were used for each scenario, and the coverage of the 95% bootstrap CI, estimated through the percentile method, was calculated [24].

3.1.4. Simulation results. The results of simulations are summarized in Table I. Our martingale residual-based method performs better in terms of bias than standard analysis, for all simulation scenarios. Our estimates are uniformly unbiased regardless of the true HR (scenarios 1–3), strength of unmeasured confounding (scenarios 4 and 5), and violation of surrogacy (scenarios 6–11). Neither reduction of the validation sample size, with as few as 50 or 25 observed events (scenarios 12–15), nor a combination of administrative and random censoring (scenario 19) appear to affect the bias of the estimates obtained with martingale residual-based imputation, although our method displays a small increase in bias relative to the baseline scenario 1 in fully random censoring scenarios (scenarios 20 and 21).

For both scenarios 16 and 17, when selection into the validation sample depends on, respectively, a measured and an unmeasured confounder, there is no notable change in performance from the baseline scenario 1 for any method (Table I). Similarly, when the validation sample was assumed to represent a high-risk subpopulation, with event rate double that of the main database, the bias of each method is not substantially different from the baseline scenario (Table I, scenario 18). Regardless of the specific reasons for the violation of the MCAR assumption, and their possible impact on the accuracy of the estimates, in all scenarios 16–18 the martingale residual-based imputation performed better than or at least as well as all other methods (Table I).

In most scenarios, martingale residual-based imputation performs almost identically to imputation including the log survival time and censoring, except in scenarios with completely random censoring and with different underlying hazard distributions, where the bias of log(t) imputation was 64-75% larger (scenarios 20 and 21). As expected, the bias of PSC estimates depends on whether and to what extent the surrogacy assumption is violated. When surrogacy holds (scenarios 1–5), PSC yields almost unbiased estimates and performs better than standard adjustment. However, when surrogacy is even moderately violated, the PSC estimates are seriously biased and may even yield more biased estimates than standard adjustment (e.g., scenarios 6, 7, 13, and 15). In contrast, the martingale residual-based imputation estimates are unbiased even when surrogacy is seriously violated. Furthermore, martingale residual-based estimates are slightly less biased than PSC estimates even when surrogacy holds (scenarios 1–5).

Table I also shows the ratios of (i) SD and (ii) RMSE, relative to martingale residual-based imputation, for the three other methods: standard adjustment, PSC, and log(t) imputation. In all scenarios, SDs of the martingale residual-based method are comparable with SDs of log(t) imputation (relative SD 0.99–1.04 for all scenarios). Similarly, for most simulation scenarios, martingale residual-based imputation SDs are similar to those for PSC. However, PSC has larger SD than martingale residual-based imputation in cases where (i) the validation sample was smaller (scenario 14, relative SD 1.42; and 15, relative SD 1.48), (ii) the validation sample was external (scenario 18; relative SD 1.24), and (iii) more random censoring occurred (scenario 20, relative SD 1.44; and 21, relative SD 1.49). Our method has larger variation than standard analysis in all scenarios, with particularly important variance inflation as the size of the validation sample decreases (scenarios 14 and 15).

Despite some variance inflation, the proposed martingale residual-based imputation method offers better bias/variance trade-off than PSC and standard analysis. Indeed, for almost all scenarios, our method yields lowest RMSE compared with these two methods, often lower by more than 50% (e.g., scenarios 4, 6, 7, and 11). In a single scenario, scenario 8, both PSC and standard analysis have slightly lower RMSE than our method, by <10%. This occurs mostly because of a combination of (i) inflation of the SD of our estimates, discussed earlier, and (ii) very weak unmeasured confounding, which implies minimal bias of the estimates obtained with all methods (Table I). For the same reasons, in scenarios with only weak to moderate unmeasured confounding, our method does not work notably better than either standard adjustment or PSC (scenarios 5 and 9). In comparison with log(t) imputation, our method performs similarly in terms of RMSE (relative RMSE 0.99–1.05), except in the case of random censoring, where log(t) imputation has 12–13% higher RMSE because of its higher bias (scenarios 20 and 21).

Coverage rates of the bootstrap CIs for the martingale residual-based estimates are presented in Table II. Because of the computational intensity of bootstrap for all simulation replications, only six simulation scenarios are presented. As expected, the coverage of the conventional covariance matrix-based 95% CI for single martingale residual-based imputation is consistently low for all scenarios, despite of the absence of bias. Multiple martingale residual-based imputation with 10 imputations improves the coverage in all scenarios only slightly and is not sufficient to increase coverage to the nominal 95% level. In contrast, the bootstrap, with CI based on the percentiles of the distribution of the estimates across the 300

Table I. Time-invariant simulation results: bias, relative SD, and relative RMSE (relative to martingale residual-based imputation) for the log hazard ratio for exposure for different methods: (i) standard analysis (adjusting only for fully measured confounders), (ii) propensity score calibration (PSC), (iii) imputation with log survival time and censoring indicator (log(t) imputation), and (iv). martingale residual-based imputation (MR-based imputation).

				Bias				Relative SD			Relative RMSE		
	Parameter changed			Standard analysis	PSC	Log(t) imputation	MR-based imputation	Standard analysis	PSC	Log(t) imputation	Standard analysis	PSC	Log(t) imputation
1				0.164	−0.032	0.006	0.006	0.846	1.089	1.011	2.109	1.152	1.011
	True HR												
2	1.2			0.162	−0.032	0.004	0.004	0.843	1.054	1.002	2.086	1.118	1.002
3	1.5			0.157	−0.040	−0.001	0.000	0.842	1.088	0.990	2.092	1.193	0.990
	Strength of unmeasured confounding												
4	Strong			0.363	−0.042	0.018	0.019	0.760	1.085	1.001	4.039	1.156	0.998
5	Weak			0.047	−0.022	0.002	0.002	0.956	1.061	0.998	1.131	1.100	0.998
	Surrogacy violated (direction of confounding)												
	U1	*U2*	*C1*										
6	$\downarrow^+$	$\downarrow^+$		−0.186	−0.308	−0.006	−0.006	0.890	0.941	0.986	1.922	2.984	0.986
7	$\downarrow$	$\downarrow$		−0.045	−0.091	0.000	0.000	0.969	1.030	1.000	1.099	1.473	1.000
8		$\downarrow^+$	$\downarrow^+$	−0.040	−0.024	0.002	0.002	0.879	0.877	1.005	0.957	0.907	1.005
9		$\downarrow$	$\downarrow$	0.035	−0.034	0.007	0.006	0.910	0.955	1.003	0.982	1.019	1.004
10			$\downarrow^+$	0.166	0.137	0.004	0.005	0.880	0.905	1.001	2.180	1.883	1.001
11			$\downarrow$	0.161	0.018	−0.001	−0.001	0.842	0.993	0.990	2.101	1.015	0.990
	Size of validation sample												
12	500 (surr ok)			0.165	−0.053	0.008	0.008	0.721	1.173	1.001	1.822	1.286	1.001
13	500 (surr viol.)			−0.182	−0.360	−0.001	−0.002	0.696	1.008	0.994	1.569	2.950	0.994
14	250 (surr ok)			0.164	−0.080	0.007	0.007	0.539	1.421	0.997	1.405	1.556	0.998
15	250 (surr viol.)			−0.187	−0.405	−0.007	−0.006	0.583	1.477	1.016	1.311	2.943	1.016

Table I. Continued.

		Bias				Relative SD			Relative RMSE		
	Parameter changed	Standard analysis	PSC	Log(t) imputation	MR-based imputation	Standard analysis	PSC	Log(t) imputation	Standard analysis	PSC	Log(t) imputation
	Selection into validation sample										
16	Depends on C	0.161	−0.004	0.003	0.003	0.835	1.008	1.015	1.995	1.009	1.015
17	Depends on U	0.163	−0.007	0.012	0.011	0.855	1.038	1.000	2.128	1.032	1.002
	External validation sample										
18	Event rate in VS	0.165	−0.027	−0.009	0.005	0.886	1.236	1.042	2.721	1.302	1.047
	Underlying hazard, censoring										
19	Random & admin censoring	0.164	−0.031	0.004	0.004	0.854	1.070	0.991	2.162	1.134	0.990
20	Random censoring	0.165	−0.007	0.036	0.022	0.737	1.439	0.925	3.628	1.276	1.130
21	Weibull hazard, random censoring	0.164	−0.008	0.035	0.020	0.725	1.486	0.921	3.586	1.350	1.118

↓ / ↓$^{+}$: moderate / strong confounding of true effect in negative direction (in baseline scenario, all confounding effects are positive.)

Table II. Coverage results of the 95% confidence interval for selected scenarios, from our method (MR-based imputation), multiple imputation using our method with 10 imputations (multiple MR-based imputation), and using bootstrap.

	MR-based imputation		Multiple MR-based imputation		Bootstrap	
	Bias	Cover (%)	Bias	Cover (%)	Bias	Cover (%)
1	0.006	89.1	0.006	90.2	0.006	94.6
2	0.005	87.5	0.005	88.8	0.008	94.6
3	0.001	88.7	0.000	90.1	0.006	95.6
4	0.019	85.1	0.019	86.6	0.018	94.8
5	0.002	94.2	0.002	94.1	0.002	96.4
6	0.004	90.0	0.004	90.7	0.002	95.0

resamples, produces coverage rates uniformly very close to the nominal level (94–96%). These results suggest that bootstrap resampling, with all steps of the procedure described in Section 2.2 replicated in each resample, is necessary in order to obtain accurate CIs for our estimates.

3.2. Extensions: time-dependent confounders and external validation sample

3.2.1. Assumptions and rationale. Having assessed the performance of martingale residual-based imputation in time-invariant simulations with an internal validation sample, we now extend and assess its performance in simulations with the data structure similar to our application (described in Section 4), where (i) both exposure and confounders are time-varying and (ii) the validation sample is an external sample, drawn from a different source population, (e.g., from a different geographical location). Similar to time-invariant simulations, we consider a hypothetical study of an association between drug use and the risk of an adverse event, and generate two measured and two unmeasured confounders, but with both confounders and the binary exposure (current use of a drug) being time-varying. For these simulations and the following application, we make two additional assumptions. Firstly, we assume that the time-varying confounders do not mediate the effects of the previous treatment, so that conventional adjustment is appropriate and there is no need to use (for example) marginal structural models. Secondly, we assume that both (i) the clinical characteristics of the subjects in the validation sample, and (ii) the relationships between exposure, unmeasured confounders, and event are similar to those in the main database (same disease, similar reasons for treatment). However, we do assume that the underlying event rate in the validation sample may be different from that in the main database. For example, subjects for whom additional confounders are measured may represent a higher risk subpopulation enrolled in a clinical, hospital-based research study. Details of the data generation process are given in the following subsection, and in Appendix B of the Supporting Information.

Although PSC has not previously been assessed in time-varying data, use of the time-varying propensity score in longitudinal data settings has been shown to effectively control for measured time-varying confounders, which are not affected by prior treatment [25, 26]. In particular, Ray *et al.* (2015) showed that, if a model which adjusts for time-varying treatment and covariates is appropriate (i.e., there is no mediation), then a model which conditions on a time-varying propensity score can also be expected to yield unbiased conditional estimates. Thus, we also compare our method with a time-varying extension of PSC [26].

3.2.2. Data generation. For these simulations, we generated time-varying data for exposure and confounders, assuming their values may be updated at up to 10 equally spaced time intervals during the follow-up period. As in the time-invariant simulations in Section 3.1, for both the main database ($n = 10,000$) and, separately, the external validation sample ($m = 1000$), we first generated two measured $\{C_1(t), C_2(t)\}$ and two unmeasured confounders $\{U_1(t), U_2(t)\}$.

For each subject, the 10 subsequent values of each continuous confounder (C_1, C_2, U_1) were simulated from a subject-specific multivariate normal distribution with autoregressive correlation structure. The 10 individual values of the unmeasured binary confounder, U_2, were generated from a binomial distribution, with subject-specific probability of $U_2 = 1$, randomly selected from a uniform distribution and assumed to remain constant over time.

We then generated exposure based on the confounders. Time-dependent exposure was allowed to change only at randomly selected intervals, at which time it was regenerated conditional on the current values of measured and unmeasured confounders.

Finally, we generated the event times, conditional on exposure and confounders. The marginal distribution of event times was taken to be exponential, with a higher event rate in the validation sample than in the main database. Once the marginal times were generated, the permutational algorithm, specifically developed and validated for simulating event times conditional on time-varying covariates [27], was used to match the current values of exposure and all four confounders to events at the corresponding times. To do this, we used the PermAlgo package in R [28]. This matching is performed so as to generate data consistent with the parameters of a multivariable proportional hazards model (Equation (7)). Administrative censoring, due to the hypothetical end of the study at time $t = 10$, was set.

In simulation, we considered scenarios in which, respectively: surrogacy is preserved; surrogacy is violated with strong and with moderate unmeasured confounding; the true hazard ratio is changed; and informative censoring occurs through an entirely unmeasured risk predictor. Further details and parameter values for each of the scenarios can be found in Appendix B of the Supporting Information.

3.2.3. Simulation results. The results of simulations are presented in Table III. Martingale residual-based imputation yields unbiased estimates for all scenarios, in contrast to seriously biased standard estimates. Accordingly, similar to time-invariant simulations, martingale residual-based imputation substantially reduces the RMSE relative to both standard analysis and PSC, especially when the strength of unmeasured confounding is increased (last two columns of Table III). Although informative censoring does not affect the performance of any of the methods (Table III, scenarios 5 and 6), we investigated only a relatively simple informative drop-out mechanism. Future work could explore more elaborate non-random censoring mechanisms, such as those in which censoring is simultaneously affected by treatment and unmeasured covariates, or different reasons for censoring may be differently affected by the treatment and/or confounders. Because impact of informative censoring may depend both on how the exposure of interest is modeled and on the underlying mechanism [29], it is possible that relative performance of different methods to deal with confounders measured only in the validation sample may also depend on these factors.

Similar to time-invariant simulations, use of time-varying PSC reduces bias, relative to standard analysis, in the scenarios where surrogacy holds (Table III: scenarios 1 and 4–6), but increases bias, and therefore RMSE, where surrogacy is violated (scenarios 2 and 3). In all time-varying simulations, martingale residual-based estimates are at least as or more accurate than the corresponding estimates obtained with time-varying PSC (Table III). These results suggest that martingale residual-based imputation may eliminate or largely reduce bias in analyses with time-varying confounders and exposure, even if imputation is based on an external validation sample with different (here higher) event rates than in the main database.

Table III. Simulation results for time-dependent simulations for martingale residual-based imputation. Time-varying simulation results: bias, relative SD, and relative RMSE (relative to martingale residual-based imputation) for the log hazard ratio for exposure for different methods: (i) standard analysis (adjusting only for fully measured confounders), (ii) propensity score calibration (PSC), and (iii) martingale residual-based imputation (MR-based imputation).

				Bias			Relative SD		Relative RMSE	
	True HR	Unmeasured confounder	Surrogacy	Standard analysis	PSC	MR-based imputation	Standard analysis	PSC	Standard analysis	PSC
1	1	Moderate	Holds	0.215	0.042	0.002	0.884	1.044	2.911	1.176
2	1	Moderate	Violated	−0.234	−0.455	−0.004	0.905	1.119	2.693	5.048
3	1	Strong	Violated	−0.414	−0.727	−0.006	0.802	1.092	4.389	7.649
4	1.5	Moderate	Holds	0.212	0.049	0.002	0.887	1.070	2.943	1.249
	Informative censoring									
5	1	Moderate	Holds	0.210	0.034	−0.005	0.882	1.083	2.820	1.166
6	1.5	Moderate	Holds	0.207	0.043	−0.001	0.887	1.052	2.771	1.187

4. Use of glucocorticoids and type II diabetes mellitus

To illustrate the implementation of the proposed method, we apply it to reassess the potential increase in risk of type II diabetes mellitus (DM) in patients being treated with glucocorticoid (GC) therapy for rheumatoid arthritis (RA) in the UK-based Clinical Practice Research Datalink (CPRD) [30]. GC therapy is a very common treatment for RA, and has been shown to be very effective in slowing disease progression [31, 32]. However, its use has been associated with a range of potentially serious adverse effects [33]. Type II DM is considered a possible important side-effect of GC therapy [34, 35], but the published results are ambiguous. Randomized controlled trials of GC therapy have not found an increase in the risk of diabetes, even in meta-regression [36]. However, randomized controlled trials are often not large or long-term enough to adequately assess risk of infrequent adverse events [1]. Observational studies have reported highly varied results, with either an increased risk of diabetes associated with GC therapy [37, 38], or a lower incidence of diabetes among GC users [39].

This application involves use of an external validation sample with measurements of some additional potential confounders, not available in the CPRD. The CPRD contains electronic medical records from around 11 million patients across the UK; and for this analysis, those with RA diagnosis between 1992 and 2009 were identified by a validated algorithm [30]. Available information includes patient demographics, medical diagnoses, and drug prescriptions. However, some potential confounders of the GC–DM relationship, including BMI, disability level, and comorbidity index were not available for our CPRD analyses. On the other hand, these potential confounders are systematically measured in the National Data Bank for Rheumatic Diseases (NDB), a longitudinal observational study of patients with RA from the US [40]. Because the NDB was designed with the study of rheumatic diseases in mind, it includes more rich information on potential confounders relevant for studying the effects of RA drugs. Importantly for our application, the confounders available in CPRD were also available in NDB. For these reasons, we will use the NDB as an external validation dataset in our analysis. Details of cohort inclusion, data collection and variable definitions for each dataset are described elsewhere [30].

For the purpose of this illustrative example, in our CPRD analyses we considered, as measured confounders, only a subset of all covariates available in the database: sex, baseline age, prior nonsteroidal anti-inflammatory drug (NSAID) use (at cohort entry), and time-varying indicators of current use of two main disease modifying anti-rheumatic drugs: methotrexate and hydroxychloroquine. These confounders were found to have statistically significant associations with the hazard of DM in preliminary analyses of CPRD (data not shown). Furthermore, the CPRD analyses included only those 16,898 individuals who met the inclusion criteria [30] and had no missing data on any of the earlier-measured confounders. These subjects were followed up on average for about 6.5 years (2365 days). In the NDB analyses, we have information on 8253 individuals with a mean follow-up time of just over 4.5 years (1670 days). In addition to the measured confounders adjusted for in the CPRD analysis, we include in the full multivariable model the following potential confounders (unmeasured in CPRD), which were identified in preliminary NDB analyses as statistically significant risk factors for diabetes: Health Assessment Questionnaire (HAQ) disability score [41], comorbidity index [42], and BMI. For the sake of simplicity, we restrict our analysis to complete cases (those with complete data on all variables required in their respective database). Because both the exposure (GC) and several confounders (methotrexate and hydroxychloroquine use among the measured confounders, and all three unmeasured confounders) were time-varying (see Section 4.1 for details), we use time-dependent martingale residual-based imputation, as outlined and assessed in simulations in Section 3.2. We also restrict the analysis to only those individuals who had not used GCs in the three years prior to their start of follow-up, in order to capture as close to a treatment naïve population as possible.

In each database, patients were followed up from first date of RA diagnosis within the study window, to either first DM diagnosis or loss to follow-up. In the main databases, current exposure to GC is represented by a binary time-varying covariate. Accordingly, we divide individual patient follow-up into consecutive time intervals during which the exposure status does not change, based on the date and duration of GC prescriptions. Values of all time-varying confounders are established at the beginning of each period and assumed to be constant until the end of that period.

4.1. Results

Table IV summarizes information on outcomes, exposure, and confounders in both cohorts. The frequency of events was slightly higher in the CPRD database (9.9% vs. 6.4%), although the mean length

Table IV. Characteristics of the main database (CPRD) and validation data (NDB). Rows contain total (percentage) for binary variables, and mean of individual means (over follow-up time) for continuous variables.

	NDB (n = 8253)	CPRD (n = 16898)
Incident type II diabetes mellitus	529 (6.4%)	1665 (9.9%)
Follow-up (days): mean (SD)	1670.0 (1359.0)	2364.8 (1643.5)
Event rate (/100 person-years) (95% CI)*	1.40 (1.28, 1.53)	1.52 (1.45, 1.60)
Used glucocorticoid during follow-up: n (%)	2358 (28.6%)	5864 (34.7%)
Days exposed amongst users: median (IQR)	304.0 (92.0–732.8)	219.0 (36.0–831.2)
Incidence rate of new exposure (/100 person-years) (95% CI)*	8.34 (8.00, 8.68)	7.22 (7.04, 7.41)
Mean % time exposed since first exposure	50.75%	38.75%
Sex = male: n (%)	1554.0 (18.8%)	4953.0 (29.3%)
Baseline age: mean (SD)	58.6 (13.3)	58.1 (14.6)
History of NSAID use at cohort entry: n (%)	6126 (74.2 %)	14335 (84.8%)
Methotrexate use in follow-up: n (%)	4280 (51.9%)	3530 (20.9%)
Hydroxychloroquine use in follow-up: n (%)	2329 (28.2%)	1110 (6.6%)
Baseline HAQ disability score: mean (SD)	0.905 (0.697)	–
Baseline comorbidity index: median (IQR)	1 (0–2)	–
Baseline BMI: mean (SD)	27.89 (6.42)	–

*Poisson CI.

Table V. Results for the fully adjusted and reduced models in the NDB (validation) database (left-hand side) and CPRD (main) database (right-hand side).

	NDB				CPRD			
	Full model		Conventional model (confounded)		MR-based imputation		Conventional model (confounded)	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Glucocorticoid use	1.48	(1.17, 1.86)	1.81	(1.45, 2.27)	1.15	(0.99, 1.33)	1.38	(1.19, 1.59)
Sex (male = 1)	1.00	(0.79, 1.27)	0.91	(0.73, 1.15)	1.56	(1.40, 1.73)	1.29	(1.16, 1.43)
Baseline age	0.99	(0.99, 1.00)	0.99	(0.99, 1.00)	1.02	(1.01, 1.02)	1.02	(1.01, 1.02)
NSAID use (before cohort)	0.73	(0.59, 0.90)	0.73	(0.59, 0.91)	1.15	(1.01, 1.31)	1.07	(0.94, 1.22)
Methotrexate use	0.89	(0.75, 1.06)	0.81	(0.68, 0.96)	1.28	(1.12, 1.48)	1.22	(1.06, 1.40)
Hydroxychloroquine use	0.73	(0.59, 0.90)	0.70	(0.57, 0.87)	0.91	(0.68, 1.22)	0.86	(0.64, 1.15)
HAQ disability score	1.15	(1.01, 1.31)			1.52	(1.41, 1.64)		
Comorbidity index	1.24	(1.17, 1.31)			1.36	(1.31, 1.41)		
BMI	1.18	(1.10, 1.26)			1.28	(1.21, 1.35)		
BMI^2 (per 50 units)	0.93	(0.89, 0.98)			0.89	(0.86, 0.92)		

of follow-up was also longer (2365 days vs. 1670 days) and so the two event rates were comparable (CPRD = 1.5 vs. NDB = 1.4 events per 100 person-years). Exposure to both methotrexate and hydroxychloroquine was more frequent in NDB than in CPRD. This is likely due to the differences in prescribing preferences of physicians in the two countries, as well as possible differences in disease severity.

Incidence rates of start of GC therapy were similar (8.3 patients initiated GC therapy per 100 person-years in NDB (95% CI 8.0 to 8.7) versus 7.2 in CPRD (95% CI 7.0 to 7.4)). It is also evident that exposure varied considerably over time. The mean percent of time exposed amongst users, from first use until end of follow-up, was 51% in NDB and 39% in CPRD, with lower percentage likely related to longer follow-up of the latter cohort. Thus, a simple time-fixed exposure definition would not be sufficient for these analyses, and time-dependent exposures must be considered.

The left-hand columns of Table V compare the results of two models in the NDB database where, respectively, (i) all important risk factors are included (columns 2–3) and (ii) only those important risk factors available in the CPRD database are included (columns 4–5). There is evident bias in the reduced model, with an exposure hazard ratio of 1.81 (95% CI 1.45 to 2.27) versus the fully adjusted hazard ratio of 1.48 (95% CI 1.17 to 1.86). In contrast, the effects of all the confounders adjusted for are almost identical between the two models estimated in the NDB database (left part of Table V), while the effect of exposure differs greatly. This pattern of results indicates that the additional covariates, not available for our CPRD

analyses and, thus, excluded from the reduced model, may act as confounders for the association between GC exposure and incidence of diabetes. In particular, the fact that the HR for exposure is inflated in the reduced model suggest that risk factors not measured in CPRD tend to be associated with both (i) more frequent use of GCs and (ii) higher type II DM risks, reflects a case of potential confounding by indication [4].

The results of the final model in the CPRD with martingale residual-based imputed unmeasured confounders are shown in the second last column of Table V. As expected based on NDB results, adjusting for the martingale residual-based imputed unmeasured confounders results in a reduction in the hazard ratio associated with current GC use, relative to that estimated adjusting only for confounders available in CPRD (from 38% increase in hazard to 15%).

Some limitations of our illustrative analyses have to be recognized and may make it difficult to directly compare our estimates with those obtained from a recent comprehensive investigation of the same association [30]. Of course, the accuracy of adjustment for additional confounders, not measured in CPRD, depends on the untestable assumptions that their associations with GC exposure and DM risk are similar in the two databases. Furthermore, we were not able to account for some additional potential confounders, such as disease severity, which was not recorded in either NDB or CPRD, so that some residual confounding of all estimates in both datasets cannot be excluded [30]. If distributions of further, completely unmeasured confounders were to differ between the two data sources, this may partly explain the difference between the final estimate in the full model in NDB and the final estimate in CPRD, obtained using our martingale residual-based method. It is also possible that some differences in the way some covariates were defined or measured in the respective datasets could contribute to the observed differences in their estimated effects. In addition, we considered only the simplest time-varying exposure metric, representing current GC exposure, while more complex exposure metrics, possibly involving cumulative effects of past doses, may account better for the impact of GC exposure [30, 43]. Finally, restriction to complete cases might have induced some selection bias and possibly affected our estimates. However, in spite of such limitations, this example illustrates the ability of the proposed martingale residual-based imputation to correct for the potential impact of unmeasured confounders, which are available in an internal or an external validation subsample.

5. Discussion

Unmeasured confounding may be considered an Achilles' heel of observational research on real-life studies of the safety or effectiveness of treatments [4, 8]. While no general solution to this problem is ever likely to be proposed, even those methods that can reduce or eliminate the resulting bias only under some plausible assumptions are of considerable interest [9, 44, 45]. In this spirit, we have proposed a new method for adjusting for unmeasured confounding bias in time-to-event analyses, when the confounders not measured in the main database are available in an internal or external validation sample. The method incorporates the martingale residuals into a model estimated in the validation sample, which is then used to impute the confounders unmeasured in the main database. In real-life applications, the validation sample could be an internal subset of the same source population included in the main database, or may be external, that is, drawn from a different population known or expected to have similar clinical characteristics, and we illustrate how the method can be applied in either case. Interestingly, sometimes an inherently internal validation sample may have to be analyzed as external, because the individuals cannot be linked between the two data sources due to data confidentiality concerns.

Although several methods have previously been proposed to use additional confounder measurements available in validation samples, such as BayesPS [9], two-stage calibration [10], and PSC [7]; only PSC has been extended to analysis of time-to-event data [7]. We have attempted to fill this gap in the current literature by developing and validating a new method, designed specifically for time-to-event analyses, that addresses the issue of unmeasured confounding while avoiding the restrictive surrogacy assumption required by PSC [7].

In simulations, our martingale residual-based method yielded almost uniformly unbiased estimates of the exposure/treatment association, regardless of the underlying assumptions about the true hazard ratio, strength and direction of unmeasured confounding, violation of surrogacy, and size of validation sample. Sensitivity analyses assessed the impact of different reasons for possible violations of the MCAR assumption with respect to the relationship between the validation sample and the main study database, where the additional confounders were not measured, that is, 'missing'. Differences in the distribution of the measured or unmeasured confounder, or in the event rate, had no marked impact on the perfor-

mance of any of the methods (Table I, scenarios 16–18). Accordingly, similar to main simulations, in all sensitivity analyses the martingale residual-based imputation performed better than, or at least as well as, all other methods considered. In many scenarios, its performance was similar to imputation with inclusion of separate terms for log of survival time and event indicator [22]. The fact that both martingale residual-based and log(t) imputations performed systematically better than simpler methods may be explained by the fact that both methods incorporate information about both survival time and event status in a mathematically similar manner. On the other hand, the results of our simulated scenarios 20 and 21 (Table I) suggest that martingale residual-based approach may be somewhat more flexible in accounting for more complex distributions of censoring and/or event times. In cases where the baseline hazard is irregular, some other transformation of t, rather than the log, may be more appropriate, while the martingale residual takes the estimated baseline hazard into account without the need for any assumptions about its analytical form or shape (Equation (1)). Furthermore, the martingale residual integrates all information on individual subjects' follow-up duration and status at end of follow-up, while the log(t)-based approach assumes their associations with the unmeasured confounder are additive, that is, independent of each other. Finally, the martingale residual accounts for the values of observed covariates and exposure, which may partly explain the observed outcome. Together, the aforementioned properties may help the martingale residual to discriminate slightly better than log(t)-based imputation between, for example, subjects who were censored early with low versus with high risk. Future studies should assess and compare the performance of these two methods across a wider range of assumptions about the underlying data structure.

As expected, our complex, multi-stage estimation procedure induced some moderate variance inflation, especially in scenarios where the validation sample was relatively small, with as few as 25 to 50 observed events. A bias-variance trade-off is typical of many complex bias reduction methods, such as IPTW methods for marginal structural models [46, 47] and instrumental variable methods [48]. Such methods focus on bias elimination, but the multi-step estimation procedures introduce further uncertainty about the final parameter estimates. However, because both PSC and standard analysis, selected for comparison because of their applicability to survival analyses, yielded often substantially biased estimates, our martingale residual-based estimates had a systematically better overall trade-off between bias and variance, with lower RMSE in all scenarios except one (with only minimal unmeasured confounding). In the main simulations, to provide proof of concept, we assumed that both confounders and exposure were time-invariant. Then, we also validated our method in more complex simulations, with time-varying measures of both confounders and exposure, and an external validation sample with event rate twice as high as that in the main database.

In real-life analyses of the association between GC therapy and risk of developing type II DM, our method reduced the hazard ratio associated with current GC use by more than 50%, relative to standard adjustment for only those confounders that were recorded in the main database.

Our method has some limitations and imposes some restrictions. Firstly, whereas our method performed well even with a fairly small size of validation sample (250 subjects and only 25 events), it should not be applied in the case of extremely small validation samples. In such situations the resulting estimates may not be stable, to the extent that their mean squared error may be possibly higher than that of the biased but considerably more stable standard estimates, which are obtained from the large main database and, thus, adjusted only for fully measured confounders. The same limitation likely applies to other methods that attempt to use validation samples, such as PSC. Secondly, although our preliminary simulations suggest a good performance of the method even when applied in studies with an external validation sample having a different event rate than the main database, and in cases where selection into the validation sample depends on measured or unmeasured confounders such that confounder distributions may differ between the two data sources, it remains to be investigated how further differences concerning confounder distributions or exposure–confounder associations would affect the accuracy of our estimates.

In the current implementation of our method, each confounder is imputed independently of all other confounders. In the situation where some confounders are highly correlated with each other, each of the independently imputed values of individual correlated confounders may separately account for the same underlying sources of the confounding bias. Depending on the pattern of multivariate correlations between the relevant confounders, exposure, and the outcome, this may possibly result in over- or under-adjustment for the joint impact of several correlated confounders. Future research should consider this

more complex setting, with multiple correlated unmeasured confounders, and assess the potential benefits of replacing separate imputation of individual confounders by imputation of an aggregate confounder score akin to the disease risk score [49].

In time-varying settings, we assumed that both (i) current exposure (treatment assignment) and (ii) current hazard (outcome) depend only on the current values of measured and/or unmeasured confounders. Accordingly, we used only current values of measured confounders, time-varying exposure and martingale residual, observed at time t, to impute the concurrent values of unmeasured time-varying confounders $U_j(t)$. In some applications, the user may prefer to assume some lag (L) in the relationship between previous values of $U_j(t-L)$ and current exposure $X(t)$, and/or current hazard $\lambda(t)$. In such applications, the methods outlined in Section 2.2 are still generally applicable, but the data analyst will have to decide how to anchor the exposure and/or martingale residual measurements in time, and may, for example use $X(t+L)$ and $\hat{M}(t+L)$ to impute $U_j(t)$. Future research should investigate the performance of our martingale residual-based method in such, more complex, settings and assess the robustness of the results with respect to choice of the lag L.

In this paper, we have described the method when both confounders and exposure are time-invariant or both are time-varying. The time-invariant case is straightforward, and the simple time-varying case is described in Section 3.2 and illustrated in Section 4 followed easily. However, extensions of the method to more complex time-varying exposure metrics, which reflect, for example, possible cumulative effects and to marginal structural Cox models, which account for time-varying variables that act as both confounders and mediators of treatment effects [50–52], will require further analytical developments and validation studies. Although we investigated only imputation of continuous and binary confounders, with, respectively, linear and logistic regression, it is possible to impute categorical or ordinal confounders using usual models for imputation of such variables, such as multinomial logistic regression or the proportional odds model [53].

Unmeasured confounding remains an important problem in many observational studies, and particularly in those which rely on large administrative databases. Our method may help reducing its impact in those time-to-event analyses where additional information on confounders is available in a smaller validation sample. We hope that our results may both improve the accuracy of real-life observational studies of different treatments or exposures and stimulate further methodological developments in this challenging but important area of statistical research.

Appendix A: A formal rationale for the use of martingale residuals in the imputation model

Under a Cox PH model with both exposure X and measured confounders C fully measured and time-invariant (for this proof, let $\{X, C\}$ be jointly represented by Z), and unmeasured confounders U

$$\lambda(t|X, C, U) = \lambda(t|Z, U) = \lambda_0(t)\exp\{\beta_Z Z + \beta_U U\}, \tag{9}$$

the log-likelihood can be written as

$$\begin{aligned}\log f_i(t) &= \delta \log \lambda(t) - \Lambda(t)\\ &= \delta(\log \lambda_0(t) + \beta_Z Z + \beta_U U) - \Lambda_0(t)e^{\beta_Z Z + \beta_U U}.\end{aligned} \tag{10}$$

Using usual notation, δ is the indicator for event (1 if subject had an event, 0 if censored), $\Lambda_0(t)$ is the cumulative baseline hazard.

Using Bayes' theorem and (10)

$$\begin{aligned}\log f(U|t, \delta, Z) &= \log \frac{f(t, \delta|Z, U).f(U|Z)}{f(t, \delta|Z)}\\ &= \log f(t, \delta|Z, U) + \log f(U|Z) + \text{const.}(t, \delta, Z)\\ &\propto \log f(U|Z) + \delta(\log \lambda_0(t) + \beta_Z Z + \beta_U U) - \Lambda_0(t)e^{\beta_Z Z + \beta_U U}.\end{aligned} \tag{11}$$

The martingale residual is generally defined as follows:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)e^{\beta' X_i(s)} d\Lambda_0(s). \tag{12}$$

However, for the case with no time-varying covariates, this can be simplified, and the estimated martingale residual written as

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(\tau_i)e^{\hat{\beta}'X_i} \tag{13}$$

where $\hat{\Lambda}_0$ is the Breslow estimate of cumulative baseline hazard, and τ_i is the event time for subject i.

We are particularly interested in the martingale residual from the reduced model, which excludes U, that is,

$$\hat{M}_i^* = \delta_i - \hat{\Lambda}_0^*(\tau_i)\exp\left\{\hat{\beta}^{*\prime}Z_i\right\}. \tag{14}$$

We consider the case of a single unmeasured confounder, U, but deal separately with a binary versus a continuous U.

A.1. (1) Binary U

Assume the true model for $U|Z$ is $\text{logit}P[U=1|Z] = \alpha_0 + \alpha_1 Z$. Then using (11)

$$\begin{aligned}\text{logit}P[U=1|\tau,\delta,Z] &= \log P[U=1|\tau,\delta,Z] - \log P[U=0|\tau,\delta,Z]\\ &\propto \log P[U=1|Z] + \delta(\log\lambda_0(\tau) + \beta_Z Z + \beta_U) - \Lambda_0(\tau)\exp\{\beta_Z Z + \beta_U\}\\ &\quad - \log P[U=0|Z] - \delta(\log\lambda_0(\tau) + \beta_Z Z) + \Lambda_0(\tau)\exp\{\beta_Z Z\}\\ &= \alpha_0 + \alpha_1 Z + \delta\beta_U - (\exp\{\beta_U\} - 1)\Lambda_0(\tau)\exp\{\beta_Z Z\}.\end{aligned} \tag{15}$$

If we can assume that $\hat{\beta}_Z^*$ (from the reduced model in Equation (2)) is a reasonable estimate for β_Z (from the full model), and $\hat{\Lambda}_0$ a reasonable estimate for Λ_0, then the last term of (15) is similar to the second term of the martingale residual in Equation (14). Also included in both is a linear term for δ.

Furthermore, for small to moderate log HRs for the effect of U on the hazard β_U, $(-0.5 < \beta_U < 0.5)$: $e^{\beta_U} - 1 \approx \beta_U$. If this is the case, then (15) $\approx \alpha_0 + \alpha_1 Z + \beta_U(\delta - \Lambda_0(\tau)\exp\{\beta_Z Z\})$, and the last term is almost identical to the martingale residual. In Section 3, we assess the impact of increasing β_U on the accuracy of our estimates.

A.2. (2) Normally distributed U

Assume the true model for U is $E[U|Z] = \alpha_0 + \alpha_1 Z$, and therefore $\log f(U|Z) = -(U - \alpha_0 - \alpha_1 Z)^2/2\sigma^2 + \text{const}$.

From (11):

$$\begin{aligned}\log f(U|\tau,\delta,Z) &\propto -(U-\alpha_0-\alpha_1 Z)^2/2\sigma^2 + \delta(\log\lambda_0(\tau) + \beta_Z Z + \beta_U U)\\ &\quad - \Lambda_0(\tau)\exp\{\beta_Z Z + \beta_U U\}\\ &= \delta\beta_U U - \Lambda_0(\tau)\exp\{\beta_Z Z + \beta_U U\} - (U-\alpha_0-\alpha_1 Z)^2/2\sigma^2\end{aligned} \tag{16}$$

This cannot be written as a normal PDF due to the $\exp\{\beta_U U\}$ term. However, if we take a Taylor expansion of order 1 around (U_i, Z_i):

$$\exp\{\beta_Z Z + \beta_U U\} \approx \exp\{\beta_Z Z_i + \beta_U U_i\}(1 + \beta_Z(Z - Z_i) + \beta_U(U - U_i)), \tag{17}$$

and (16) becomes:

$$\begin{aligned}&\delta\beta_U U - \Lambda_0(\tau)\exp\{\beta_Z Z_i + \beta_U U_i\}(1 + \beta_Z(Z - Z_i) + \beta_U(U - U_i)) - (U - \alpha_0 - \alpha_1 Z)^2/2\sigma^2\\ &= \beta_U U(\delta - \Lambda_0(\tau)\exp\{\beta_Z Z_i + \beta_U U_i\}) - (U - \alpha_0 - \alpha_1 Z)^2/2\sigma^2 + const.\\ &= -\left(U - \left(\alpha_0 + \alpha_1 Z + \sigma^2\beta_U\left(\delta - \Lambda_0(\tau)\exp\{\beta_Z Z_i + \beta_U U_i\}\right)\right)\right)^2/2\sigma^2 + const.\end{aligned} \tag{18}$$

Therefore $U|\tau,\delta,Z$ is approximately distributed as $N(\alpha_0 + \alpha_1 Z + \sigma^2\beta_U(\delta - \Lambda_0(\tau)\exp\{\beta_Z Z_i + \beta_U U_i\}), \sigma^2)$. The mean is, therefore, a function of measured covariates and a term very similar to the full model martingale residual in Equation (12). If we assume that the reduced model martingale residual in Equation (14) is a good approximation for the full model martingale residual, then its inclusion in the imputation model would be expected to increase the accuracy of the imputation of U.

Acknowledgements

This work was funded by the Government of Canada through the Canadian Institutes of Health Research/Drug Safety and Effectiveness Network (CIHR/DSEN; grant no. 298283), and by the Natural Sciences and Engineering Research Council (NSERC) grant no. 228203. Computations were made on the supercomputer Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA, and the Fonds de recherche du Québec – Nature et technologies (FRQ-NT). The authors thank Drs. William Dixon and Kaleb Michaud for providing the motivating example and for their comments on the manuscript.

References

1. Skegg DC. Evaluating the safety of medicines, with particular reference to contraception. *Statistics in Medicine* 2001; **20**(23):3557–3569.
2. Abrahamowicz M, Tamblyn R. *Drug Utilization Patterns* (2nd edn.), Encyclopedia of Biostatistics. John Wiley and Sons, Ltd.: Chichester (UK), 2005.
3. Wolfe F, Zwillich SH. The long-term outcomes of rheumatoid arthritis: a 23-year prospective, longitudinal study of total joint replacement and its predictors in 1,600 patients with rheumatoid arthritis. *Arthritis & Rheumatism* 1998; **41**(6): 1072–1082.
4. Walker AM. Confounding by indication. *Epidemiology* 1996; **7**(4):335–336.
5. McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiology and Drug Safety* 2003; **12**(7):551–558.
6. Schneeweiss S, Wang PS. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. *Journal of Clinical Psychopharmacology* 2004; **24**(6):632–638.
7. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology* 2005; **162**(3):279–289.
8. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology* 2005; **16**(1):17–24.
9. McCandless LC, Richardson S, Best N. Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association* 2012; **107**(497):40–51.
10. Lin HW, Chen YH. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *American Journal of Epidemiology* 2014; **180**(3): 308–317.
11. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration – a simulation study. *American Journal of Epidemiology* 2007; **165**(10):1110–1118.
12. Chen Y, Chen H. A unified approach to regression analysis under doublesampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2000; **62**(3):449–460.
13. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* 2009; **338**:b2393.
14. Barlow WE, Prentice RL. Residuals for relative risk regression. *Biometrika* 1988; **75**(1):65–74.
15. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika* 1990; **77**(1): 147–160.
16. Breslow N. Covariance analysis of censored survival data. *Biometrics* 1974; **30**(1):89–99.
17. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 2006; **59**(10):1092–1101.
18. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**(5): 615–625.
19. R Core Team. *R: A Language Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2014.
20. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723.
21. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**(6):681–694.
22. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine* 2009; **28**(15):1982–1998.
23. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons: New York, 1987.
24. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
25. Achy-Brou AC, Frangakis CE, Griswold M. Estimating treatment effects of longitudinal designs using regression models on propensity scores. *Biometrics* 2010; **66**(3):824–833.
26. Ray WA, Liu Q, Shepherd BE. Performance of time dependent propensity scores: a pharmacoepidemiology case study. *Pharmacoepidemiology and Drug Safety* 2015; **24**(1):98–106.
27. Sylvestre MP, Abrahamowicz M. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine* 2008; **27**(14):2618–2634.
28. Sylvestre MP, Evans T, Mackenzie T, Abrahamowicz M. Permalgo: permutational algorithm to generate event times conditional on a covariate matrix including time-dependent covariates, 2010. R package version 1.0.
29. Howe CJ, Cole SR, Lau B, Napravnik S, Eron Jr JJ. Selection bias due to loss to follow up in cohort studies. *Epidemiology* 2016; **27**(1):91–97.

30. Movahedi M, Beauchamp ME, Abrahamowicz M, Ray DW, Michaud K, Pedro S, Dixon WG. Risk of incident diabetes associated with dose and duration of oral glucocorticoid therapy in patients with rheumatoid arthritis. *Arthritis & Rheumatology* 2016; **68**(5):1089–1098.
31. Caplan L, Wolfe F, Russell AS, Michaud K. Corticosteroid use in rheumatoid arthritis: prevalence, predictors, correlates, and outcomes. *The Journal of Rheumatology* 2007; **34**(4):696–705.
32. Caporali R, Todoerti M, Sakellariou G, Montecucco C. Glucocorticoids in rheumatoid arthritis. *Drugs* 2013; **73**(1):31–43.
33. Saag KG. Short-term and long-term safety of glucocorticoids in rheumatoid arthritis. *Bulletin of the Hospital for Joint Diseases* 2011; **70**:21–25.
34. Kwon S, Hermayer KL. Glucocorticoid-induced hyperglycemia. *The American Journal of the Medical Sciences* 2013; **345**(4):274–277.
35. Perez A, Jansen-Chaparro S, Saigi I, Bernal-Lopez MR, Miñambres I, Gomez-Huelgas R. Glucocorticoid-induced hyperglycemia. *Journal of Diabetes* 2014; **6**(1):9–20.
36. Tarp S, Furst D, Kirwan J, Boers M, Bliddal H, Woodworth T, Bartel EM, Danneskiold-Samsoe B, Kristensen L, Thirstrup S, Rasmussen M, Kaldas M, Christensen R. Short to medium term safety of glucocorticoid therapy in rheumatoid arthritis: a systematic review and dose–response analysis of randomized controlled trials. *Arthritis & Rheumatism* 2012; **64** (10 (supplement)):S917.
37. Panthakalam S, Bhatnagar D, Klimiuk P. The prevalence and management of hyperglycaemia in patients with rheumatoid arthritis on corticosteroid therapy. *Scottish Medical Journal* 2004; **49**(4):139–141.
38. Raúl AAC, Barile-Fabris LA, Frati-Munari AC, Baltazár-Montufar P. Risk factors for steroid diabetes in rheumatic patients. *Archives of Medical Research* 1997; **29**(3):259–262.
39. Di Comite G, Rossi CM. Risk of diabetes in patients with rheumatoid arthritis taking hydroxychloroquine. *Journal of the American Medical Association* 2007; **298**(20):2367–2370.
40. Wolfe F, Michaud K. The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank. *Rheumatology* 2011; **50**(1):16–24.
41. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. *Arthritis & Rheumatism* 2000; **43**(12):2751–2761.
42. Michaud K, Wolfe F. Comorbidities in rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology* 2007; **21**(5):885–906.
43. Dixon WG, Abrahamowicz M, Beauchamp ME, Ray DW, Bernatsky S, Suissa S, Sylvestre MP. Immediate and delayed impact of oral glucocorticoid therapy on risk of serious infection in older patients with rheumatoid arthritis: a nested case–control analysis. *Annals of the Rheumatic Diseases* 2012; **71**(7):1128–1133.
44. Brookhart MA, Wang P, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**(3):268.
45. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**(4):512.
46. Cole SR, Hernàn MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**(6):656–664.
47. Xiao Y, Moodie EE, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* 2013; **2**(1):1–20.
48. Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiology and Drug Safety* 2009; **18**(7):562–571.
49. Miettinen OS. Stratification by a multivariate confounder score. *American Journal of Epidemiology* 1976; **104**(6):609–620.
50. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570.
51. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
52. Xiao Y, Abrahamowicz M, Moodie EE, Weber R, Young J. Flexible marginal structural models for estimating the cumulative effect of a time-dependent treatment on the hazard: reassessing the cardiovascular risks of didanosine treatment in the Swiss HIV cohort study. *Journal of the American Statistical Association* 2014; **109**(506):455–464.
53. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 2011; **30**(4):377–399.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.