SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2019, Vol. 28(2) 357–371
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280217726800
journals.sagepub.com/home/smm
SAGE

Adjustment for time-dependent unmeasured confounders in marginal structural Cox models using validation sample data

Rebecca M Burne and Michal Abrahamowicz

Abstract

Large databases used in observational studies of drug safety often lack information on important confounders. The resulting unmeasured confounding bias may be avoided by using additional confounder information, frequently available in smaller clinical "validation samples". Yet, no existing method that uses such validation samples is able to deal with unmeasured time-varying variables acting as both confounders and possible mediators of the treatment effect. We propose and compare alternative methods which control for confounders measured only in a validation sample within marginal structural Cox models. Each method corrects the time-varying inverse probability of treatment weights for all subject-by-time observations using either regression calibration of the propensity score, or multiple imputation of unmeasured confounders. Two proposed methods rely on martingale residuals from a Cox model that includes only confounders fully measured in the large database, to correct inverse probability of treatment weight for imputed values of unmeasured confounders. Simulation demonstrates that martingale residual-based methods systematically reduce confounding bias over naïve methods, with multiple imputation including the martingale residual yielding, on average, the best overall accuracy. We apply martingale residual-based imputation to re-assess the potential risk of drug-induced hypoglycemia in diabetic patients, where an important laboratory test is repeatedly measured only in a small sub-cohort.

Keywords

Unmeasured confounding, survival analysis, marginal structural models, martingale residuals, simulations

1 Introduction

Accurate estimation of the causal effects of treatments or exposures on a specific outcome is a major challenge of epidemiological and clinical research. In longitudinal cohort studies of time-varying treatments, this is further complicated if current treatment may affect subsequent values of certain time-varying risk factors, whose previous values might have influenced previous treatment decisions, acting as potential confounders.[1–3] Indeed, in such situations, the treatment effect might be partly mediated through the resulting changes in such time-varying confounders. If so, then including such time-varying confounders/mediators as adjustment variables in the conventional multivariable regression model will bias the treatment effect estimate, which will not account for the mediated (indirect) effect.[4] Marginal structural models (MSMs) with inverse probability of treatment weighting (IPTW) are a popular method to account for such time-varying confounders/ mediators.[4,5]

However, whereas IPTW removes bias due to *measured* time-varying confounders/mediators,[4,6] it does *not* protect against confounding due to factors not recorded in the study database.[7,8] Indeed, one of the fundamental assumptions underlying MSM methodology is that there are no unmeasured confounders.[9] Yet, unmeasured confounding is recognized as the Achilles heel of observational cohort studies of the intentional or unintentional (adverse) effects of treatments.[10–12] Indeed, large databases typically used in

Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Canada

Corresponding author:
Rebecca M Burne, Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montréal, QC, Canada H3A 1A1.
Email: rebecca.burne@mail.mcgill.ca

pharmacoepidemiological studies of drug safety or comparative effectiveness do not record several clinical variables, lifestyle characteristics and laboratory test results that may change during the follow-up, and may affect both the treatment decisions and the outcome.[13]

Interestingly, the paramount challenge of reducing the impact of unmeasured confounding in the specific context of MSMs has received relatively little attention in statistical literature. Indeed, we have identified only two methodological articles that attempted to address selected aspects of unmeasured confounding in MSMs. Brumback et al.[7] presented a framework for sensitivity analysis to assess the impact of potential unmeasured confounding on MSM estimates of the causal effects of time-varying treatments. The method requires specifying a functional form for the unmeasured confounding bias, defined as the difference between potential outcomes between subjects receiving the alternative treatments, conditional on all measured covariates, and in many real-life applications prior knowledge may not be sufficient to ensure plausibility of such complex assumptions. Moodie et al.[5] considered adapting the inverse probability of censoring weighting (IPCW) approach to a longitudinal setting when time-varying confounders may be measured only until a certain time point during the follow-up, which may vary across the subjects.[8] Their approach involves censoring a subject at the first time-point when the values of the confounder(s) are missing, and multiplying the conventional IPT weights by time-varying IPCW-like weights, calculated at a corresponding time point.[8] However, in simulations, the resulting estimates were slightly more biased, and markedly more variable, than those obtained through conventional multiple imputation.[8] Moreover, their method cannot be applied when for most study subjects some important time-varying confounders are not measured at all, which likely occurs in most studies based on large administrative health databases.[13]

The current study is motivated by our belief that, in some real-life MSM analyses, the impact of unmeasured confounding can be reduced or even eliminated by adapting other methods, recently proposed and validated for conventional multivariable (unweighted) regression modeling. Specifically, we focus on the applications where, in addition to a large main study database, the researchers have access to a smaller clinical validation sample, with information on additional time-varying confounders/mediators, not recorded in the main data.[14,15] Indeed, in the last decade several alternative methods have been proposed to make an efficient use of such validation samples, including propensity score calibration (PSC),[16,17] BayesPS,[18] and two-stage calibration.[19] All the aforementioned methods use the measurements of additional confounders in the validation sample data to correct the propensity score (PS) in the main database. However, most of these methods have not been extended to time-to-event analyses with time-varying exposure and confounders, and none has been adapted to MSMs. Recently, we have proposed a new method, designed specifically for time-to-event analyses, which relies on the martingale residuals (MRs) from a multivariable Cox model, estimated from the validation sample, to directly impute the unmeasured confounders for all subjects in the main database.[20] In simulations, this MR-based imputation largely reduced unmeasured confounding bias, and yielded more accurate exposure effect estimates than either PSC or conventional multiple imputation. Thus, the MR-based imputation performed very well in all standard (unweighted) Cox proportional hazards model analyses, both for time-fixed exposure and confounders, and in the case where both exposure and confounders were time-varying but with no mediation.[20] However, the method uses the imputed confounder data as a standard adjustment variable in the multivariable Cox model, which makes it inappropriate for marginal structural modeling analyses of studies where time-varying confounder(s) act also as mediators of the indirect effect of previous exposures and, thus, cannot be directly adjusted for.[4]

To address the aforementioned challenge, in this paper, we propose different methods to correct MSM time-to-event analyses for additional time-varying confounder(s)/mediator(s), available only in a relatively small validation sample. The alternative methods adapt either PSC[16] or our recent MR-based imputation approach[20] to correct estimation of the IPT weights in the large main database. We compare and validate these methods under a variety of simulation scenarios. Then, to illustrate the performance of the selected methods, we re-assess the potential association between current use of dipeptidyl peptidase (DPP)-4 inhibitors and the hazard of hospitalization due to hypoglycemia in patients with type II diabetes mellitus, using a large cohort where an important laboratory test was repeatedly measured only for a small sub-cohort.

2 Methods

2.1 Overview of the Cox MSM

We consider a cohort study with time-to-event outcome, where some of the time-varying confounders $L(t_k)$, measured at each visit $k = 1, \ldots, p$ at time t_k, may both affect current assignment of a time-varying treatment $A(t_k)$, and be affected by past treatment $A(t_{k-1})$. In this case, simply conditioning on time-varying characteristics $L(t_k)$, that act as both confounders and mediators, in a standard (unweighted) multivariable model, e.g. Cox proportional hazards model, will not account for the part of the effect of the treatment $A(t_k)$ that is mediated through changes in the values of these characteristics. This will result in a biased estimate for the causal effect of the treatment on the outcome.[4] This challenge can be addressed by MSMs, with inverse probability of treatment (IPT) weights.[3,4] Using the notation of Hernán et al.,[4] a Cox MSM can be written as:

$$\lambda_{T_{\bar{a}}}(t|V) = \lambda_0(t)\exp\{\beta_1 a(t) + \beta_2 \mathbf{V}\} \tag{1}$$

where individual observations are weighted by the IPT weights to create a pseudo-population in which treatment is not correlated with time-varying confounders $\mathbf{L}(t_k)$. $T_{\bar{a}}$ denotes the event time had an individual, possibly counterfactually, received treatment history $\bar{a}$, and $\mathbf{V}$ are baseline covariates. As in the MSM literature, overbars denote history (for example, $\bar{L}(t_k) = \{L(t_1), \ldots, L(t_k)\}$).

To reduce the variance of the MSM estimates, we re-weight individual subject-by-time observations using stabilized IPT weights[4,9,21]:

$$sw_i(t_k) = \prod_{k=0}^{t_k} \frac{P[A(t_k) = a_i(t_k)|\bar{A}_i(t_{k-1}) = \bar{a}_i(t_{k-1}), V_i = v_i]}{P[A(t_k) = a_i(t_k)|\bar{A}_i(t_{k-1}) = \bar{a}_i(t_{k-1}), \bar{L}_i(t_k) = \bar{l}_i(t_k)]} \tag{2}$$

The resulting estimate in the weighted pseudo-population is un-confounded by $\mathbf{L}(t_k)$, and, therefore, yields an unbiased estimate for the causal effect of treatment.[4]

2.2 Impact of unmeasured time-varying confounders

Suppose, however, that, among all relevant time-varying confounders $\mathbf{L}(t_k)$, in the main database only a subset $\mathbf{C}(t_k)$ is available, while the remaining confounders $\mathbf{U}(t_k)$ are not measured, as illustrated in the directed acyclic graph (DAG) in Figure 1. (For simplicity, we assume that all baseline fixed-in-time confounders $\mathbf{V}$, generally easier and less expensive to measure, are measured for all subjects in the main database.) If the subset $\mathbf{U}(t_k)$ is *not* empty, then the standard MSM assumption of no unmeasured confounding is violated.[4,5] If, however, $\mathbf{U}(\mathbf{t})$ is a true confounder, then this strong assumption will be clearly violated. Then, in the pseudo-population created by using the stabilized weights in equation (2) with denominators based on only the fully measured confounders $\mathbf{C}$, the treatment groups will not be balanced with respect to values of $U(t)$, resulting in biased estimates of the treatment effect.

2.3 Proposed extensions of MSMs to account for time-varying confounders measured only in a validation sample

We focus on the data set-up where the confounders $\mathbf{U}(t_k)$ not measured for most subjects in the study cohort are available in a relatively small validation sample, such as a clinical database collected for research purposes.[14,16] In such a situation, it would be useful to employ methods which make efficient use of the additional confounder information in the validation sample to correct the MSM analyses of the main database.

We propose to extend the existing recent methods to control for time-dependent confounders ($U(t)$), measured only in a validation sample, to Cox MSM analyses with IPT weights. Each of the proposed methods employs a different approach to account for $U(t)$ when estimating stabilized time-varying IPTWs in equation (2). Then,

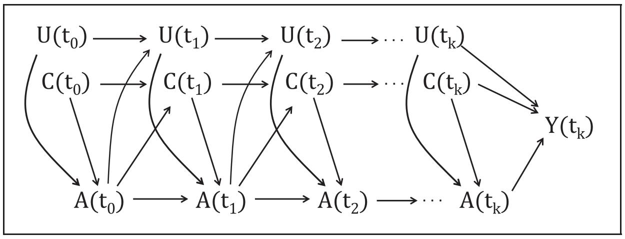

Figure 1. Directed acyclic graph (DAG) showing the assumed structure of confounding.

the resulting stabilized weights are used to weight individual (person-by-time) observations in the final Cox MSM, similar to relevant previous approaches.[4,9,6,22]

1A: Regression calibration of the PS (RC). The first proposed method adapts the PSC approach of Stürmer et al.[16,17] to IPTW estimation. PSC formulates the unmeasured confounding, rather than as a "missing data" problem, as a problem of measurement error in the PS.[16] It considers the PS that adjusts for only the fully measured confounders as an error-prone PS, in contrast to the true "gold standard" PS, which adjusts additionally for confounders measured only in the validation sample. PSC uses a well-known measurement error method, regression calibration,[23] to correct the error-prone PS in the main database, using information from the validation sample.[16] We adapt this approach and propose to correct the PS in the denominator of the stabilized IPT weight in equation (2) by using regression calibration. Specifically, the denominator for the fully adjusted "gold standard" weight is made up by $PS_{GS} = P(A(t_k) = 1|\bar{a}_i(t_{k-1}), \bar{\mathbf{c}}_i(t_k), \bar{\mathbf{u}}_i(t_k))$ for those currently treated ($A(t) = 1$), and $1 - PS_{GS}$ for those untreated ($A(t) = 0$). On the other hand, the error-prone weight, that fails to account for U, has in the denominator a product of $PS_{EP} = P(A(t_k) = 1|\bar{a}_i(t_{k-1}), \bar{\mathbf{c}}_i(t_k))$ if treated, or $1 - PS_{EP}$ if untreated. By using the validation sample data to correct the error-prone PS, through regression calibration, we can approximate the IPT weights that would be based on the gold standard PS and, thus, attempt to adjust for the unmeasured confounders $\mathbf{U}$. To this end, we use regression calibration in a similar manner to Stürmer et al.,[16] as a single imputation from the linear measurement error model:

$$E[PS_{GS}|PS_{EP}, A] = \alpha_0 + \alpha_1 PS_{EP} + \alpha_2 A \tag{3}$$

The above method relies on the PS, which does not use information on the outcomes observed for individual subjects. Yet, in un-weighted regression settings, accounting for the observed outcomes of individual subjects improves the imputation of missing covariates associated with the outcome.[24,25] Therefore, our remaining proposed methods attempt to enhance the way we correct the IPT weights for unmeasured $U(t)$ by incorporating information on the outcomes. However, this task is less straightforward in the context of survival analyses, where the outcome is bivariate, and composed of both follow-up time and event indicator.[25] We have recently proposed to overcome this challenge by including the (possibly time-varying) MR in an imputation model for the unmeasured confounder, which is estimated using the validation sample data.[20] The reason behind use of the MR is two-fold. Firstly, the MR accounts simultaneously for both follow-up time and the status (event or censoring)[26]:

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\left\{\widehat{\boldsymbol{\beta}}' \mathbf{Z}_i(s)\right\} \mathrm{d}\hat{\Lambda}_0(s) \tag{4}$$

where $Y_i(s)$ is an indicator function denoting whether individual i is at risk at time $s \leq t$, $N_i(t)$ 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,[27] and $\mathbf{Z}_i(s)$ is the vector of possibly time-dependent covariates for individual i at time s.

Secondly, the MR from the reduced model, which accounts only for confounders measured in the main database ($C(t)$), measures the lack of fit of that model.[26] This is important because we hypothesize that this lack of fit may be directly informative about the value of the unmeasured confounder(s). For example, a highly positive MR, which indicates survival time shorter than expected based on the values of all measured covariates,[26] may suggest that the subject has some high-risk value(s) of other, unmeasured prognostic factors (e.g. is a heavy smoker or has severe hypertension). Indeed, we have demonstrated that, both in time-invariant and time-varying simulations with no mediation, MR-based imputation largely reduced or eliminated unmeasured confounding bias.[20] We now adapt this approach to the setting where Cox MSM analyses, with IPT stabilized weights, are employed to account for mediation.

1B: MR-*enhanced regression calibration of the PS (RC+MR)*. The MR-based proposed methods differ in how the MRs are used to improve the accuracy of the IPTW estimation. In particular, the second method combines the regression calibration approach of the first method with MRs, which are expected to account for the relationship of the PS (through the confounders) to the outcome. We use a two-stage approach to implement this method. At the first stage of analyses, we include a term for the MR, $\alpha_3 M$, in the linear measurement error model of the first method (equation (3)):

$$E[PS_{GS}|PS_{EP}, A] = \alpha_0 + \alpha_1 PS_{EP} + \alpha_2 A + \alpha_3 M \tag{5}$$

At the second stage, we use the resulting expected PS_{GS} values from equation (3) to calculate the corrected time-varying IPT weights, which are then employed in the Cox MSM analyses to weight all observations, for all subjects in the main database

2A: MR-*based multiple imputation (MR-based MI)*. Our third proposed method, instead of focusing on the correction of the PS used to calculate the IPT weights, involves direct imputation of the unmeasured confounder(s). In contrast to the first two methods, the third method does not rely on regression calibration at all. Instead, we first impute the unmeasured confounder(s) using MR-based imputation, previously proposed for imputation of confounders to be used as standard adjustment variable(s) in unweighted Cox PH models.[20] This method is implemented via the following steps:

Step 1: Fit an unweighted Cox PH model in the main database and in the validation sample, using data only on exposure and fully measured confounders:

$$\lambda(t) = \lambda_0(t)\exp\{\gamma_1{}'X(t) + \gamma_2{}'\mathbf{C}(t) + \gamma_3{}'\mathbf{V}\} \tag{6}$$

From this fitted model, obtain the MR $\hat{M}_i(t_k)$ at each visit time t_k for each individual i.

Step 2: The resulting time-varying MRs are then used in imputation models for each of the unmeasured confounders $U_1(t), U_2(t), \ldots, U_p(t)$. To this end, in the validation sample, where unmeasured confounders are measured, we fit separate models for each unmeasured confounder (dependent variable), using treatment $A(t)$ and the fully measured confounders as independent variables. If $U_j(t)$ is continuous, then a linear regression model may be used (perhaps with a suitable transformation of $U_j(t)$ to ensure normality), while for a binary variable a logistic regression model is employed.[20]

Step 3: Based on the models estimated in Step 2, the value of the unmeasured confounder is imputed for each observation (time-by-subject) in the main database. If $U_j(t)$ is continuous, this imputation involves sampling from a normal distribution with the mean obtained from the linear regression model and variance approximated by the prediction variance from this model. If $U_j(t)$ is binary, we sample from a Bernoulli distribution with the probability of "success" equal to the corresponding logistic regression model estimate. Details of Steps 2 and 3 are described in our recent paper.[20]

Step 4: For each subject in the main database, we then simply use the imputed time-specific values of the time-varying confounders $U(t)$, together with the fully measured confounders $C(t)$, to estimate the time-varying stabilized IPT weights in equation (2). Multiple imputation is used, with 10 imputations for each replication, and results combined using Rubin's rules.[28] Finally, the resulting time-varying IPT weights are used to weight individual subject-by-time observations in the Cox MSM analyses.

2B: MR-*based multiple imputation including prior exposure (MR-based MI+ A_{-1})*. We also consider a minor modification of the above (third) method. We hypothesize that prior exposure $A(t_{k-1})$ may be informative about the value(s) of unmeasured confounder(s), since in the assumed structure of confounding (Figure 1) it affects current values of both measured $C(t_k)$ and unmeasured $U(t_k)$ time-varying confounders. Thus, including exposure at the previous interval in the imputation model in Step 2 above may improve the imputation.

3 Simulations

3.1 Simulation design and data generation

In order to assess and compare the performance of the proposed methods, we simulated data from a hypothetical cohort study in which subjects are followed for up to 10 equally spaced visits $k = 1, 2, \ldots, 10$ at times $t_1, \ldots, t_{10}$. Below, we summarize the underlying assumptions and outline the essential elements and the methods used for data-generation. All simulations were run using R software.[29]

We considered a case with a binary time-varying exposure $A(t)$ (e.g. current use of a drug), two time-invariant covariates, measured at the cohort entry (V_1, V_2) and three time-varying ($C_1(t), C_2(t), U_1(t)$), covariates. Current value of exposure, from visit k to (k+1), was generated from a binomial distribution, with the probability of being exposed defined as:

$$\begin{aligned}\text{logit}(P(A(t_k) = 1|V, C, U)) = \;& \alpha_0 + \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 C_1(t_k) + \alpha_4 C_2(t_k) \\ & + \alpha_5 U_1(t_k) + \alpha_6 A(t_{k-1})\end{aligned}$$

In addition to the main database that includes all cohort members, we assumed a much smaller validation sample, which, in different simulated scenarios, may or may not represent a random sample of the full cohort. At each visit at time t_k, updated values of all time-varying covariates $C_1(t_k), C_2(t_k), U_1(t_k)$ are measured in the validation sample subjects, while only $C_1(t_k), C_2(t_k)$ are measured in the main database, where $U_1(t)$ is an unmeasured covariate. For most scenarios (unless explicitly indicated otherwise) all fully measured covariates $V_1, V_2, C_1(t_k)$ and $C_2(t_k)$ were continuous, whereas unmeasured $U_1(t_k)$ was binary.

We then generated the outcome data, by adapting the method originally proposed by Young et al. specifically for MSM analyses of time-to-event data.[30] This ensures that the data structure is consistent with a marginal structural Cox model, and allows us to quantify the true strength of the total (marginal) causal effect of the time-varying exposure on the hazard, which is then used as a benchmark to assess potential bias of the estimators obtained with different analytical methods.[22] We first generated a counterfactual event time T_0 based on a pre-specified marginal distribution, and then generated confounders with pre-specified relationships to this counterfactual event time. The value of each time-varying confounder at t_k was also assumed to be affected by prior values, at t_{k-1}, of the same confounder and exposure. For example, for binary $U_1(t_k)$:

$$\operatorname{logit}(P(U_1(t_k) = 1)) = \gamma_0 + \gamma_1 T_0 + \gamma_2 A(t_{k-1}) + \gamma_3 U(t_{k-1}) \tag{8}$$

Exposure at t_k, $A(t_k)$, was then generated based on prior exposure $A(t_{k-1})$ and current values of all confounders $\mathbf{V}$, $\mathbf{C}(t_k)$, and $U(t_k)$ (equation (7)). In this way the pattern of confounding is consistent with the DAG in Figure 1, with $U(t)$ (a) dependent on previous exposure, (b) affecting the outcome, and (c) influencing current exposure. Thus, unmeasured $U(t)$ acts here as an unmeasured time-varying confounder (through a combination of (b) and (c)) and, at the same time, a time-varying mediator (through (a) and (c)). Finally, in order to maintain the exposure's relationship with event time, consistent with the MSM, event indicator $Y(t_k)$ (with value 1 indicating that the event occurred between time t_{k-1} and t_k) is generated, at each interval k, according to the algorithm proposed by Young et al.[30]:

- if $T_0 > \int_0^{t_k+1} e^{\beta A(t_j)} \mathrm{d}j$ then $Y(t_{k+1}) = 0$ (no event in the interval (t_k, t_{k+1})), or
- if $T_0 \leq \int_0^{t_k+1} e^{\beta A(t_j)} \mathrm{d}j$ then $Y(t_{k+1}) = 1$ and the event time $T = t_k + (T_0 - \int_0^{t_k} \exp\{\beta A(t_j)\} \mathrm{d}j) \exp\{-\beta A(t_k)\}$.

Here, β is the target parameter of interest—the log hazard ratio for the causal effect of exposure on the hazard of the event of interest. More details on this data generation process are available in Young et al.[30]

In most simulation scenarios, the above procedure was employed identically to generate data for both 10,000 subjects in the main database, and 1000 (or fewer, in certain scenarios) additional subjects in the validation sample. In this way, in the vast majority of our simulations, the validation sample was assumed to be drawn from the same population. Then, when analyzing the simulated data in the main database, all values of the confounder U_1 were "deleted". In a few scenarios certain assumptions were altered, as described below.

Across simulated scenarios, different parameters were manipulated in order to investigate the sensitivity of the alternative methods to the corresponding assumptions. (i) The log odds ratio for the association between $U_1(t_k)$ and exposure $A(t_k)$ was varied to explore the effect of the direction of unmeasured confounding. (ii) The log odds ratio for the relationship between U_1 and the counterfactual event time, T_0, (γ_1 in equation (8)) was altered to make U_1 a stronger risk factor. (iii) An informative censoring scenario was also considered, where censoring probability depended on the counterfactual event time T_0. (iv) In different scenarios, the size of the validation sample was decreased to either 500 or 250, with 50 or 25 events, respectively. (v) Finally, in two scenarios the event rate in the validation sample was increased, by 50% or 100%, respectively, relative to the main database, to assess if and how the performance of the proposed estimators changes if the validation sample arises from a different (higher-risk) population than the main database.[20] Parameters used for the simulation scenarios are given in Table 1. More details on the data generation process, distributions of relevant variables, and parameters used, are available in online Appendix A.

3.2 Analysis of simulated datasets

In total, we investigated eight different simulation scenarios, corresponding to parameter sets shown in Table 1. For each scenario, we analyzed 1000 generated datasets. Each sample was analyzed with the following five MSM Cox models, including the methods proposed in Section 2.3: (1A) regression calibration of PS (RC); (1B) regression calibration of PS, including the MR in the measurement error model (RC + MR); (2A) MR-based

Table 1. Parameters used for the simulation scenarios.

	Scenario	$T_0 - U_1$ relationship	*Log* (OR_{U1})	$n_{V\ S}$	Event rate (VS)
1	Base	−0.1053	−1.0986	1000	0.10
2	U bias +	−0.1053	1.0986	1000	0.10
3	U stronger risk factor	−0.2231	−1.0986	1000	0.10
4	VS size 500	−0.1053	−1.0986	500	0.10
5	VS 1.5 × event rate	−0.1053	−1.0986	1000	0.15
6	VS 2 × event rate	−0.1053	−1.0986	1000	0.20
7	Informative censoring	−0.1053	−1.0986	1000	0.10
8	Continuous U	−0.1000	−0.0500	1000	0.10

multiple imputation of unmeasured confounders (MR-based MI); and (2B) a slight modification of the MR-based MI method (2A) above, where the prior exposure was included in the imputation model (MR-based MI + A_{-1}). The last method (equation (3)) simply relied on a naïve marginal structural Cox model, in which IPT weights were based only on the confounders $C_1(t)$ and $C_2(t)$ fully measured for all subjects in the main database, without any attempt to address potential confounding due to $U(t)$ (Naïve).

IPTW may be sensitive to very large weights, which may result if "improbable" treatment histories are observed in some individuals,[9] which is likely to occur, at least a few times, in large-scale simulations.[21] Extremely large weights, which reflect the near-violation of the positivity assumption, may lead to highly inflated variance of the resulting MSM estimates[9] and, especially if associated with actual event in time-to-event analyses, may produce point estimates that diverge dramatically from the true effect.[21] One method of reducing this variance is through truncation of large weights, although there is little consensus over what truncation criteria should be used.[21] While truncation is a valuable tool for reducing the variance of IPTW estimates, it may also sometimes introduce bias.[9] We therefore implement, for each of the MSMs, two alternative truncations, one at an arbitrarily selected absolute cut-off of 30 and another at the 99.9th percentile of the sample-specific empirical distribution of the "corrected" time-varying IPT weights, across all subjects and all time points. We then assess the impact of either truncation strategy on the bias, variance and root mean square error (RMSE) of the estimates.

We then compared the performance of the alternative estimators of the marginal log HR for the exposure, in terms of absolute bias, calculated as the difference between the mean of the estimates from 1000 simulated samples and the true log HR, used to generate the data; standard deviation of the 1000 estimates (SD); and RMSE.

3.3 Simulation results

In preliminary simulations, we analyzed the simulated data using the ideal (unrealistic) model, in which—for all subjects in the main database—the IPT weights were estimated using all confounders C_1, C_2, and U. In that case, as expected, this "gold standard" MSM Cox yielded unbiased estimates of the marginal hazard ratio for the time-varying exposure (Table A2 in online Appendix B), which provides indirect evidence of the validity of our data generation procedures, outlined in Section 3.1.

The results from the eight more "realistic" simulation scenarios, with U measured only in the validation sample, are shown in Tables 2 and 3 (see online Appendix A for detailed assumptions and values of the relevant parameters for each scenario). Truncation of IPT weights at either the pre-specified cut-off of 30 or the 99.9th percentile of their sample distribution yielded very similar results, and, therefore, only those results obtained using either truncation at the 99.9th percentile or un-truncated weights are shown. (Results of truncation at 30 are shown in online Appendix B.)

In all scenarios, for both truncated and un-truncated weights, MR-based multiple imputation (2A: MR-based MI) and MR-based MI including prior exposure (2B: MR-based MI + A_{-1}) substantially reduced bias compared with a naïve MSM with IPT weights based only on C_1, C_2 (Table 2). In contrast, regression calibration (1A: RC) of the weights without the MR slightly *increased* bias, for scenarios 1, and 3 to 8 (Table 2). This may be related to possible violation of surrogacy assumption, known to strongly affect the accuracy of PSC estimates,[16] but its implications in the context of more complex, time-varying confounding remains to be explored. RC estimates were less biased than the naïve estimates only in scenario 2, where all other proposed methods eliminated the bias (Table 2).

Table 2. Simulation results: bias of the estimates for the log hazard ratio for exposure from each of the five methods: (1) naïvely weighted MSM (Naïve), (2) regression calibration of the propensity score (RC), (3) regression calibration including the martingale residual (RC + MR), (4) martingale residual -based multiple imputation of unmeasured confounders (MR-based MI), (5) MR-based MI including prior exposure in the imputation model (MR-based MI + A_{-1}).

	Scenario	Trunc.	Naïve	1A RC	1B RC + MR	2A MR-based MI	2B MR-based MI + A_{-1}
1	Base	–	−0.240	−0.278	−0.146	−0.069	−0.064
		99.9%	−0.240	−0.278	−0.148	−0.073	−0.070
2	U bias +	–	0.195	0.150	0.003	0.006	0.019
		99.9%	0.175	0.141	−0.005	−0.008	−0.001
3	U stronger	–	−0.388	−0.425	−0.141	−0.092	−0.087
	risk factor	99.9%	−0.388	−0.422	−0.173	−0.097	−0.095
4	VS size	–	−0.239	−0.279	−0.145	−0.066	−0.063
	500	99.9%	−0.244	−0.281	−0.150	−0.074	−0.071
5	VS 1.5×	–	−0.242	−0.282	−0.184	−0.118	−0.111
		99.9%	−0.244	−0.283	−0.188	−0.121	−0.118
6	VS 2×	–	−0.245	−0.282	−0.205	−0.145	−0.141
	event rate	99.9%	−0.247	−0.283	−0.208	−0.149	−0.147
7	Informative	–	−0.243	−0.283	−0.139	−0.060	−0.056
	censoring	99.9%	−0.244	−0.283	−0.143	−0.063	−0.061
8	Continuous U	–	−0.156	−0.193	−0.098	−0.039	−0.051
		99.9%	−0.169	−0.200	−0.107	−0.063	−0.069

Results from models with un-truncated weights, and with weights truncated at 99.9th percentile, are shown.

Table 3. Simulation results: relative standard deviation (SD) and relative root mean squared error (RMSE), relative to naïve untruncated MSM.

			SD relative to naïve untruncated					RMSE relative to naïve untruncated				
	Scenario	Trunc.	Naïve	1A RC	1B RC+ MRMR	2A MR-based MI	2B MR-based MI + A_{-1}	Naïve	1A RC	1B RC+MR	2A MR-based MI	2B MR-based MI + A_{-1}
1	Base	–	1.00	0.65	0.69	1.00	1.01	1.00	1.07	0.63	0.53	0.52
		99.9%	0.67	0.57	0.62	0.71	0.72	0.94	1.06	0.62	0.43	0.42
2	U bias +	–	1.00	0.81	0.89	1.06	1.09	1.00	0.78	0.37	0.45	0.47
		99.9%	0.79	0.75	0.83	0.87	0.87	0.88	0.73	0.35	0.37	0.37
3	U stronger	–	1.00	0.66	1.88	1.02	1.01	1.00	1.06	0.69	0.39	0.39
	risk factor	99.9%	0.63	0.55	0.65	0.70	0.71	0.97	1.04	0.47	0.33	0.32
4	VS size 500	–	1.00	0.66	0.76	1.05	1.03	1.00	1.08	0.64	0.55	0.53
		99.9%	0.68	0.59	0.69	0.78	0.78	0.96	1.08	0.64	0.45	0.45
5	VS 1.5×	–	1.00	0.66	0.68	0.97	0.99	1.00	1.07	0.74	0.62	0.61
	event rate	99.9%	0.66	0.57	0.59	0.68	0.68	0.94	1.07	0.74	0.54	0.54
6	VS 2×	–	1.00	0.59	0.61	0.98	0.97	1.00	1.04	0.79	0.71	0.69
	event rate	99.9%	0.60	0.51	0.52	0.61	0.62	0.93	1.04	0.78	0.61	0.60
7	Informative	–	1.00	0.64	0.69	0.98	0.99	1.00	1.05	0.60	0.54	0.54
	censoring	99.9%	0.63	0.54	0.58	0.67	0.68	0.92	1.04	0.59	0.41	0.41
8	Continuous U	–	1.00	0.57	0.60	0.92	0.93	1.00	0.94	0.61	0.70	0.72
		99.9%	0.49	0.41	0.45	0.52	0.52	0.82	0.92	0.57	0.47	0.49

In all scenarios, the bias in the RC estimates was reduced, often by about 50% or more, by including the MR in the calibration model (1B: RC + MR). However, with the exception of scenario 2, a direct MR-based multiple imputation of the unmeasured confounder $U(t)$ (2A: MR-based MI) systematically outperformed RC + MR yielding lower bias (Table 2). Finally, inclusion of prior exposure in the imputation model (2B: MR-based MI + A_{-1}) did not further reduced the (small) bias of MR-based MI estimates (two last columns of Table 2). In most scenarios, IPT weight truncation did not affect the bias of the estimates from any of the five methods, but improved their numerical stability.

The MR-based MI estimates showed appreciable bias (although substantially lower than naïve and both RC-based estimators) only in scenarios 5 and 6, where the validation sample had 50% or 100% higher event rate than the main database (Table 2). In contrast, the performance of the RC estimator was not affected by the difference in the event rates (column 4), because it does *not* use information on the events in the validation sample.[16,17] From this perspective, the fact that even with doubled event rate, the MR-based estimates had bias lower by about 30% than the RC + MR estimate (scenario 6) is encouraging.

Standard deviations and RMSEs of the different estimates, relative to the un-truncated naïve estimate, are shown in Table 3. As expected, truncation systematically reduced the SD of the estimates (lower vs upper row, for each method in the left part of Table 3). Overall, RC yielded the lowest SD, whereas MR-based MI estimates had SDs mostly similar to the naïve estimate.

In terms of the overall accuracy of the estimates, as quantified by RMSE, for truncated weights both MR-based MI approaches (last two columns of Table 3) performed almost uniformly optimally. Only in scenario 2 the RC + MR method yielded minimally lower RMSE of 0.35 vs. 0.37 for MR-based MI.

Truncation at the 99.9th percentile of weights did reduce RMSE for both MR-based MI methods for all scenarios, but it yielded somewhat higher RMSE for RC estimates in all scenarios, and for RC + MR estimates in some scenarios. The fact that truncation benefited MR-based MI methods can be attributed to their relatively higher variance of IPT weights, compared to the RC methods. Although there is little consensus over what cut-off for IPT weight truncation to use, it does appear to be generally beneficial to perform some truncation for these methods.[21]

4 Application

We illustrate an application of these methods in a real-life analysis of a cohort (derived from the Truven Health MarketScan® Research Databases) of diabetic patients initiating DPP-4 inhibitor therapy, after having failed second-line therapy (with metformin and sulfonylurea). Specifically, we investigate whether patients who continue DPP-4 inhibitor therapy have reduced risk of hospitalization due to hypoglycemia, compared to those who stopped using the drug.[31–33] DPP-4 inhibitors are a relatively new class of drugs used as a treatment for diabetes, which have been shown to be effective in lowering blood glucose levels. Hypoglycemia is a relatively common adverse effect of glucose-lowering therapies for diabetes, and has been associated with a range of negative health outcomes.[34] Different observational studies have either found an increased risk of hypoglycemia at high values of HbA1c,[35] or at low values,[36] or at both high and low values of HbA1c.[34]

The Truven Health MarketScan Commercial Claims and Encounters database provides information on patient characteristics recorded at the time of each medical visit, as well as demographic information (age, sex, employment status, etc.). Additional data on repeated measures of laboratory tests are available for only for a relatively small subset of patients, in the Truven Health MarketScan Lab Database. Among these laboratory tests, hemoglobin A1C (HbA1c) may be considered a potentially important time-varying confounder/mediator for the DPP-4 inhibitors effect on hypoglycemia. Indeed, the most recent HbA1c value indicates glycemic control (or lack thereof)[32] and may therefore both affect the treatment decision and be related to the outcome. Furthermore, use of DPP-4 inhibitors is expected to affect HbA1c level, making it a potential time-varying mediator of the treatment effect.[37] For this example, therefore, we treat the subset of patients for whom HbA1c measurements are available as the validation sample, while using all eligible type II diabetes patients in the large, main Marketscan database for the analyses of the possible association between DPP-4 inhibitors and the hazard of hospitalization for hypoglycemia.

4.1 Data preparation and cohort definition

The cohort included all patients with type II diabetes, who were considered "new users" of DPP-4 inhibitors, i.e. had at least one prescription for DPP-4 inhibitors between 1 January 2011 and 30 December 2014, and no history of use in the 12 months prior to the first prescription, which defined the date of cohort entry. Inclusion criteria included prior use of both metformin and sulfonylurea in the 12 months before cohort entry, so that DPP-4 inhibitors were a third-line therapy for all cohort members.[33,38] Type II diabetes was identified by at least one specific ICD-9 code for type II diabetes mellitus between one year before and one month after cohort entry, and no code for type 1 diabetes in the same period. Online Appendix C lists all ICD-9 codes used as inclusion criteria. Patients with any record of pregnancy or gestational diabetes were excluded. Patients were followed up until occurrence of the outcome (hospitalization due to hypoglycemia), or censoring due to (1) death, (2) end of study period (31 December 2014), (3) loss of coverage, or (4) initiation of insulin, considered alternative third-line therapy[33]).

Inclusion and exclusion criteria for the validation sample were the same but, in addition, these subjects were required to have at least one HbA1c measurement in the 6 months prior to, and/or up to 3 months after cohort entry. In the analyses, the most recent value of HbA1c for a given subject was carried forward until a new, updated value was recorded.

4.1.3. Exposure definition. We constructed the binary time-varying indicator of current exposure to a DPP-4 inhibitor based on the subject's history of the dates of start and durations of all DPP-4 inhibitors prescriptions. All subjects were exposed at their respective dates of entry into the cohort. At each subsequent visit to a healthcare professional at which either (i) a new prescription to DPP-4 inhibitors or another antidiabetic drug of a different class was given, or (ii) a lab test was performed, the current DPP-4 inhibitor exposure status was either changed or kept the same. Details on time-varying exposure definition are available in online Appendix D.

4.2 IPT weights

We calculated IPT weights for each healthcare professional visit during which the decision to prescribe DPP-4 inhibitors may or not may have occurred. Details are described in online Appendix D.

As in simulations (Section 3), stabilized IPT weights were used to avoid variance inflation. Logistic regression models were used to calculate the PS for the numerator of the stabilized weights, conditional on all baseline covariates, while the probabilities in the denominator were based on both baseline and time-varying covariates. Baseline covariates included: age, sex, employment status, Charlson Comorbidity Index,[39] and the binary indicators of the following measures of healthcare services utilization, in a 1-year period before the cohort entry: ≥ 1 visit to the emergency department; ≥ 1 hospitalization; ≥ 20 physician visits; and prior hypoglycemic event. A potential time-varying confounder/mediator, available for all subjects in the main database, was a binary indicator of a current prescription for another anti-diabetic agent (Metformin, Sulfonylurea, Thiazolidinedione). In contrast, time-varying measures of HbA1c test results were available only for subjects in the validation sample (see above), but unmeasured in the main database. Squared and cubic terms were considered while modeling the effect of HbA1c value on the logit of the propensity score PS (in the validation sample), to account for potential non-linearities.

4.3 Analyses

The analyses relied on the methods described in Section 2.3. Similarly to the analyses of the simulated data, we considered alternative models: (1A) regression calibration of the weights (RC), (1B) regression calibration of the weights, including the MR (RC + MR), and (2A) MR-based multiple imputation of the unmeasured confounder, HbA1c (MR-based MI), (3) a naïve MSM, with weights based only on the fully measured covariates (Naïve). However, because the validation sample is a subset of the main database, contrary to the analyses of the simulated data, in the final analysis of the main database we included all subjects in the validation sample, and used their actual observed HbA1c values.

4.4 Results

Table 4 compares the characteristics of the 47,964 patients in the main data for whom lab tests were not available, with the 2341 in the validation sample, for whom at least one HbA1c measurement was recorded. As expected, the characteristics of both groups are quite similar (Table 4). The mean follow-up time in both groups was around 1.6 years, with the average time exposed to DPP-4 being around 36–42% of that time (249 days in the main database, and 216 days in the validation sample). During the follow-up, a total of 2867 events were observed in the main database excluding those subjects in the validation sample and 130 in the validation sample, yielding similar event rates of 3.7/100 person-years and 3.4/100 person-years, respectively.

Table 4. Descriptive values for the cohort of patients in the main data and in the validation sample, from the MarketScan databases.

	Main data ($N = 47{,}964$)	Validation sample ($n = 2341$)
Event (hospitalization due to hypoglycemia)	2867 (6%)	130 (5.6%)
Follow-up (days): mean (SD)	591.9 (360.9)	592.9 (372.4)
Days exposed to DPP-4 inhibitors: mean (SD)	248.9 (231.3)	216.3 (199.7)
Baseline characteristics		
Age (years): mean (SD)	58 (11.1)	56.5 (10.1)
Charlson index: mean (SD)	0.4 (0.9)	0.4 (0.9)
Female: n (%)	19348 (40.3%)	1011 (43.2%)
Employed: n (%)	17875 (37.3%)	570 (24.3%)
Characteristics in year prior to cohort entry:		
Emergency department visit: n (%)	10000 (20.8%)	539 (23%)
Hospitalization: n (%)	4528 (9.4%)	175 (7.5%)
$\geq$ 20 physician visits: n (%)	5495 (11.5%)	223 (9.5%)
Prior hypoglycemic event: n (%)	1925 (4%)	113 (4.8%)
Time-varying characteristics		
Days exposed to other anti-diabetic agents: mean (SD)	475.6 (361)	458.8 (365.2)
Baseline HbA1c value: mean (SD)	–	8.7 (1.7)
Number of HbA1c tests: mean (SD)	–	2.3 (1.8)

Table 5. Mean and standard deviation (SD) for weights (truncated at 99.7th percentile) for each of the methods.

Method	Mean (SD)	Maximum
Naïve	1.0 (2.5)	36.6
Regression calibration	0.9 (1.5)	17.4
Regression calibration + MR	0.9 (1.5)	17.4
MR-based imputation (single)	1.0 (2.5)	37.2

Table 6. Results (hazard ratio, and 95% confidence interval (CI) based on robust standard errors) for each of the methods in the applied example.

Method	HR	95% CI
Naïve	0.62	(0.47, 0.82)
Regression calibration	0.69	(0.59, 0.82)
Regression calibration + MR	0.69	(0.59, 0.82)
MR-based multiple imputation	0.74	(0.57, 0.97)
Gold standard (validation sample only)	0.84	(0.55, 1.27)

The "gold standard" weight denominator model included square and cubic terms for HbA1c, which were found to be significant in the validation sample. The estimated function, shown in online Appendix E, indicates that patients with both intermediate and very high recent HbA1c are more likely to be prescribed DDP-4 inhibitors.

To avoid extreme weights, for all four estimators, we truncated the corresponding IPT weights[9] at the 99.7th percentile of their distribution, pooled across all subjects and all visits. As a result, the maximum truncated weights for each method is less than 40, and their mean is close to 1 (Table 5).

Final estimates of marginal hazard ratio, with 95% CI, obtained with each method are shown in Table 6, in addition to the estimate from the gold-standard weighted model for the validation sample (last row). For all methods, current exposure to DPP-4 inhibitors was found to be associated with a statistically significant decrease in the hazard of hospitalization due to hypoglycemia, by 26–38% (Table 6: HR = 0.62 (Naïve) to 0.74 (MR-based

MI)). The MR-based MI method suggested that the protective effect of current DPP-4 inhibitor use against hypoglycemia (26% reduction: HR = 0.74; 95% CI 0.57–0.97) may be about 30% weaker than that suggested by the naïve analysis that ignored potential unmeasured confounding by HbA1c values (HR = 0.74; 95% CI 0.57–0.97). The MR-based MI point estimate was also the closest to the "gold standard" estimate obtained using only the validation sample and taking into account actual Hb1Ac values for individual subjects in that sample (Table 6). Finally, both regression calibration methods produced identical estimates, that suggested a weaker protective effect (HR = 0.69; CI: 0.57–0.82) than the naïve estimate but somewhat stronger than the MR-based MI estimate.

On the other hand, the largely overlapping confidence intervals, and similar point estimates, suggest that the impact of correcting the weight for the unmeasured HbA1c is rather minor. To determine why this was so, we investigated the association of HbA1c with the hazard of hypoglycemia hospitalization in a conditional Cox model, fit in the validation sample. Contrary to initial expectations, we found the most recent HbA1c value not to be strongly associated with the outcome (HR for 1 SD increase in Hb1Ac = 0.95, 95% CI 0.78–1.17, p = 0.65). Therefore, although it was an important predictor of DPP-4 inhibitor exposure (with a significant non-linear relationship shown in online Appendix E), HbA1c is not an important confounder of the treatment-outcome association.

Performing a "complete case" analysis in the validation sample, where the denominator for the weight included linear, square, and cubic terms for HbA1c, we get a similar point estimate (Table 6, HR = 0.84), but, as expected, a much higher standard error, resulting in the estimate not being statistically non-significant, making it difficult to interpret (95% CI 0.55–1.27). This finding provides an *a posteriori* support for our decision not to merely rely on complete case analysis of the validation sample, and suggests the proposed MR-based MI may approximate the resulting "gold standard" estimate while improving considerably its numerical stability and, thus, the power to detect an effect.

5 Discussion

In many observational studies, which aim to uncover the causal effect of an exposure on some outcome, unmeasured confounding remains an important problem, and the development of methods which may help address this problem is a highly active area of research. Methods, such as instrumental variables,[40] high-dimensional PS,[41] missing cause,[42] and bias sensitivity analyses,[43] have been developed in order to reduce the impact of potential unmeasured confounding under specific assumptions about the data structure. Recently, alternative methods have been proposed for studies in which information on potential confounders not measured in the main database is available in a relatively small validation sample.[16,18–20] However, the range of analytical models to which these methods are applicable is currently limited. In particular, little progress has been achieved in the last decade in dealing with the unmeasured confounding problem in the context of MSMs.[7] Indeed, both methodological papers that develop new MSM models, and real-life applications of the MSM methods, typically rely on the usually un-testable assumption that all important confounders are measured for all study subjects.[4,3,22,44,45] In this paper, we proposed and investigated the performance of several methods, developed particularly to address the challenge of time-varying unmeasured confounding in the context of Cox MSMs. Specifically, we have focused on situations where some time-varying confounder(s)/mediator(s) are measured only in a relatively small validation sample, drawn from the same population.

Simulations indicated the satisfactory performance of the proposed method, which adapted our recent MR (MR)-based approach, originally developed for conventional unweighted Cox model,[20] to IPT weight estimation in Cox MSM. Most importantly, across the simulated scenarios, our MR-based method substantially reduced the bias in the exposure effect obtained from fitting a Cox MSM, compared to a similar MSM which used only confounders fully measured in the main database to estimate IPT weights. Both MR-based multiple imputation methods performed uniformly better, in terms of bias and RMSE, than an alternative method which adapted a regression calibration-like approach to correct the PS used for calculation of the weights. On the other hand, the bias and RMSE of regression calibration of the PS were improved in all scenarios considered by inclusion of the MR in the calibration equation, confirming the advantages of using MRs to account for the outcomes observed for individual subjects.[20]

In most simulated scenarios, the validation sample was assumed to simply be a random sample of the main database. In scenarios 5 and 6 where this assumption was avoided and the validation sample had a higher event rate than the main data, the MR-based imputation methods yielded somewhat biased estimates of the marginal HR for the exposure, even if their bias was still substantially lower than that of the naïve MSM estimates and of the RC-based estimates. This suggests that, as expected, our method works best when the validation sample is drawn from a population similar to the main database, and further investigation is required to assess how particular differences between the two populations may affect the accuracy of our MR-based imputation estimates.

To illustrate a real-life application of our method, we re-assessed the safety of DPP-4 inhibitors in a cohort of type II diabetes patients, where repeated measurements of HbA1c were available only for a small subset of patients. Although this laboratory test did not turn out to be associated with the hazard of the hospitalization for hypoglycemia and, thus, did not act as an important confounder of its association with DPP-4 inhibitors use, the results are clinically useful. Indeed, because recent HbA1c values are important predictors of the decision to continue, start or interrupt DPP-4 inhibitor treatment, and could be expected to mediate some of its effect on the outcome, the validity of the results of the conventional analyses of the full Marketscan database, which did not account for this potential time-varying mediator, could be questioned. On the other hand, the complete case analyses, limited to a small subsample for whom HbA1c measurements were available, did not have sufficient power, as reflected in a statistically non-significant estimate, with wide confidence intervals (HR = 0.84 95% CI: 0.55–1.27). In contrast, our MR-based imputation of HbA1c values allowed including them in the calculation of the IPT weights for all subjects in the full cohort, and confirmed a statistically significant and clinically important protective effect of continuing DPP-4 inhibitor therapy, which may reduce the hazard of hypoglycemia hospitalization by more than 25% (HR = 0.74, 95% CI: 0.57–0.97).

Similarly to other methods developed for the specific situations where unmeasured confounders are available in a validation sample, MR-based imputation should not be considered a Panaceum for addressing the paramount challenge of unmeasured confounding bias, and should be used with appropriate caution. Firstly, when the validation sample had a substantially higher event rate, our method did not completely eliminate bias even if, in our simulations, it still yielded more accurate estimates than the alternative methods we considered. This may occur if the validation sample arose from a sicker or more vulnerable population, which is plausible if additional time-varying confounders / mediators are available only for participants of a clinical research study, limited to patients from university-based specialized clinics. Further investigation is required to establish how potential differences between the validation sample and the main database, related either to distributions of relevant variables (exposure, measured and unmeasured confounders, event times) or the way they are associated with each other, may affect the accuracy of our MR-based imputation Cox MSM estimates, both in absolute terms and relative to alternative approaches. Furthermore, while our method aims to reduce the impact of unmeasured confounding by variables available only in the validation sample, it relies on an implicit assumption of no *further* unmeasured confounding besides these variables. For applications where it is expected that some additional, important confounders may not be measured even in the validation sample, future research should assess the possibility of combining our MR-based imputation with methods that do not require specifying the unmeasured confounders, such as IVs[40] or "missing cause".[42] However, from this perspective, it may be considered a refinement over the naïve MSM that relies on the stronger assumption that all relevant confounders are available for all subjects in the large main database.[4,44]

Future research may involve comparisons of our approaches with alternative methods for imputing the values of time-varying confounders/mediators measured only in a small validation sample. However, at present, it is unclear which among the existing methods could improve the accuracy of the MSM estimates over that achieved by the proposed MR-based imputation. Indeed, any "conventional" multiple imputation method that does not account for the outcome is expected to yield less accurate imputation of confounders, that—by definition—have important associations with the outcome.[24] On the other hand, in time-to-event analyses the outcome is bivariate, combining information on the follow-up duration and on the status at the end of follow-up. This makes it more difficult to represent the "outcome" in the model used for imputation, resulting in a limited range of options, of which using the logarithm of the follow-up duration, of a given subject, seems most popular.[25] Yet, in our previous simulations, in the context of the conventional un-weighted Cox regression, without mediation, we found that our method yielded corrected treatment effect estimates that were either equally accurate (in terms of bias and mean squared error) or, especially with wider variation of censoring times, more accurate than those based on "log t" imputation.[20]

The development of methods which reduce the impact of unmeasured confounding is likely to be a growing area of research for many years to come, with new methods attempting to address this paramount challenge for particular data structures and/or a particular class of statistical models. Our hope is that the proposed MR-based imputation may assist in improving the accuracy of the estimated causal effects of time-varying exposures in MSM time-to-event analyses of large prospective or retrospective cohorts, where data on time-varying unmeasured confounders are available in a smaller validation sample.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Government of Canada through the Canadian Institutes of Health Research (CIHR) grant #PJT-148946, the Natural Sciences and Engineering Research Council (NSERC) grant no. 228203, and by a studentship from the Research Institute of the McGill University Health Centre.

References

1. Choi HK, Hernán MA, Seeger JD, et al. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *Lancet* 2002; **359**: 1173–1177.
2. Cole SR, Hernán MA, Robins JM, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol* 2003; **158**: 687–694.
3. Roumie CL, Greevy RA, Grijalva CG, et al. Association between intensification of metformin treatment with insulin vs sulfonylureas and cardiovascular events and all-cause mortality among patients with diabetes. *J Am Med Assoc* 2014; **311**: 2288–2296.
4. Hernán MA, Brumback B and Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**: 561–570.
5. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
6. Xiao Y, Abrahamowicz M and Moodie EE. Accuracy of conventional and marginal structural Cox model estimators: a simulation study. *Int J Biostat* 2010; **6**: 1–28.
7. Brumback BA, Hernán MA, Haneuse SJ, et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat Med* 2004; **23**: 749–767.
8. Moodie EE, Delaney JA, Lefebvre G, et al. Missing confounding data in marginal structural models: a comparison of inverse probability weighting and multiple imputation. *Int J Biostat* 2008; **4**: 1–23.
9. Cole SR and Hernàn MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–664.
10. Avorn J. In defense of pharmacoepidemiology – embracing the yin and yang of drug research. *N Engl J Med* 2007; **357**: 2219.
11. Walker AM. Confounding by indication. *Epidemiology* 1996; **7**: 335–336.
12. Patorno E, Garry EM, Patrick AR, et al. Addressing limitations in observational studies of the association between glucose-lowering medications and all-cause mortality: a review. *Drug Saf* 2015; **38**: 295–310.
13. Wolfe F and 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 Rheum* 1998; **41**: 1072–1082.
14. Franklin JM, Eddings W, Schneeweiss S, et al. Incorporating linked healthcare claims to improve confounding control in a study of in-hospital medication use. *Drug Saf* 2015; **38**: 589–600.
15. Wood ME, Frazier JA, Nordeng HM, et al. Prenatal triptan exposure and parentreported early childhood neurodevelopmental outcomes: an application of propensity score calibration to adjust for unmeasured confounding by migraine severity. *Pharmacoepidemiol Drug Saf* 2016; **25**: 493–502.
16. Stürmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005; **162**: 279–289.
17. Stürmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration – a simulation study. *Am J Epidemiol* 2007; **165**: 1110–1118.
18. McCandless LC, Richardson S and Best N. Adjustment for missing confounders using external validation data and propensity scores. *J Am Stat Assoc* 2012; **107**: 40–51.
19. Lin HW and Chen YH. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *Am J Epidemiol* 2014; **180**(3): 308–317.
20. Burne RM and Abrahamowicz M. Martingale residual-based method to control for confounders measured only in a validation sample in time-to-event analysis. *Stat Med* 2016; **35**: 4588–4606.
21. Xiao Y, Moodie EE and Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiol Methods* 2013; **2**: 1–20.
22. Xiao Y, Abrahamowicz M, Moodie EE, et al. 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. *J Am Stat Assoc* 2014; **109**: 455–464.

23. Rosner B, Willett W and Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989; **8**: 1051–1069.
24. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; **59**: 1092–1101.
25. White IR and Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009; **28**: 1982–1998.
26. Therneau TM, Grambsch PM and Fleming TR. Martingale-based residuals for survival models. *Biometrika* 1990; **77**: 147–160.
27. Breslow N. Covariance analysis of censored survival data. *Biometrics* 1974; **30**: 89–99.
28. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons, 1987.
29. R Core Team. *R: a language environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
30. Young JG, Hernän MA, Picciotto S, et al. Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Anal* 2010; **16**: 71–84.
31. Canadian Diabetes Association Clinical Practice Guidelines Expert Committee. Pharmacologic management of type 2 diabetes. *Can J Diabetes* 2013; **37**: S61–S68.
32. Nathan DM, Buse JB, Davidson MB, et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy a consensus statement of the American Diabetes Association and the European Association for the study of diabetes. *Diabetes Care* 2009; **32**: 193–203.
33. Tricco AC, Antony J, Khan PA, et al. Safety and effectiveness of dipeptidyl peptidase-4 inhibitors versus intermediate-acting insulin or placebo for patients with type 2 diabetes failing two oral antihyperglycaemic agents: a systematic review and network meta-analysis. *BMJ Open* 2014; **4**: e005752.
34. Lipska KJ, Warton EM, Huang ES, et al. Hba1c and risk of severe hypoglycemia in type 2 diabetes the diabetes and aging study. *Diabetes Care* 2013; **36**: 3535–3542.
35. McCoy RG, Van Houten HK, Ziegenfuss JY, et al. Increased mortality of patients with diabetes reporting severe hypoglycemia. *Diabetes Care* 2012; **35**: 1897–1901.
36. Miller CD, Phillips LS, Ziemer DC, et al. Hypoglycemia in patients with type 2 diabetes mellitus. *Arch Internal Med* 2001; **161**(13): 1653–1659.
37. Dicker D. Dpp-4 inhibitors impact on glycemic control and cardiovascular risk factors. *Diabetes Care* 2011; **34**(Suppl 2): S276–S278.
38. Lozano-Ortega G, Goring S, Bennett H, et al. Network meta-analysis of treatments for type 2 diabetes mellitus following failure with metformin plus sulfonylurea. *Curr Med Res Opin* 2016; **32**: 807–816.
39. Charlson M, Szatrowski TP, Peterson J, et al. Validation of a combined comorbidity index. *J Clin Epidemiol* 1994; **47**: 1245–1251.
40. Brookhart MA, Wang P, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**: 268–275.
41. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**: 512.
42. Abrahamowicz M, Bjerre LM, Beauchamp ME, et al. The missing cause approach to unmeasured confounding in pharmacoepidemiology. *Stati Med* 2016; **35**: 1001–1016.
43. Groenwold RHH, Nelson DB, Nichol KL, et al. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *Int J Epidemiol* 2010; **39**: 107–117.
44. Daniel RM, Cousens S, De Stavola BL, et al. Methods for dealing with timedependent confounding. *Stat Med* 2013; **32**: 1584–1618.
45. Bembom O, van der Laan M, Haight T, et al. Leisure-time physical activity and all-cause mortality in an elderly cohort. *Epidemiology* 2009; **20**: 424–430.