Multiple imputation for systematically missing confounders within a distributed data drug safety network: a simulation study and real-world example

Running head: Multiple imputation in distributed drug safety networks

Authors*: Matthew H. Secrest[1], Robert W. Platt[1,2,3], Pauline Reynier[1], Colin R. Dormuth[4], Andrea Benedetti[2,5], Kristian B. Filion[1,2,6]

[1] Centre for Clinical Epidemiology, Lady Davis Research Institute, Jewish General Hospital, McGill University, Montreal, Quebec, Canada

[2] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

[3] Department of Pediatrics, McGill University, Montreal, Quebec, Canada

[4] Department of Anesthesiology, Pharmacology, and Therapeutics, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada

[5] Respiratory Epidemiology and Clinical Research Unit, McGill University Health Center, Montreal, QC

[6] Division of Clinical Epidemiology, Department of Medicine, McGill University, Montreal, Quebec, Canada.

[*] MHS is currently employed at IQVIA (Cambridge, MA) and is no longer employed by the Centre for Clinical Epidemiology.

Address for Correspondence:
Kristian B. Filion, PhD
Assistant Professor and William Dawson Scholar
Departments of Medicine and of Epidemiology, Biostatistics, and Occupational Health, McGill University
Centre for Clinical Epidemiology, Lady Davis Research Institute, Jewish General Hospital
3755 Cote Ste-Catherine, H-410.1
Montreal, Quebec H3T 1E2 Canada
Tel: (514) 340-8222 x 28394
Fax: (514) 340-7564
E-mail: kristian.filion@mcgill.ca

Keywords: bias, cohort study, confounding, distributed data network, missing data, multiple imputation, simulation study

Key points:

- Residual confounding may bias pooled effect estimates in studies conducted in distributed data drug safety networks if important confounders are systematically missing at a data site

- Multiple imputation can be used to impute data that are systematically missing at a data site, allowing for statistical adjustment for these missing variables

Word Count (text): 4,016; **Word Count** (abstract): 248; **Tables**: 4; **Figures**: 1

Prior posting and presentation: This work is the sole product of the authors. It was presented as an abstract at the 2018 International Conference for Pharmacoepidemiology and Therapeutic Risk Management in Prague, Czech Republic:

Secrest MH, Platt RW, Reynier P, Dormuth CR, Benedetti A, Filion KB. Multi-database multiple imputation within a distributed data drug safety network: A simulation study and real-world data example. In PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2018 Aug 1 (Vol. 27, pp. 29-30). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.

Research sponsor: This research is funded by the Canadian Network for Observational Drug Effects Studies (CNODES), a collaborating center of the Drug Safety and Effectiveness Network (DSEN), funded by the Canadian Institutes of Health Research (Grant Number DSE-146021).

Conflicts of interest: RWP has received personal fees from Amgen, Biogen, Merck, Eli Lilly and Pfizer. The authors declare no other conflict of interest.

ABSTRACT

Purpose:

In distributed data networks, some data sites may be systematically missing important confounders that are captured by other sites in the network (e.g., body mass index [BMI]). Multiple imputation may help repair bias in these scenarios. However, multiple imputation has not been described for distributed data networks where data access restrictions prevent centralized analysis.

Methods:

We conducted a simulation study and a real-world analysis using the UK's Clinical Practice Research Datalink to evaluate multiple imputation for confounders that are systematically missing from a subset of data sites in mock distributed data networks. The simulation study addressed univariate missing data, while the real-world analysis addressed multivariate missing data. Both studies were designed as retrospective cohort studies of the effect of current statin use on the risk of myocardial infarction among patients with newly-treated type 2 diabetes.

Results:

In our simulation study, multiple imputation repaired bias from missing BMI in all scenarios, with a median bias reduction of 118% in the default scenario. In our real-world study, the multiply-imputed analysis (hazard ratio [HR]: 0.86; 95% confidence interval [CI]: 0.69-1.08) was closer to the analysis that considered the true confounder values (HR: 0.85; 95% CI: 0.66-1.10) than the analysis that ignored them (HR: 0.93; 95% CI: 0.73-1.20).

Conclusions:

Multiple imputation adapted to distributed data settings is a feasible method to reduce bias from unmeasured but measurable confounders when at least one database contains the variables of interest. Further research is needed to evaluate its validity in real distributed data networks.

INTRODUCTION

Distributed data drug safety networks such as the Canadian Network for Observational Drug Effect Studies (CNODES) play an important role in evaluating the safety and effectiveness of post-marketed medications through routine surveillance.[1] Because of data access restrictions, these networks commonly use a two-step approach[2] to pooling results: first, a standard analysis plan is implemented at each database, and then the aggregate effect estimates from each site are pooled using meta-analysis. One intrinsic problem with meta-analyzing results from different jurisdictions is discrepant capture of important confounding variables such as body mass index (BMI) and smoking status. These unmeasured, but measurable, confounders can bias the summary estimate to the extent that they are missing from databases in the network.

Several reports from the individual patient data (IPD) meta-analysis literature have explored multiple imputation as a means to repair bias from missing data in multi-level settings similar to distributed data networks.[3-8] Resche-Rignon and colleagues[8] first applied multiple imputation by chained equations (MICE)[9,10] to address systematically missing data (*i.e.,* data that are missing for all individuals as a feature of the database or study design) in these settings. The authors have since expanded their MICE method to incorporate both data that are missing systematically (for all individuals in a database or data site as a feature of the data collection methods) and sporadically (for only a subset of individuals in the database or data site).[7] While these methods sufficiently address the needs of IPD meta-analyses, they are not applicable to distributed data drug safety networks such as CNODES because they require access to data from each study site at a central site to apply the MICE algorithm. We propose an application of multiple imputation in distributed data networks with data access restrictions that prevent IPD from leaving the study site.

In this paper, we first describe our proposed application of multiple imputation. We then explore the effects of varying different study parameters on the validity of multiple imputation for a univariate missing data problem in a simulation study that mimics a distributed data drug safety network wherein all but one database are missing BMI, a variable commonly missing from administrative health databases such as the provincial health databases in CNODES. Finally, we evaluate multiple imputation in a mock distributed data network based on real-world data from the UK's Clinical Practice Research Datalink (CPRD). Our real-world example investigates a multivariate missing data problem with four structurally missing variables that are commonly absent from administrative health databases such as those in CNODES but are present in electronic health databases such as the CPRD: BMI, smoking status, glycated hemoglobin (HbA1c), and cholesterol. Both simulated and real-world studies evaluate the effect of statins on the risk of myocardial infarction (MI) in patients with newly-treated type 2 diabetes, an association that has been previously described[11] and for which we can identify confounders[12,13] and predict their effects[14-16]. An analysis of statins on the risk of MI that does not consider smoking, obesity, HbA1c, and cholesterol would likely be biased towards the null.

METHODS

Multiple Imputation

In multiple imputation, missing values are randomly sampled from a posterior predictive distribution X times to create X distinct datasets, where $X \geq 2$.[17] These X datasets are analyzed separately and the resulting X effect estimates are pooled using Rubin's rules.[17] Our proposed method applies multiple imputation to a multi-level distributed data setting in which data cannot leave a site. In our method, a database containing all variables of interest, hereafter referred to as the "validation database", is used to produce predictive distributions from which draws are taken to impute values in the "missing databases", those databases structurally missing the variables of interest.

First, a posterior predictive distribution is developed in the validation database that predicts the missing variables of interest. The case of a single structurally missing variable is straightforward. Assume that a variable Y is observed for all individuals in the validation database and is structurally missing for all individuals in the missing databases. Further assume that V is a set of predictor variables of Y observed for all individuals in all databases. A single regression model at the validation database can be used to generate a posterior predictive distribution of Y conditional on V, or $F(Y \mid V)$, where prior distributions can either be specified or assumed to be non-informative. Conditional draws of Y can then be made for individuals in the missing databases for the purposes of multiple imputation using the distribution $F(Y \mid V)$ estimated by regression in the validation database. Operationally, posterior prediction model parameters must be communicated from the validation database to the missing databases.

For multivariate missing data, we propose approximating the joint posterior predictive distribution with a single application of chained equations. Consider two related variables of interest, Y_1 and Y_2, which are structurally missing in all databases except the validation database. The set of variables V is observed in all databases and can be used to predict Y_1 and Y_2. The joint posterior predictive distribution of Y_1 and Y_2 is therefore denoted as $F(Y_1, Y_2 | V)$, which is the product of conditional distributions $F(Y_1 | V)$ and $F(Y_2 | V, Y_1)$. That is, $F(Y_1, Y_2|V) = F(Y_1|V) \times F(Y_2|Y_1, V)$. Therefore, to approximate $F(Y_1, Y_2 | V)$, one can first estimate the distribution of a single missing variable, $F(Y_1 | V)$, using an appropriate regression model in the validation database. One can then use regression methods in the validation database to estimate Y_2 conditional on V and Y_1, $F(Y_2 | V, Y_1)$. Conditional draws of Y_1 can be made for individuals in the missing databases using the model-estimated distribution, $F(Y_1, Y_2 | V)$. The drawn values of Y_1 can then be used to produce conditional draws of Y_2 in the missing databases using the model-estimated distribution, $F(Y_2 | V, Y_1)$.

For our purposes, we assume that systematically missing data are missing completely at random (MCAR)[17]. For data that are MCAR, missingness is independent of all observed and unobserved covariates. This assumption is unreasonable in many situations, but it may be applicable to systematically missing data in distributed data network settings where missingness is due to administrative reasons unrelated to patient, physician, and other clinical characteristics. Further, the MCAR assumption is reasonable in situations where patients are subject to rigorous eligibility criteria. Thus, we argue that a single application of chained equations should produce unbiased conditional draws. This differs from the multiple imputation methods used in IPD meta-analyses[3-8], in which several iterations of the MICE algorithm are applied until convergence.[9,10]

Once all missing values have been imputed, the parameter of interest (e.g., log hazard ratio [HR]) is estimated at each site. The process of sampling values and generating parameter estimates is performed X times, where $X \geq 2$. The mean parameter estimate for X imputations is $\bar{\beta}_X =$

$\frac{1}{X}\sum_{x=1}^{X}\hat{\beta}_x$, where $\hat{\beta}_x$ is the estimate for each imputed dataset.[17] The variance of the summary estimate (T_X) is a function of two components, the average within-imputation variance ($\bar{W}_X$) and the between-imputation variance (B_X), given by $\bar{W}_X = \frac{1}{X}\sum_{x=1}^{X} W_x$ and $B_X = \frac{1}{X-1}\sum_{x=1}^{X}(\hat{\beta}_x - \bar{\beta}_X)^2$, respectively. T_X is determined as $T_X = \bar{W}_X + \frac{X+1}{X}B_X$.[17] Once estimated, $\bar{\beta}_X$ and T_X can be input into traditional meta-analytic models, as in studies with no missing data.

Simulation Study

Variables

We assumed a distributed data network of *D* databases with *n* new users of antihyperglycemic agents at each database. In this population, we assumed several baseline covariates are available in all databases: *Age* (age at cohort entry), *Dur* (duration since diabetes diagnosis), *Male* (male sex), and *Smoke* (smoking status [current, former, never]). *BMI* was assumed to be available in all but one database, the validation database. In addition, we assumed that a dichotomous, time-varying exposure variable *E* denoting statin use is recorded four times (E_1, E_2, E_3, E_4; Supplemental Section I) throughout each patient's follow-up. We assumed patients are censored for the outcome *MI* or at 20 years follow-up. We assumed no measurement error.

According to these assumptions, we simulated, for each individual in each database, *Age, BMI, Dur, Male, Smoke,* and E_1 (baseline exposure status) from a multivariate normal distribution with a pre-defined correlation matrix and marginal distributions, categorizing variables from their respective normal distributions, as appropriate, to meet pre-specified marginal distributions (Supplemental Tables I-II)[18]. *BMI* was simulated for patients in all databases, not just the validation database, allowing us to compare the 'true' simulated values of *BMI* with the imputed values. We simulated time-dependent statin use over the course of a 20-year follow-up period (Supplemental Section I).[19] We used an increasing Weibull distribution to represent the cumulative hazard of *MI* (Supplemental Section I). We estimated time to *MI* according to each patient's covariate/exposure history (Supplemental Section I).

Parameters

To evaluate the effect of study parameters on our application of multiple imputation, we varied several of them uniformly across sites (Table 1; Supplemental Table III): *D*, *n*, the scale parameter of the Weibull function (λ), the HR of statin use, the nonlinear HR of *BMI* (Supplemental Figures I-III), the continuous or categorical nature of *BMI* in outcome models, the correlation between *BMI* and E_1, the multivariate baseline correlation matrix (Supplemental Table IV), and the number of imputations (*X*).

In real distributed data networks, some heterogeneity of effects is expected across sites. For this reason, a parameter b_d drawn from the distribution N(mean=0, standard deviation [SD]=0.2) was created to confer site-specific effects of statins on *MI*. We lowered the SD of b_d in one scenario (Table 2; Supplemental Table III). Inter-site heterogeneity was also explored by varying the marginal distributions of *BMI* and the effect of *BMI* on *MI* between databases (Table 2; Supplemental Table III). Finally, the correlations between baseline study covariates were varied for each data site in each simulation (Supplemental Table IV).

We conducted 200 simulations in each of 20 parameter scenarios (Supplemental Table III).

Imputation Models

We fit a model predicting *BMI* in the validation database using multivariable linear regression, with *Age*, *Male*, *Dur*, *Smoke*, a variable indicating whether a patient was ever exposed to statins (E_{ever}), and *MI* as model inputs. Beta splines with six degrees of freedom for *Age* and *Dur* were fitted, with the extreme knot boundaries set at the maximum and minimum values across databases. We bootstrapped parameters and their 95% CIs with 200 samples (to address spline over-fitting). *BMI* values were imputed *X* times per database (*X*=20 in the default scenario).

Statistical Analyses

We first conducted a "naïve" analysis wherein missing *BMI* values were ignored, then an "imputed" analysis that used multiple imputation. In both analyses, the effect of statins on *MI* was estimated using Cox proportional hazards models bootstrapped with 200 samples. Current statin use was treated as time-varying with no exposure lag. All other covariates were adjusted for baseline values. *BMI* was modeled using B-splines with 6 degrees of freedom. The summary HR was estimated using random-effects models with the restricted maximum-likelihood estimator.

Our simulation study was evaluated by several metrics. The % bias attributable to multi-dataset imputation was calculated as $\%\, bias = \frac{\widehat{\beta_E}-\beta_E}{\beta_E} \times 100\%$. We also calculated the true parameter's 95% coverage rate as the proportion of simulations whose 95% CIs included the true parameter. The type II error rate was also calculated as the proportion of simulations whose 95% CIs included the null hypothesis (HR=1.0).

All analyses were done using R (3.1.2) with the aid of the Guillimin computing cluster.

Real-World Analysis

Study Population

We obtained data from the CPRD, a database of >700 general practitioner databases in the UK that contains a wealth of diagnostic, procedural, and laboratory-based data.[20] A subpopulation of the CPRD may be linked to the Hospital Episode Statistics (HES) database, which contains inpatient diagnoses and procedures, and to the Office of National Statistics (ONS) vital statistics database, which contains causes of death. Among patients >18 years old with at least one year of observation time in the CPRD, HES, and ONS, we identified all initiators of non-insulin antihyperglycemic agents between April 1, 1998 and March 31, 2016. The date of cohort entry was the date of first antihyperglycemic agent prescription.

We excluded patients who: 1) had a previous recorded prescription for insulin (a marker of advanced diabetes) prior to cohort entry; 2) were previously diagnosed with polycystic ovary syndrome at any time before cohort entry; 3) were diagnosed with gestational diabetes in the year prior to cohort entry; 4) had a history of statin use in the year before cohort entry; and/or 5) had a history of cardiovascular disease (Supplemental Section II). We then paired the 10 English practice regions[20] into five adjacent mock distributed databases (Supplemental Section II). The mock database "West Midlands and Southwest" was chosen a priori as the validation database.

Patients were followed until incident MI, death, end of study (March 31st, 2016), end of follow-up (four years), end of CPRD registration, date of last data collection, or one year since last recorded physician contact or prescription, whichever occurred first.

Exposure and Outcome

Patients were classified into mutually-exclusive, time-dependent exposure groups: currently exposed or currently unexposed to statins. A patient was considered exposed to statins if the date of the risk set overlapped with a statin prescription's intended duration of use + 30 days (Supplemental Section II).

The primary outcome was time to fatal or nonfatal MI (International Classification of Disease 10th Revision [ICD-10] codes I21.x-I22.x; ICD-9 code 410.x), excluding perioperative MIs. Outcomes were ascertained in the primary or secondary position using HES and ONS data, with the earliest code determining the event date.

Imputation Models

We used single imputation at each mock study site to replace sporadically missing data (i.e., data that was missing from a subset of individuals for unknown reasons) with a conditionally random draw for patients missing BMI, smoking status, HbA1c, and/or serum cholesterol in the five years before cohort entry. For systematically missing data, samples of $F(smoking, BMI, HbA1c, cholesterol \mid X)$ were drawn from sequential samples of $F(smoking \mid X)$, $F(BMI \mid smoking, X)$, $F(HbA1c \mid BMI, smoking, X)$, and $F(cholesterol \mid BMI, smoking, HbA1c, X)$, which were estimated by multinomial (BMI, HbA1c, cholesterol) or binomial (smoking) logistic regression at the validation data site. We used 20 imputations in the primary analysis.

Statistical Analysis

All site-specific analyses used baseline-adjusted Cox proportional hazard models with time-dependent statin use. We adjusted for the missing variables (smoking status, BMI, HbA1c, serum cholesterol) except in the naïve analyses. We also adjusted for age, sex, history of alcohol abuse, year of cohort entry, year of CPRD registration, and other important potential confounders (Supplemental Section II).

We performed three analyses: 1) the "naïve" analysis wherein smoking status, BMI, HbA1c, and serum cholesterol values were ignored; 2) the "imputed" analysis wherein imputed values were considered for each database; and 3) the "true" analysis that considered the measured (or singly-imputed) values for these covariates. In all analyses, we pooled results with random-effects models according to the restricted maximum likelihood approach. The results of the three meta-analyses were qualitatively compared.

Several sensitivity analyses were pre-specified. First, we repeated our analyses with one missing variable (smoking status or cholesterol). Second, we censored patients at two years of follow-up (instead of four). Third, we limited our study population to patients with no previous statin use. Fourth, we used a different validation database. Fifth, we used linear imputation models for the continuous variables (BMI, HbA1c, and serum cholesterol). Sixth, we applied stepwise variable selection for our imputation models, starting with the variables in our primary analysis. Seventh, we reversed the order of our chained equations.[21] Eighth, we excluded patients with a missing value for any of: BMI, smoking status, HbA1c, or serum cholesterol. Finally, we reduced the number of imputations from 20 to five.

All analyses were done using R (3.1.2) and SAS (9.4).

Ethics

Our study was approved by the Independent Scientific Advisory Board of the CPRD (protocol number 17_133R) and by the research ethics board of the Jewish General Hospital in Montreal, Canada.

RESULTS

Simulation Study

Default Parameterization

In the default scenario, the naïve analyses were substantially biased, with a median bias of 120% across the 200 simulations (Table 3; Supplemental Figure IV). The median pooled HR of 1.05 (2.5th quantile: 0.92, 97.5th quantile: 1.21) was attenuated compared to the true effect of 0.80.

Application of multiple imputation substantially reduced bias from unmeasured *BMI* (Table 3; Supplemental Figure IV). The median bias for the imputed simulations was 5%, and the median bias reduction was 118%. The median pooled HR estimate was 0.81 (2.5th quantile: 0.71, 97.5th quantile: 0.93), close to the true effect of 0.80.

Parameter Changes across Databases

Multiple imputation reduced bias from unmeasured *BMI* in all parameterizations (Table 3; Supplemental Figure IV). Imputed analyses on fewer data sites (D=3 or 5) in the network yielded greater type II error rates and similar 95% coverage rates compared to the default scenario (D=7). With fewer patients in each database, the type II error rates increased; in these analyses, the 95% coverage rates were augmented. Stronger protective effects of E on *MI* greatly lowered the type II error rates of the imputed analyses. Our method was robust to changes in the shape of the HR of *BMI* curve and produced similar parameter estimates, 95% coverage rates, and type II error rates across specifications. Specifying *BMI* as a categorical variable reduced the ability of our method to correct for unmeasured confounding (median bias after imputation: 98%). Lowering the correlation between E_1 and *BMI* reduced the bias in the naïve analyses, but had minimal impact on the imputed results. Reducing the number of imputations, X, from 20 to five had little impact on our results. Finally, our method was robust to random variations in the correlation matrix.

Parameter Changes between Databases

Our method of imputing values from one validation database was robust to all evaluated differences in study parameters across databases in the simulated distributed network (Table 4; Supplemental Figure IV), including when each database had a unique correlation matrix of study variables (median bias reduction: 105%) and when the HR of *BMI* was different for each database (median bias reduction: 76%).

Real World Analysis

Study Population

We identified 59,957 patients meeting our inclusion criteria, of whom 14,924 were in the validation database (Supplemental Figure V; Supplemental Table V). The mean age across databases was 56.8 years (SD: 14.3 years). Average BMI was 32.1 kg/m^2 (SD: 7.2 kg/m^2); 48.2% of patients were obese (BMI≥30 kg/m^2). HbA1c levels varied at baseline: the mean was 9.0% (SD: 2.2%). Cholesterol was broadly distributed (mean: 216.6 mg/dL; SD: 49.8 mg/dL). A total of 48.3% of patients had at least one sporadically missing value for BMI (15.4% missing), cholesterol (20.5% missing), HbA1c (34.8% missing), or smoking status (13.1% missing): 25.1%, 14.3%, 5.6%, and 3.3% were sporadically missing one, two, three, and four variables, respectively. No notable discrepancies in baseline patient characteristics were seen across mock databases (Supplemental Table V).

Primary Analysis

In the full study population, 625 MIs were observed over 181,777 person-years of follow-up (incidence rate of 3.4 per 1000 person-years). This rate was slightly lower in patients currently exposed statins compared with those unexposed to statins (3.2 and 3.6, respectively, per 1000 person-years).

In the “naïve” analysis, the summary HR for the effect of statins on MI was 0.93 (95 % CI: 0.73-1.20) (Figure 1). In the “true” analysis, we observed a summary HR of 0.85 (95% CI: 0.66-1.10). The “imputed” analysis produced a summary HR of 0.86 (95% CI: 0.69-1.08).

Sensitivity Analyses

When smoking status was the only missing covariate, there was little bias in the naïve analysis, rendering imputation unnecessary (Supplemental Figure VI). In contrast, the naïve analysis was biased when only total cholesterol was missing (Supplemental Figure VII); this bias was repaired with multiple imputation. With a follow-up duration of two years, the naïve and true HRs were qualitatively different than the primary analysis at 1.36 (95% CI: 1.07-1.73) and 1.28 (95% CI: 0.95-1.73), respectively (Supplemental Figure VIII). The imputed analysis nevertheless directed the pooled effect estimate towards the true analysis (HR: 1.29; 95% CI: 1.01-1.65). The results of all other sensitivity analyses were similar to those of the primary analysis (Supplemental Figures IX-XV).

DISCUSSION

Main Results

In both simulated and real-world data structures, our proposed method of multiple imputation reduced confounding bias across all parametrizations and sensitivity analyses. In the simulation study, our method was effective even in the presence of heterogeneity between data sites, such as different correlations between baseline variables and different relationships between the *BMI* and *MI*. Indeed, multiple imputation performed no worse in these circumstances than in the default scenario.

In our real-world example, the true analysis (HR: 0.85; 95% CI: 0.66-1.10) was consistent with a Cochrane review of trials (risk ratio: 0.75; 95% CI: 0.70-0.81)[11]. In this naïve analysis, bias due to missing confounders (HbA1c, BMI, cholesterol, smoking) was greatly reduced by multiple imputation. In sensitivity analyses, our method was robust to the order of chained equations, the specific validation database chosen, the type of imputation models (linear versus multivariate logistic regression), and a lower number of imputations. The precision of the imputed estimate was improved relative to the true analysis as evinced by the width of the 95% CIs, which may have resulted from patient outlier values being replaced by imputed values closer to the expected mean.

The results of our real-world analyses were qualitatively different when follow-up was limited to two years: a substantial increased risk of MI among statin users was observed. We speculate that residual confounding by indication is responsible for this inversion. In both two- and four-year follow-ups, MI events occurred disproportionately in early periods following a statin prescription (Supplemental Figures XVI-XVII)[22], possibly because many patients presented with deteriorating cardiovascular markers that signalled an impending MI and served as a statin indication. Despite the potential for residual confounding in our analysis limited to two years, our method repaired bias from the specified unmeasured confounders.

Strengths and Limitations

Strengths of our study include application to both a controlled and real-world setting, exhaustive parameter constellations and sensitivity analyses, the use of a time-varying exposure to mimic analyses frequently conducted in distributed data networks, and our choice of substantive example, which allowed us to predict the direction of unmeasured confounding.

Our study is not without limitations. The results of our simulation study may not be generalizable because the assumptions in our simulated data may not have been realistic—for instance, the default correlation between baseline statin use (E_1) and *BMI* may have been unrealistically high (0.86 in the multivariate normal distribution). Our real-world data example was designed to address these concerns about generalizability. Nevertheless, our application of multiple imputation to a mock distributed data setting using real-world data from the CPRD also may have limited generalizability. In particular, data from real distributed settings may be more likely to depart from the MCAR assumption because predictor variables and their relations to missing variables differ across databases even after application of rigorous eligibility criteria. Discrepant variable capture and coding schemes between real distributed databases may also render our method less valid. In real distributed database situations with >1 validation database, our method would also need to be adjusted to accommodate additional information from other databases, though we anticipate research in CNODES may have a single validation database for important clinical or demographic variables best captured in the CPRD. Finally, our study addressed a missing data problem wherein the missing variables can be predicted from observed variables; the performance of our method in scenarios where the missing variables cannot be easily predicted is unknown.

CONCLUSIONS

Multiple imputation adapted to distributed data settings is a feasible method to reduce bias from unmeasured but measurable confounders when at least one database contains the variables of interest. Further research is needed to evaluate its validity in real-world distributed data networks.

ACKNOWLEDGEMENTS

This research is funded by CNODES, a collaborating center of the Drug Safety and Effectiveness Network (DSEN), funded by the Canadian Institutes of Health Research (Grant Number DSE-146021). Dr. Platt is the Albert Boehringer I Chair in Pharmacoepidemiology at McGill University. Dr. Filion is supported by a salary support award from the *Fonds de recherche du Québec – santé* (Quebec Foundation for Health Research) and holds a William Dawson Scholar award from McGill University.

CONFLICT OF INTEREST STATEMENT

RWP has received personal fees from Amgen, AbbVie, Pfizer, and Novartis. The other authors declare no conflicts of interest.

REFERENCES

1. Suissa S, Henry D, Caetano P, *et al.* CNODES: the Canadian Network for Observational Drug Effect Studies. *Open Med* 2012;6:134-40.
2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj* 2010;340:c221.
3. Burgess S, White IR, Resche-Rigon M, *et al.* Combining multiple imputation and meta-analysis with individual participant data. *Statistics in medicine* 2013;32:4499-514.
4. Jolani S, Debray TP, Koffijberg H, *et al.* Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in medicine* 2015;34:1841-63.
5. Koopman L, van der Heijden GJ, Grobbee DE, *et al.* Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. *American journal of epidemiology* 2008;167:540-5.
6. Quartagno M, Carpenter J. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in medicine* 2016;35:2938-54.
7. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research* 2018;27:1634-49.
8. Resche-Rigon M, White IR, Bartlett JW, *et al.* Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* 2013;32:4890-905.
9. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, *et al.* Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation* 2006;76:1049-64.
10. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377-99.
11. Taylor F, Huffman MD, Macedo AF, *et al.* Statins for the primary prevention of cardiovascular disease. *The Cochrane Library* 2013.
12. National Institute for Health and Care Excellence (UK). Cardiovascular disease: risk assessment and reduction, including lipid modification 2014.
13. Stone NJ, Robinson JG, Lichtenstein AH, *et al.* 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Circulation* 2014;129:S1-S45.
14. Yusuf S, Hawken S, Ôunpuu S, *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364:937-52.
15. Stratton IM, Adler AI, Neil HAW, *et al.* Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000;321:405-12.
16. UK Prospective Diabetes Study Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. *BMJ* 1998:703-13.
17. Little RJ, Rubin DB. *Statistical analysis with missing data*: John Wiley & Sons; 2014.
18. MacCallum RC, Zhang S, Preacher KJ, *et al.* On the practice of dichotomization of quantitative variables. *Psychological methods* 2002;7:19.

19. Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in medicine* 2012;31:3946-58.
20. Herrett E, Gallagher AM, Bhaskaran K*, et al.* Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827-36.
21. Hughes RA, White IR, Seaman SR*, et al.* Joint modelling rationale for chained equations. *BMC medical research methodology* 2014;14:28.
22. Bhatnagar S, Turgeon M, Saarela O*, et al.* casebase: fitting flexible smooth-in-time hazards and risk functions via logistic and multinomial regression. R package version 0.1.0 ed: CRAN; 2017.

Table 1. Default and alternative parameters varied uniformly in all databases

Parameters		Values	
Notation	**Meaning**	**Default**	**Alternatives**
D	Number of datasets (including validation)	7	3, 5
n	Number of patients per dataset	10,000	1,000, 5,000
λ	Scale parameter for Weibull function	2E-06	1E-06, 5E-07
HR(*E*)	HR for MI for current statin use vs non-use	0.8	0.3, 0.5
HR(*BMI*)	HR for MI compared to reference value of 22 kg m^{-2}	Model 1*	Model 2, Model 3*
cor(*BMI*-E_1)	Correlation between BMI and baseline statin exposure	0.86	0.30, 0.50
corr	Correlation matrix	Matrix 1†	Matrix 2†
X	Number of imputations	20	5
BMI	Body mass index in outcome models	~ N(mean = 33, SD = 3)	7 categories: [0, 25), [25,30), [30, 32.5), [32.5, 35), [35, 37.5), [37.5, 40), [40, 1)

Abbreviations: HR, hazard ratio; SD, standard deviation

*Supplementary Figures I-III

†Supplementary Tables I and III. Matrix 2 samples random correlations between variables, and replaces the resulting matrix with the nearest positive definite. All *D* databases use the same correlation matrix in these analyses.

Table 2. Default and alternative parameters varied between databases

Parameters		Values	
Notation	**Meaning**	**Default**	**Alternatives**
BMI_d	Body mass index at each database	BMI_d ~ N(mean = 33, SD = 3)	BMI_d ~ N(mean_j = N(mean = 33, SD = 2), SD = 3)
$corr_d$	Correlation matrix	Matrix 1*	Matrix 2*
b_d	Beta coefficient for random effects of statin exposure	b_d ~ N(mean = 0, SD = 0.2)	b_d ~ N(mean = 0, SD = 0.1)
HR$(BMI)_d$	HR for MI compared to reference value of 22 kg m^{-2}	Model 1: $HR(BMI)_d = 4 \times 10^{-5} \times (BMI - 22)^4 + 1$	Model 4: $HR(BMI)_d$ $= \begin{cases} unif(1 \times 10^{-5}, 6 \times 10^{-5}) \times (BMI - unif(20{,}24))^4 + 1 & \text{if } d \text{ is 1 or 4} \\ unif(0.27, 0.75) \times \lvert BMI - unif(20{,}24) \rvert + 1 & \text{if } d \text{ is 2 or 5} \\ unif(5 \times 10^{-3}, 3 \times 10^{-2}) \times (BMI - unif(20{,}24))^2 + 1 & \text{if } d \text{ is 3 or 6} \\ unif(5 \times 10^{-5}, 2 \times 10^{-3}) \times \lvert BMI - unif(20{,}24) \rvert^3 + 1 & \text{if } d \text{ is 7} \end{cases}$

Abbreviations: HR, hazard ratio; SD, standard deviation

* Matrices 1 and 2 can be found in Supplemental Tables I and IV. Matrix 2 samples random correlations between variables and replaces the resulting matrix with the nearest positive definite. Each *D* database uses a distinct correlation matrix in these analyses.

Table 3. Effect of different simulation study parameters varied uniformly in all databases

		Naïve analysis[¶]								Imputed analysis[‖]								
		HR			% Bias					HR			% Bias					
			Quantile			Quantile					Quantile			Quantile				
Parameter[*]	Scen	Med	2.5	97.5	Med	2.5	97.5	Type II error (%)	95% coverage (%)	Med	2.5	97.5	Med	2.5	97.5	Type II error (%)	95% coverage (%)	Med % Bias Reduction[**]
D																		
7	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
3	2	0.98	0.77	1.23	89	-15	194	96.0	72.5	0.80	0.64	1.02	0	-103	107	64.5	98.0	88
5	3	1.03	0.87	1.24	113	38	195	98.0	36.5	0.81	0.68	0.98	6	-70	91	46.5	97.0	109
n																		
10000	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
1000	4	1.04	0.79	1.26	117	-4	205	98.0	52.0	0.81	0.66	0.94	7	-84	72	80.0	100.0	113
5000	5	1.04	0.90	1.23	116	52	192	99.0	24.0	0.80	0.68	0.93	1	-70	68	39.0	99.5	118
λ																		
2 × 10^-6	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
5 × 10^-7	6	1.06	0.92	1.28	127	61	210	99.5	21.5	0.81	0.69	0.94	7	-65	72	57.5	99.0	126
1 × 10^-6	7	1.07	0.93	1.22	129	68	190	100.0	19.0	0.81	0.70	0.93	5	-57	68	40.0	98.0	122
HR(E)																		
0.8	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
0.3	8	0.41	0.35	0.48	25	14	39	0.0	17.5	0.30	0.26	0.35	1	-12	14	0.0	98.0	24
0.5	9	0.67	0.58	0.78	42	21	64	1.5	14.5	0.51	0.44	0.58	2	-19	22	0.0	98.0	41
HR(*BMI*)†																		
Model 1	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
Model 2	10	0.90	0.77	1.03	50	-15	113	69.0	66.0	0.81	0.69	0.92	4	-65	65	26.0	95.5	48
Model 3	11	0.97	0.84	1.12	87	21	152	97.5	36.0	0.81	0.70	0.92	5	-60	65	27.0	97.5	86
BMI specification																		
Continuous	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
Categorical	12	1.06	0.93	1.21	127	69	186	99.5	11.5	0.99	0.87	1.18	98	36	174	100	54.0	27
cor(*BMI*-E_1)																		
0.86	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
0.30	13	0.87	0.73	1.01	35	-40	104	53.5	78.5	0.81	0.70	0.93	5	-61	68	30.0	97.5	31
0.50	14	0.92	0.80	1.07	60	1	128	81.0	58.5	0.81	0.71	0.93	4	-56	68	28.0	98.0	58
Correlation matrix‡																		
Matrix 1	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
Matrix 2	15	1.03	0.84	1.23	114	23	192	96.0	25.0	0.82	0.68	0.95	8	-73	75	34.0	94.0	105
X																		
20	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
5	16	1.04	0.92	1.21	119	61	185	99.0	17.0	0.81	0.69	0.94	5	-66	70	29.0	97.5	117

Abbreviations: BMI, body mass index; HR, hazard ratio; Med, median; MI, myocardial infarction; Scen, scenario

[*] D refers to the number of databases; n is the number of patients; λ is the scale parameter of the Weibull function; HR(E) is the HR of statin use; HR(*BMI*) is the relationship between *BMI* and the HR of *MI*; cor($BMI\text{-}E_1$) is the correlation between *BMI* and the first statin exposure period, E_1; "correlation matrix" refers to the multivariate normal correlation matrix used to determine baseline covariates; X is the number of imputations

[†] See Supplemental Figures I-III

[‡] See Supplemental Tables I and IV

[¶] In the naïve analysis, missing confounders were not considered in outcome models

[‖] In the imputed analysis, missing confounders were considered in the outcome models using multiple imputation

[**] Absolute reduction in % bias

Table 4. Effect of different simulation study parameters varied between databases

		Naïve analysis[¶]								Imputed analysis[‖]								
		HR			% Bias Reduction					HR			% Bias Reduction					
			Quantile			Quantile					Quantile			Quantile				
Parameter*	Scen	Med	2.5	97.5	Med	2.5	97.5	Type II error (%)	95% coverage (%)	Med	2.5	97.5	Med	2.5	97.5	Type II error (%)	95% coverage (%)	Med % Bias Reduction[**]
BMI$_d$																		
N(mean = 33, SD = 3)	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
N(meanj = N(mean = 33, SD =2), SD =3)	17	1.03	0.90	1.20	115	54	183	98.0	21.5	0.81	0.70	0.93	3	-62	66	31.5	97.5	115
Correlation matrix†																		
Matrix 1	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
Matrix 2	18	1.02	0.89	1.18	110	50	175	99.0	22.0	0.81	0.70	0.95	7	-57	79	33.5	95.5	105
b$_d$																		
N(mean=0, SD=0.2)	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
N(mean=0, SD=0.1)	19	1.04	0.97	1.14	120	85	157	100.0	1.5	0.80	0.74	0.87	1	-33	38	8.0	100.0	119
HR(*BMI*)$_d$‡																		
Model 1	1	1.05	0.92	1.21	120	61	185	99.0	17.5	0.81	0.71	0.93	5	-53	70	30.0	97.5	118
Model 4	20	0.96	0.83	1.09	80	15	138	90.5	42.5	0.81	0.70	0.92	4	-61	62	23.5	98.0	76

Abbreviations: BMI, body mass index; HR, hazard ratio; Med, median; MI, myocardial infarction; Scen, scenario; SD, standard deviation

[*] *BMI*$_d$ refers to the distribution of BMI at each *d* database of *D* total databases; "correlation matrix" refers to the multivariate normal correlation matrix used to determine baseline covariates; *b*$_d$ is the log-HR for the random-effect of statins on MI in database *d* of *D* total databases; HR(*BMI*)$_d$ is the relationship between *BMI* and the HR of *MI* in database *d* of *D* total databases

[†] See Supplemental Tables I and IV

[‡] See Table 2

[¶] In the naïve analysis, missing confounders were not considered in outcome models

[‖] In the imputed analysis, missing confounders were considered in the outcome models using multiple imputation

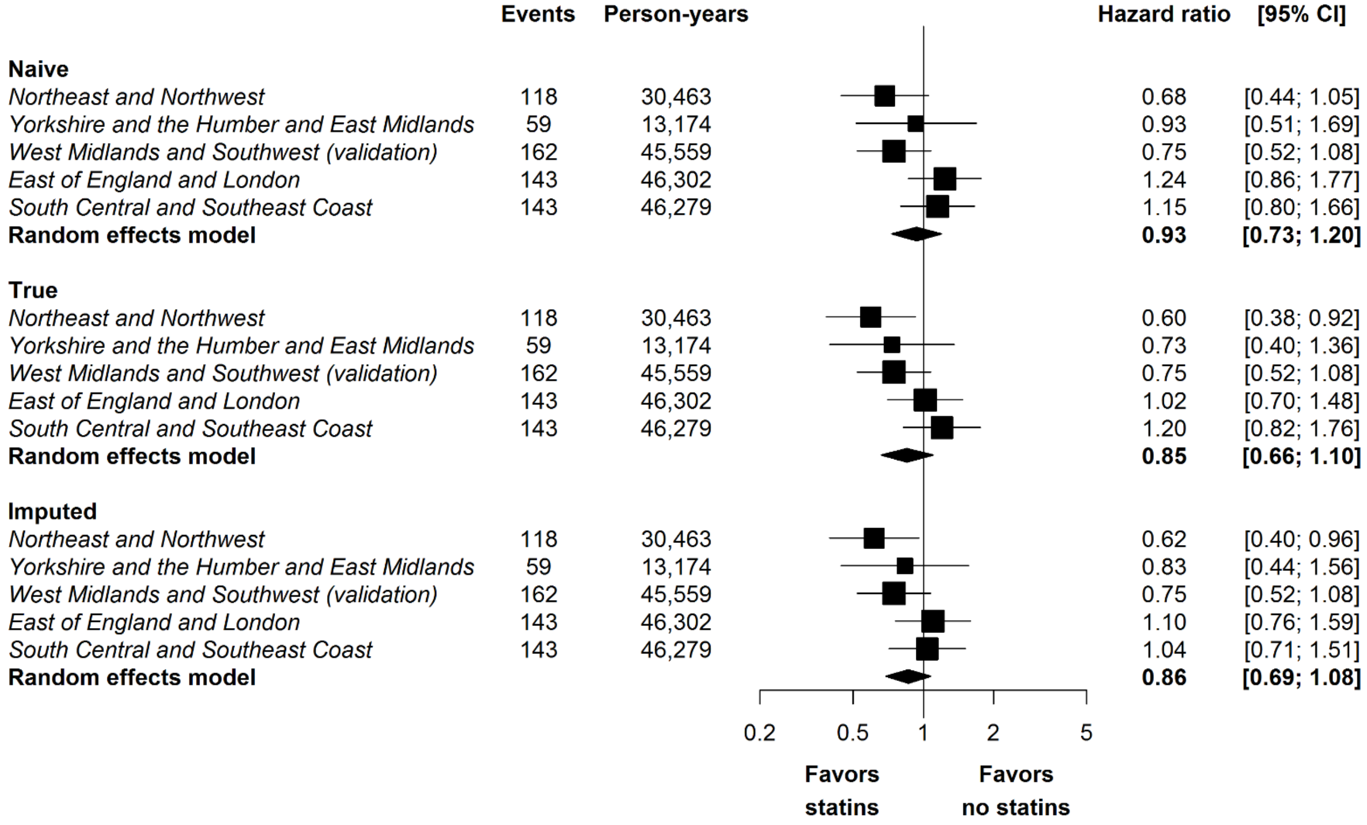
[**] Absolute reduction in % bias

FIGURE LEGEND

Figure 1. **Effect of statins on myocardial infarction in databases with missing data (naïve), no missing data (true), and imputed data (imputed)**[*]

[*] In this analysis, the mock database "West Midlands and Southwest" served as the validation database. In the "naïve" analysis, the pre-specified confounders (HbA1c, smoking, cholesterol, BMI) were ignored in all databases but the validation database. In the "true" analysis, the observed or singly-imputed values of the pre-specified confounders were used in all databases. In the "imputed" analysis, multiple imputation was applied to consider values for the pre-specified confounders in the missing databases.

Figure 1.



	Events	Person-years	Hazard ratio	[95% CI]
Naive				
Northeast and Northwest	118	30,463	0.68	[0.44; 1.05]
Yorkshire and the Humber and East Midlands	59	13,174	0.93	[0.51; 1.69]
West Midlands and Southwest (validation)	162	45,559	0.75	[0.52; 1.08]
East of England and London	143	46,302	1.24	[0.86; 1.77]
South Central and Southeast Coast	143	46,279	1.15	[0.80; 1.66]
Random effects model			**0.93**	**[0.73; 1.20]**
True				
Northeast and Northwest	118	30,463	0.60	[0.38; 0.92]
Yorkshire and the Humber and East Midlands	59	13,174	0.73	[0.40; 1.36]
West Midlands and Southwest (validation)	162	45,559	0.75	[0.52; 1.08]
East of England and London	143	46,302	1.02	[0.70; 1.48]
South Central and Southeast Coast	143	46,279	1.20	[0.82; 1.76]
Random effects model			**0.85**	**[0.66; 1.10]**
Imputed				
Northeast and Northwest	118	30,463	0.62	[0.40; 0.96]
Yorkshire and the Humber and East Midlands	59	13,174	0.83	[0.44; 1.56]
West Midlands and Southwest (validation)	162	45,559	0.75	[0.52; 1.08]
East of England and London	143	46,302	1.10	[0.76; 1.59]
South Central and Southeast Coast	143	46,279	1.04	[0.71; 1.51]
Random effects model			**0.86**	**[0.69; 1.08]**