Correcting Hazard Ratio Estimates for Outcome Misclassification Using Multiple Imputation with Internal Validation Data

Jiayi Ni

Department of Epidemiology, Biostatistics and Occupational Health

McGill University, Montreal, Quebec, Canada

February 2015

Supervisor

Elham Rahme, PhD

Co-supervisor

Nandini Dendukuri, PhD

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Biostatistics

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Supplemental Files

LIST OF FIGURES

Chapter 1

Chapter 1

LIST OF SUPPLEMENTAL FILES

ABSTRACT

Background: Multiple imputation (MI) has been used to account for exposure misclassification and unmeasured confounders in the estimation of hazard ratios (HRs) in survival analyses and has been shown to effectively remove bias due to outcome misclassification from effect estimates in logistic regression when validation data are available for a subgroup of study participants, but similar research that accounts for outcome misclassifications in the estimation of HRs is lacking.

Research Design and Methods: We assessed through various simulation scenarios the performance of MI using internal validation data to account for outcome misclassification in estimating HR from Cox regression models. Mean squared errors (MSEs) and coverage of the confidence intervals (CIs) were used to assess the results. We then applied this method to a health administrative database study of the association between statin use and risk of new-onset diabetes in a stratified random sample of 6,247 Quebec individuals among whom about half responded to a survey on diabetes status and about a quarter also provided fasting blood samples for glucose testing. We used self-reported diabetes and/or elevated fasting plasma glucose (≥7 mmol/L for diabetes and 6.1-6.9 mmol/L for impaired fasting glucose [IFG]) as the gold-standard outcome when available. We first compare time to develop diabetes among individuals initiated on statin treatment matched on age and sex to statin naïve individuals. We then used a 2-level MI technique to first impute undiagnosed diabetes from the sample that provided blood samples to the survey sample, and second impute diabetes status from the survey sample to the total random sample. To estimate the time of diabetes occurrence for the corrected cases, we selected a random time from the set of available ones. We compared the HR estimates of statin users versus non-users obtained from using the administrative data alone and from using the same data corrected for outcome misclassification.

Results: MI accounting for misclassification in time-to-event outcomes yielded less biased HR estimates and had appropriate coverage for both non-differential and differential misclassification and under all scenarios explored. A non-representative validation subgroup with low proportion of participation resulted in estimates with large variances. Using MSE as a criterion, the bias correction was sometimes outweighed by the uncertainty introduced by the unknown time of occurrence of the event. The HR comparing statin users to non-users in the random sample was 1.61 (95% CI 1.09-2.38) when using physician-diagnosed diabetes, 1.49 (0.95-2.34) when correcting for self-reported diabetes or undiagnosed diabetes, and 1.36 (0.92-2.01) when correcting for self-reported diabetes or IFG/undiagnosed diabetes.

Conclusions: MI performs well in removing bias due to outcome misclassification from HR estimates when internal validation data are available for representative subgroups. By using this method, we found that the HR associated with statin treatment and diabetes occurrence was overestimated when misclassification in diabetes status assessed using physician diagnosis was ignored. Our findings highlight the importance of accounting for these misclassifications to prevent erroneous results in studies based on administrative databases.

ABRÉGÉ

Le contexte: L'imputation multiple réduit les biais dus aux erreurs de classification de l’exposition dans les modèles de régression de Cox ainsi que ceux dus aux erreurs de classification de l’issue dans les modèles de régression logistique lorsque des données de validation interne sont disponibles. Mais, aucune recherche similaire sur l’effet de l’imputation multiple sur la réduction du biais dus aux erreurs de classification de l’issue dans l'estimation des ratios de hasard (RH) à partir de modèles de Cox n’a été recensée.

Conception et méthodes de recherche: En utilisant des données de validation interne, nous avons évalué la performance de la méthode d’imputation multiple via divers scenarios de simulation pour estimer l’impact de l’erreur de classification dans l’issue sur les RH de par les modèles de régression de Cox. L'erreur quadratique moyenne (EQM) et la couverture des intervalles de confiance (IC) ont été utilisées pour évaluer les résultats. Cette méthode a été appliquée à une étude de cohorte rétrospective qui visait à évaluer l'association entre l'utilisation des statines et le risque d’apparition de diabète à partir des données administratives de la Régie de l’assurance maladie du Québec d’un échantillon aléatoire stratifié de la population comprenant 6247 individus. Environ la moitié de cet échantillon avait répondu à une enquête sur le statut de diabète, et environ le quart d'entre eux avaient de plus fourni un échantillon de sang à jeun pour le test de glucose. Le diabète auto-déclaré et/ou un niveau élevé de glucose plasmatique à jeun (≥7 mmol/L pour le diabète et 6.1-6.9 mmol/L pour l'altération de la glycémie à jeun [AGJ]) étaient considérés comme l'étalon-or lorsque disponibles. Les nouveaux utilisateurs de statine étaient appariés aux non-utilisateurs pour l'âge et le sexe. L'imputation multiple à 2 niveaux a été utilisée pour imputer le diabète non diagnostiqué à toute la cohorte. En premier, le statut de diabète a été imputé à partir des résultats de l'analyse de sang à tous ceux qui ont participé à l'enquête. En deuxième, le statut de diabète a été imputé à partir de l'échantillon qui a participé à l'enquête à l'échantillon aléatoire total. Pour estimer le temps de l'apparition du diabète pour les cas corrigés, un temps aléatoire a été choisi à partir de l'ensemble des temps des cas disponibles. Les RH obtenus en utilisant les données administratives ont été comparés à ceux obtenus en utilisant les données corrigées.

Résultats: Les RH résultant des données corrigées pour l'erreur de classification quand différentielle ou non-différentielle entre les exposés et les non-exposés étaient moins biaisés que les RH des données non corrigées et la couverture des intervalles de confiance était appropriée pour tous les scénarios explorés. Un sous-groupe de validation non représentatif avec une faible proportion de participation a conduit à des estimations avec de grandes variances. La correction du biais a été parfois dépassée par l'incertitude introduite par le temps inconnu de l’apparition de l'évènement. Le RH obtenu à partir de l'échantillon aléatoire était de 1,61 (IC à 95% 1,09-2,38), 1,49 (0,95-2,34) lors de la correction pour le diabète auto-déclaré ou non diagnostiqué, et 1,36 (0,92-2,01) lors de la correction pour le diabète auto-déclaré ou AGJ/diabète non diagnostiqué.

Conclusions: L'imputation multiple est utile pour la réduction des biais des estimations de RH dus à l’erreur de classification de l’issue lorsque les données de validation internes sont disponibles pour un sous-groupe représentatif. Le RH de l’apparition du diabète associé au traitement par statine est surestimé lorsque l’erreur de classification de l'apparition du diabète évaluée en utilisant le diagnostic du médecin est ignorée. Nos résultats soulignent l'importance de tenir compte de ces erreurs de classification pour éviter des résultats erronés dans les études basées sur les bases de données administratives.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my research supervisor, Dr. Elham Rahme and co-supervisor, Dr. Nandini Dendukuri, who had supported me throughout my graduate study. I thank them for reviewing my thesis and providing countless pieces of advice. This thesis could not be accomplished without their patient guidance.

I would like to thank the Régie de l’Assurance Maladie du Québec (RAMQ) for providing the administrative data and the personnel from the Institut de la Statistique du Quebec (ISQ) who oversaw the survey.

I also thank the department of Epidemiology, Biostatistics and Occupational Health for a friendly academic atmosphere.

I would like to thank my parents for their love and care. They educated me to have a positive attitude to face difficulties. They encouraged me whenever I needed it and supported me on every decision I made. Without them, I would not have achieved this goal.

Finally, I thank all my friends who have helped and supported me through my time in Montreal.

CONTRIBUTION OF AUTHORS

This thesis consists of two manuscripts linked by connecting sections. The first manuscript (Chapter 1) is entitled "Correcting Hazard Ratio Estimates for Outcome Misclassification Using Multiple Imputation with Internal Validation Data – A Simulation Study". The author list for the first paper is: Jiayi Ni, Nandini Dendukuri, Kaberi Dasgupta, Aaron Leong and Elham Rahme.

The second paper (Chapter 2) is entitled "Correcting Hazard Ratio Estimates for Outcome Misclassification in Physician-diagnosed Diabetes in Health Administrative Database Using Internal Validation Data". The author list for the second paper is: Jiayi Ni, Aaron Leong, Kaberi Dasgupta, Jean-Louis Chiasson, Nandini Dendukuri and Elham Rahme.

I am the first author on both papers and have contributed substantially to this body of work. Specifically, I conducted the literature search and summarized the literature, formulated the study questions, performed the simulations and data analyses, and reported the results.

AL and KD are authors on both manuscripts and played a major role in manuscript revisions. KD contributed to the formulation of the study designs, and was involved in all aspects of results reporting, manuscript preparation and revisions.

JLC is the author on the second manuscript (Chapter 2). He contributed to the formulation of the study design of the second manuscript. He oversaw the collection and analysis of blood samples in his institution, Saint-Luc Hospital and provided clinical expertise in assessing the confounders and interpreting the results.

ND is the author and co-supervisor on both manuscripts. she advised the statistical analyses, the formulation of the study questions, and the conduct of the simulation. She also reviewed the final manuscripts.

ER is senior author on both manuscripts. She led the formulation of study designs and coordinated the data-linkage between health administrative data, survey data and blood samples data. She supervised all aspects of data analyses, results reporting, and manuscript preparation and revisions. The study was funded by a grant from the Canadian Institutes of Health Research held by ER (Principal Investigator), and co-investigators KD and JLC.

LIST OF ABBREVIATIONS

MI	Multiple Imputation
ML	Maximum Likelihood
RC	Regression Calibration
PSC	Propensity Score Calibration
HR:	Hazard Ratio
OR:	Odds Ratio
MSE:	Mean Square Error
SD:	Standard Deviation
DAG	Directed Acyclic Graph
CIHR:	Canadian Institute of Health Research
QSI/ISQ:	Québec Statistical Institute/Institut de la statistique de Québec
RAMQ:	Régie de l’assurance maladie du Québec
ICD:	International Statistical Classification of Diseases and Related Health Problems
FPG:	Fasting Plasma Glucose
IFG:	Impaired Fasting Glucose

INTRODUCTION

Misclassification of a variable can occur in any type of epidemiologic study design, and in some situations it may produce serious bias that may yield invalid results. Some authors have suggested collecting additional data on a subsample to correct for misclassification through various statistical techniques including multiple imputation (MI) (1-15). MI techniques have been shown in simulation studies to remove or reduce the bias in odds ratios (ORs) estimates associated with misclassification in the outcome variable in logistic regression models (5), and in hazard ratios (HRs) associate with misclassification in the exposure variable or confounders in survival analysis (4) when internal validation data are available for a subgroup of the main study sample. But little research has been done to account for outcome misclassification in time-to-event analyses and reduce the associated bias in the estimated HRs.

This thesis illustrates an approach that uses MI to account for outcome misclassification in time-to-event analyses using internal validation data. It contains two manuscripts that were prepared to submit independently for publication. Each manuscript is presented as a separate chapter with connecting texts that provide overall conclusions to the preceding chapter and logical bridges between chapters. Chapter 1 provides details of the proposed approach, and the methodology that was evaluated with simulations under various representative circumstances. Limitations of the use of the proposed method in studies of exposure-disease associations are discussed.

In Chapter 2 we apply the method studied in Chapter 1 to a study using the Quebec health administrative databases obtained from the Régie de l'assurance maladie du Québec (RAMQ). We assess the association of statin use and new-onset type 2 diabetes mellitus, here on diabetes, in a random sample of 6,247 individuals from the Quebec population 20 years of age and older. Survey data were also available for 3,322 (53.2%) from this sample who also agreed for data linkage with administrative data. Among these 3,322 individuals who participated in the survey, 1,599 (48.1%) also provided a fasting capillary blood glucose sample for fasting blood glucose assessment. Therefore two subsamples of the original sample were available where additional data provided the opportunity for correcting the outcome misclassifications, here diabetes. Estimates of the association between statin use and new-onset diabetes obtained from a Cox regression model were corrected for potential misclassification in diabetes status using MI techniques to impute missing values of self-reported diabetes obtained from the survey, as well as missing values of elevated fasting blood glucose obtained from the blood tests.

LITERATURE REVIEW

Bias due to misclassification in studies of exposure-disease associations

Effect estimates in epidemiologic research are subject to bias from confounding by indication, selection bias, and misclassification. While confounding by indication and selection bias can be, at least in part, accounted for by adjusting for baseline characteristics, misclassification biases are usually not easy to control by researchers because these rely on the accuracy of record-keeping in the database.

In cohort studies, the bias due to misclassification is primarily dependent on the sensitivity and specificity of the classification scheme, disease frequency, and exposure frequency (16). It has been shown that misclassification of the same magnitude and in the same direction in two compared groups (non-differential misclassification) tends to bias the effect estimates towards the null value (17). When differential misclassification occurs the bias can be in either direction, and may be great (18).

In studies of the exposure-disease association, misclassifications of disease status will affect not only the outcome, but also subject inclusion at baseline. Inclusion of individuals with undiagnosed disease before exposure in the study may lead to a false estimate of the exposure effect, when the disease is later discovered.

An example of potential misclassification bias can be found in the study of the association between statin treatment and diabetes, where statins are prescribed to reduce the cholesterol level in patients with hypercholesterolemia. Statins have been shown to reduce the risk of cardiovascular events in numerous trials of primary and secondary prevention populations, providing an apparently low risk approach to improve cardiovascular health (19). However, evidence from recent large randomized trials has revealed a likely causal relationship between statin therapy and new-onset diabetes. A meta-analysis including 13 statin trials with a total of 91,140 participants showed that statin therapy was associated with a 9% increased risk for incident diabetes (95% CI 2-17%) (20). The results of another meta-analysis including 5 randomized trials indicated that high-dose statin therapy was associated with improved cardiovascular outcomes, but also a 12% (4-22%) increased risk of new-onset diabetes, compared with moderate-dose statin therapy (21). The risk of new-onset diabetes reported by observational studies that compared statin users versus non-users were even higher. In an observational study of 161,808 postmenopausal women aged 50-79 years and followed for over 1,004,466 person-years (average 6.2 years per person), baseline statin use was found to be associated with a 1.5 times higher risk of diabetes after adjusting for potential confounders (22). Another analysis of electronic medical records conducted in the U.K. showed that statin use was associated with a 14% increased risk of diabetes (23).

Although these studies included information on important confounders and new-onset diabetes was rigorously assessed in those that were prospectively designed to look for diabetes, the recruitment of patients was based in general on previous diagnosis records and glucose levels were not assessed at baseline. Therefore, there is a possibility of misclassification in diabetes onset in some patients enrolled in the clinical trials which may have biased the observed association between statin use and diabetes occurrence. This problem is much more frequent in observational studies because most risk factors for hypercholesterolemia, such as obesity, unhealthy eating and sedentary lifestyle are also the major risk factors for diabetes. Individuals with high glucose levels that are not yet diagnosed by a physician for diabetes will be included into the study. Diabetes is later discovered in these individuals and a biased association is incorrectly inferred between statin and the risk of diabetes even after adjusting for all measured confounders.

Account for misclassification in data analyses

Researchers have proposed different methods to account for bias due to misclassification. All of these methods use information that involves modeling the relation between observed measurements and unobserved true values of the misclassified variable. The modeling could be based on a validation study employing an essentially error-free classification criterion (a gold standard). Examples include Barren's matrix formula (1) that corrects point estimates of effect for non-differential misclassification in two-by-two tables; Greenland and Kleinbaum's generalization of this formula (7) that allows it to be applicable to situations involving differential misclassification, matched data, and arbitrary two-way tables; inverse-variance weighted estimation (9), which combines results from the validation subsample and the remainder by weighting in inverse proportion to their variances; regression calibration (RC) (3,14), a 2-step approach which uses the validation study to predict true measurements for all main study subjects based on the estimated regression model for the gold standard and the misclassified variable, and then run the regression of the outcome of interest on the predicted true measurements in the main study; maximum likelihood (ML) (6,11), which includes data from all participants, with those in the validation subgroup providing data on the correct outcome and those not in the validation subgroup providing data on the misclassified outcome; and propensity score calibration (PSC) (15), which combines propensity scores (13) and RC developed to correct for measurement error. On the other hand, when validation data are not available, alternative methods include sensitivity analysis (8) that assesses the direction of the bias and robustness of the results in subgroups where misclassification is less likely, and a Bayesian approach (2,10,12) that allows for the incorporation of subjective prior information on misclassification (perhaps suggested by external literatures or expert opinions) in the models.

The problem of misclassification may be viewed as arising from missing data (24). In the case of outcome misclassification, an outcome that is subject to error is observed for every participant, while data are missing on the true outcome for some or all participants. Treating outcome misclassification as a missing-data problem allows the bias to be addressed by easily employed methods, such as MI, for handling missing data (25-27). The MI method replaces each missing value by a vector composed of M (>2) possible values. Each set of possible values of the vectors for the missing values are used to create one completed data set which is analyzed using standard complete-data methods (28). The relation between the misclassified and the true outcome may be estimated either from a subset of data (a validation sub-study) with values for both the misclassified outcome and a gold standard assumed to be equal to the true outcome or from external information relating the misclassified outcome to the gold standard. The MI technique has been applied to correct for exposure misclassification in estimating HRs (4), and has been shown to effectively remove bias due to outcome misclassification in effect estimates in logistic regression (5), using internal validation data available for a subgroup of the study sample.

The choice between the methods to account for misclassification will depend largely upon available data sources, performance, ease of implementation, and the objectives of the analyses. When internal validation data are available one may use simpler and more efficient correction formulas based on predictive values (for example ML, RC and MI). When such data are not available, the correction formulas based on sensitivity analyses are preferred because predictive values heavily depend on the unknown true values, which can vary largely across studies (8,27), while the sensitivity and specificity of a classification criterion are more likely stable across similar populations.

In the present study we focus on the internal validation design, in which all ML, RC and MI have been shown to work well (29). We chose to use MI because although the ML method has an expected small advantage in efficiency compared to RC, it must be programmed explicitly using a procedure, such as the SAS NLMIXED procedure, that is able to obtain ML estimates given a general likelihood expression (5). The MI method is easily implemented for researchers familiar with missing data methods. It has advantages regarding flexibility in the choice of analysis models which enables its extension to account for misclassification of non-binary outcomes or measurement error of continuous outcomes by altering the imputation and analysis models. Another advantage of the MI approach is that it allows the adjustment for different variables in the imputation model and the analysis models; this avoids the problem of conditioning on variables influencing only the relationship between the observed and gold-standard variable in the final analysis model (5).

RESEARCH DESIGN AND METHODS

Simulation Study

We used simulation to assess the performance of the MI approach, under internal validation designs (where validation data were available for a subgroup of the study sample), in removing bias from the HR estimates of an event (that we will refer to by disease) in group1 (refer to by exposed group) versus group 2 (unexposed group) obtained from a Cox proportional hazard model constructed with known outcome misclassification. We simulated one exposure variable, measured without error; one true event indicator and one true time-to-event outcome within a specified time period; one observed event indicator and one observed time-to-event outcome, both measured with error. Truncating the start of follow up to time zero, we generated data for a sample of size 10,000, from a Cox proportional hazard model with a constant baseline hazard. We assumed that three data sources were available for the disease outcome. The first source (observed data) provided outcome information for all study individuals that contained three sorts of errors: 1) disease is present, but has not been previously diagnosed (undiagnosed and unknown to the patient) and, therefore was not recorded in the observed data; 2) disease is present and has been diagnosed previously (known to the patient), but was not recorded in the observed data; and 3) individual is not diseased, but has been classified as diseased in the observed data. A validation study was conducted at the end of the study follow-up, assuming that the second source of data containing only the first sort of error (undiagnosed and unknown to the patient) was available for a subsample 1, and an additional third source of error free data was obtained for a subsample 2 of subsample 1 (Figure 1 of Chapter 1).

In a two-level correction procedure, subsample 2 was used in a MI procedure to impute the correct data for those individuals that were in subsample 1 but not in subsample 2. The corrected subsample 1 was then used to impute the corrected outcome status for individuals that were not in subsample 1. The steps of data generation are detailed in Supplemental File 1.

We based our simulations on the following assumptions: (i) the information from additional data source is missing at random for individuals who were not in the validation subsamples; (ii) misclassification affects only the event outcome, and errors in the time of onset occur only if the event status is misclassified; (iii) in subsamples 1 and 2 the additional information available on the true state of the disease is collected at the end of the follow-up; and (iv) there is no confounding in the data from which the estimates are calculated.

Several sets of data were simulated under various scenarios, each of which represented different values of key parameters: the true HR, sensitivity, specificity of observed disease status and size of the validation subsamples. One set of scenarios was designed to explore the efficiency of correction in terms of true HRs and degrees of misclassification at a validation participant rate of 30% that is commonly seen in surveys. We focus on high specificity (90%), as false positive disease records are assumed to be less likely to occur, and varied the sensitivity (30%-90%). Another set of simulations was to assess the performance of MI in terms of the size of validation participation subgroups for moderate misclassifications (sensitivity 50%-70% and specificity 90%) under both non-differential and differential misclassification. We focused on participation rate below 30% and differential participation rate across exposure groups.

For each of the 100 simulations across each scenario explored, validation data was first corrected to include undiagnosed disease status and time based on available error free data as mentioned before. The onset time of undiagnosed disease in subsample 1 was corrected to the time obtained based on data from subsample 2. The corrected subsample 1 was then used to correct disease status and time of onset in the full sample. More specifically, false positive observed disease status were corrected and time of onset were changed to time until censoring; for false negative cases in subsample 1, the observed time of onset was changed according to the corrected data from subsample 1; for false negative cases that were not in subsample 1, missing disease onset dates were randomly selected with replacement from the available dates stratifying on exposure status. Five imputations were performed for each incomplete data set, creating a total of 25 complete copies of data in the two steps.

We estimated the HR of the exposed versus non-exposed using Cox proportional hazard model for both the observed data and the corrected data and compared them to the true HR used in the simulations. The maximum likelihood estimates of the HR were obtained for each imputed complete data set, and the results were combined using the SAS MIANALYZE procedure (24). To compare the estimates within each scenario, we calculated the bias of the HR estimate, defined as the difference between the average estimate and the true HR; the 95% confidence limit coverage for HR, computed as the percentage of simulations in which the estimated 95% confidence limits included the true value; and the mean squared error (MSE), calculated as the sum of the squared bias and the variance of estimates.

Administrative database study

We applied the proposed approach used in our simulations to the data of a stratified random sample of 6,247 individuals obtained from the Quebec health services administrative databases of Régie de l'assurance maladie du Québec (RAMQ) and hospital abstract summary database (Med-Echo) to assess the association between statin use and new onset diabetes. Validation data were collected through a survey and fasting capillary blood glucose measurement of those who agreed to participate from the random sample mentioned above. Therefore, two subsamples were available for the validation study: a subsample 1 who self-reported diabetes status and a subsample 2 of subsample 1 who also provided blood samples for glucose testing. Potential misclassifications in diabetes status were corrected first in subsample 1 based on data from subsample 2 and then in the full sample based on the corrected data from subsample 1. This method corrects for diabetes status at a specific point in time but does not provide the time of diabetes onset. To assess this time for individuals who were misclassified on the diabetes status, we selected at random a time of diabetes onset from the pool of times of the individuals for whom times were available. For effect estimation, we matched statin users and non-users by sex and age (±2 years) at a user-to-nonuser ratio of 1:3 for both the uncorrected and corrected data. Results from Cox proportional hazard model adjusted for other selected baseline characteristics were compared.

CHAPTER 1 Correcting Hazard Ratio Estimates for Outcome Misclassification Using Multiple Imputation with Internal Validation Data – A Simulation Study

Jiayi Ni[1,2], MA; Nandini Dendukuri, PhD[1,2,3]; Kaberi Dasgupta, MD, MSc[2,3]; Aaron Leong, MD, MSc[2]; Elham Rahme, PhD[2,3]

Affiliations:

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

[2] Research Institute, McGill University Health Centre, Montreal, Canada;

[3] Department of Medicine, McGill University, Montreal, Canada;

Correspondance: Elham Rahme, Division of Clinical Epidemiology, 687 Pine Avenue West, V Building, Montreal, Quebec H3A 1A1; tel: (514) 934-1934 ext. 44724; fax: (514) 934-8293; email:elham.rahme@mcgill.ca:

Word count: 3,418

Number of figures: 3

Number of tables: 2

ABSTRACT

Objective: Multiple imputation (MI) has been shown to effectively remove bias due to outcome misclassification from effect estimates in logistic regression when validation data are available for a subgroup of study participants, but similar research that accounts for outcome misclassification in the estimation of hazard ratios (HRs) is lacking. We conducted a simulation study to assess the performance of a MI approach under different simulation scenarios to adjust for the misclassification in the event status and time of event onset in survival analyses using internal validation data.

Research Design and Methods: We based our simulations on the expectations of a cohort study aiming to assess exposure-disease association using Cox regression models for a time-to-event analysis. We assumed that the event status was misclassified for some individuals in the study sample and we simulated data for two subsamples where in subsample 1, misclassification was partially corrected, and subsample 2 of subsample 1 was error free. We then conducted a two-level validation study where in level 1, we applied the monotone logistic method of MI to correct subsample 1 and in level 2, we used the corrected subsample 1 to correct the full sample. The HR estimates from both the observed data and the corrected data were compared under different combinations of true HRs, proportions of misclassification and proportions of the validation subsamples.

Results: MI accounting for misclassification in time-to-event outcomes yielded less biased HR estimates and had appropriate 95% confidence limits coverage for both non-differential and differential misclassification between the exposed and unexposed groups under all scenarios explored, compared to estimates using the observed data only. The mean squared error (MSE) of the corrected estimates decreased as the proportion of the validation subgroups increased. Unrepresentative validation subgroups with low proportion of participation resulted in estimates with large variances. Using MSE as a criterion, the bias correction was sometimes outweighed by the added imprecision that arose from the multiple assumptions that were required to impute the data.

Conclusions: MI performs well to account for bias in HR estimates due to misclassification in time-to-event outcomes when validation data are available from representative subgroups with sufficient size. This approach is useful in addressing the true association between predictors and time-to-event outcomes and can be applied to a wide range of public health questions.

INTRODUCTION

Effect estimates in any type of observational epidemiologic research are subject to bias from misclassification. Even relatively modest degrees of misclassification can produce large amounts of bias (18,30) and may invalidate the results. An incorrect inference concerning the association between exposures and diseases can have negative psychological and financial implications, and may cause potentially serious health problems for the patient related to treatment choices and risk management.

Researchers have developed various methods to account for potential biases due to misclassification in the analysis of observational data (7,29,31-34). Methods to account for misclassification rely on information relating the observed outcome to the gold standard outcome. A validation study may be conducted, for a subgroup of the main study sample when resources are limited, to estimate the relation between the observed outcome and the gold-standard outcome. In this case, outcome misclassification can be treated as a missing data problem where the gold-standard outcome is observed for participants in the validation subgroup and is missing for all other participants (4).

Multiple imputation (MI) (25,27) has been applied to correct for exposure misclassification in estimating HRs (4), and has been shown to effectively remove bias due to outcome misclassification in effect estimates in logistic regression (5). But little research has been done to deal with misclassification of time-to-event outcomes in the analysis of exposure-outcome association.

In this paper we described an approach to account for misclassification in time-to-event outcomes using a MI approach that combines information from a large cohort with that from two validation subgroups embedded within the cohort. Whether or not to adjust for possible misclassification depend on many factors, including for instance the degree of misclassification, sample size, and the primary analytic question of interest (35,36). Since conducting validation studies are generally costly, the size of the validation subsample needed for an efficient correction also deserves consideration. With these concerns, we evaluated this approach with simulations under various representative circumstances.

RESEARCH DESIGN AND METHODS

Clinical settings

For our simulations, we assumed that observed data on exposure and disease status and time of onset were available on the sample of individuals (say from a health services database) and that the disease status of some of these individuals was misclassified in these data. We also assumed that for a subsample 1 additional data on the disease status was available to partially correct the disease status, say for example, that a subsample 1 of the original sample was interviewed and the disease status (if previously diagnosed) was correctly self-reported. However, in some patients of subsample 1 the disease was present but not diagnosed yet, so unknown to the patient. Therefore, we also assumed that a clinical test was undertaken on a subsample 2 of subsample 1 and the disease status was completely known (error free) (Figure 1). Therefore, in the full sample, 3 types of errors may have occurred: 1) disease is present, but has not been previously diagnosed (undiagnosed and unknown to the patient) and was not recorded in the database; 2) disease is present and has been diagnosed previously (known to the patient), but was not recorded in the database (false negative); and 3) individual is not diseased, but has been classified as diseased in the database (false positive).

Simulations

We based our simulations on a cohort study of an exposure-disease association over 3 years of follow-up, in a hypothetical population among which 16% of the individuals were exposed. A sample of 10,000 individuals was selected with probability sampling from the target population. The follow up period for each study participant was from the date of enrollment in the study to the end of the study. The time of enrollment for all participants was truncated to zero. Setting aside a 3% from the sample that had the disease before the start of follow up, the time from the start of follow up until the true disease onset or censoring for the rest of the sample was generated with a constant baseline hazard of 1/5000 and an normally distributed error term with zero mean and standard deviation of 30 days, yielding an approximately 10% prevalence of disease at the baseline level of the exposure variable (the unexposed). A proportion of the true disease-positive patients was assumed to be undiagnosed. For simplicity we fixed this proportion at 30% for the exposed, and 40% for the unexposed, assuming that the exposed would have seen a physician and therefore had more opportunity to be diagnosed at the time of exposure than the non-exposed. The choice of the true prevalence of disease and undiagnosed disease was motivated by diabetes research (37-39).

Information about the observed disease (event type and the time to event) was obtained for each participant in the study sample. We assumed that the validation data was available for a subgroup from the study sample (self-report validation group). Participants of this subgroup reported true information of diagnosed disease. We further assumed that a second level subgroup was selected from the self-report validation group to undergo a comprehensive clinical examination, from which the undiagnosed disease was determined. Time-to-event outcomes were of the same value in the true, observed and self-reported data if event types were the same. False positive observed time-to-event outcome for the unexposed was assumed to be uniformly distributed within the follow up period and that for the exposed followed a Beta distribution which favored events towards the start of follow up. This assumed that exposed individuals had a higher opportunity to be diagnosed for the disease because they saw a physician at the time of exposure. The steps of data generation are detailed in Supplemental File 1.

This approach assumes that the information from nonparticipants of either validation subgroups is missing at random – that is, the validated participants are a random sample of the study participants or a stratified random sample with stratification on observed variables (26). Three simplifying assumptions are also made here in generating the results. The first is that misclassification affects only the event outcome and errors in time-to-event variable present only if the event type is misclassified. The second assumption is that both self-report and clinical validations are conducted at the end of the study follow-up. A third assumption is that there is no confounding in the data from which the estimates are calculated.

Multiple imputation to correct for disease misclassification

We conducted a two-level validation study where in level 1, we applied the monotone logistic method of MI (40) to correct the self-report validation data for undiagnosed diabetes based on available clinical validation data. Clinical validated disease status was modeled using binary logistic regression on self-reported disease status and exposure status. Based on the fitted regression model, new parameters were drawn from the posterior predictive distribution of the estimated parameters and were used to predict the expected probability of having clinically validated disease for each participants in the self-report validation subsample who were absent from the clinical validation, followed by a Bernoulli draw with that probability producing imputed indicator for clinical validated disease. The onset time of undiagnosed disease was corrected to the time of the clinical validation, which was at the end of the follow up. In the second step, the corrected self-report validation data was used as a gold standard and modeled on observed disease status and exposure status to impute missing self-report data in the full study sample. False positive observed time-to-events were corrected to time until censoring. For false negative cases in the self-report validation group, the observed time-to-event was changed according to the corrected self-report data; for false negative cases by imputation, missing time-to-event outcome was obtained by a random selection with replacement from the non-missing outcome values stratified on exposure status, and shifted according to the difference in the start of follow up.

Five imputations were performed for each incomplete data set, creating a total of 25 complete copies of data in the two steps. The exposure effect on the disease was estimated using Cox proportional hazard model. We compared the true HR with results from: (a) the naïve analyses of the observed outcome; (b) the analysis of the gold-standard outcome in the whole study sample based on the 2-level MI to account for outcome misclassification; and (c) an ideal analysis of the gold-standard outcome when validation data are available for the whole study sample (complete-data). In each model, subjects having the disease before the start of follow up were excluded from the analysis. The maximum likelihood estimates of HRs were estimated for each imputed complete data set, and the results were combined using the SAS MIANALYZE procedure (40). We used the bias in the HR estimate, defined as the difference between the average estimate and the true HR; 95% confidence limits coverage for HR, computed as the percentage of simulations in which the estimated 95% confidence limits included the true value; and the mean squared error (MSE), calculated as the sum of the squared bias and the variance of estimates to assessed the efficiency of the proposed approach under different simulation scenarios, each of which represented different values of key parameters: the true HR, sensitivity (defined as 1 minus the proportion of false negative), specificity (defined as 1 minus the proportion of false positive), and the size of validation subsamples.

One set of scenarios was designed to explore the efficiency of correction in terms of true HRs and degrees of misclassification at a validation participant rate of 30% that is commonly seen in surveys. Another set of simulations was to assess the performance of MI in terms of the size of validation participation subgroups for moderate misclassifications (sensitivity 50%-70% and specificity 90%) under both non-differential and differential misclassification. In the later set of scenarios we focused on participation rate below 30% and different participation rate across exposure groups, and showed through two examples that adjustment for differential participation is necessary to prevent worsening the bias caused by an erroneous correction. Participation rate in the clinical validation was fixed at 50% in all scenarios. In each scenario, the HR for the effect of exposure on developing disease was estimated and summarized over 100 simulations.

Data simulations and analyses in this study were performed with SAS version 9.3 statistical software (SAS Institute Inc., USA).

RESULTS

Efficiency of correction in terms of the true hazard ratios and the degrees of misclassification in observed disease

Correction with MI markedly reduced bias due to outcome misclassification in HR estimates for every scenario examined (Table 1). Effect estimates of the observed data were biased in both non-differential and differential misclassification, with bias (-0.40 to 0.33) varying with the change of true HR and sensitivity. In contrast, the correction based on MI yielded estimates with less bias (-0.07 to 0.07) than the naïve analysis of the observed outcome in all combinations of true HR and degrees of misclassification explored.

The ideal analysis with no missing information of the gold-standard outcome did not yield totally unbiased HR estimates. Bias in that analysis arose with the assumption concerning the unknown disease onset date for the undiagnosed cases determined by clinical validation, a slight underestimate of the true association was observed in all scenarios explored. Bias in HR estimates by MI was similar to bias in estimates obtained in the ideal analysis.

Confidence limits from the analysis based on MI maintained high coverage (94-99%) for all scenarios explored, while that from the observed data analyses showed poor coverage (1-86%) varied as a function of the true HR and sensitivity.

For a study sample of 10,000 subjects and the same validation participation proportions of 30% in both exposure groups, the bias-corrected HR estimates generally had equal or smaller mean squared errors than the naïve estimates only when the naïve estimates suffered severe bias. In many scenarios the reduction of bias by correction was offset by added imprecision.

Efficiency of correction in terms of the size of validation subgroups

Table 2 presents the results summarized from 100 samples with self-report validation participation of 10%, 20% and 30% in both exposure groups, as well as different participation with 20% in one group, 40% in the other and vice versa. Results from the complete data scenario of having full participation in both self-report and clinical validation groups are also listed.

For all examined combinations of validation percent, the corrected HR estimates showed notable reduction in bias and appropriate coverage (96-100%) under both non-differential and differential misclassification. With low self-report validation participation of 10%, the correction method yielded estimates with larger variation (MSE 0.13 for the corrected data versus 0.10 and 0.03 for the uncorrected data under non-differential and differential misclassification, respectively). The MSE of the corrected HR estimates decreased as the validation percent increased. With validation participation of 20% and above, the corrected estimates based on MI (bias -0.07 to -0.06) were similar to the estimates from the ideal analysis of complete data (bias -0.06).

Under differential participation, the bias-correction method yielded estimates with less bias (-0.02) when the exposed group had higher validation percent than the unexposed, compared to the situation when the exposed group had lower validation percent (bias was -0.10 to -0.09). The resulting HR estimates from differential participation varied around the estimates from the complete data analysis.

When the participation percent of validation studies was not the same across exposure groups, our results showed that MI should account for this difference – that is, the exposure status must be included as a covariate in the prediction model of imputations. Figure 1 and Figure 2 compared the corrected HR estimates with and without adjustment of exposure status to the estimates of observed data under non-differential and differential misclassification, respectively. The corrected estimates adjusted for different validation participation appeared less biased than the estimates from observed data under all 4 scenarios explored, while those ignoring participation differences were sometimes more biased than the uncorrected estimates.

DISCUSSIONS

MI performed well at removing bias due to outcome misclassification from HR estimates in the scenarios explored through simulation with the two-step validation setting. The MI approach combined the complete-data analyses to obtain an estimate that was corrected for missing validation data and at the same time accounted for the uncertainty of the whole validation procedure. Estimates obtained from the MI procedure were similar in magnitude to estimates from the complete data using the gold-standard outcome.

The two-step design that we used in our approach is applicable in a wide range of research areas. In practice, self-report validation data could be obtained through survey interviews or chart reviews. Clinical validation can be in various formats depending on the disease of interest, for instance, a blood pressure test for hypertension or a blood glucose test for diabetes.

One key requirement for any method using a validation study to account for misclassification is to obtain a representative validation subgroup that allows correct estimation of the relation between the gold-standard and misclassified variable. If this relation is inconsistent between validation subgroups and the whole sample, then the estimates of the exposure-outcome association based on MI may be biased. As noted in the simulations, the bias in HR estimates was markedly reduced and similar to that from the complete data analysis when the validation participation proportion was 20% and more. For a sample size of 10,000 individuals, 20% self-report validation participation seems sufficient for yielding relatively stable estimates, as the mean squared error of the correction method were smaller than the one obtained from lower participation proportions and did not vary a lot when participation proportion was further increased. For studies with smaller sample size, an increase of the required validation proportion is necessary in order to obtain a sufficient number of subjects and reduce the bias in the HR estimates. In our setting of a two-level validation study, poor participation in the self-report validation would lead to an even smaller subgroup in the second level clinical validation, thus the maximum likelihood estimates may not even exist because of complete or quasi-complete separation outcome of the logistic regression in the monotone logistic method of imputation. In this case, alternative estimation methods such as the Firth's penalized likelihood should be applied to reduce the small-sample bias and to produce finite and more consistent estimates (41-43).

Because participation in the validation study determines if the gold-standard outcome is missing or not, the probability of participation must be independent of that participant's gold-standard outcome given the observed outcome and the covariates that may intrude participation. When collection of self-report validation information is intruded by particular attributes, subject refusal may result in a non-random sample in the validation study. We showed through simulations that in such situations, corrections can even worsen the bias of estimates if differential participation was ignored during imputations. The proposed MI approach was based on the assumption that the information from the validation studies were missing at random. Therefore in the case of a non-random selection of the validation participants, adjustment on observed covariates that may affect the participation in the validation studies must be considered to ensure that the validated participants are a stratified random sample with stratification on the observed variables.

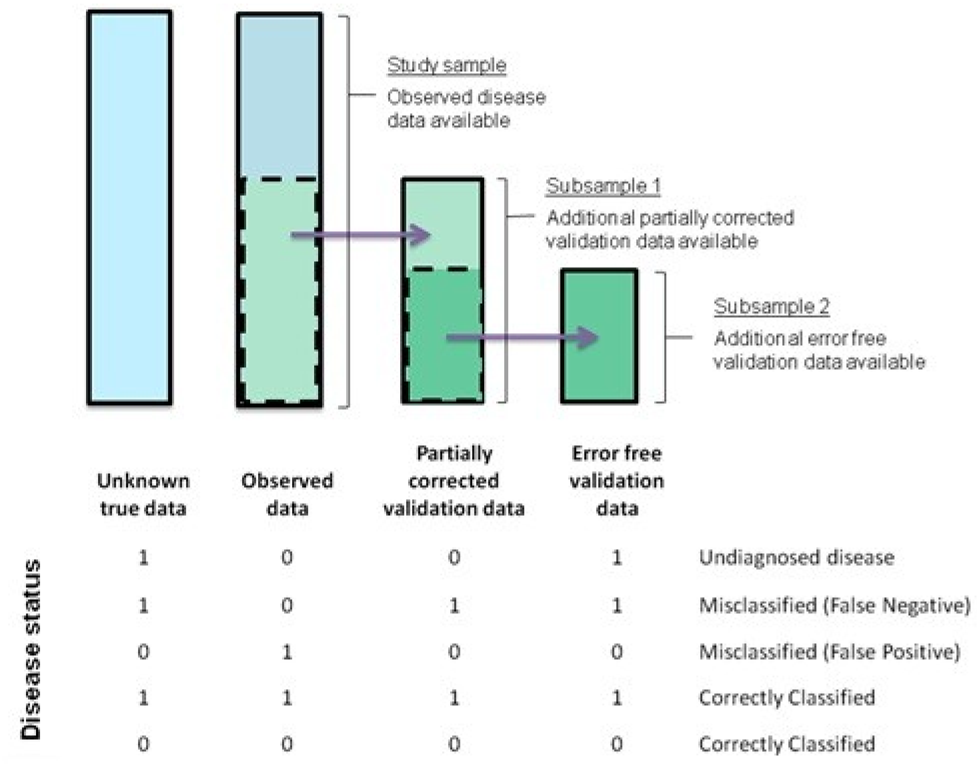
It has been pointed out that adjusting for misclassification of a binary outcome in logistic regression is not always beneficial because although the adjustment can reduces the bias, it can also inflates the variability, yielding an estimator with a larger MSE than an unadjusted method (42). Our findings in Cox proportional hazard model agreed with this previous finding. As shown by the simulations, if the primary objective was to estimate the strength of an exposure-outcome association, the corrected estimates can perform less well than the uncorrected one because of the added-in variance that overwhelmed the reduction in bias, especially under poor validation participation proportions or when the uncorrected estimates do not have severe bias. On the other hand, the corrected HR estimates always acquired large gain in the confidence limits coverage. Therefore, adjusting for possible misclassification in time-to-event outcome is necessary if the primary goal is CI construction or hypothesis testing regarding the parameters.

The limitation of any validation for undiagnosed disease is that the disease status can only be observed by the time when the validation is conducted, while the disease may have occurred at any time before that. The proposed correction method related the observed outcome to the gold-standard outcome measure, which was the self-reported disease corrected for undiagnosed disease. However, the corrected self-report validation data was an imperfect gold standard because it did not reflect the exact true time of the disease onset for every subject. Minor bias still presented even when the validation information was available for all individuals in the study sample. The slight underestimate of the true HR as noted in the simulations was likely to arise with the assumption that the disease onset date for undiagnosed cases was at the date of validation studies for both exposure groups.

Because MI relies on valid estimates of the relation between the gold standard and misclassified variable, concern was raised that the gold-standard variable itself was not a true gold standard and was subject to unknown errors. For example, self-report validation data can introduce recall bias due to inaccurate or incomplete report by study participants regarding events or experiences from the past (44). In such situations, correction methods relying on the validation data may bias the estimates and yield false results.

In addition, the proposed method did not account for the delay in diagnosis, in which case the event indicator is not misclassified but the observed time-to-event outcome presents measurement errors. An example is the delay in diagnosis of diabetes, which can occur because of either lack of medical visits or glucose measurement, or clinical inertia (45). Difference in the probability of delay in diagnosis between the exposure groups can cause bias towards either direction. On the other hand when a disease has not yet been suspected or diagnosed is associated with an exposure that actually results from early signs and symptoms of the disease, the delay in diagnosis will create another problem of protopathic bias which may misleadingly suggest causation (46-48).

Despite these potential limitations, the proposed approach supports the fact that effect estimates in Cox proportional hazard model can be corrected for outcome misclassification using data from validation studies on subgroups of the main study sample. Under the assumptions of this study, MI performs well to account for bias in HR estimates under the scenarios explored through simulation. This approach is useful in addressing the true association among predictors and time-to-event outcomes and can be applied to a wide range of public health questions.


Study sample
Observed disease data available
Subsample 1
Additional partially corrected validation data available
Subsample 2
Additional error free validation data available

	Unknown true data	Observed data	Partially corrected validation data	Error free validation data	
Disease status	1	0	0	1	Undiagnosed disease
	1	0	1	1	Misclassified (False Negative)
	0	1	0	0	Misclassified (False Positive)
	1	1	1	1	Correctly Classified
	0	0	0	0	Correctly Classified

Figure 1: Clinical settings for the simulation study.

Study sample: observed data on disease status and time of onset were available; 3 types of errors may have occurred: undiagnosed disease, false negative and false positive diagnosis. Subsample 1: additional data on the disease status was available to partially correct the disease status; only the undiagnosed disease may have occurred. Subsample 2: embedded within subsample 1 and error free data on disease status was available.

Table 1: Bias, 95% confidence limit coverage and mean squared error for a self-report validation percent of 30%, 100 samples of 10,000 individuals under 18 scenarios in terms of true hazard ratio and degrees of misclassification

True Hazard		Specificity	Sensitivity	Observed Data			Corrected Data			Complete Data		
Ratio				Bias[b]	Cover[c]	MSE[d]	Bias	Cover	MSE	Bias	Cover	MSE
1.5	Non-differential	0.9	0.9	-0.22	35	0.06	-0.05	97	0.04	-0.06	100	0.02
	Misclassification		0.6	-0.30	19	0.10	-0.06	97	0.05			
			0.3	-0.40	10	0.17	-0.07	95	0.05			
	Differential	0.9	(0.95, 0.9)	-0.18	51	0.04	-0.05	97	0.04			
	Misclassification[a]		(0.9, 0.8)	-0.16	61	0.03	-0.06	97	0.04			
			(0.7, 0.5)	-0.14	67	0.03	-0.06	96	0.05			
1.12	Non-differential	0.9	0.9	-0.02	86	0.01	0.01	96	0.03	-0.03	100	0.01
	Misclassification		0.6	-0.03	81	0.01	0.01	94	0.04			
			0.3	-0.08	72	0.01	0.00	96	0.04			
	Differential	0.9	(0.95, 0.9)	0.04	86	0.01	0.02	94	0.03			
	Misclassification		(0.9, 0.8)	0.07	77	0.01	0.01	96	0.03			
			(0.7, 0.5)	0.11	63	0.02	0.01	95	0.04			
0.8	Non-differential	0.9	0.9	0.21	15	0.05	0.07	98	0.02	0.01	100	0.01
	Misclassification		0.6	0.22	22	0.05	0.07	97	0.03			
			0.3	0.22	25	0.06	0.06	97	0.02			
	Differential	0.9	(0.95, 0.9)	0.23	10	0.06	0.07	99	0.02			
	Misclassification		(0.9, 0.8)	0.26	7	0.07	0.07	97	0.02			
			(0.7, 0.5)	0.33	1	0.12	0.07	98	0.03			

[a] *Sensitivity differs by exposure groups: presented as (sensitivity for the exposed, sensitivity for the unexposed).*

[b] *Bias is the difference between the average estimated hazard ratio and the true hazard ratio.*

[c] *Confidence limit coverage is defined as the percentage of the simulations that the estimated 95% confidence limits included the true value.*

[d] *MSE, mean square error, is the sum of squared bias and the variance of estimated hazard ratios.*

Table 2: Bias, 95% confidence limit coverage and mean squared error for a true hazard ratio of 1.5, 100 samples of 10,000 individuals under 10 Scenarios in terms of self-report validation participation percent

				Observed Data			Corrected Data		
	Specificity	Sensitivity	Self-report Validation Percent[b]	Bias[c]	Cover[d]	MSE[e]	Bias	Cover	MSE
Non-differential Misclassification	0.9	0.6	10	-0.30	19	0.10	0.01	99	0.13
			20	-0.30	19	0.10	-0.07	96	0.07
			30	-0.30	19	0.10	-0.06	97	0.04
			(20, 40)	-0.30	19	0.10	-0.09	98	0.05
			(40, 20)	-0.30	19	0.10	-0.02	98	0.05
			100[f]	-0.30	19	0.10	-0.06	100	0.02
Differential Misclassification[a]	0.9	(0.7, 0.5)	10	-0.14	67	0.03	0.01	99	0.13
			20	-0.14	67	0.03	-0.07	98	0.07
			30	-0.14	67	0.03	-0.06	97	0.05
			(20, 40)	-0.14	67	0.03	-0.10	97	0.05
			(40, 20)	-0.14	67	0.03	-0.02	98	0.05
			100[f]	-0.14	67	0.03	-0.06	100	0.02

[a] *Sensitivity differs by exposure groups: presented as (sensitivity for the exposed, sensitivity for the unexposed).*

[b] *Percent of participants in the self-report validation subgroup. Different self-report validation proportions in exposure groups are presented as (validation percent for the exposed, validation percent for the unexposed)*

[c] *Bias is defined as the difference between the average estimated hazard ratio and the true hazard ratio.*

[d] *Confidence limit coverage is defined as the percentage of simulations that the estimated 95% confidence limits included the true value.*

[e] *MSE, mean square error, is the sum of squared bias and the variance of estimated hazard ratios.*

[f] *Ideal analysis of 100% participation in both self-report and clinical validation (complete data); clinical validation percent was 50% in other scenarios.*

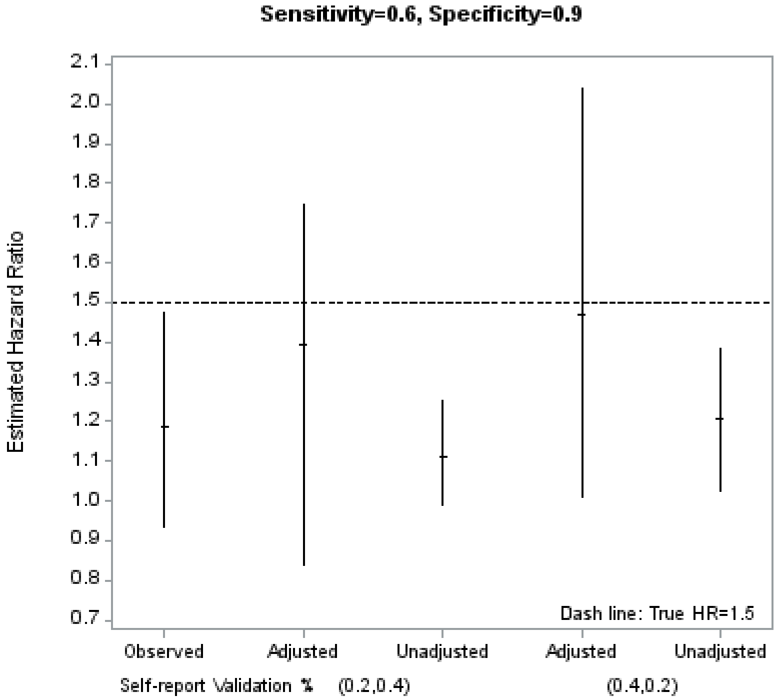


Figure 2: Vertical bars showing the ranges and means of hazard ratio estimates with and without adjustment of differential validation participation for a true hazard ratio of 1.5, 100 samples of 10,000 individuals under 2 Scenarios of non-differential misclassification

Sensitivity is 0.6 and specificity is 0.9 for both exposure groups. Adjusted: corrected estimates based on multiple imputation with adjustment for differential validation participation proportion between exposure groups; Unadjusted: corrected estimates based on multiple imputation without adjustment for differential validation participation proportion between exposure groups.

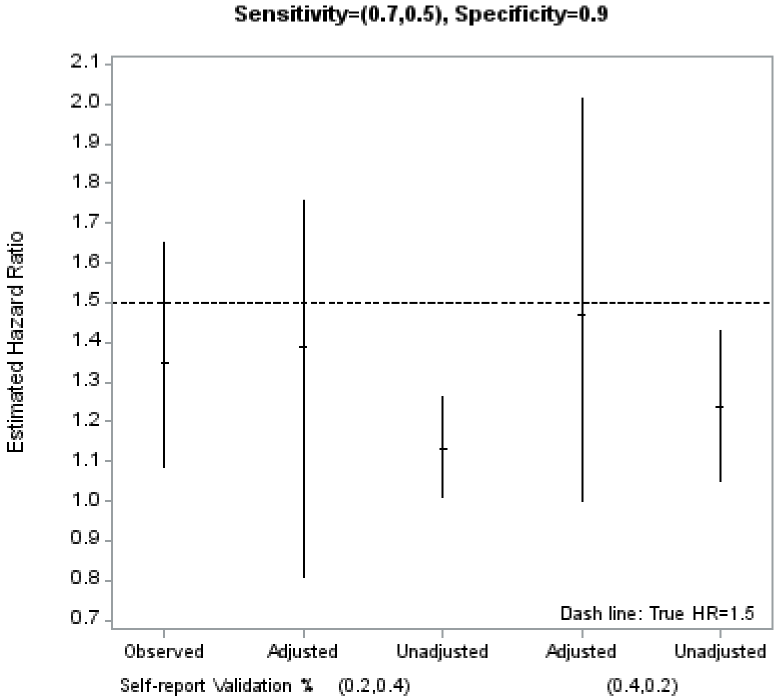


Figure 3: Vertical bars showing the ranges and means of hazard ratio estimates with and without adjustment of differential validation participation for a true hazard ratio of 1.5, 100 samples of 10,000 individuals under 2 Scenarios of differential misclassification

Sensitivity is 0.7 for the exposed group and 0.5 for the unexposed group; specificity is 0.9 for both exposure groups. Adjusted: corrected estimates based on multiple imputation with adjustment for differential validation participation proportion between exposure groups; Unadjusted: corrected estimates based on multiple imputation without adjustment for differential validation participation proportion between exposure groups.

CHAPTER 1 CONCLUSION

Viewing misclassification as being due to missing information of the gold-standard variable allows the use of well-developed missing data methods to account for the inherit bias. The multiple imputation approach described in chapter 1 is applicable in any setting in which a primary study sample provides variables measuring health outcomes and/or risk factors, but validation data are available only for a subsample representative of the primary study sample. For situations in which the event type is known to be subject to error but the exact event date cannot be observed, the proposed method provides an option to account for misclassification in time-to-event outcome.

This approach was developed as part of a study of the association between statin use and new-onset diabetes on a random sample from the Quebec health services administrative databases. For simplicity, we assumed in the simulation study that all covariates were balanced across exposure groups. Yet in practice, retrospective studies of exposure-disease association may not be adequate to assess temporal relationships because of various confounders. While confounding by indication can be, at least in part, accounted for by adjusting for baseline characteristics, misclassification or information bias are usually not easy to control by researchers because these rely on the accuracy of record-keeping in the database. Misclassifications in disease status will affect not only the outcome, but also subject inclusion at baseline. Inclusion of individuals with undiagnosed disease before exposure in the study may lead to a false estimate of the exposure effect, when the disease is later discovered.

In chapter 2, we applied this multiple imputation approach to account for misclassification in administrative database definition of physician-diagnosed diabetes using data from an internal validation study involving a population-based survey and a blood glucose measurement. Effect estimates of statin initiation on physician-diagnosed diabetes and on the gold-standard outcome (self-reported diabetes and/or elevated fasting plasma glucose) were compared.

CHAPTER 2 Correcting for Outcome Misclassification in Health Administrative Database Study Using Internal Validation Data

Jiayi Ni[1,2], MA; Aaron Leog, MD, MSc[2]; Kaberi Dasgupta, MD, MSc[2,3]; Jean-Louis Chiasson, MD[4,5]; Nandini Dendukuri, PhD[1,2,3]; Elham Rahme, PhD[2,3]

Affiliations:

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

[2] Research Institute of the McGill University Health Centre, Montreal, Canada

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

[4] Department of Medicine, University of Montreal, Montreal, Canada

[5] Hotel-Dieu Hospital, Hospital Centre of the University of Montréal, Montreal, Canada

Correspondance: Elham Rahme, Division of Clinical Epidemiology, 687 Pine Avenue West, V Building, Montreal, Quebec H3A 1A1; tel: (514) 934-1934 ext. 44724; fax: (514) 934-8293; email:elham.rahme@mcgill.ca.

Word count: 3,216

Number of figures: 1

Number of tables: 5

ABSTRACT

Objective: We aimed to assess the association of statin use and new-onset diabetes using data obtained from the health services administrative databases on a random sample of the Quebec population and to correct this effect estimate for outcome misclassification using internal validation data from self-report and glucose measurement.

Research Design and Methods: A stratified random sample of 6,247 individuals from the province of Quebec were surveyed and asked to provide self-collected fasting capillary blood samples. By using multiple imputation (MI), self-reported diabetes in the validation subgroup was updated to include elevated fasting plasma glucose (≥7 mmol/L for undiagnosed diabetes and 6.1-6.9 mmol/L for impaired fasting glucose - IFG), and used as a reference standard to correct the misclassification in administrative database definition of physician-diagnosed diabetes. Association between statin treatment and new-onset diabetes was estimated using Cox proportional hazard model and corrected for bias due to misclassification in physician-diagnosed diabetes.

Results: The survey included 3,322 participants who consented to administrative record linkage. Among these, 1,599 (48.1%) participants provided analyzable blood samples. Prevalence of physician-diagnosed diabetes among the primary study sample was 10.6% (95% CI 9.8-11.4%). When self-reported diabetes or undiagnosed diabetes was considered as the gold standard for diabetes case assessment, estimated proportion of positive cases was 14.7% (13.8-15.6%) among the entire primary study sample. The hazard ratio (HR) comparing statin users with non-users on the gold-standard outcome was 1.49 (0.95-2.34) compared to 1.61 (1.09-2.38) on physician-diagnosed diabetes. When IFG was also included into the gold standard, the estimated proportion of positive cases increased to 29.0% (27.9-30.1%) and the HR was 1.36 (0.92-2.01).

Conclusions: The association between statin use and diabetes occurrence was overestimated when misclassification in physician-diagnosed diabetes was ignored. Our findings highlight the importance of accounting for these misclassifications to prevent erroneous results in studies based on administrative database.

INTRODUCTION

Statins have been shown to reduce the risk of cardiovascular events in numerous trials of primary and secondary prevention populations, providing an apparently low risk approach to improve cardiovascular health (19). However, evidence from recent large randomized trials has revealed an association between statin therapy and new-onset diabetes (20,21). A meta-analysis including 13 placebo-controlled randomized trials with a total of 91,140 participants showed a 9% higher odds of incident diabetes (95% CI 2-17%) in the statin arm (20). Another meta-analysis of 5 secondary prevention trials that compared intensive- versus moderate-dose statin treatment reported a 12% (4-22%) excess risk of new-onset diabetes in the intensive-dose regimen arm (21). One analysis of electronic medical records conducted in the U.K. also showed an increased risk of diabetes (hazard ratio, HR 1.14 [1.10-1.19]) among statin users compared with non-users, and indicated that such risk cannot be explained by longer survival among statin users (23). Another observational study in postmenopausal women also reported a positive association between baseline statin use and subsequent risk of diabetes (22).

Although these studies included information of important confounders and new-onset diabetes was rigorously screened for in studies prospectively designed to look for diabetes, the recruitment of patient was based on previous diagnosis records. Part of the observed association between stain use and the risk of diabetes might be attributable to the bias due to misclassification of diabetes status that cannot be fully controlled for in the analyses. This arises with the situation in which statin is prescribed to subjects to reduce the risk of cardiovascular events. The treated subjects may be at higher risk of diabetes because of multiple cardiovascular risk factors that are also risk factors for diabetes, such as obesity and sedentary lifestyle, or even have undiagnosed diabetes or prediabetes (49) before receiving statin therapy. When individuals with diabetes that has not been brought to medical attention are included in the study and diabetes is later discovered, an association may be incorrectly inferred between statin and the risk of diabetes even after adjusting for all measured confounders (Figure 1). Neither the randomized trials nor the observational studies addressed this possibility.

Misclassification of diabetes status would affect both subject inclusion at baseline and the outcome occurrence during follow-up. In this paper, we proposed an approach using internal validation data to account for such misclassification in estimating the association between statin use and new-onset diabetes on a random sample from the Quebec health services administrative databases.

RESEARCH DESIGN AND METHODS

The primary study

Data source

A stratified random sample of 6,247 individuals was generated from the database of the Quebec drug insurance program administered by the Régie de l'assurance maladie du Québec (RAMQ). Physician and prescription drug claims, hospital abstract and demographic records for the period of 1 January 1997 to 31 December 2010 were obtained for these individuals. The physician claims database and the hospital discharge abstract database cover the entire population in Quebec, while coverage of the prescription drug database includes all residents 65 years or older, those receiving social assistance and individuals in the working force who do not have collective private drug insurance. During the study period, the prescription drug database provided information on about 92%

of the Quebec population aged 65 years or older and 32% of those aged under 65 years (50). These databases are linkable by a unique patient identifier. Permission to link the data was obtained from the Provincial Ethics Board, the Commission d'accès à l'information.

Study cohort

Statin users from 1 January 1999 to 2 October 2010 (90 days before the end of the study period) were identified from our study sample and matched to non-users by age (±2 years) and sex at a user-to-nonuser ratio of 1:3. A non-user was an individual who neither used statin within one year prior to the date of statin initiation of his/her matched user, nor started using stain within 90 days after this date. We included individuals who were covered by the drug insurance plan in the past year of the matched statin initiation date and in the whole follow-up period as defined below, did not have any physician claims that fulfilled the International Classification of Diseases 9 (ICD-9) codes for diabetes or hospitalization with a principal or secondary diagnosis of diabetes within the two-year period prior to the matched statin initiation date, and did not use anti-diabetic drugs within one year prior to the matched statin initiation date. We excluded individuals (cases or potential controls) who died within 90 days of the matched statin initiation date.

The outcome of interest was the time from the matched statin initiation date until diabetes onset. A diabetes case was defined by one or more physician claims for diabetes and/or one or more hospitalizations with a primary or secondary diagnosis of diabetes and/or use of anti-diabetic drugs. The date of diabetes onset was the earliest date of physician claims or hospital admissions for diabetes or prescriptions of anti-diabetic drugs. Subjects were followed until diabetes onset, death, the end of the study period, or statin initiation for non-users, whichever came first.

Statistical Analyses

A Cox proportional hazards model, stratified by the matched sets, was used to estimate the hazard ratio (HR) of new-onset diabetes in statin users versus non-users during the follow-up. Potential confounders in addition to age and sex included hospitalization, hypertension, ischemic heart disease, heart failure and cancer identified in the past year of the matched statin initiation date, as well as a social deprivation index (scored from 1 to 5 on the basis of quintiles of six area-based socio-economic indicators that were know for their relations with health (51);). These variables were available from the administrative database and were thought to affect both statin initiation and the risk of developing diabetes (52-54).

The validation study

Self-report validation data were collected from a survey that we conducted with the assistance of the Institut de la Statistique du Quebec (ISQ). The individuals in the primary study sample were surveyed via telephone or mail from March 31 to July 14 2009 to inquire about their diabetes status with the survey question: "Q1: Have you ever been told by a doctor or another health professional that you had diabetes?" (other than during pregnancy was added for women), followed by the question "Q1a: How old were you when you were first diagnosed with diabetes?" if the answer to the previous question was "Yes". Self-reported date of diabetes onset was calculated as carrying backwards the date of the survey by the difference between participants' current age and their age at the first diagnosis, or assigned to be the date of the survey if the answer to Q1a was missing. We assumed that participants reported their true diabetes status or were aware of their borderline glucose level. Survey participants were also asked to provide a self-collected fasting capillary blood sample using the materials and instructions provided by ISQ via mail. The blood samples were analyzed for glucose measurement at the central laboratory at Saint-Luc Hospital in Montreal. More details on the test procedure have been published elsewhere (55).

Survey and glucose results were linked to the administrative data of the primary study. For simplicity, we assumed that both the survey and the blood test were conducted on March 31, 2009 for all participants, because the date of the interview and blood test mailing were not recorded. We treated misclassification as a missing data problem and applied multiple imputation (MI) (26) in two steps.

Subjects with elevated fasting plasma glucose (FPG) level were identified from the survey participants that provided analyzable blood samples. Two thresholds were applied: ≥7 mmol/L for diabetes and 6.1-6.9 mmol/L for impaired fasting glucose (IFG) based on the guidelines for diagnosis of diabetes and prediabetes (56). Elevated FPG was modeled on self-reported diabetes and selected subject baseline characteristics, including sex, age, social deprivation index, as well as hospitalization, hypertension, ischemic heart disease, heart failure and cancer in the past year of the survey to impute missing indicator of elevated FPG for survey participants without glucose measurements. Positive cases by self-report or glucose measurement excluded diabetes occurring in women only during pregnancy (gestational diabetes) according to the information collected in the survey. Five imputations were performed using logistic regression method for monotone missing data (26). For participants who self-reported no diabetes yet had elevated FPG (IFG or undiagnosed diabetes according to the threshold applied), diabetes onset date was corrected to the date of the blood test.

Survey response and/or elevated FPG were then used as a gold standard to correct for misclassification in physician-diagnosed diabetes. For survey participants who consented to RAMQ linkage, self-reported diabetes and/or elevated FPG was modeled on physician-diagnosed diabetes before the survey and subject characteristics to impute missing gold standard values for survey non-participants using monotone logistic regression method. Five imputations were performed for each imputed dataset from the previous step, creating a total of 25 copies of complete datasets in the two steps. The time-to-event outcome for false positive physician-diagnosed diabetes was corrected as time until censoring. For false negative cases, the date of diabetes diagnosis was changed to the date by self-report or elevated FPG for survey participants, and imputed based on subject characteristics using Markov Chain Monte Carlo method for survey non-participants.

We then repeated the matching and data analyses as in the primary study for each of the 25 complete datasets that were corrected for undiagnosed diabetes and misclassification in physician-diagnosed diabetes. The results were combined using the SAS MIANALYZE procedure (40). The analyses in this study were performed with the SAS version 9.3 statistical software (SAS Institute Inc., Cary, USA).

RESULTS

Primary and validation study samples

The response rate of the survey among the 6,247 individuals in the primary study sample was 56.1%, comprising 3,506 participants among whom 95.8% (n=3,322 [53.2% of primary study sample]) agreed to a record linkage between survey and health administrative data. Among survey respondents who consented to record linkage, 48.1% (n=1,599 [25.6% of primary study sample]) provided analyzable blood samples.

Individuals in the primary study sample had a mean age of 49.7 (standard deviation [SD] 16.4) years and were equally distributed between males and females (Table 1). The proportion of physician-diagnosed diabetes was highest among survey respondents who provided analyzable blood samples (14.3% [95% CI 12.6-16.1%]) followed by overall survey respondents (11.8% [10.7-12.9%]) and the primary study sample (10.6% [9.8-11.4%]). Survey respondents and participants who provided analyzable blood sample were comparable with the primary study sample for other baseline characteristics.

Correction for misclassification in physician-diagnosed diabetes

The proportion of self-reported diabetes was 11.3% (95% CI 9.8-12.9%) among survey participants who provided analyzable blood samples (Table 2). By including participants with elevated FPG, the proportion of self-reported diabetes or undiagnosed diabetes increased to 16.9% (15.1-18.8%); MI methods adjusting for baseline characteristics yielded an estimate of 15.4% (14.2-16.7%) among all survey respondents (Table 3). Proportion of self-reported diabetes or IFG/undiagnosed diabetes was 31.2% (29.0-33.5%); MI yielded an estimate of 29.7% (28.1-31.2%) among survey respondents.

Prevalence of physician-diagnosed diabetes among survey respondents was 11.8% (10.7-12.9%) (Table 2). When self-reported diabetes or undiagnosed diabetes was the reference standard, MI yielded an estimate of 14.7% (13.8-15.6%) true diabetes among the entire primary study sample (Table 3). When IFG was also included into the reference standard, the estimated proportion of true diabetes was increased to 29.0% (27.9-30.1%).

Association between statin initiation and new-onset diabetes

Table 4 presents the length of follow-up, rate of new-onset diabetes and baseline characteristics by statin initiation status for the matched data under each classification criteria. The cohort included 1,612 individuals, 403 of whom were statin users and 1,209 were the matched. Individuals in the matched dataset had a mean age of 63.9 (SD 11.3) years at matched statin initiation date. A larger proportion of statin users had hospitalization, ischemic heart disease, or used anti-hypertensive drugs in the past year. Cancer in the past year and social deprivation index were evenly distributed among users and non-users.

During 7,720 person-years of follow-up, there were 148 new cases of physician-diagnosed diabetes. By comparison with non-users, statins appeared to accelerate the average time to diagnosis of diabetes by 6 months (97 [SD 78.4] months on users versus 103 [89.6] months on non-users). The crude rate of new-onset diabetes was 28.9 per thousand person-years among users and 15.6 among non-users. An elevated risk of new-onset diabetes was observed among statin users (HR 1.61 [95% CI

1.09-2.38]) compared to non-users, adjusted for selected baseline characteristics (Table 5).

Using self-reported diabetes or undiagnosed diabetes as the gold-standard outcome, the baseline characteristics averaged over 25 matched dataset did not change importantly. The crude rate of new-onset diabetes increased to 39.5 and 24.2 per thousand person-years among users and non-users, respectively. No significant association was observed between statin initiation and new-onset diabetes (HR 1.49 [0.95-2.34]). The estimated HR was further reduced to 1.36 (0.92-2.01) when IFG was also classified as a positive case.

DISCUSSIONS

Our study demonstrates that a substantial proportion of diabetes cases were misclassified by administrative data compared to self-report and glucose measurements. The association between statin treatment and diabetes was overestimated when misclassification in administrative definition of physician-diagnosed diabetes was ignored (HR 1.61 [95% CI 1.09-2.38] for physician-diagnosed diabetes versus 1.49 [0.95-2.34] corrected by self-reported diabetes or undiagnosed diabetes). Association was less strong when impaired fasting glucose was also included as a positive case (HR 1.36 [0.92-2.01]).

In administrative claims data, it is usually impossible to determine the exact timing of diabetes onset. As with many other diseases in observational research, the date of the first encounter with the health service system with the record of the disease occurrence is used as a proxy for the timing of disease onset. Statin users usually have more frequent physician visit and diabetes are more likely to be discovered earlier in them than in non-users. Part of the observed overestimation of the association between statin and diabetes based on administrative database definition may be explained by different frequency of physician visit.

We applied two different thresholds to define elevated fasting glucose. These thresholds are aligned with current clinical practice guidelines for diabetes diagnosis (≥7 mmol/L) and impaired fasting glucose (6.1-6.9 mmol/L), respectively, by fasting plasma glucose (56-58). One third (90 among 270) of the true diabetes cases among participants who provided analyzable blood samples did not report physician diagnosis of diabetes and were captured by glucose test results (FPG≥7 mmol/L). Using self-reported diabetes or undiagnosed diabetes as the "gold" standard, MI yielded an estimate of 14.7% (95% CI 13.8-15.6%) for true diabetes in the primary study sample. Lowering the threshold to 6.1 mmol/L increased the number of missed positive cases by self-report, thus consequently doubled the estimated proportion of positive cases in the primary sample to 29.0 (27.9-30.1).

Differential misclassification between statin users and nonusers was likely to arise from the fact that statin was prescribed for subjects with multiple cardiovascular risk factors such as high blood pressure or abnormal cholesterol and triglyceride levels. These patients are likely routinely screened for diabetes at physician visits or may be under continuous self-monitoring of blood glucose as suggested by diabetes prevention guidelines (59). Therefore, diabetes cases among these individuals are less likely to be missed in physician diagnosis compared to low-risk individuals.

Rigorous definitions for identifying diabetes patients in administrative claims require multiple physician visits over time, in order to exclude cases where diabetes was not confirmed in subsequent testing. The definition of diabetes case by the National Diabetes Surveillance System (NDSS) is having two or more physician billings for diabetes and/or one or more hospitalizations for diabetes within a 2-year period from administrative database (60-62). The administrative data of our study only covered one year after the survey; this somehow limited tracking for new diabetes cases based on the NDSS definition. Therefore in this paper, we applied a more flexible algorithm requiring only a single physician billing. Single physician claim definition decreases the number of claims required for case definitions, thus has improved sensitivity but reduced specificity when compared to the gold-standard outcome (60,63). Nevertheless, corrections based on MI rely on the correct specification of the model relating the gold-standard outcome to the observed outcome. As long as the relationship between the gold-standard and the observed outcomes (either the 1-claim or 2-claims definition) conditioned on other confounders are transportable from the validation subgroup to the whole study sample, a correct imputation of the missing values of the gold-standard outcome could be obtained.

There are potential limitations for the reference standards used to define diabetes. Some clinical practice guidelines recommend verifying screening test results with a second glucose test, before making a clinical diagnosis (57,58,64). In our study, a single mailed-in fasting blood sample test has potential sources of measurement error. Participants may not follow written instructions completely (eg: collecting non-fasting blood samples or incorrect hand washing). The time elapsed from sampling to receipt of the mailed-in sample could affect measurement precision. Information bias could be introduced in the survey by participants' inaccurate recall, lacking comprehension of survey questions, or incomplete knowledge of diagnoses. Surveys and self-collected blood sample tests may also be subject to selection bias as they allow subjects to self-select into the subgroup (65). We observed a higher prevalence of physician-diagnosed diabetes cases among survey respondents and among those who provided mailed-in blood samples, suggesting that higher-risk individuals were generally more willing to participate. The validity of self-report is also affected by demographic characteristics such as age and sex (66,67). We adjusted for potential selection bias by basing MIs on the baseline characteristics selected from administrative data that were available for both respondents and non-respondents. We acknowledge that residual bias may have remained due to unmeasured confounders that differed between respondents and non-respondents.

In addition, our survey collected data on the age at the first physician diagnosis of diabetes; this information was later used to calculate the self-reported date of diabetes onset. The resulting date was a rough estimation with respect to the date of the survey and subjects' age at the survey. Therefore, to avoid introduction of more recall bias, we based our approach on the assumption that no error presented in the date of diagnosis if diabetes status was correctly classified. Yet there is possibility of delay in diagnosis, in which case the event indicator is not misclassified but the observed time-to-event outcome presents measurement error. This can arise from lack of regular medical visits or glucose monitoring. Different probability of delay in diagnosis between statin users and non-users can cause bias towards either direction. Whereas it is difficult to obtain a precise date by self-report, alternative source of validation information such as medical chart review could also suffer from the same limitation.

Our estimates of HRs may also have been affected by unmeasured confounders such as body mass index, smoking status, and family history of diabetes which were not included in the RAMQ database.

Despite these potential limitations, our results uncovered the effect of potential misclassification of physician-diagnosed diabetes in administrative database on the results of the association between statin use and the occurrence of diabetes. Such misclassification bias in effect estimates cannot be corrected completely by adjusting for baseline risk factors. Our findings highlight the importance of accounting for these misclassifications to prevent erroneous results in studies based on administrative database.

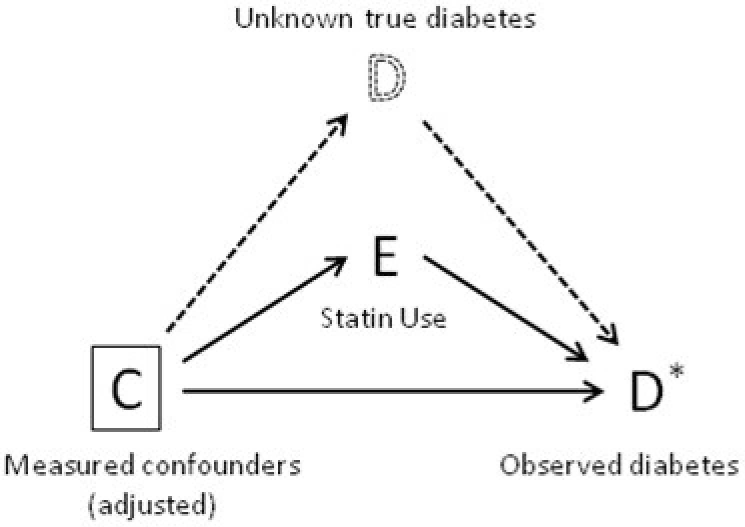


Figure 1: Simplified directed acyclic graph (DAG) illustrating the potential bias introduced by differential misclassification of diabetes status.

Dash arrows and letters represent unobserved paths and variables. Differential misclassification of diabetes status: both statin use (E) and the true diabetes status (D) are associated with the observed diabetes status (D*). In a DAG, the only sources of marginal association between variables are the open paths between them. Observed data: the two open paths from E to D* are E-D* and E-C-D*; adjusting for measured confounders (C) blocks the path E-C-D*, but the path E-D* remains open; association between E and D*; true data: the open path from E to D is E-C-D; this path is blocked by adjusting for C.

Table 1: Selected baseline characteristics of validation study from the Régie de l'assurance maladie du Québec (RAMQ) administrative databases

	Primary study sample	Survey respondents[a]	Participants who provided analyzable blood samples
Participants (n)	6247	3322	1599
Physician diagnosed diabetes before survey	661 (10.6)	391 (11.8)	228 (14.3)
Self-reported diabetes[b]	–	304 (9.2)	180 (11.3)
Age (years)	49.7 (16.4)	51.2 (15.1)	52.4 (14.4)
Male	3041 (48.7)	1555 (46.8)	754 (47.2)
Hospitalization in the past year[c]	624 (10.0)	357 (10.8)	186 (11.6)
Hypertension in the past year	808 (12.9)	457 (13.8)	240 (15.0)
Ischemic heart disease in the past year	267 (4.3)	153 (4.6)	79 (4.9)
Heart failure in the past year	175 (2.8)	89 (2.7)	43 (2.7)
Cancer in the past year	324 (5.2)	179 (5.4)	100 (6.3)
Social deprivation index[e]			
1	1111 (17,8)	630 (19.0)	326 (20.4)
2	1201 (19.2)	681 (20.5)	320 (20.0)
3	1198 (19.2)	671 (20.2)	342 (21.4)
4	1250 (20.0)	627 (18.9)	295 (18.5)
5	1265 (20.3)	606 (18.2)	263 (16.5)
Missing	222 (3.6)	107 (3.2)	53 (3.3)

Data are mean (standard deviation) or n (%).

[a] *Survey respondents are the individuals who participated in the survey and agreed to the linkage of their responses and biochemical data linked to RAMQ information.*

[b] *Self-reported diabetes included previous diagnosis of diabetes or known borderline diabetes.*

[c] *Hospitalization in the past year was that the period of any hospitalization overlapped the past year of the survey.*

[d] *Disease in the past year was having any physician claim or hospitalization with corresponding diagnosis codes of that disease.*

[e] *Social deprivation index was scored from 1 to 5 on the basis of quintiles of six area-based socio-economic indicators.*

Table 2: Two-by-two table for self-reported diabetes against elevated fasting plasma glucose and physician-diagnosed diabetes against self-reported diabetes and/or elevated fasting plasma glucose

		Elevated FPG (≥7 mmol/L)[a]		Elevated FPG (≥6.1 mmol/L)[a]		
		Yes	No	Yes	No	Total
Self-reported diabetes	Yes	78	102	113	67	180 (11.3%)[b]
	No	90 (5.6%)[c]	1329	319 (19.9%)[d]	1100	1419
	Total	168	1431	432	1167	1599
		Self-reported diabetes or undiagnosed diabetes[e]		Self-reported diabetes or IFG/undiagnosed diabetes[e]		
		Yes	No	Yes	No	Total
Physician-diagnosed diabetes	Yes	267	124	294	97	391 (11.8%)[f]
	No	244	2687	691	2240	2931
	Total	511 (15.4%)[g]	2811	985 (29.7%)[h]	2337	3322

FPG, fasting plasma glucose; IFG, impared fasting glucose.

[a] *Two-by-two tables among the 1599 survey respondents who provided analyzable blood samples.*

[b] *Proportion of diabetes by self-report was 180/1599=11.3%.*

[c] *Proportion of undiagnosed diabetes was 90/1599=5.6%.*

[d] *Proportion of IFG/undiagnosed diabetes was 319/1599=19.9%.*

[e] *Two-by-two tables among the 3322 survey respondents from the average of 5 imputation.*

[f] *Prevalence of diabetes by physician diagnosis was 391/3322=11.8%.*

[g] *Proportion of self-reported diabetes and/or FPG≥7 mmol/L was 511/3322=15.4%.*

[h] *Proportion of self-reported diabetes and/or FPG≥6.1 mmol/L was 985/3322=29.7%.*

Table 3: Proportion of positive cases by physician diagnosis and validation study in the primary study sample and validation subgroups

	Participants who provided analyzable blood sample	Survey respondents	Primary study sample
Participants (n)	1599	3322	6247
Physician-diagnosed diabetes	14.3 (12.6-16.1)	11.8 (10.7-12.9)	10.6 (9.8-11.4)
Self-reported diabetes	11.3 (9.8-12.9)	9.2 (8.2-10.2)	–
Elevated FPG (≥7 mmol/L)	10.5 (9.1-12.1)	–	–
Elevated FPG (≥6.1mmol/L)	27.0 (24.9-29.3)	–	–
Self-reported diabetes or undiagnosed diabetes	16.9 (15.1-18.8)	15.4 (14.2-16.7)[a]	14.7 (13.8-15.6)[b]
Self-reported diabetes or IFG/undiagnosed diabetes	31.2 (29.0-33.5)	29.7 (28.1-31.2)[a]	29.0 (27.9-30.1)[b]

Data are % (95% confidence limit).

[a] *Data are average of 5 imputed datasets.*

[b] *Data are average of 25 imputed datasets.*

Table 4: Baseline characteristics by statin use for matched data with physician diagnosis and self-report and/or elevated fasting plasma glucose as classification criteria

	Physician-diagnosed diabetes		Self-reported diabetes or undiagnosed diabetes[a]		Self-reported diabetes or IFG/undiagnosed diabetes[a]	
	Statin users	Statin nonusers	Statin users	Statin nonusers	Statin users	Statin nonusers
Person-years of follow-up	7720		7880		7214	
New-onset diabetes after matched statin initiation date	60	88	84	139	125	254
Time to diabetes diagnosis (months)	97 (78.4)	103 (89.6)	115 (90.4)	107 (83.2)	119 (89.3)	115 (84.1)
Rate of new-onset diabetes (per thousand person years)	28.9	15.6	39.5	24.2	64.6	48.1
Age at matched statin date (years)	63.9 (11.4)	63.9 (11.3)	63.9 (11.1)	63.9 (11.1)	63.7 (11.2)	63.7 (11.2)
Male	195 (48.4)	585 (48.4)	191 (47.1)	572 (47.1)	182 (47.7)	547 (47.7)
Hospitalization in the past year[b]	105 (26.1)	99 (8.2)	109 (26.9)	116 (9.5)	101 (26.7)	109 (9.5)
Use anti-hypertensive drugs in the past year	252 (62.5)	388 (32.1)	260 (64.1)	385 (31.7)	242 (63.4)	352 (30.7)
Ischemic heart disease in the past year[c]	97 (24.1)	39 (3.2)	98 (24.1)	44 (3.6)	92 (24.0)	39 (3.4)
Heart failure in the past year	42 (10.4)	29 (2.4)	42 (10.3)	26 (2.2)	39 (10.3)	25 (2.2)
Cancer in the past year	36 (8.9)	97 (8.0)	33 (8.2)	86 (7.1)	31 (8.1)	83 (7.3)
Social deprivation index[d]						
1	68 (16.9)	197 (16.3)	67 (16.7)	208 (17.1)	62 (16.2)	198 (17.3)
2	70 (17.4)	230 (19.0)	70 (17.5)	233 (19.2)	67 (17.5)	217 (19.0)
3	77 (19.1)	242 (20.0)	79 (19.6)	245 (20.2)	77 (20.1)	229 (20.0)
4	87 (21.6)	249 (20.6)	84 (20.9)	241 (19.8)	80 (20.9)	233 (20.3)
5	83 (20.6)	246 (20.4)	83 (20.6)	248 (20.5)	78 (20.4)	230 (20.1)
Missing	18 (4.5)	45 (3.7)	19 (4.8)	40 (3.3)	19 (4.9)	39 (3.4)

Data are mean (standard deviation) or n (%). [a]Data are the average of the 25 imputations from the first and the second stage; the counts had been rounded to the nearest integer. [b]Hospitalization in the past year was that the period of any hospitalization overlapped the past year of the matched statin initiation date. [c]Disease in the past year was having any physician claim or hospitalization with corresponding diagnosis codes of that disease. [d]Social deprivation index was scored from 1 to 5 on the basis of quintiles of six area-based socio-economic indicators.

Table 5: Hazard ratios (95% confidence limit) for the effect of statin use on diabetes adjusted for baseline characteristics

Classification criteria	HR (95% CI) Statin users vs. nonusers
Physician-diagnosed diabetes	1.61 (1.09-2.38)
Self-reported diabetes or undiagnosed diabetes	1.49 (0.95-2.34)
Self-reported diabetes or IFG/undiagnosed diabetes	1.36 (0.92-2.01)

HR, hazard ratio; CI, confidence interval.

DISCUSSIONS

Our simulation study in chapter 1 showed that the MI approach reduced bias in HR estimates for both non-differential and differential misclassification for a large study sample with more than 20% participation in the self-report validation subgroup, and a fixed 50% participation in the clinical validation subgroup. The ISQ survey described in Chapter 2 had a 53.2% response rate with agreement to health service record linkage and a 48.1% that provided analysable blood samples, creating a sufficient size for the validation subgroups, although the total sample size was smaller than the one used in the simulations (6,247 instead of 10,000).

We applied two different thresholds to define elevated fasting glucose in Chapter 2. These widely-accepted thresholds are aligned with current clinical practice guidelines for diabetes diagnosis (≥7 mmol/L) and impaired fasting glucose (6.1-6.9 mmol/L) by FPG (56-58). The World Health Organization in 1999 (68) proposed a threshold of 6.1 mmol/L on fasting whole blood glucose for diabetes diagnosis. While we obtained fasting whole blood and not plasma glucose measurements in our study, glucose measurements on capillary whole blood can be up to 15% lower than plasma due to the influence of hematocrit, therefore we performed analyses with both thresholds and compared the results.

Appropriate probability sampling techniques are preferred for selection of the validation subgroup. When collection of self-report validations information is intruded by particular attributes, potential selection biases of validation subgroups must be accounted for to avoid worse biased effect estimates in adjusted analyses (Figure 1 and 2 of chapter 1). In chapter 2, we found that older individuals and those with physician-diagnosed diabetes were more inclined to participate in the survey and to return an analyzable blood sample. We corrected for selection bias by performing MI based on the baseline information from administrative data. While survey respondents and participants with analyzable blood sample did not differ greatly from non-respondents on important baseline characteristics (Table 1 of chapter 2), there could be residual confounding from unmeasured confounders that were not available in the administrative databases.

The goal of validation studies is to correct for misclassification of observed disease status which is influenced by flaws in data collection procedures or infrequent use of health services by some patients. However, the primary limitation of validation studies is misclassification by reference standards. Underreporting of diabetes in surveys may arise with an individual's poor understanding of survey questions or lack of knowledge of their disease. Potential sources of measurement error from a single capillary fasting blood glucose sample, including inadequate fasting period and time lapse from home sampling to laboratory testing have been described in chapter 2.

In addition, lack of precise information on diabetes onset date limits survey responses and glucose testing to form a perfect gold standard. As a result, the proposed MI approach does not fully remove the bias due to misclassification in time-to-event outcome. We discussed in chapter 1 that assuming a same disease onset date for undiagnosed cases in both exposure groups may lead to a slightly underestimate of the exposure-disease association. Nevertheless, the correction method always yielded less biased estimates than the naïve analysis. In chapter 2, analysis of the gold-standard outcome (self-reported diabetes and/or elevated fasting plasma glucose) yielded a smaller HR estimate, pointing to an overestimate of the effect of statin treatment on the risk of diabetes based on administrative data alone.

Despite these limitations, this research work demonstrates that MI can be used to correct the effect estimates in Cox proportional hazard models for outcome misclassification using data from validation studies on subgroups of the main study sample. While in practice, conducting a large validation study on the whole study sample is time consuming and costly and hindered by limited participation, this approach is useful in addressing the true association among predictors and time-to-event outcomes and can be applied to a wide range of public health questions.

Reference List

1. Barron BA: The effects of misclassification on the estimation of relative risk. Biometrics 33:414-418, 1977
2. Bollinger, CR and van Hasselt, M. A Bayesian Analysis of Binary Misclassification: Inference in Partially Identified Models. 2009.
3. Carroll R, Ruppert D, Stefanski L, Crainiceanu C: Measurement Error in Nonlinear Models: A Modern Perspective. Boca Raton, FL, Chapman and Hall, 2006
4. Cole SR, Chu H, Greenland S: Multiple-imputation for measurement-error correction. Int J Epidemiol 35:1074-1081, 2006
5. Edwards JK, Cole SR, Troester MA, Richardson DB: Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. Am J Epidemiol 177:904-912, 2013
6. Espeland M, Hui S: A general approach to analyzing epidemiologic data that contain misclassification errors. Biometrics 43:1001-1012, 1987
7. Greenland S, Kleinbaum DG: Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol 12:93-97, 1983
8. Greenland S: Basic methods for sensitivity analysis of biases. Int J Epidemiol 25:1107-1116, 1996
9. Greenland S: Variance estimation for epidemiologic effect estimates under misclassification. Stat Med 7:745-757, 1998
10. Greenland S: Multiple bias modeling for analysis of observational data . J R Statist Soc A 168:267-306, 2005
11. Lyles R: A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics 58:1034-1036, 2002
12. MacLehose R, Olshan A, Herring A, Honein M, Shaw G, Romitti P: Bayesian Methods for Correcting Misclassification An Example from Birth Defects Epidemiology. Epidemiology 20:27-35, 2009

13. Rosenbaum P, Rubin D: The central role of the propensity score in observational studies for causal effects. Biometrika 70:41-55, 1983

14. Spiegelman D, Carroll R, Kipnis C: Efficient and regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med 20:139-160, 2001

15. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ: Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 162:279-289, 2005

16. Newell D: Errors in the interpretation of errors in epidemiology. Am J Public Health 52:1925-1928, 1962

17. Bross I: Misclassiflcation in 2 x 2 tables. Biometrics 10:478-486, 1954

18. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH: Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 105:488-495, 1977

19. Baigent C, Blackwell L, Emberson J, et al: Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomized trials. Lancet 376:1670-1681, 2010

20. Sattar N, Preiss D, et al: Statins and risk of incident diabetes: a collaborative meta-analysis of randomized statin trials. Lancet 375:735-742, 2010

21. Preiss D, et al: Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis. JAMA 305:2556-2564, 2011

22. Culver AL, Ockene IS, Balasubramanian R, Olendzki BC, et al: Statin use and risk of diabetes mellitus in postmenopausal women in the women's health initiative. Arch Intern Med 172:144-152, 2012

23. Danaei G, Rodriguez LAGR, Cantero OFC, Hernan MA: Statins and risk of diabetes - an analysis of electronic medical records to evaluate possible bias due to differential survival. Diabetes Care 36:1236-1240, 2013

24. Rubin D: Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47:1213-1234, 1991

25. Rubin D: Inference and missing data. Biometrika 63:581-592, 1976

26. Rubin D: Multiple Imputation for Nonresponse in Surveys. New York, John Wiley & Sons, 2008

27. Little R, Rubin D: Statistical Analysis with Missing Data. New York, John Wiley & Sons, 2002

28. Rubin, DB. An Overview of Multiple Imputation. 2015.

29. Messer K, Natarajan L: Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. Stat Med 27:6332-6350, 2008

30. Greenland S: The effect of misclassification in the presence of covariates. Am J Epidemiol 112:564-569, 1980

31. Gustafson P: Measurement error and misclassification in statistics and epidemiology. Boca Ration, FL, Chapman and Hall, 2003

32. Fox MP, Lash TL, Greenland S: A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 34:1370-1376, 2005

33. Greenland S, Lash TL: Bias Analysis . In Modern epidemiology. 3rd ed. Philadelphia, Lippincott Williams & Wilkins, 2008, p. 345-380

34. Maldonado G: Adjusting a relative-risk estimate for study imperfections. J Epidemiol Community Health 62:655-663, 2008

35. Luan X, Pan W, Gerberich SG, Carlin BP: Does it always help to adjust for misclassification of a binary outcome in logistic regression? Stat Med 24:2221-2234, 2005

36. Magder L, Hughes J: Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol 146:195-203, 1997

37. Public Health Agency of Canada. Diabetes in Canada: Highlights. 2003.

38. Public Health Agency of Canada. Diabetes in Canada: Facts and figures from a public health perspective. 2011.

39. Leong A, Dasqupta K, Chiasson JL, Rahme E: Estimating the population prevalence of diagnosed and undiagnosed diabetes. Diabetes Care 36:3002-3008, 2013

40. SAS Institut Inc. SAS/STAT (R) 9.2 User's Guide. 2nd. 2014.

41. Firth D: Bias reduction of maximum likelihood estimates. Biometrika 80:27-38, 1993

42. Heinze G, Schemper M: A solution to the problem of separation in logistic regression. Stat Med 21:2409-2419, 2002

43. Heinze G: A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med 25:4216-4226, 2006

44. Last JM: A dictionary of epidemiology. In A dictionary of epidemiology. 5th ed. Last JM, Ed. Oxford University Press, 2013, p. 1

45. Phillips LS, Branch WT, Cook CB, Doyle JP, El-Kebbi IM, Gallina DL, Miller CD, Barnes CS: Clinical inertia. Ann Intern Med 135:825-834, 2001

46. Tamim H, Monfared AA, LeLorier J: Application of lag-time into exposure definitions to control for protopathic bias. Pharmacoepidemiol Drug Saf 16:250-258, 2007

47. Malkasian GD Jr, McDonald TW, Pratt JH: Carcinoma of the endometrium: Mayo clinic experience. Mayo Clin Proc 52:175-180, 1977

48. Horwitz RI, Feinstein AR: The problem of "protopathic bias" in case-control studies. Am J Med 68:255-258, 1980

49. Canadian Diabetes Association Clinical Practice Guidelines Expert Committee: Screening for type 1 and type 2 diabetes. Canadian Journal of Diabetes 37:S12-S15, 2013

50. RAMQ. Données et statistiques. 9-18-2012.

51. Pampalon R, Hamel D, Gamache P, Philibert MD, Raymond G, Simpson A: An Area-based Material and Social Deprivation Index for Public Health in Quebec and Canada. Can J Public Health 103:S17-S22, 2012

52. Griffin S, Little P, Hales C, Kinmonth A, Wareham N: Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes Metab Res Rev 16:164-171, 2000

53. Lindstrom J, Tuomilehto J: The diabetes risk score: a practical score to predict risk of type two diabetes. Diabetes Care 26:725-731, 2003

54. Kanaya A, Fyr C, de Rekeneire N, Shorr R, Schwartz A, Goodpaster B, et al: Predicting the development of diabetes in older adults: the derivation and validation of a prediction rule. Diabetes Care 28:404-408, 2005

55. Leong A, Chiasson J-L, Dasgupta K, Rahme E: Estimating the population prevalence of diagnosed and undiagnosed diabetes. Diabetes Care 36:3002-3008, 2013

56. Canadian Diabetes Association Clinical Practice Guidelines Expert Committee: Definition, classification and diagnosis of diabetes, Prediabetes and Metabolic Syndrome. Canadian Journal of Diabetes 37:S8-S11, 2013

57. Canadian Task Force on Preventive Health C: Recommendations on screening for type 2 diabetes in adults. Canadian Medical Association Journal 34:1687-1696, 2012

58. American Diabetes Association: Diagnosis and classification of diabetes mellitus. Diabetes Care 34:S62-S69, 2011

59. American Diabetes Association: Executive Summary: Standards of Medical Care in Diabetesd-2013. Diabetes Care 36:S4-S10, 2013

60. Hux J, Ivis F, Flintoft V, Bica A: Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care 25:512-516, 2002

61. Responding to the challenge of diabetes in Canada: first report of the National Diabetes Surveillance System (NDSS). Health Canada . 2003. Ottawa.

62. Saydah S, Geiss L, Tierney E, Benjamin S, Engelgau M, Brancati F: Review of the performance of mehtods to identify diabetes cases among

vital statistics, administrative, and survey data. Ann Epidemiol 14:507-516, 2004

63. Leong A, Dasgupta K, Bernatsky S, Lacaille D, Avina-Zubieta A, Rahme E: Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. PLoS One 8:e75256, 2013
64. Canadian Task Force on Preventive Health C: Recommendations on screening for type 2 diabetes in adults. Canadian Medical Association Journal 184:1696, 2012
65. Schonlau M, et al: A comparison between responses from a propensity-weighted web survey and an identical rdd survey. Social Science Computer Review 22:128, 200
66. Martin L, Leff M, Calonge N, Garrett C, Nelson D: Validation of selfreported chronic conditions and health services in a managed care population. Am J Prev Med 18:215-218, 2000
67. Kriegsman D, Penninx B, van Eijk J, Boeke A, Deeg D: Selfreports and general practitioner information on the presence of chronic diseases in community dwelling elderly - A study on the accuracy of patients' self-reports and on determinants of inaccuracy. Journal of Clinical Epidemiology 49:1407-1417, 1996
68. The World Health Organization. Definition, diagnosis, classification of diabetes mellitus and it's complications. 2-27-0013.

SUPPLEMENTAL FILES

Supplemental File 1: Simulation details

We used simulations to assess the performance of multiple imputation (MI) estimators for a Cox proportional hazard model constructed with misclassification in time-to-event outcome, using internal validation designs. We simulated one exposure variables X, measured without error; one variable T for time to true disease onset. Truncating the start of follow up to time zero, we generated a sample of size 10,000, from a Cox proportional hazard model with constant baseline hazard λ_0 as in equation (1):

$$T|X \sim Exponential(\eta),\ \ with\ \ ln(\eta) = ln(\lambda_0) + \beta X \qquad (1)$$

where β is the regression coefficient and e^{β} is the true hazard ratio of exposed versus unexposed in developing the disease. An error term with a Normal distribution with zero mean is added to the time until true disease onset.

We generate one censoring variable C as:

$$C \sim Uniform(0, b),\ \ where\ b\ is\ the\ length\ of\ study\ follow\ up \qquad (2)$$

Generate one indicator W for having true disease before the start of study follow up; for subjects with W = 0, a true disease indicator D_{true} and a time-to-event outcome Y_{true} for the ith individual are generated as in equation (3). Let Y_{true} = 0 and D_{true} = 1 for those who have true disease before the start of study follow up.

$$d_{true}(i) = \begin{cases} 1, & if\ w(i) = 1\ or\ t(i) \le c(i) \\ 0, & if\ t(i) > c(i) \end{cases} \tag{3}$$

$$y_{true}(i) = \begin{cases} t(i), & if\ w(i) = 0\ and\ t(i) \le c(i) \\ c(i), & if\ w(i) = 0\ and\ t(i) > c(i) \\ 0, & if\ w(i) = 1 \end{cases}$$

Generate one participation indicator variable $V1$ for the self-report validation, and another participation indicator $V2$ for the clinical validation subgroup as:

$$V1 \sim Bernoulli(p_{1|X})$$

$$V2 \sim \begin{cases} Bernoulli(p_{2|X}), & if\ V1 = 1 \\ 0, & if\ V1 = 0 \end{cases}$$

where $\mathrm{p}_{1|\mathrm{X}}$ and $\mathrm{p}_{2|\mathrm{X}}$, respectively, are the proportion of participation in the self-report validation and the clinical validation based on different exposure status.

Generate one undiagnosed disease indicator U as:

$$U \sim Bernoulli(p_{3|X})$$

where $\mathrm{p}_{3|\mathrm{X}}$ is the proportion of undiagnosed disease within each exposure group.

The degree of misclassification in the observed disease is measured by sensitivity and specificity, which can differ across exposure groups. Generate one misclassification indicator for self-report validation participants as:

$$M \sim \begin{cases} Bernoulli(1 - Sensitivity|X), & if\ U = 0\ and\ D = 1 \\ Bernoulli(1 - Specificity|X), & if\ U = 0\ and\ D = 0 \\ Missing, & if\ U = 1 \end{cases}$$

We then generated corresponding disease indicators $D_{observed}$ for observed data, D_{v1} for self-report validation, and D_{v2} for clinical validation as:

$$D_{observed} = \begin{cases} 0, & if\ U = 1 \\ |D_{true} - M|, & if\ U = 0 \end{cases}$$

$$D_{v1} = \begin{cases} 0, & if\ V1 = 1\ and\ U = 1 \\ D_{true}, & if\ V1 = 1\ and\ U = 0 \\ Missing, & otherwise \end{cases}$$

$$D_{v2} = \begin{cases} 1, & if\ V2 = 1\ and\ U = 1 \\ D_{true}, & if\ V2 = 1\ and\ U = 0 \\ Missing, & otherwise \end{cases}$$

The time of disease onset of the undiagnosed cases determined by the clinical validation was assumed to be at the time of the validation, i.e. at the end of the study follow up. We generated one variable T^{*} for the surrogate observed time until disease onset based on the following assumptions:

$$T^{*} \sim \begin{cases} Uniform(C, b - C), & if\ X = 0 \\ Beta(\alpha 1, \beta 1), & if\ X = 1\ and\ W = 0 \\ Beta(\alpha 2, \beta 2), & if\ X = 1\ and\ W = 1 \end{cases}$$

Time-to-event outcomes in observed and validation data are then generated as follows:

$$Y_{observed} = \begin{cases} C, & if\ U = 1\ or\ (M = 1\ and\ D_{observed} = 0) \\ T^{*}, & if\ M = 1\ and\ D_{observed} = 1 \\ Y_{true}, & if\ U = 0\ and\ M = 0 \end{cases}$$

$$Y_{v1} = \begin{cases} C, & if\ V1 = 1\ and\ U = 1 \\ Y_{true}, & if\ V1 = 1\ and\ U = 0 \\ Missing, & otherwise \end{cases}$$

$$Y_{v2} = \begin{cases} C, & if\ V2 = 1\ and\ U = 1 \\ Y_{true}, & if\ V2 = 1\ and\ U = 0 \\ Missing, & otherwise \end{cases}$$

The event types and time-to-event outcomes by the definitions of undiagnosed disease, misclassified disease and correctly classified disease are summarized in Supplemental Table 1.

Supplemental Table 1: Definition of undiagnosed disease and misclassification according to event type[a] in observed data and validation data

	True		Observed		Self-report Validation		Clinical Validation	
	D_{true}[a]	Y_{true}[b]	$D_{observed}$[a]	$Y_{observed}$[b]	D_{v1}[a]	Y_{v1}[b]	D_{v2}[a]	Y_{v2}[b]
Undiagnosed	1	T	0	C	0	C	1	C[c]
Misclassified (false negative)	1	T	0	C	1	T	1	NA[d]
Misclassified (false positive)	0	C	1	T^*	0	C	0	C
Correctly classified	1	T	1	T	1	T	1	NA
Correctly classified	0	C	0	C	0	C	0	C

[a] *Event type is 1 = Disease onset and 0 = No disease onset.*

[b] *Time-to-event outcomes are defined as the time from start of follow-up until disease onset, death or censoring.*

[c] *Clinical validation can determine only the event type; the time of disease onset for undiagnosed cases is assumed to be at the time when the validation is conducted.*

[d] *The time of disease onset is not observed in the clinical validation; and no assumption was made for subjects who self-reported true disease onset date.*

In the disease model, we set $\lambda_0 = 1/5000$ and length of follow up b = 3 years, yielding a disease prevalence of ~10% at the baseline level of the exposure variable. For simplicity we also fixed the exposure rate to be 16%; the proportion of the clinical validation participation $p_{2|X}$ to be 50% of the self-report validation group; the proportion of true disease prevalence before the start of study follow up to be 3%; the proportion of undiagnosed disease $p_{3|X=1} = 0.3$, $p_{3|X=0} = 0.4$; specificity = 0.9 for both exposure groups; $\alpha_1 = \alpha_2 = 2$, $\beta_1 = 3$, $\beta_2 = 5$, so that the distribution of observed time to disease onset skews to the start of follow up. We varied the remaining parameters: the proportion of self-report validation participation $p_{1|X}$, the sensitivity of observed disease status in each exposure group.