International Journal of Epidemiology, 2018, 605–616
doi: 10.1093/ije/dyx253
Advance Access Publication Date: 14 December 2017
Original article

Methods

A comparison of methods to correct for misclassification bias from administrative database diagnostic codes

Carl van Walraven

Departments of Medicine and Epidemiology & Community Medicine, University of Ottawa, ASB1–003 1053, Carling Ave, Ottawa ON, K1Y 4E9, Canada. E-mail: carlv@ohri.ca

Editorial decision 6 November 2017; Accepted 6 December 2017

Abstract

Background: In administrative database research, misclassification bias can result from diagnostic codes that imperfectly represent the condition being studied. It is unclear how to correct for this bias.

Methods: Severe renal failure and Colles' fracture status were determined in two distinct cohorts using gold standard methods. True disease prevalence and disease association with other covariables were measured and compared with results when disease status was determined using diagnostic codes. Differences ('misclassification bias') were then adjusted for using two methods: quantitative bias analysis (QBA) with bias parameters (code sensitivity and specificity) of varying accuracy; and disease status imputation using bootstrap methods and disease probability models.

Results: Prevalences of severe renal failure ($n = 50\ 074$) and Colles' fracture ($n = 5680$) were 7.5% and 37.0%, respectively. Compared with true values, important bias resulted when diagnostic codes were used to measure disease prevalence and disease-covariable associations. QBA increased bias when population-based (vs strata-specific) bias parameters were used. QBA's ability to account for misclassification bias was most dependent upon deviations in code specificity. Bootstrap imputation accounted for misclassification bias, but this depended on disease model calibration.

Conclusions: Extensive bias can result from using inaccurate diagnostic codes to determine disease status. This bias can be addressed with QBA using accurate bias parameter measures, or by bootstrap imputation using well-calibrated disease prediction models.

Key words: Misclassification bias, information bias, observation bias, bootstrap, health administrative data

Key Messages

- The use of diagnostic codes to determine disease status results in important misclassification bias when measuring disease prevalence or disease-covariable associations.
- Using quantitative bias analysis (QBA) will not necessarily decrease bias.
- QBA is very dependent upon the accuracy of its bias parameters (in particular the accuracy of code specificity) when addressing bias. Therefore, researchers must pay particular attention to determining values for bias parameters that are used for QBA. In particular, they should strongly consider using values that are actually measured on the population used in the study (or one that is very similar to that in the study).
- The ability of bootstrap imputation to address misclassification bias deteriorated when the scaled Brier Scores for prediction models exceeded 60%.

Introduction

A large majority of administrative database research studies use diagnostic or procedural codes to identify patient cohorts, exposures or outcomes.[1] Commonly, these codes are assigned to patient encounters by health records analysts' review of patient medical records; as such, codes may inaccurately indicate true disease status for many reasons, including unclear physician documentation, misinterpretation of clinical data, or incorrect diagnoses.[2] Since these codes are never perfectly accurate, their use will produce study results that deviate from the truth. These deviations have been termed misclassification bias,[3,4] information bias[5] or observation bias.[5]

The extent of misclassification bias in code-based administrative database research is rarely, if ever, determined. Quantitative bias analysis (QBA) is a collection of calculations that attempts to correct for bias from misclassification and other sources.[6] To address misclassification bias, QBA first measures the sensitivity and specificity of the surrogate marker (which, in administrative database research, is the diagnostic code) for the entity it represents. It uses these bias parameters to calculate the expected number of patients in each cell of the disease-covariable contingency table (Appendix A). This permits the calculation of measures of disease prevalence and disease-covariable association which are corrected for misclassification bias. We have recently shown that misclassification can also be successfully addressed using bootstrap methods to impute disease status using disease probability estimates that are generated from a multivariate model.[7]

The capabilities of QBA and bootstrap imputation methods to correct for misclassification bias have not been directly compared. This study compared the correction for misclassification bias due to the use of administrative database codes using these two methods.

Methods

Study cohorts

This study included two distinct patient cohorts in whom disease status was determined using gold standard criteria.

The severe renal failure cohort was created to determine if severe renal failure status could be accurately determined using covariables from administrative datasets. The cohort included 100 000 randomly selected adults admitted to a multi-institutional, tertiary care teaching hospital between 2002 and 2008.[8] Glomerular filtration rate (GFR) was estimated in each person using the abbreviated Modified Diet in Renal Disease formula using each in-hospital serum creatinine.[9] From definitions from recognized guidelines on chronic kidney disease,[10,11] patients having two or more consecutive GFRs less than 30 mL/min/1.73 m^2 were classified as having severe kidney disease. Patients with only one serum creatinine measured during their admission were classified with severe kidney disease if its GFR was less than 30 mL/min/1.73 m^2. All other patients, including those with no creatinine measures, were classified as having no severe kidney disease. Patients were randomly divided into a measurement ($n = 49\ 926$) and an analysis ($n = 50\ 074$) group.

The Colles' fracture cohort was created to study treatment and outcomes of patients with Colles' fracture. The cohort included all patients assessed in the emergency department (at the same hospital as the severe renal failure cohort) between 1 January 2006 and 31 December 2014, who underwent plain radiography of the forearm ($n = 11\ 233$). Patients in 2010 were used to derive the Colles' Fracture Model (described below) and were excluded from the current analysis. Text reports of all radiographs were manually reviewed to determine the presence or absence of a Colles' fracture (acute fracture of the distal radius or ulna, with fractures of the proximal radius or ulna, the carpal bones or the metacarpal bones being excluded). Patients were randomly divided into a measurement ($n = 5553$) and an analysis ($n = 5680$) group.

Administrative database codes

Patients were deemed to have been coded with severe renal failure or Colles' fracture if they were assigned any of the International Classification of Diseases 10 (ICD10) codes listed in Appendix B during their hospital encounter.

Models for disease status

Bootstrap implementation requires a model which generates a probability that an individual patient is truly diseased. For the severe renal failure cohort, this was accomplished using the Severe Renal Failure Model. This model was derived (in the measurement group of the severe renal failure cohort) and internally validated (in the analysis group) in a previous study.[8] All predictor variables considered for the model came from the hospital's discharge abstract database and included patient factors (age, sex and all Elixhauser comorbidities using ICD codes cited by Quan[12]), hospitalization factors (admission urgency, admitting service, intensive care unit treatment, surgical procedures, hospital survival status and length of stay) and renal failure-specific codes (dialysis-related diagnoses and procedures, the most common acute diagnoses causing renal dysfunction, and manifestations of renal dysfunction). A macro from Sauerbrai was used to create the models using fractional polynomial methods for continuous variables and forward variable selection.[13] This model was used in the validation group to estimate each patient's probability of having severe renal failure.

For the Colles' fracture cohort, disease probability was determined using the Colles' Fracture Model. This model was derived using text classification methods to identify a Colles' fracture in 493 radiology reports.[14] In an internal validation population of 258 reports, the model had a sensitivity of 95.5% and a specificity of 92.9%. This algorithm was applied to all text reports of forearm radiographs in the Colles' fracture validation cohort to return the probability of a Colles' fracture.

Covariables

A total of 43 covariables for the renal failure cohort were abstracted from the Discharge Abstract Database and were listed in the 'Models for disease status' section. Covariables for the Colles' fracture cohort were abstracted from the National Ambulatory Care Reporting System (which records all emergency room visits). These nine variables included patient age and sex, Charlson comorbidity score (based on coded comorbidities), diabetes status, year of presentation, presentation during winter months, triage location, presence of Colles' fracture-related procedural codes [Canadian Classification of Intervention codes of 1.UB.73* (Reduction, wrist joint using closed approach) or 1.TV.73 (Reduction, radius and ulna using closed approach)] and patient disposition.

Analysis

The primary outcome for the study was the amount of bias in the measurement of: (i) disease prevalence; and (ii) the association of the disease with covariables when they were calculated using true disease status and diagnostic codes (Appendix B). True disease prevalence [with exact 95% confidence intervals (CIs)] was measured in patient strata defined by the presence or absence of covariates in each cohort (with continuous variables dichotomized by their median values). Logistic regression was used to determine the true association of disease status with each covariable (measured using odds ratios with 95% CI), again with continuous variables dichotomized by their median values. All analyses were limited to the analysis group ($n = 50\ 074$ in the severe renal failure cohort, $n = 5680$ in the Colles' fracture cohort).

Measures of disease prevalence and disease-covariable associations were repeated after disease status was determined using diagnostic codes (Appendix B). To quantify bias in prevalence estimates, these results ('surrogate values') were subtracted from true values. Differences between true and surrogate values were expressed in both absolute (true-surrogate) and relative (|true-surrogate|/true) values. As a qualitative measure of the extent of the difference between the true and surrogate values, the proportion of the surrogate prevalence estimates within the 95% confidence intervals of the true prevalence estimates was also calculated. To measure disease-covariable associations, logistic regression models were created using the surrogate disease status. Absolute differences between parameter estimates from these models and those using true disease status were calculated and then exponentiated to facilitate interpretation. Parameter estimates that are identical have absolute differences of 0 and exponentiated values of 1; in contrast, absolute differences less than 0 have exponentiated values less than 1, and differences exceeding 0 have exponentiated values exceeding 1. The proportion of surrogate odds ratios within the 95% confidence intervals of true odds ratios was also calculated. Finally, overall bias for both prevalence estimates and disease-covariable associations was summarized using the mean squared error. This was calculated as the sum of squared difference of the true value and the surrogate estimate divided by the total number of groups:

$$\frac{\sum (true\,valve - surrogate\,estimate)^2}{\#\,groups}$$

The direction of the bias was not reported, since this can be influenced by a large number of factors.[15]

These statistics were repeated after correction for misclassification bias using quantitative bias analysis (QBA) or bootstrap imputation. The QBA methods for addressing misclassification bias described by Lash, Fox and Fink[6] were used (Appendix A). This method requires the sensitivity and specificity of the code for the disease (Appendix A, section 1). These 'bias parameters' permit the calculation of cell counts in the covariable-code for disease contingency tables (Appendix A, section 2) that are corrected for misclassification using the equations that are specified in Appendix A, section 3.

The initial QBA iteration used bias parameters from patients in the measurement group of each cohort (the 'overall bias parameters' analysis). We also conducted an analysis using 'strata-specific bias parameters' that used bias parameters (again determined in the measurement group) specific to each patient stratum. Finally, analyses were repeated using bias parameters that were measured in the analysis group to illustrate results with completely accurate sensitivity and specificity values ('perfect bias parameters' analysis).

For the bootstrap imputation, 1000 bootstrap samples with replacement were created from the validation patients in each patient stratum.[8] Each bootstrap sample had the same sample size as the original cohort. For each patient within each bootstrap sample, a uniformly distributed number between 0 and 1 was randomly selected; disease status was then imputed as present if the randomly selected number was below the estimated probability of disease (from the Severe Renal Failure Model for the renal failure cohort or the Colles' Fracture Model for the Colles' fracture cohort) for that particular patient. Disease prevalence and disease-covariable associations were then measured on each bootstrap sample. The final point estimate for disease prevalence or disease-covariable odds ratio was the median value of all 1000 bootstrap samples.

Finally, the factors influencing the capability of QBA and bootstrap imputation to address bias for disease prevalence estimation were explored. Bias was quantified as the mean squared error (calculated using the equation above) within each patient stratum. For QBA, this was plotted against the relative difference in code sensitivity and specificity (calculated in each stratum as: |bias parameter$_{measurement}$ – bias parameter$_{analytical}$|/ bias parameter$_{analytical}$). For bootstrap imputation, bias was plotted against the scaled Brier score measuring calibration of the models for disease status. The scaled Brier score measures agreement between predicted and actual binary outcomes on a scale ranging from 0% (perfect agreement between observed and predicted) and 100% (perfect disagreement).[16]

Results

The severe renal failure cohort analysis included 100 000 patients (measurement cohort $n = 49\ 926$, analysis cohort $n = 50\ 074$). The Colles' fracture cohort included 11 233 patients (measurement cohort $n = 5553$, analysis cohort $n = 5680$). Table 1 describes the analysis patients of both cohorts by disease status. In the severe renal failure cohort, 3748 patients (7.5%) had the disease. Renal failure was notably more common as patients aged and in patients with dialysis-related diagnoses or procedures, those who were treated in the intensive care unit or died in the hospital and those assigned codes for causes for, or manifestations of, renal failure. In the Colles' fracture cohort, 2102 (37.0%) had the disease, with fractures being more common in females and those coded with a Colles'-related procedure. Patients in the measurement and analysis populations of both study cohorts were essentially equivalent (Appendix C).

Accuracy of ICD codes

In the measurement groups of the severe renal failure cohort ($n = 49\ 926$) and Colles' fracture cohort ($n = 5553$), the ICD codes used to determine disease status (Appendix B) had overall sensitivities of 71% and 72%, respectively, and specificities of 96% and 98%, respectively. These values are the 'overall bias parameters' for each disease. However, these values varied extensively when measured in distinct patient strata. In the renal failure cohort, ranges of code sensitivities and specificities in the 86 patient strata were 0.3–0.93 and 0.72–0.99, respectively; in the Colles' fracture cohort, ranges in code sensitivities and specificities in the 18 distinct patient strata were 0.62–0.90 and 0.78–0.99, respectively.

There were large differences in bias parameter values when measured in the entire cohort vs specific patient strata. In the analysis population, relative differences in code sensitivity when measured in the entire population vs specific strata ranged from 0.1% to 88.7% (median 4.4%) in the renal failure cohort and 0.03% to 20.4% (median 2.4%) in the Colles' fracture cohort. Corresponding values for code specificity were smaller: 0.1–37.4% (median 1.7%) for the renal failure cohort and 0.005–11.9% (median 0.2%) for the Colles' fracture cohort. There were also notable

Table 1. Description of study cohorts

Severe renal failure cohort	Value	No severe renal failure ($n = 46326$)	Severe renal failure ($n = 3748$, 7.5%)	Overall ($n = 50074$)
Mean age (SD)		53.6 ± 19.8	68.8 ± 15.3	54.8 ± 19.9
Female		27066 (58.4%)	1768 (47.2%)	28834 (57.6%)
Coded with possible causes of renal failure		878 (1.9%)	312 (8.3%)	1190 (2.4%)
Coded with manifestations of renal failure		634 (1.4%)	509 (13.6%)	1143 (2.3%)
Dialysis-related diagnoses		94 (0.2%)	258 (6.9%)	352 (0.7%)
Patient admitted emergently		27010 (58.3%)	3255 (86.8%)	30265 (60.4%)
Patient admitted from emergency		17238 (37.2%)	2199 (58.7%)	19437 (38.8%)
Patient came in by ambulance		10513 (22.7%)	1687 (45.0%)	12200 (24.4%)
Patient admitted to surgical service		11373 (24.5%)	632 (16.9%)	12005 (24.0%)
Patient had operation during admission		6921 (14.9%)	852 (22.7%)	7773 (15.5%)
Patient in ICU during admission		1663 (3.6%)	634 (16.9%)	2297 (4.6%)
Patient had dialysis-related procedure		1553 (3.4%)	1372 (36.6%)	2925 (5.8%)
Patient died in hospital		1382 (3.0%)	718 (19.2%)	2100 (4.2%)
Median hospital length of stay (IQR)		3.0 (2.0–7.0)	9.0 (4.0–19.0)	3.0 (2.0–7.0)
Colles' fracture cohort		No fracture ($n = 3578$)	Fracture ($n = 2102$, 37.0%)	Overall ($n = 5680$)
Mean age (SD)		67.3 ± 13.1	67.5 ± 12.3	67.4 ± 12.8
Patient > 65		1799 (50.3%)	1067 (50.8%)	2866 (50.5%)
Patient is male		1552 (43.4%)	443 (21.1%)	1995 (35.1%)
Visit year	<2010	1562 (43.7%)	987 (47.0%)	2549 (44.9%)
	≥2010	2016 (56.3%)	1115 (53.0%)	3131 (55.1%)
CTAS score	1–2	579 (16.2%)	336 (16.0%)	915 (16.1%)
	3+	2999 (83.8%)	1766 (84.0%)	4765 (83.9%)
Colles' fracture-related procedure coded		8 (0.2%)	744 (35.4%)	752 (13.2%)
Discharged home		3068 (85.7%)	1735 (82.5%)	4803 (84.6%)
Charlson score	0	3227 (90.2%)	1994 (94.9%)	5221 (91.9%)
	> 0	351 (9.8%)	108 (5.1%)	459 (8.1%)

SD, standard deviation; ICU, intensive care unit; CTAS, Canadian Triage and Acuity Scale.

differences in strata-specific bias parameter values in the measurement vs the analysis population: for code sensitivity, relative differences ranged 0.06–26.9% (median 1.4%) in the renal failure cohort and 0.003–8.3% (median 1.9%) in the Colles' fracture cohort. Corresponding values for specificities were smaller (0.002–4.8%, median 1.0% for the renal failure cohort; 0.02–11.1%, median 0.2% for the Colles' fracture cohort).

Accuracy of disease prediction models

In the analysis population, the Severe Renal Failure Model was highly discriminative (c-statistic 0.937) and very well calibrated (Figure 1; scaled Brier score 0.43). The Colles' Fracture Model had a greater discrimination (c-statistic 0.981) but its calibration suffered from a systematic overestimation of risk (Figure 1; scaled Brier score 0.77). Model calibration varied between patient strata most notably in the renal failure cohort; scaled Brier scores ranged from 0.18 to 0.59 (median 0.43) between strata in the severe renal failure cohort and 0.71 to 0.81 (median 0.77) in the Colles' fracture cohort.

Bias measuring disease prevalence

The median prevalence of severe renal failure in the 86 patient strata was 7.5%, range 2.5–73.3% (Table 2A, column A). Measuring severe renal failure prevalence using the ICD codes overestimated disease prevalence with a median absolute difference with true values of -1.2% [interquartile range (IQR) -3.4% to -1.0%, Table 2, column B] and a median relative difference of 16.6% (IQR 15.5%-25.3%). Only 13.8% of prevalence estimates based on codes were within the 95% confidence intervals of the true measure.

Misclassification bias actually increased when QBA with overall bias parameters was used, with increases in relative differences of renal failure prevalence and mean squared error (Table 2A, column C). QBA using strata-specific bias parameters notably decreased bias, with the

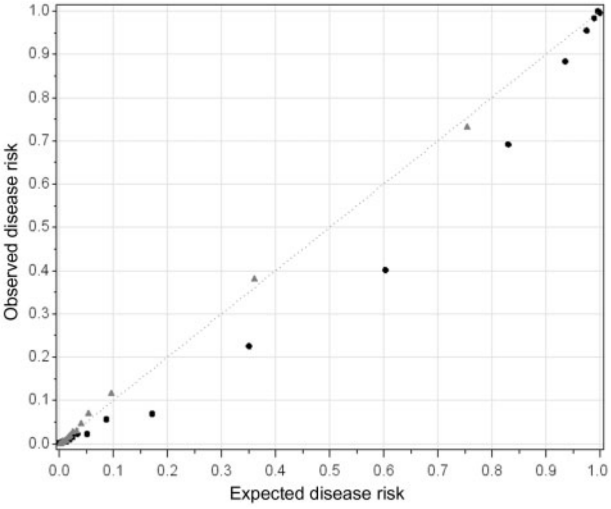

Figure 1. Calibration of the Severe Renal Failure Model and Colles' Fracture Model.

This graphic plots the observed disease risk (vertical axis) against the expected disease risk (horizontal axis). Severe renal failure (triangles) was defined as a glomerular filtration rate less than 30 ml/min/1.73 m^2 with the expected disease risk determined with the Severe Renal Failure Model.[8] Colles' fracture (circles) was defined as an acute fracture of the distal radius or ulna on plain radiography with the expected disease risk determined with the Colles' Fracture Model.[14] The scaled Brier scores of the Severe Renal Failure Model and Colle's Fracture Model were 0.43 and 0.78, respectively.

distribution of renal failure prevalence being almost identical to true values (Table 2A, column D, row 1), and a notable drop in relative differences (Table 2A, column D, row 3). Misclassification bias was essentially eliminated with QBA when perfectly accurate bias parameters were used (Table 2A, column E). Bootstrap imputation methods resulted in the next smallest amount of misclassification bias, with 94.2% of estimates falling with the 95% confidence intervals of true values and a mean squared error of 6.9 (Table 2A, column F).

Patterns were different in the Colles' fracture cohort (Table 2B). Disease prevalence was higher (median 36.9%, range 22.2–98.9%, Table 2B, column A) but prevalence using the ICD codes underestimated disease prevalence [median relative difference 9.5% (IQR 8.9%–10.1%)]. No prevalence estimates using codes were within the 95% confidence intervals of the true measures. Misclassification bias decreased notably with QBA using overall bias parameters (Table 2B, column C) and decreased further still with QBA using strata-specific bias parameters (Table 2B, column D). Again, QBA using perfectly accurate bias parameters eliminated misclassification bias (Table 2B, column E). In contrast to the severe renal failure cohort, misclassification bias using the bootstrap imputation method was notably higher than QBA methods (Table 2B, column F).

Bias measuring disease-covariable associations

True associations between severe renal failure and the 43 covariables ranged from 0.47 and 36.3 (median 2.71, Table 3A, column A). Using ICD codes to determine renal failure status, disease-covariable associations were biased away from the null (median odds ratio 3.01; Table 3A, column B). However, estimates of association were exaggerated even further away from the null using QBA with overall bias parameters (Table 3A, column C, row 1) with an increase in the mean squared error (Table 3A, row 6, columns C vs B). Misclassification bias using QBA with strata-specific bias parameters was similar to that using ICD codes alone (Table 3A, column D). QBA with perfectly accurate bias parameters eliminated bias (Table 3A, column E). Bias was also very small when bootstrap imputation was used (Table 3A, column F).

Misclassification bias measuring disease-covariable associations with Colles' fracture (Table 3B) was eliminated using QBA with perfectly accurate bias parameters (Table 3B, column E). However, misclassification bias did not vary extensively between the other surrogate methods used.

Factors influencing capability of QBA and bootstrap imputation to address bias

Since QBA using perfect bias parameters resulted in essentially no misclassification bias (Table 2 and Table 3, column E), differences in bias parameters (code sensitivity and specificity) in the measurement group relative to values in the analysis groups were calculated and plotted against the mean squared error (Figure 2). Compared with relative differences in code sensitivity, relative differences in code specificity appeared to be strongly associated with mean squared error. In the bootstrap imputation analysis, mean squared error appeared to increase as values for the scaled Brier score increased.

Discussion

Misclassification bias affects any administrative database research study in which codes are used to identify patient cohorts, exposures or outcomes. This study found that the use of diagnostic codes resulted in extensive and clinically important bias when estimating disease prevalence or its association with covariables. Quantitative bias analysis (QBA) methods can successfully remove misclassification bias, but its success depended on bias parameter accuracy. Imputing patient disease status using bootstrap methods can also address misclassification bias, but requires a well-calibrated disease prediction model.

This study made several important findings. First, the use of database codes to determine disease status is

Table 2. Misclassification bias in disease prevalence

		Surrogate disease status				
			Method used to account for misclassification bias			
	A. True disease status	B. Disease status from ICD codes	C. QBA (overall bias parameters[a])	D. QBA (strata bias parameters[a])	E. QBA (perfect bias parameters[b])	F. Bootstrap imputation
Severe renal failure cohort (86 patient strata)						
Disease prevalence						
Mean (SD)	0.124 (0.108)	0.147 (0.119)	0.159 (0.152)	0.121 (0.108)	0.124 (0.108)	0.123 (0.108)
Median (IQR)	0.075 (0.071–0.142)	0.087 (0.082–0.178)	0.073 (0.065–0.229)	0.073 (0.069–0.143)	0.075 (0.071–0.142)	0.073 (0.069–0.153)
Range	0.025–0.733	0.028–0.739	0.013–0.694	0.018–0.711	0.025–0.733	0.023–0.728
Absolute difference prevalence (true-surrogate)						
Mean (SD)	–	−0.023 (0.025)	−0.038 (0.07)	0.003 (0.018)	0 (0)	0.001 (0.009)
Median (IQR)	–	−0.012 (−0.034–0.01)	0.002 (−0.09–0.006)	0.002 (0.001–0.004)	0 (0–0)	0.002 (0.001–0.003)
Range	–	−0.118–0.045	−0.249–0.063	−0.086–0.058	0–0	−0.034–0.027
Relative difference prevalence [absolute value(true-surrogate)/true]						
Mean (SD)	–	0.205 (0.114)	0.304 (0.258)	0.073 (0.111)	0 (0)	0.042 (0.044)
Median (IQR)	–	0.166 (0.155–0.253)	0.272 (0.047–0.487)	0.030 (0.017–0.082)	0 (0–0)	0.026 (0.018–0.044)
Range	–	0.008–0.745	0.004–0.936	0.004–0.616	0–0	0–0.213
% Estimates in 95% CI true Value		13.8 (7.4, 23.1)	14.1 (7.3, 23.8)	76.7 (66.4, 85.2)	100 (95.8, 100)	94.2 (87.0, 98.1)
Mean squared error		1.11	6.09	0.35	0	0.08
Colles' fracture cohort (18 patient strata)						
Disease prevalence						
Mean (SD)	0.388 (0.164)	0.293 (0.159)	0.348 (0.074)	0.381 (0.165)	0.388 (0.164)	0.425 (0.148)
Median (IQR)	0.369 (0.336–0.387)	0.274 (0.242–0.283)	0.363 (0.328–0.374)	0.365 (0.339–0.375)	0.369 (0.336–0.387)	0.409 (0.379–0.431)
Range	0.222–0.989	0.159–0.895	0.199–0.486	0.214–0.996	0.222–0.989	0.276–0.965
Absolute difference prevalence (true-surrogate)						
Mean (SD)	–	0.095 (0.022)	0.01 (0.021)	0.008 (0.013)	0 (0)	−0.037 (0.018)
Median (IQR)	–	0.095 (0.089–0.101)	0.006 (0.002–0.015)	0.006 (−0.001–0.009)	0 (0–0)	−0.039 (−0.044–0.034)
Range	–	0.059–0.156	−0.026–0.072	−0.01–0.043	0–0	−0.064–0.024
Relative difference prevalence [absolute value(true-surrogate)/true]						
Mean (SD)	–	0.26 (0.056)	0.042 (0.045)	0.028 (0.027)	0 (0)	0.117 (0.056)
Median (IQR)	–	0.257 (0.25–0.279)	0.024 (0.013–0.055)	0.019 (0.007–0.037)	0 (0–0)	0.109 (0.096–0.127)
Range	–	0.095–0.373	0.004–0.172	0–0.104	0–0	0.025–0.287
% Estimates in 95% CI true value		0 (0, 18.5)	75.0 (47.6, 92.7)	88.9 (65.3, 98.6)	100 (82.5, 100)	11.1 (1.4, 34.7)
Mean squared error		9.48	0.51	0.22	0	1.63

The true prevalence of disease (severe renal failure or Colles' fracture) was determined in 86 and 18 strata, respectively (column A). Disease prevalence was also measured using five surrogate methods (columns B-F). Bias in prevalence measures with each surrogate method was quantified using four statistics: the absolute difference between the true value and the surrogate; the relative difference between the true value and the surrogate; the percentage of surrogate strata-specific prevalence estimates that were within the 95% confidence intervals of the true value; and the mean squared error [Σ(surrogate-true value)^2/number of strata].

[a] From measurement patient group.

[b] From analysis patient group.

Table 3. Misclassification bias in measures of association and its correction by classical quantitative bias analysis (QBA) or the bootstrap method

		Surrogate disease status				
			Method used to account for misclassification bias			
	A. True disease status	B. Disease status from ICD codes	C. QBA (overall bias parameters[a])	D. QBA (strata bias parameters[a])	E. QBA (perfect bias parameters[b])	F. Bootstrap imputation
Severe renal failure cohort (43 covariables)						
Association with covariables (odds ratio)						
Median (IQR)	2.71 (1.6–4.26)	3.01 (1.57–4.73)	4.85 (1.87–8.69)	2.45 (1.37–3.95)	2.71 (1.6–4.26)	2.67 (1.46–4.74)
Range	0.47–36.31	0.5–31.41	0.16–36.64	0.23–33.27	0.47–36.36	0.38–36.71
Exponentiated difference in parameter estimates (true value-surrogate method)						
Median (IQR)		0.92 (0.8–1.03)	0.53 (0.44–0.77)	1.03 (0.9–1.11)	1 (1–1)	0.99 (0.95–1.05)
Range		0.63–2.04	0.24–2.93	0.63–2.73	1 – 1	0.77–1.24
Exponentiated difference in parameter estimates (absolute value)						
Median (IQR)		1.16 (1.08–1.26)	2.03 (1.62–2.27)	1.11 (1.05–1.2)	1 (1–1)	1.05 (1.03–1.13)
Range		1.01–2.04	1.02–4.11	1.01–2.73	1–1	1–1.29
% Estimates invalid results (95% CI)		0 (0, 8.2)	9.3 (2.6, 22.1)	2.3 (0.1, 12.3)	0 (0, 8.2)	0 (0, 8.2)
% Estimates in 95% CI of true		53.5 (37.7, 68.8)	7.7 (1.6, 20.9)	59.5 (43.3, 74.4)	100 (91.8, 100)	95.4 (84.2, 99.4)
Mean squared error		0.054	0.600	0.073	0	0.010
Colles' fracture cohort (9 covariables)						
Association with covariables (odds ratio)						
Median (IQR)	0.88 (0.58–1.02)	0.92 (0.64–1.08)	0.89 (0.53–1.07)	0.88 (0.61–1.09)	0.88 (0.58–1.02)	0.84 (0.6–1.01)
Range	0.35–244.47	0.37–38.92	0.3–2.02	0.34–650.23	0.35–244.48	0.44–57.67
Exponentiated difference in parameter estimates (true value-surrogate method)						
Median (IQR)		0.95 (0.92–0.99)	0.99 (0.9–1.09)	0.96 (0.92–1.03)	1 (1–1)	1.04 (0.97–1.05)
Range		0.73–6.28	0.72–1.16	0.38–1.12	1–1	0.79–4.24
Exponentiated difference in parameter estimates (absolute value)						
Median (IQR)		1.09 (1.05–1.1)	1.09 (1.02–1.18)	1.08 (1.04–1.12)	1 (1–1)	1.05 (1.04–1.08)
Range		1.01–6.28	1.01–1.39	1.01–2.66	1–1	1.02–4.24
% Estimates invalid results (95% CI)		0 (0, 33.6)	11.1 (0.3, 48.2)	0 (0, 33.6)	0 (0, 33.6)	0 (0, 33.6)
% estimates in 95% CI of true		77.8 (40, 97.2)	62.5 (24.5, 91.5)	77.8 (40.0, 97.2)	100 (66.4, 100)	77.8 (40, 97.2)
Mean squared error		0.390	0.023	0.114	0	0.240

The association of covariates with severe renal failure or Colles' fracture was measured using gold standard methods to determine disease status (column A). Associations were also measured after disease status was determined with International Classification of Diseases (ICD) codes (column B). Four methods were then used to attempt to correct misclassification bias: classical quantitative bias analysis (QBA) with overall bias parameters (column C): QBA with strata-specific bias parameters (column D); QBA with perfectly accurate bias parameters (column E); and bootstrap imputation methods imputing disease status with model-based probability estimates (column E). The extent of misclassification bias was examined using six measures: comparison of distribution of odds ratios with the gold standard; exponentiated differences in logistic regression model parameter estimates associating disease and covariable; exponentiated absolute differences in logistic regression model parameter estimates; percentage of estimates with invalid results; percentage of estimates within 95% confidence intervals of true value; and mean squared error.

[a] From measurement patient population.

[b] From analysis patient population.

associated with an important amount of misclassification bias in both disease prevalence estimates and disease-covariable associations. Administrative database researchers must be aware that using codes can result in meaningfully biased results. More importantly, the 'direction' of this bias (i.e. disease prevalence estimates increasing or decreasing and associations moving towards or away from null values) was unpredictable. Second, the use of classical quantitative bias analysis (QBA) can (but does not always) decrease misclassification bias. The success of QBA to adjust for misclassification bias depends upon the accuracy of the bias parameters used; in particular, the limited analysis presented here indicated that the extent to which misclassification bias is decreased by QBA appears to be especially dependent upon the accuracy of specificity values. Administrative database researchers who use classical QBA to address misclassification bias must ensure that the bias parameters used are both measured accurately using valid methods and are generalizable to the population being studied.[17] Ideally, bias parameters would be measured in a large sample of the study population to ensure applicability. Given the sensitivity of QBA results to bias parameter values, the use of probabilistic quantitative bias analysis—in which a distribution of potentially valid sensitivity and specificity values are used[18]—should be considered to illustrate the extent of uncertainty around estimates. Finally, this study showed that bootstrap imputation methods were capable of generating results with very little misclassification bias, but only when the model used to predict disease probability was well calibrated. Because of the time required for resampling, this method of accounting for misclassification is computationally more intensive than classical QBA. However, an important potential advantage of the bootstrap imputation over QBA is its ability to be done using multivariate models. As such, it can account for misclassification bias when measuring association adjusted for covariables. Therefore, developing accurate models to determine disease probability is an important step to optimizing administrative database research.

Some issues regarding the study should be kept in mind. First, the study addressed two conditions only, from a single centre. It is important that this study's methods be replicated in other conditions at other centres, to determine if the results seen here are replicated. Second, further research is required to determine the conditions under which a researcher can be confident that misclassification bias has

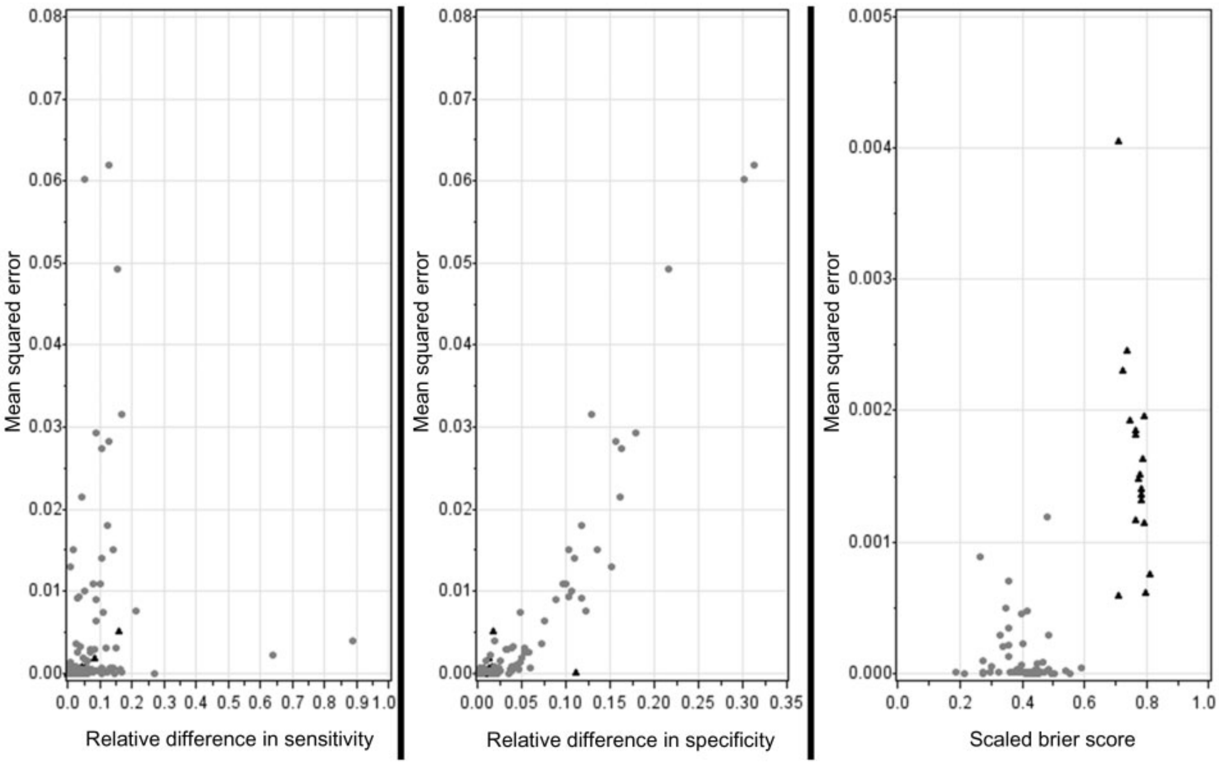

Figure 2. Influence of bias parameter accuracy and model calibration on misclassification bias in the estimation of disease prevalence. Misclassification bias was quantified as the mean squared difference between estimated and actual disease prevalence. This was plotted against the: relative difference of sensitivity values used in the quantitative bias analysis [QBA] and those in the analysis cohort (left plot); relative difference of specificity values used in the QBA and those in the analysis cohort (middle plot); and the scaled Brier score, measuring calibration of disease prediction models used in bootstrap imputation (right plot). (Grey circles = severe renal failure strata; black triangles = Colles' fracture strata).

been addressed using QBA or bootstrap imputation. This study found that the relative difference in code specificity appeared to be linearly related to misclassification bias (Figure 2b). It also found that bias appeared to be related to model calibration in bootstrap imputation (Figure 2c). A more exhaustive examination of this area is necessary, so that researchers will know the conditions under which they can confidently use these methods to address misclassification bias. Third, this study focused on misclassification bias resulting from the use of administrative database codes to identify patient cohorts. However, the results are equally applicable to other situations in which disease status is imputed using inaccurate methods. Finally, using probabilistic bias analysis—in which errors around the estimates for sensitivity and specificity are recognized using Monte Carlo resampling techniques to generate a distribution of corrected estimates—would not have changed the results of the study, since it would use the same bias parameters as those that were used in the current analysis (i.e. bias parameters from the measurement section of both disease cohorts). Such a PBA—assuming a large enough resampling size—would essentially return the same point estimates as those presented here.

In summary, this study explored misclassification bias resulting from the use of diagnostic codes to identify disease status and how it might be addressed. Future studies are needed to determine if these results are replicated and to clarify the conditions in which these methods will reliably produce results that are less biased than those generated using administrative database codes alone.

Funding

C.vW. is supported by a University of Ottawa Department of Medicine Clinician Scientist Chair.

Conflict of interest: None declared.

References

1. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently uses validated diagnostic or procedural codes. *J Clin Epidemiol* 2011;**64:**1054–59.
2. Nicholls SG, Langan SaM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. *CMAJ* 2017;**189:**E1054–55.
3. Colman S, Joyce T, Kaestner R. Misclassification bias and the estimated effect of parental involvement laws on adolescents' reproductive outcomes. *Am J Public Health* 2008;**98:**1881–85.
4. Mohler B, Earls F. Trends in adolescent suicide: misclassification bias? *Am J Public Health* 2001;**91:**150–53.
5. Last JM. *A Dictionary of Epidemiology*, third edn. New York, NY: Oxford University Press, 1995.
6. Lash TL, Fox MP, Fink AK. Misclassification. *Applying Quantitative Bias Analysis to Epidemiologic Data*. London: Springer, 2010.
7. van Walraven C. Bootstrap imputation with a disease probability model minimizes bias from misclassification due to administrative database codes. *J Clin Epidemiol* 2017;**84:** 114–20.
8. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A, Forster AJ. The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. *J Clin Epidemiol* 2010;**63**:1332–41.
9. Manjunath G, Sarnak MJ, Levey AS. Prediction equations to estimate glomerular filtration rate: an update. *Curr Opin Nephrol Hypertens* 2001;**10:**785–92.
10. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification.[see comment]. *Am J Kidney Dis* 2002;**39**(Suppl 1)**:** S1–266.
11. Levey AS, Eckardt KU, Tsukamoto Y *et al*. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int* 2005;**67:**2089–100.
12. Quan H, Sundararajan V, Halfon P *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;**43:**1130–39.
13. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Comput Stat Data Anal* 2006;**50:**3464–85.
14. de Bruijn B, Cranney A, O'Donnell S, Martin JD, Forster AJ. Identifying wrist fracture patients with high accuracy by automatic categorization of X-ray reports. *J Am Med Inform Assoc* 2006;**13:**696–98.
15. Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol* 2005;**34:**680–87.
16. Steyerberg EW. Evaluation of Performance. *Clinical Prediction Models*. New York, NY: Springer, 2010.
17. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;**64:**821–29.
18. Lash TL, Fox MP, Fink AK. Probabilistic Bias Analysis. *Applying Quantitative Bias Analysis to Epidemiologic Data*. London: Springer, 2010.

Appendix A Quantitative bias analysis using sensitivity and specificity of surrogate indicator for true disease status.

1. Code-disease contingency table

		Disease	
		+	−
Code	+	a	b
	−	c	d

Sensitivity (of disease code for true disease status) = a/(a + c).
Specificity (of disease code for true disease status) = d/(b + d).

2. Covariable-disease code contingency table

		Disease code	
		+	−
Covariable	+	A	B
	−	C	D

3. Cell values in covariable-disease contingency table corrected for misclassification bias (given observed values in covariable-disease code contingency table in Section 2)

a (corrected)	[A-(A + C)*(1-specificity))/(sensitivity-(1-specificity)]
b (corrected)	[B-(B + D)*(1-specificity))/(sensitivity-(1-specificity)]
c (corrected)	(A + C)-a(corrected)
d (corrected)	(B + D)-b(corrected)
Corrected disease prevalence	[a(corrected) + c(corrected)]/ [a(corrected) + b(corrected) + c(corrected) + d(corrected)]
Corrected odds ratio	a(corrected)*d(corrected)/b(corrected)* c(corrected)

Appendix B International Classification of Diseases (ICD), 10th revision codes used to identify conditions

Severe renal failure

I12.0:	Hypertensive chronic kidney disease, stage 5
I13.1:	Hypertensive heart and chronic kidney disease without heart failure
N18.X:	Chronic kidney disease
N19.X:	Unspecified kidney failure
N25.0:	Renal osteodystrophy
Z49.0– Z49.2:	Preparatory care for renal dialysis
Z94.0:	Kidney transplant status
Z99.2:	Dependence on renal dialysis
K76.7:	Hepatorenal syndrome
D59.3:	Haemolytic-uraemic syndrome
R39.2:	Extra-renal uraemia
O08.4:	Renal failure following ectopic pregnancy
N99.0:	Post-procedural kidney failure
N17.X:	Acute kidney failure
N14.X:	Acute tubular necrosis from toxins

Colles' fracture

S52.50:	Colles' fracture, closed
S52.58:	Other fracture of lower end of radius, closed
S52.59:	Unspecified fracture of lower end of radius, closed
S52.60:	Fracture of lower end of both ulna and radius, closed

Appendix C Description of patient cohorts in the measurement and the analysis patient groups

Severe renal failure cohort	Measurement population	Analysis population	Overall
	$n = 49926$	$n = 50074$	$n = 100000$
Severe renal failure present	3613 (7.2%)	3748 (7.5%)	7361 (7.4%)
Mean age (SD)	54.8 ± 19.9	54.8 ± 19.9	54.8 ± 19.9
Female	28934 (58.0%)	28834 (57.6%)	57768 (57.8%)
Coded with possible causes of renal failure	1126 (2.3%)	1190 (2.4%)	2316 (2.3%)
Coded with manifestations of renal failure	1082 (2.2%)	1143 (2.3%)	2225 (2.2%)
Dialysis-related diagnoses	342 (0.7%)	352 (0.7%)	694 (0.7%)
Patient admitted emergently	30260 (60.6%)	30265 (60.4%)	60525 (60.5%)
Patient admitted from emergency	19271 (38.6%)	19437 (38.8%)	38708 (38.7%)
Patient came in by ambulance	12075 (24.2%)	12200 (24.4%)	24275 (24.3%)
Patient admitted to surgical service	12061 (24.2%)	12005 (24.0%)	24066 (24.1%)
Patient had operation during admission	7816 (15.7%)	7773 (15.5%)	15589 (15.6%)
Patient in ICU during admission	2274 (4.6%)	2297 (4.6%)	4571 (4.6%)
Patient had dialysis-related procedure	2913 (5.8%)	2925 (5.8%)	5838 (5.8%)
Patient died in hospital	2063 (4.1%)	2100 (4.2%)	4163 (4.2%)
Median length of hospital stay (IQR)	3.0 (2.0–7.0)	3.0 (2.0–7.0)	3.0 (2.0–7.0)

Colles' fracture cohort		Measurement population	Analysis population	Overall
		$n = 5553$	$n = 5680$	$n = 11233$
Colles' fracture present		2029 (36.5%)	2102 (37.0%)	4131 (36.8%)
Mean age (SD)		67.2 ± 12.9	67.4 ± 12.8	67.3 ± 12.8
Patient aged > 65		2770 (49.9%)	2866 (50.5%)	5636 (50.2%)
Patient is male		1929 (34.7%)	1995 (35.1%)	3924 (34.9%)
Visit year	<2010	2499 (45.0%)	2549 (44.9%)	5048 (44.9%)
	≥2010	3054 (55.0%)	3131 (55.1%)	6185 (55.1%)
CTAS score	1–2	907 (16.3%)	915 (16.1%)	1822 (16.2%)
	3+	4646 (83.7%)	4765 (83.9%)	9411 (83.8%)
Colles' fracture-related procedure coded		733 (13.2%)	752 (13.2%)	1485 (13.2%)
Discharged home		4741 (85.4%)	4803 (84.6%)	9544 (85.0%)
Charlson score > 0	>0	477 (8.6%)	459 (8.1%)	936 (8.3%)