BMJ Open

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006858
Article Type:	Research
Date Submitted by the Author:	07-Oct-2014
Complete List of Authors:	Lix, Lisa; University of Manitoba, Community Health Sciences Yao, Xue; Winnipeg Regional health Authority, Kephart, George; Dalhousie University, Quan, Hude; University of Calgary, Smith, Mark; Manitoba Centre for Health Policy, Kuwornu, John; University of Manitoba, Manoharan, Nitharsana; Institute for Clinical Evaluative Sciences, Kouokam, Wilfrid; Université de Bretagne-Sud, Sikdar, Khokan; University of Calgary,
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Health informatics, Research methods
Keywords:	STATISTICS & RESEARCH METHODS, PUBLIC HEALTH, EPIDEMIOLOGY

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

Lisa M. Lix[1], Xue Yao[2], George Kephart[3], Hude Quan[4], Mark Smith[5], John Paul Kuwornu[1], Nitharsana Manoharan[6], Wilfrid Kouokam[7], Khokan Sikdar[4]

[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, CANADA

[2]Winnipeg Regional Health Authority, Winnipeg, CANADA

[3]Department of Community Health Sciences, Dalhousie University, Halifax, CANADA

[4]Department of Community Health Sciences, University of Calgary, Calgary, CANADA

[5]Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, CANADA

[6]School of Public Health, University of Saskatchewan, Saskatoon, CANADA

[7]Faculty of Sciences and Engineering Sciences, Université de Bretagne-Sud, Vannes, FRANCE

Address for Author Correspondence:

Lisa M. Lix, PhD

Department of Community Health Sciences, University of Manitoba

S113-750 Bannatyne Avenue

Winnipeg, MB R3E 0W3

Phone: 204-789-3573; Fax: 204-789-3905

e-mail: lisa.lix@med.umanitoba.ca

Keywords. regression analysis; data quality; medical records; chronic disease; prevalence

Word Count (excluding title page, abstract, references, figures/tables): 2918

Abstract

Background. Electronic physician claims databases are widely used for chronic disease research and surveillance, but quality of the data may vary with a number of physician characteristics, including payment method. This research uses a population-based observational design to develop a prediction model for the number of prevalent diabetes cases in fee-for-service (FFS) electronic physician claims databases and apply it to estimate cases amongst non-fee-for-service (NFFS) physicians, for whom claims data are often incomplete.

Methods. Physician claims, physician registry, insured resident registry, and hospitalization records for one Canadian province were linked to ascertain a cohort with diagnosed diabetes. A generalized linear model with a gamma distribution was used to model the number of diabetes cases per FFS physician as a function of physician characteristics. The expected number of cases for NFFS physicians was estimated. The model was internally and externally validated.

Results. The diabetes case cohort consisted of 31,714 individuals; the mean cases per FFS physician was 75.5 (median = 49.0). Sex and years since specialty licensure were significantly associated ($p < .05$) with the number of cases per physician. Applying the prediction model to NFFS physician registry data resulted in an estimate of 18,546 cases; only 411 were observed in claims data. The model demonstrated face validity in an independent dataset.

Conclusions. Comparing observed and predicted disease cases is a useful and generalizable approach to assess the quality of electronic databases for population-based research and surveillance.

Strengths and limitations of this study

- This study developed a prediction model to estimate the completeness of non-fee-for-service electronic physician claims for capturing services to regional populations.
- The prediction model developed in this study is an efficient and potentially generalizable tool for routine estimation of the magnitude of data completeness.
- This study emphasizes that incomplete electronic physician claims data should be supplemented with other data sources, including electronic medical records, to ensure comprehensive data for chronic disease research and surveillance.
- The study focuses on completeness of electronic physician claims databases for diabetes only; the research should be extended to other chronic diseases to ensure its generalizability.

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

INTRODUCTION

Electronic administrative health databases are widely used for population-based health research and surveillance.[1;2] The popularity of these databases has arisen for several reasons: they are available in a timely manner, provide information about large numbers of individuals, and are relatively inexpensive to access and use. Physician claims electronic databases, which contain information on outpatient healthcare contacts, capture information on a larger proportion of the population than inpatient hospital records, but quality of claims databases tends to be poorer than that of hospital records for which standards for data collection and coding exist.[3;4] Studies about the quality of claims databases are therefore essential to evaluate and improve their accuracy. However, most studies about physician claims databases have focused only on the validity of diagnosis codes,[5-8] while other elements of data quality that could impact on the usefulness of these data for research and surveillance have infrequently been examined.[9]

Incompleteness of physician claims databases, which can result in substantially biased estimates of disease prevalence and healthcare utilization, may arise for a number of reasons. The information in these databases is used to remunerate physicians for services provided to patients, usually on a fee-for-service (FFS) basis. However, physicians not remunerated by FFS methods may inconsistently record patient encounters in these databases. Specifically, non-FFS (NFFS) physicians, who are often paid via salaries and contracts, are not always required to use the same claims submission processes as FFS physicians,[10] a process known as shadow billing. Incomplete capture of NFFS physician claims can have serious consequences; previous research has demonstrated substantial underestimation of diabetes prevalence associated with a lack of shadow billing.[11]

Methods to estimate completeness of electronic administrative databases [12-16] include: (a) comparing observed to expected numbers of cases, where expected cases are estimated from a parametric or non-parametric model, (b) comparing the number of cases ascertained in administrative databases to cases ascertained from a validated database, (c) using capture-recapture models, and (d) conducting database audits. These techniques have primarily been applied to cancer registry and hospital records, but not to physician claims databases. Therefore, the purpose of this study was to develop a population-based model to predict prevalent diabetes cases from FFS physician claims and apply it to estimate cases amongst NFFS physicians, for whom claims data may be incomplete. We focus on diabetes because administrative health databases have demonstrated good sensitivity and specificity for case identification using electronic administrative databases and surveillance of diabetes is of interest worldwide.[6]

METHODS

Data Sources for Prediction Model

Data to construct the prediction model were from the eastern Canadian province of Newfoundland and Labrador (NL), which has a population of approximately 515,000 according to the 2011 Statistics Canada Census. NL physicians remunerated by NFFS methods do not submit shadow-billed claims to the provincial ministry of health,[17] while physicians remunerated by FFS methods submit all of their claims to the ministry. NL has a larger proportion of NFFS physicians than most other Canadian provinces.[18]

Physician claims, physician registry records, hospital discharge abstracts, and insured resident registry records from April 1, 2002 to March 31, 2004 were used to conduct the study. We selected these years because the NL physician registry contains comprehensive information on all registered physicians in this time period but is incomplete in later years; the registry includes information about physician remuneration methods, sex, age, specialty and year it was

received, year that the medical degree was obtained, and practice region. Each physician claim contains a single three-digit diagnosis code recorded using the International Classification of Diseases, 9th revision (ICD-9) and date of service. Hospital discharge abstracts contain dates of admission and discharge and up to 20 ICD-9 and ICD-10-CA diagnosis codes. The resident registry contains dates of health insurance coverage, sex, date of birth, and health region for all residents eligible for health insurance benefits. Physician claims, hospital separation abstracts, and insured resident registry records are linkable using a unique, anonymized patient identifier. Physician claims and the physician registry are also linkable using an anonymized physician identifier.

Study Cohort for Prediction Model

The diabetes case cohort comprised all individuals who met a validated case definition, which requires one hospitalization or two physician billing claims (ICD-9 code 250; ICD-10-CA code E10-E14) within a 730-day period.[5;19] Individuals less than 20 years of age or without health insurance coverage at the date of the case-qualifying diagnosis were excluded. For cases ascertained from hospital discharge abstracts, the date of the case-qualifying diagnosis was the date of hospital admission; for cases ascertained from physician claims, the date of the case-qualifying diagnosis was the date of the physician claim for the second diagnosis within the 730-day period. Diabetes cases were classified into three mutually exclusive groups: (a) cases ascertained only from hospital discharge abstracts, (b) cases ascertained from physician claims for which the case-qualifying diagnosis was from a FFS physician, and (c) cases ascertained from physician claims for which the case-qualifying diagnosis was from a NFFS physician. The last group is comprised of cases from the claims of a small number of NFFS physicians who receive a portion of their remuneration by FFS payments.

The physician cohort included all members of the physician registry who had at least one claim for an individual in the diabetes case cohort. Each physician was assigned to each member of the diabetes case cohort in the second and third groups based on the physician identification number found on the billing claim for the case-qualifying diabetes diagnosis.

Statistical Analyses for Prediction Model

The diabetes case and physician cohorts were described using means, standard deviations, medians, frequencies, and percentages. The mean and median number of diabetes cases per physician was estimated and stratified by physician cohort characteristics.

A multivariable generalized linear regression model with a gamma distribution was fit to the number of diabetes cases for each FFS physician.[20] The model covariates were years since specialty licensure (quartiles; reference = lowest quartile), physician sex (reference = female), region of practice (reference = Labrador, a rural/remote region of NL), and specialty (reference = specialist). Years since specialty licensure was highly correlated with years since medical licensure and age ($r \geq 0.80$), hence the latter two variables were excluded. A main effects model was compared to a model with main and two-way interaction effects.[20] Penalized goodness-of-fit measures, including the Akaike Information Criterion (AIC),[21] were used to select the best fit model. The ratio of the deviance to degrees of freedom was used to assess model dispersion.

Model Validation

We selected the Canadian province of Manitoba (MB) for external validation, which has a population of 1.2 million according to the 2011 Statistics Canada Census. NFFS physicians in this province submit shadow-billed claims to the provincial ministry of health. Watson et al.[22] reported that amongst family physicians practicing in Winnipeg, the only major urban centre in

Manitoba (680,000+ population), up to 90% of physicians remunerated by NFFS methods submit claims for services provided to patients. However, rates of shadow billing are expected to be lower in other regions of the province.

The same data sources were available in MB as in NL, with minor differences in database characteristics. Specifically, physician claims in MB contain diagnosis codes based on ICD-9-CM (i.e., Clinical Modification).[23] The MB physician registry does not contain information on year of medical licensure. Five health regions, defined by the ministry of health for planning the delivery of healthcare services, were used to identify patient residence and physician practice locations.

Internal validation was conducted for both the NL and MB models. Measures of prediction accuracy, which included bias, mean absolute error (MAE), and root mean square error (RMSE),[24] were calculated based on 10-fold cross-validation.[25;26]

Model Prediction

The final fitted model for NL was used to predict the number of prevalent diabetes cases per NFFS physician. However, given that not all NFFS physicians provide services to diabetes patients, we used the ratio of FFS physicians in the physician cohort to the total number of FFS physicians in the province[27] to select a random prediction sample. A similar process was used predict the number of cases from the MB data. In MB we also compared the predicted number of diabetes cases for NFFS physicians to the observed number of cases from the shadow-billed claims of NFFS physicians.

The total number of prevalent diabetes cases in each province was estimated as the sum of: (a) observed cases ascertained from hospital discharge abstracts only, (b) observed cases ascertained from claims of FFS physicians, (c) predicted cases for NFFS physicians.

Denominators of the prevalence estimates were based on 2001 Statistics Canada Census data; 95% confidence intervals were calculated using the binomial distribution.

All analyses were conducted using SAS version 9.3. Ethics approval was provided by the University of Manitoba Health Research Ethics Board and the NL Health Research Ethics Board. Data access approval was provided by the Newfoundland and Labrador Centre for Health Information and the Manitoba Health Information Privacy Committee.

RESULTS

Descriptive Analyses

A total of 31,714 prevalent diabetes cases were identified from the NL administrative data (Table 1); 91.1% (n = 28,989) of cases were identified from billing claims of physicians remunerated by FFS, while 1.3% (*n* = 411) of cases were ascertained from billing claims submitted by NFFS physicians who received a portion of their remuneration by FFS. Almost two-thirds (60.7%) of diabetes cases from FFS physician claims were residents of the Eastern health region, which contains the largest city in NL (200,000+ population); 40.5% were 65+ years.

In the MB external validation data, 51,031 prevalent diabetes cases were identified (Table 1), of which 84.1% were ascertained from the billing claims of FFS physicians. Three-quarters (75.9%) of prevalent cases ascertained from the shadow-billed claims of NFFS physicians were from rural health regions.

Table 1. Characteristics of diabetes case cohort by ascertainment source and province

Case characteristics	Cases ascertained from hospital discharge abstracts		Cases ascertained from physician billing claims for FFS physicians		Cases ascertained from physician billing claims for NFFS physicians[a]	
	n	%	*n*	%	*n*	%
Newfoundland and Labrador (*N* = 31,714)						
Total	2,405	100.0	28,898	100.0	411	100.0
Sex						
Male	1,158	48.1	13,872	48.0	217	52.8
Female	1,247	51.9	15,026	51.9	194	47.2
Age group						
<35 years	39	1.6	1,448	5.0	30	7.3
35 - 49 years	168	7.0	4,932	17.1	84	20.4
50 - 64 years	570	23.7	10,808	37.4	136	33.1
65+ years	1,628	67.7	11,710	40.5	161	39.2
Health region						
Eastern	1,201	49.9	17,547	60.7	110	26.8
Central	523	21.7	5,909	20.4	258	62.8
Western	389	16.2	4,840	16.7	7	1.7
Labrador	267	11.1	464	1.6	35	8.5
Missing	25	1.0	138	0.5	1	0.2
Manitoba (*N* = 51,031)						
Total	2,250	100.0	42,933	100.0	5,848	100.0
Sex						
Male	1,161	51.6	22,078	51.4	2,764	47.3
Female	1,089	48.4	20,855	48.6	3,084	52.7
Age group						
<35 years	71	3.2	1,952	4.6	375	6.4
35 - 49 years	236	10.5	7,636	17.8	1,358	23.2
50 - 64 years	534	23.7	15,319	35.7	2,120	36.3
65+ years	1,409	62.6	18,026	42.0	1,995	34.1
Health region						
Winnipeg	1,180	62.6	25,949	60.4	1,409	24.1
Interlake-Eastern	262	11.6	4,503	10.5	970	16.6
Northern	189	8.4	1,951	4.5	1,562	26.7
Prairie Mountain	370	16.4	6,400	14.9	1,067	18.3
Southern	249	11.1	4,130	9.6	840	14.4

[a]These cases were ascertained from the claims of NFFS physicians receiving partial FFS remuneration in Newfoundland and Labrador, and from the claims of NFFS physicians who shadow bill in Manitoba.

Table 2. Characteristics of the physician cohort by method of remuneration and province

Physician characteristics	Newfoundland and Labrador (N = 388)			
	FFS (n = 362)		NFFS[a] (n = 26)	
	n	%	*n*	%
Specialty				
General practitioner	291	80.4	22	84.6
Specialist	71	19.6	4	15.4
Sex				
Male	257	70.9	19	73.1
Female	105	29.0	7	26.9
Age group				
< 40 years	85	23.5	15	57.7
40 – 64 years	269	74.3	11	42.3
65+ years	8	2.2	0	0.0
Health region				
Eastern	258	71.3	6	23.1
Central	56	15.5	13	50.0
Western	42	11.6	3	11.5
Labrador	6	1.7	4	15.4
Medical licensure, years[b]	22.5 (10.7)	22.0	15.0 (9.7)	14.0
Specialty licensure, years[b]	17.2 (10.1)	17.0	6.8 (8.9)	3.5
	Manitoba (N = 1,229)			
	FFS (n = 989)		**NFFS** (n = 270)	
	n	%	*n*	%
Specialty				
General practitioner	770	77.9	--	--
Specialist	219	22.1	--	--
Sex				
Male	741	74.9	201	74.4
Female	248	25.1	69	25.6
Age group				
< 40 years	301	30.4	185	68.5
40 - 64 years	572	57.8	--	--
65+ years	116	11.8	--	--
Missing	0	0.0	0	0.0
Health region				
Winnipeg	659	66.6	57	21.1
Interlake-Eastern	61	6.2	40	14.8
Northern	25	2.5	63	23.3
Prairie Mountain	152	15.4	62	23.0

Southern	92	9.3	48	17.8
Specialty licensure, years[a]	12.1 (9.9)	10.0	5.2 (6.4)	3.0

[a]In Newfoundland and Labrador, NFFS physicians identified in claims data received partial FFS remuneration, while in Manitoba, NFFS physicians identified in claims data shadow bill.

[b]Reported values are mean (SD) and median; some cells cannot be reported, in accordance with Manitoba Health requirements, because of small numbers

There were 388 individuals in the NL physician cohort (Table 2). Amongst FFS physicians (93.3%), the majority were general practitioners (80.4%), and most were from the Eastern health region (71.3%). In the MB physician cohort, which contained more than 1200 physicians, 80.4% were FFS physicians. Amongst these FFS physicians, more than half (57.8%) were in the 40-64 years age group. The NFFS physicians (n = 270) were primarily less than 40 years (68.5%) and almost 80.0% practiced outside of the urban Winnipeg health region.

Table 3 describes the mean and median number of prevalent diabetes cases per FFS physician. In NL, the average number of prevalent cases per FFS physician was 75.5 and the median was 49.0. The mean and median were higher for general practitioners than for specialists and also for males than females. For MB, the average number of prevalent diabetes cases per FFS physician was 43.4 and the median was 25.0.

Prediction Model

For NL, the main effects model provided a good fit to the data, as judged by the ratio of model deviance to degrees of freedom (ratio =1.0) and the AIC was smaller for a main effects model than for one with main and two-way interaction effects (3833.1 versus 3830.4); Likelihood ratio tests revealed statistically significant main effects for sex ($p < .0001$) and years since specialty licensure ($p = .0006$).

Table 3. Mean (standard deviation) and median number of prevalent cases in the diabetes case cohort per FFS physician in the physician cohort

Physician characteristics	Mean (SD)	Median
	Newfoundland and Labrador	
Province	75.5 (84.6)	49.0
Specialty		
General practitioner	79.0 (66.2)	66.0
Specialist	61.0 (136.8)	9.0
Sex		
Male	89.3 (94.1)	75.0
Female	41.5 (37.2)	32.5
Age group		
< 40 years	54.9 (64.8)	32.5
40 – 64 years	99.9 (98.3)	91.0
65+ years	63.8 (68.6)	34.5
Health region		
Eastern	67.8 (73.1)	42.0
Central	87.8 (87.7)	59.0
Western	108.1 (129.5)	86.0
Labrador	47.6 (47.9)	38.5
	Manitoba	
Province	43.4 (74.2)	25.0
Specialty		
General practitioner	45.1 (45.7)	35.0
Specialist	37.6 (132.8)	3.0
Sex		
Male	47.7 (76.0)	33.0
Female	30.5 (67.2)	17.0
Age group		
<40 years	25.5 (34.2)	14.5
40-64 years	52.1 (90.4)	34.0
65+ years	47.1 (49.8)	34.5
Health region		
Interlake-Eastern	45.9 (37.8)	48.0
Northern	49.4 (59.4)	20.0
Prairie Mountain	42.4 (39.7)	35.0
Southern	35.1 (29.1)	28.0
Winnipeg	44.3 (86.7)	20.0

The regression analyses produced similar results in the MB external validation data; the ratio of model deviance to degrees of freedom was close to 1.0 for the main effects model. The model with main and two-way interaction effects resulted in a negligible decrease in the AIC. The main effects of sex ($p < .0001$), speciality ($p = .0021$), and years since specialty licensure ($p < .0001$) were statistically significant.

With respect to the internal cross-validation, for the NL model absolute bias estimates ranged from 0.2% to 12.9% across the ten data folds, while for the MB model the estimates ranged from 0.6% to 13.8%. The MAE ranged from 40.1 to 67.5 for the NL model and from 26.7 to 43.2 for the MB model. Finally, the RMSE ranged from 56.5 to 131.2 for the NL model and from 33.8 to 151.0 for the MB model.

Table 4. Observed and predicted average number of diabetes cases per fee-for-service (FFS) and non-fee-for-service (NFFS) physician in Manitoba's physician cohort

	FFS		NFFS	
	Observed	**Predicted**	**Observed**	**Predicted**
Entire province	43.4	43.8	21.7	32.7
Health region				
Interlake-Eastern	45.9	49.7	20.7	31.9
Northern	49.4	43.3	15.1	30.6
Prairie Mountain	42.4	44.0	16.0	39.4
Southern	35.1	36.0	17.1	21.8
Winnipeg	44.3	44.3	39.5	37.4

Using the MB model results, we compared the observed and expected number of prevalent diabetes cases per FFS and NFFS physician (Table 4) for the entire province and by health region. The provincial and regional figures were similar for FFS physicians, supporting the internal validity of the model. For NFFS physicians, the expected number of cases was 51.0% higher than the observed number for the entire province. When we examined these values by health region, we found that the expected value was 8.2% lower than the observed value for the Winnipeg (urban) health region. However, for the remaining health regions, which encompass rural and/or remote areas, the expected values were much higher than the observed values.

Figure 1 shows the percentage of diabetes cases ascertained from each data source in both provinces. In NL, the prediction model resulted in a 37.2% increase in the number of diabetes cases ascertained from the administrative databases, while in MB it resulted in a 16.3% increase. In NL, crude diabetes prevalence based on cases ascertained only from hospital data and FFS physician claims was 8.1%, while the estimate based on observed and expected cases was 13.0% (95% CI: 12.9, 13.0). In MB, the crude diabetes prevalence estimate based on cases ascertained from hospital data and FFS physician claims was 5.6%, while the estimate based on both observed and expected cases was 6.7% (95% CI: 6.7%, 6.8%).

DISCUSSION

This study developed a prediction model for linked administrative health databases to estimate the completeness of electronic physician claims data; the model was applied to estimate under-ascertainment of prevalent diabetes cases but could be applied to other chronic or acute conditions that are managed or treated in primary care settings. When the model was applied to data from the Canadian province of NL, the results revealed that close to 40% of diabetes cases

were missed because NFFS physicians do not report contacts with patients in claims data. When the model was externally validated in MB, a province in which some NFFS physicians submit some claims, the modeling results indicated than less than 20% of diabetes cases were missed, but this percentage varied substantially by region; there was less bias in an urban health region and more substantial bias in rural health regions having a higher proportion of NFFS physicians.

Data from the 2005 Canadian Community Health Survey,[28] a national survey used for regional chronic disease surveillance, revealed a crude diabetes prevalence of 6.8% for NL and 4.4% for MB for the population 12+ years, a difference of more than 50%. When we compared crude prevalence estimates for the two provinces using only FFS claims and hospital records, rates in NL were just 8.9% higher than those in MB. However, after adjustment for potential missed cases using our prediction model, crude prevalence was 45.1% higher in NL than in MB, producing a similar difference in estimates to those observed in survey data.

Incomplete capture of claims for NFFS physicians is similar to unit non-response in survey data, both of which can bias parameter estimates and increase variance estimates. Unit non-response in surveys is often difficult to adjust for, because information about non-responders is rarely available to the researcher. In fact, administrative data have been used in previous research to estimate the effect of survey non-response bias in estimates of health care use.[29] However, our study suggests that the use of administrative data for evaluating survey non-response should be adopted with caution, as administrative databases may themselves be incomplete.

While the proposed prediction model provides a useful tool to estimate bias in disease prevalence due to incomplete claims data, it is equally important to consider how other databases can be used to address gaps in these data. Electronic medical records are increasingly being

adopted in population-based chronic disease research and surveillance studies,[30] and could represent an important additional source of data for case ascertainment. Pharmacy databases have also been used for case ascertainment [31] when the medications used for disease treatment have high specificity for case capture.

Limitations of the study include the restricted set of explanatory variables available to develop the prediction model; residual confounding due to factors such as physician productivity,[10] type of practice, and even characteristics of the patients seen by a physician may affect prediction accuracy.[32] Strengths of the study include the use of a validated case definition to ascertain diabetes cases and the internal and external validation process.

Further research could examine the validity of the prediction model by applying it to other chronic diseases and in other jurisdictions; [33] the utility of the model is not limited to Canadian administrative data, as a similar approach has been proposed to evaluate the completeness of cancer registry data.[16] Simulation could also be used to assess the impact of patient, physician, and health system characteristics on estimates of completeness.[34] For example, the model assumes that physician characteristics will have the same distribution and association with the number of prevalent diabetes cases in FFS and NFFS populations, which may not be a valid assumption.[35]

In summary, this study revealed that completeness of physician claims data are associated with method of physician remuneration and that a predictive model can be used to estimate the magnitude of data incompleteness for disease surveillance. This predictive model makes use of routinely collected linked data, and therefore is feasible to implement over time and across jurisdictions.

ACKNOWLEDGEMENTS

The authors are indebted to Manitoba Health, Healthy Living, and Seniors (HIPC 2012/2013-04) and the Newfoundland and Labrador Centre for Health Information for the provision of data. The results and conclusions are those of the authors, and no official endorsement by Manitoba Health, Healthy Living, and Seniors is intended or should be inferred.

FUNDING

This research was funded by the Canadian Institutes of Health Research (Funding Reference Number 123357). The first author is supported by a Research Chair from the Manitoba Health Research Council. The funders had no involvement in the conduct of the research or in manuscript preparation.

STATEMENT OF CONTRIBUTIONS

LML, GK, HQ, MS, and KS designed the analysis and acquired the study data. JPK, XY, NM, and WK conducted the analyses. LML, JPK, and NM drafted the manuscript and all remaining authors read and revised it substantially. All authors approved the final version of the manuscript before submission.

DATA SHARING STATEMENT

No data are available.

REFERENCES

(1) Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Annu Rev Public Health* 2001;22:213-230.

(2) Dombkowski KJ, Wasilevich EA, Lyon-Callo S, et al. Asthma surveillance using Medicaid administrative data: a call for a national framework. *J Public Health Manag Pract* 2009;15:485-493.

(3) Potter BK, Manuel D, Speechley KN, et al. Is there value in using physician billing claims along with other administrative health care data to document the burden of adolescent injury? An exploratory investigation with comparison to self-reports in Ontario, Canada. *BMC Health Serv Res* 2005;5:15.

(4) Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Med Care* 2006;44:1011-1019.

(5) Hux JE, Ivis F, Flintoft V, et al. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* 2002;25:512-516.

(6) Saydah SH, Geiss LS, Tierney E, et al. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Annals of Epidemiology* 2004;14:507-516.

(7) Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension* 2009;54:1423-1428.

(8) Tu K, Campbell NRC, Chen Z-L, et al. Accuracy of administrative databases in identifying patients with hypertension. *Open Medicine* 2007;1:E3-E5.

(9) Saez M, Barcelo MA, Coll de TG. A selection-bias free method to estimate the prevalence of hypertension from an administrative primary health care database in the Girona Health Region, Spain. *Comput Methods Programs Biomed* 2009;93:228-240.

(10) Wranik DW, Durier-Copp M. Physician remuneration methods for family physicians in Canada: Expected outcomes and lessons learned. *Health Care Anal* 2009;18:35-59.

(11) Alshammari AM, Hux JE. The impact of non-fee-for-service reimbursement on chronic disease surveillance using administrative data. *Can J Public Health* 2009;100:472-474.

(12) Crocetti E, Miccinesi G, Paci E, et al. An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy. *Eur J Cancer Prev* 2001;10:417-423.

(13) Dockerty JD, Becroft DM, Lewis ME, et al. The accuracy and completeness of childhood cancer registration in New Zealand. *Cancer Causes Control* 1997;8:857-864.

(14) Schouten LJ, Straatman H, Kiemeney LA, et al. The capture-recapture method for estimation of cancer registry completeness: a useful tool? *Int J Epidemiol* 1994;23:1111-1116.

(15) Brenner H, Stegmaier C, Ziegler H. Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany. *J Epidemiol Community Health* 1995;49:426-430.

(16) Das B, Clegg LX, Feuer EJ, et al. A new method to evaluate the completeness of case ascertainment by a cancer registry. *Cancer Causes Control* 2008;19:515-525.

(17) Newfoundland and Labrador Centre for Health Information. Enhancing chronic disease surveillance in Newfoundland and Labrador: adjustment of rates based on physician payment methods. Newfoundland and Labrador Centre for Health Information. 2010. St. John's, NL, Newfoundland and Labrador Centre for Health Information.

(18) Canadian Institute for Health Information. National physician database, 2008-2009. 2010. Ottawa, Canadian Institute for Health Information.

(19) Clottey C, Mo F, LeBrun B, et al. The development of the National Diabetes Surveillance System (NDSS) in Canada. *Chronic Dis Can* 2001;22:67-69.

(20) McCulloch CE, Searle SR. Generalized, Linear, and Mixed Models. New York: Wiley; 2001.

(21) Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 1987;52:345-370.

(22) Watson DE, Katz A, Reid RJ, et al. Family physician workloads and access to care in Winnipeg: 1991 to 2001. *CMAJ* 2004;171:339-342.

(23) Lix LM, Walker R, Quan H, et al. Features of physician billing claims databases in Canada. *Chron Dis Can* 2012;32:186-193.

(24) Dunn G, Mirandola M, Amaddeo F, et al. Describing, explaining or predicting mental health care costs: A guide to regression models - Methodological review. *Br J Psychiatry* 2003;183:398-404.

(25) Austin PC, Rothwell DM, Tu JV. A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. *Health Services and Outcomes Research Methodology* 2002;3:107-133.

(26) Kuwornu JP, Lix LM, Quail J, et al. A comparison of statistical models for analyzing episode-of-care costs for chronic obstructive pulmonary disease. *Health Services and Outcomes Research Methodology* 2013;13:203-208.

(27) Canadian Institute for Health Information. The status of alternate payment programs for physicians in Canada: 2002-2003 and preliminary information for 2003-2004. 2005. Ottawa, ON, Canadian Institute for Health Information.

(28) Sanmartin C, Gilmore J. Diabetes prevalence and care practices. *Health Rep* 2008;19:59-63.

(29) Gundgaard J, Ekholm O, Hansen EH, et al. The effect of non-response on estimates of health care utilisation: linking health surveys and registers. *Eur J Public Health* 2008;18:189-194.

(30) Desai JR, Wu P, Nichols GA, et al. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care* 2012;50 Suppl:S30-S35.

(31) Maio V, Yuen E, Rabinowitz C, et al. Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. *J Health Serv Res Policy* 2005;10:232-238.

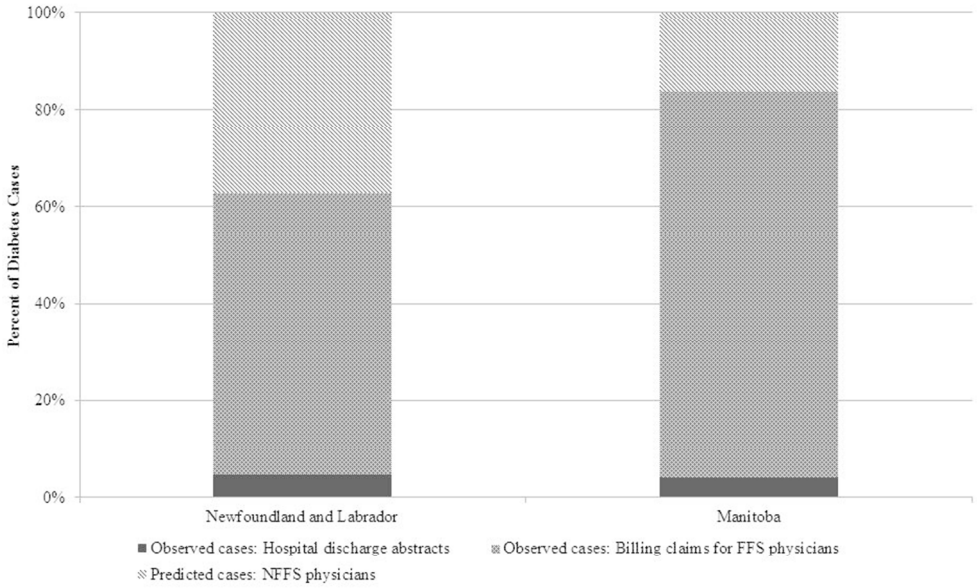
(32) Hanley JA, Dendukuri N. Efficient sampling approaches to address confounding in database studies. *Stat Methods Med Res* 2009;18:81-105.

(33) Kleinberg S, Elhadad N. Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. *AMIA Annu Symp Proc* 2013; 2013:786-795.

(34) Silcocks PB, Robinson D. Simulation modelling to validate the flow method for estimating completeness of case ascertainment by cancer registries. *J Public Health (Oxf)* 2007;29:455-462.

(35) Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-483.

LIST OF FIGURES



254x190mm (96 x 96 DPI)

STROBE Statement—checklist of items that should be included in reports of observational studies

NOTE: ALL ITEMS THAT HAVE BEEN ACHIEVED ARE HIGHLIGHTED IN YELLOW.

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy NOT APPLICABLE
		(*e*) Describe any sensitivity analyses

Continued on next page

Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time
		Case-control study—Report numbers in each exposure category, or summary measures of exposure
		Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results	18	Summarise key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006858.R1
Article Type:	Research
Date Submitted by the Author:	30-Apr-2015
Complete List of Authors:	Lix, Lisa; University of Manitoba, Community Health Sciences Yao, Xue; Winnipeg Regional health Authority, Kephart, George; Dalhousie University, Quan, Hude; University of Calgary, Smith, Mark; Manitoba Centre for Health Policy, Kuwornu, John; University of Manitoba, Manoharan, Nitharsana; Institute for Clinical Evaluative Sciences, Kouokam, Wilfrid; Université de Bretagne-Sud, Sikdar, Khokan; University of Calgary,
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Health informatics, Research methods
Keywords:	STATISTICS & RESEARCH METHODS, PUBLIC HEALTH, EPIDEMIOLOGY

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

Lisa M. Lix[1], Xue Yao[2], George Kephart[3], Hude Quan[4], Mark Smith[5], John Paul Kuwornu[1], Nitharsana Manoharan[6], Wilfrid Kouokam[7], Khokan Sikdar[4]

[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, CANADA

[2]Winnipeg Regional Health Authority, Winnipeg, CANADA

[3]Department of Community Health Sciences, Dalhousie University, Halifax, CANADA

[4]Department of Community Health Sciences, University of Calgary, Calgary, CANADA

[5]Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, CANADA

[6]Institute for Clinical Evaluative Sciences, Toronto, CANADA

[7]Faculty of Sciences and Engineering Sciences, Université de Bretagne-Sud, Vannes, FRANCE

Address for Author Correspondence:

Lisa M. Lix, PhD
Department of Community Health Sciences, University of Manitoba
S113-750 Bannatyne Avenue
Winnipeg, MB R3E 0W3
Phone: 204-789-3573; Fax: 204-789-3905
e-mail: lisa.lix@med.umanitoba.ca

Keywords. regression analysis; data quality; medical records; chronic disease; prevalence

Word Count (excluding title page, abstract, references, figures/tables): 2920

Abstract

Objectives. Electronic physician claims databases are widely used for chronic disease research and surveillance, but quality of the data may vary with a number of physician characteristics, including payment method. The objectives were to develop a prediction model for the number of prevalent diabetes cases in fee-for-service (FFS) electronic physician claims databases and apply it to estimate cases amongst non-fee-for-service (NFFS) physicians, for whom claims data are often incomplete.

Design. A retrospective observational cohort design was adopted.

Setting. Data from the Canadian province of Newfoundland and Labrador were used to construct the prediction model and data from the province of Manitoba were used to externally validate the model.

Participants. A cohort of diagnosed diabetes cases was ascertained from physician claims, insured resident registry, and hospitalization records. A cohort of FFS physicians who were responsible for the diagnosis was ascertained from physician claims and registry data.

Primary and Secondary Outcome Measures. A generalized linear model with a gamma distribution was used to model the number of diabetes cases per FFS physician as a function of physician characteristics. The expected number of diabetes cases per NFF physician was estimated.

Results. The diabetes case cohort consisted of 31,714 individuals; the mean cases per FFS physician was 75.5 (median = 49.0). Sex and years since specialty licensure were significantly associated ($p < .05$) with the number of cases per physician. Applying the prediction model to NFFS physician registry data resulted in an estimate of 18,546 cases;

only 411 were observed in claims data. The model demonstrated face validity in an independent dataset.

Conclusions. Comparing observed and predicted disease cases is a useful and generalizable approach to assess the quality of electronic databases for population-based research and surveillance.

Strengths and limitations of this study

- This study developed a prediction model to estimate the completeness of non-fee-for-service electronic physician claims for capturing services to populations.
- The prediction model developed in this study is an efficient and potentially generalizable tool for routine estimation of the magnitude of administrative data completeness.
- This study emphasizes that incomplete electronic physician claims data should be supplemented with other data sources, including electronic medical records, to ensure comprehensive data for chronic disease research and surveillance.
- The study focuses on completeness of electronic physician claims databases for diabetes; the research should be extended to other chronic diseases to ensure its generalizability.

A Prediction Model to Estimate Completeness of Electronic Physician Claims Databases

INTRODUCTION

Electronic administrative health databases are widely used for population-based health research and surveillance.[1;2] The popularity of these databases has arisen for several reasons: they are available in a timely manner, provide information about large numbers of individuals, and are relatively inexpensive to access and use. Physician claims electronic databases, which contain information on outpatient healthcare contacts, capture information on a larger proportion of the population than inpatient hospital records, but quality of claims databases tends to be poorer than that of hospital records for which standards for data collection and coding exist.[3;4] Studies about the quality of claims databases are therefore essential to evaluate and improve their accuracy. However, most studies about physician claims databases have focused only on the validity of diagnosis codes,[5-8] while other elements of data quality that could impact on the usefulness of these data for research and surveillance have infrequently been examined.[9]

Incompleteness of physician claims databases, which can result in substantially biased estimates of disease prevalence and healthcare utilization, may arise for a number of reasons. The information in these databases is used to remunerate physicians for services provided to patients, usually on a fee-for-service (FFS) basis. However, physicians not remunerated by FFS methods may inconsistently record patient encounters in these databases. Specifically, non-FFS (NFFS) physicians, who are often paid via salaries and contracts, are not always required to use the same claims submission processes as FFS physicians,[10] a process known as shadow billing. Incomplete capture of NFFS physician claims can have serious consequences; previous research has demonstrated substantial underestimation of diabetes prevalence associated with a lack of shadow billing.[11]

Possible methods to estimate completeness of electronic administrative databases [12-16] include: (a) comparing observed to expected numbers of cases, where expected cases are estimated from a parametric or non-parametric model, (b) comparing the number of cases ascertained in administrative databases to cases ascertained from a validated database, (c) using capture-recapture models, and (d) conducting database audits. These methods have primarily been applied to cancer registry and hospital records, but not to physician claims databases. Therefore, the purpose of this study was to develop a population-based model to predict prevalent diabetes cases from FFS physician claims and apply it to estimate cases amongst NFFS physicians, for whom claims data may be incomplete. We focus on diabetes because administrative health databases have demonstrated good sensitivity and specificity for case identification using electronic administrative databases and surveillance of diabetes is of interest worldwide.[6]

METHODS

Data Sources for Prediction Model

Data to construct the prediction model were from the eastern Canadian province of Newfoundland and Labrador (NL), which has a population of approximately 515,000 according to the 2011 Statistics Canada Census. NL physicians remunerated by NFFS methods do not submit shadow-billed claims to the provincial ministry of health,[17] while physicians remunerated by FFS methods submit all of their claims to the ministry. NL has a larger proportion of NFFS physicians than most other Canadian provinces.[18]

Physician claims, physician registry records, hospital discharge abstracts, and insured resident registry records from April 1, 2002 to March 31, 2004 were used to conduct the study. We selected these years because the NL physician registry contains comprehensive information on all registered physicians in this time period but is incomplete in later years; the registry

includes information about physician remuneration methods, sex, age, specialty, year the medical degree was obtained, and health region of the practice location. Each physician claim contains a single three-digit diagnosis code recorded using the International Classification of Diseases, 9th revision (ICD-9) and date of service. Hospital discharge abstracts contain dates of admission and discharge and up to 20 ICD-9 and ICD-10-CA diagnosis codes. The resident registry contains dates of health insurance coverage, sex, date of birth, and health region for all residents eligible for health insurance benefits. Physician claims, hospital separation abstracts, and insured resident registry records are linkable using a unique, anonymized patient identifier. Physician claims and the physician registry are also linkable using an anonymized physician identifier.

Study Cohort for Prediction Model

The diabetes case cohort comprised all individuals who met a validated case definition, which requires at least one hospitalization or at least two physician billing claims (ICD-9 code 250; ICD-10-CA code E10-E14) within a 730-day period.[5;19] Individuals less than 20 years of age or without health insurance coverage at the date of the case-qualifying diagnosis were excluded. For cases ascertained from hospital discharge abstracts, the date of the case-qualifying diagnosis was the date of hospital admission; for cases ascertained from physician claims, the date of the case-qualifying diagnosis was the date of the physician claim for the second diagnosis within the 730-day period. Diabetes cases were classified into three mutually exclusive groups: (a) cases ascertained only from hospital discharge abstracts, (b) cases ascertained from physician claims for which the case-qualifying diagnosis was from a FFS physician, and (c) cases ascertained from physician claims for which the case-qualifying diagnosis was from a NFFS physician. The last group is comprised of cases from the claims of a small number of NFFS physicians who receive a portion of their remuneration by FFS payments. While cases in the

latter two groups could have a hospital discharge abstract with a diabetes diagnosis, they qualified as a case based on having at least two physician billing claims with a diabetes diagnosis.

The physician cohort included all members of the physician registry who had at least one claim for an individual in the diabetes case cohort. Each physician was assigned to each member of the diabetes case cohort in the second and third groups based on the physician identification number found on the billing claim for the case-qualifying diabetes diagnosis.

Statistical Analyses for Prediction Model

The diabetes case and physician cohorts were described using means, standard deviations, medians, frequencies, and percentages. The mean and median number of diabetes cases per physician was estimated and stratified by physician cohort characteristics.

A multivariable generalized linear regression model with a gamma distribution was fit to the number of diabetes cases for each FFS physician.[20] The model covariates were years since specialty was received (quartiles; reference = lowest quartile), physician sex (reference = female), health region of practice (reference = Labrador, a remote region of NL), and specialty (reference = specialist). Years since specialty was highly correlated with years since medical degree and age ($r \geq 0.80$), hence the latter two variables were excluded. A main effects model was compared to a model with main and two-way interaction effects.[20] Penalized goodness-of-fit measures, including the Akaike Information Criterion (AIC),[21] were used to select the best fit model. The ratio of the deviance to degrees of freedom was used to assess model dispersion.

Model Validation

We selected the Canadian province of Manitoba (MB) for external validation, which has a population of 1.2 million according to the 2011 Statistics Canada Census. NFFS physicians in this province submit shadow-billed claims to the provincial ministry of health. Watson et al.[22] reported that amongst family physicians practicing in Winnipeg, the only major centre in Manitoba (680,000+ population), up to 90% of physicians remunerated by NFFS methods submit claims for services provided to patients. However, rates of shadow billing are expected to be lower in other regions of the province.

The same data sources were available in MB as in NL, with minor differences in database characteristics. Specifically, physician claims in MB contain diagnosis codes based on ICD-9-CM (i.e., Clinical Modification).[23] The MB physician registry does not contain information on year of medical degree. Five health regions, defined by the ministry of health for planning the delivery of healthcare services, were used to identify patient residence and physician practice locations.

Internal validation was conducted for both the NL and MB models. Measures of prediction accuracy, which included bias, mean absolute error (MAE), and root mean square error (RMSE),[24] were calculated based on 10-fold cross-validation.[25;26]

Model Prediction

The final fitted model for NL was used to predict the number of prevalent diabetes cases per NFFS physician. However, given that not all NFFS physicians provide services to diabetes patients, we used the ratio of FFS physicians in the physician cohort to the total number of FFS physicians in the province[27] to select a random prediction sample. A similar process was used predict the number of cases from the MB data. In MB we also compared the predicted number of

diabetes cases for NFFS physicians to the observed number of cases from the shadow-billed claims of NFFS physicians.

The total number of prevalent diabetes cases in each province was estimated as the sum of: (a) observed cases ascertained from hospital discharge abstracts only, (b) observed cases ascertained from claims of FFS physicians, (c) predicted cases for NFFS physicians. Denominators of the prevalence estimates were based on 2001 Statistics Canada Census data; 95% confidence intervals were calculated using the binomial distribution.

All analyses were conducted using SAS version 9.3. Ethics approval was provided by the University of Manitoba Health Research Ethics Board and the NL Health Research Ethics Board. Data access approval was provided by the Newfoundland and Labrador Centre for Health Information and the Manitoba Health Information Privacy Committee.

RESULTS

Descriptive Analyses

A total of 31,714 prevalent diabetes cases were identified from the NL administrative data (Table 1); 91.1% (n = 28,989) of cases were identified from billing claims of physicians remunerated by FFS, while 1.3% (*n* = 411) of cases were ascertained from billing claims submitted by NFFS physicians who received a portion of their remuneration by FFS. Almost two-thirds (60.7%) of diabetes cases from FFS physician claims were residents of the Eastern health region, which contains the largest city in NL (200,000+ population); 40.5% were 65+ years.

In the MB external validation data, 51,031 prevalent diabetes cases were identified (Table 1), of which 84.1% were ascertained from the billing claims of FFS physicians. Three-quarters

(75.9%) of prevalent cases ascertained from the shadow-billed claims of NFFS physicians were from non-Winnipeg health regions.

Table 1. Characteristics of diabetes case cohort by ascertainment source and province

Case characteristics	Cases ascertained from hospital discharge abstracts		Cases ascertained from physician billing claims for FFS physicians		Cases ascertained from physician billing claims for NFFS physicians[a]	
	n	%	*n*	%	*n*	%
Newfoundland and Labrador (*N* = 31,714)						
Total	2,405	100.0	28,898	100.0	411	100.0
Sex						
Male	1,158	48.1	13,872	48.0	217	52.8
Female	1,247	51.9	15,026	51.9	194	47.2
Age group						
<35 years	39	1.6	1,448	5.0	30	7.3
35 - 49 years	168	7.0	4,932	17.1	84	20.4
50 - 64 years	570	23.7	10,808	37.4	136	33.1
65+ years	1,628	67.7	11,710	40.5	161	39.2
Health region of residence						
Eastern	1,201	49.9	17,547	60.7	110	26.8
Central	523	21.7	5,909	20.4	258	62.8
Western	389	16.2	4,840	16.7	7	1.7
Labrador	267	11.1	464	1.6	35	8.5
Missing	25	1.0	138	0.5	1	0.2
Manitoba (*N* = 51,031)						
Total	2,250	100.0	42,933	100.0	5,848	100.0
Sex						
Male	1,161	51.6	22,078	51.4	2,764	47.3
Female	1,089	48.4	20,855	48.6	3,084	52.7
Age group						
<35 years	71	3.2	1,952	4.6	375	6.4
35 - 49 years	236	10.5	7,636	17.8	1,358	23.2
50 - 64 years	534	23.7	15,319	35.7	2,120	36.3
65+ years	1,409	62.6	18,026	42.0	1,995	34.1
Health region of residence						
Winnipeg	1,180	62.6	25,949	60.4	1,409	24.1
Interlake-Eastern	262	11.6	4,503	10.5	970	16.6
Northern	189	8.4	1,951	4.5	1,562	26.7
Prairie Mountain	370	16.4	6,400	14.9	1,067	18.3
Southern	249	11.1	4,130	9.6	840	14.4

[a]These cases were ascertained from the claims of NFFS physicians receiving partial FFS remuneration in Newfoundland and Labrador, and from the claims of NFFS physicians who shadow bill in Manitoba.

Table 2. Characteristics of the physician cohort by method of remuneration and province

Physician characteristics	Newfoundland and Labrador (N = 388)			
	FFS (n = 362)		NFFS[a] (n = 26)	
	n	%	n	%
Specialty				
General practitioner	291	80.4	22	84.6
Specialist	71	19.6	4	15.4
Sex				
Male	257	70.9	19	73.1
Female	105	29.0	7	26.9
Age group				
< 40 years	85	23.5	15	57.7
40 – 64 years	269	74.3	11	42.3
65+ years	8	2.2	0	0.0
Health region of practice				
Eastern	258	71.3	6	23.1
Central	56	15.5	13	50.0
Western	42	11.6	3	11.5
Labrador	6	1.7	4	15.4
Medical degree, years[b]	22.5 (10.7)	22.0	15.0 (9.7)	14.0
Specialty, years[b]	17.2 (10.1)	17.0	6.8 (8.9)	3.5
	Manitoba (N = 1,229)			
	FFS (n = 989)		NFFS (n = 270)	
	n	%	n	%
Specialty				
General practitioner	770	77.9	--	--
Specialist	219	22.1	--	--
Sex				
Male	741	74.9	201	74.4
Female	248	25.1	69	25.6
Age group				
< 40 years	301	30.4	185	68.5
40 - 64 years	572	57.8	--	--
65+ years	116	11.8	--	--
Missing	0	0.0	0	0.0
Health region of practice				
Winnipeg	659	66.6	57	21.1
Interlake-Eastern	61	6.2	40	14.8
Northern	25	2.5	63	23.3
Prairie Mountain	152	15.4	62	23.0

Southern	92	9.3	48	17.8
Specialty, years[a]	12.1 (9.9)	10.0	5.2 (6.4)	3.0

[a]In Newfoundland and Labrador, NFFS physicians identified in claims data received partial FFS remuneration, while in Manitoba, NFFS physicians identified in claims data shadow bill.

[b]Reported values are mean (SD) and median; some cells cannot be reported, in accordance with Manitoba Health requirements, because of small numbers

There were 388 individuals in the NL physician cohort (Table 2). Amongst FFS physicians (93.3%), the majority were general practitioners (80.4%), and most were from the Eastern health region (71.3%). The MB physician cohort contained more than 1200 physicians, of which 80.4% were FFS physicians. Amongst these FFS physicians, more than half (57.8%) were in the 40-64 years age group. The NFFS physicians (n = 270) were primarily less than 40 years (68.5%) and almost 80.0% practiced outside of the urban Winnipeg health region.

Table 3 describes the mean and median number of prevalent diabetes cases per FFS physician. In NL, the average number of prevalent cases per FFS physician was 75.5 and the median was 49.0. The mean and median were higher for general practitioners than for specialists and also for males than females. For MB, the average number of prevalent diabetes cases per FFS physician was 43.4 and the median was 25.0.

Prediction Model

For NL, the main effects model provided a good fit to the data, as judged by the ratio of model deviance to degrees of freedom (ratio =1.0) and the AIC was smaller for a main effects model than for one with main and two-way interaction effects (3833.1 versus 3830.4); Likelihood ratio tests revealed statistically significant main effects for sex ($p < .0001$) and years since specialty ($p = .0006$).

Table 3. Mean (standard deviation) and median number of prevalent cases in the diabetes case cohort per FFS physician in the physician cohort

Physician characteristics	Mean (SD)	Median
	Newfoundland and Labrador	
Province	75.5 (84.6)	49.0
Specialty		
General practitioner	79.0 (66.2)	66.0
Specialist	61.0 (136.8)	9.0
Sex		
Male	89.3 (94.1)	75.0
Female	41.5 (37.2)	32.5
Age group		
< 40 years	54.9 (64.8)	32.5
40 – 64 years	99.9 (98.3)	91.0
65+ years	63.8 (68.6)	34.5
Health region of practice		
Eastern	67.8 (73.1)	42.0
Central	87.8 (87.7)	59.0
Western	108.1 (129.5)	86.0
Labrador	47.6 (47.9)	38.5
	Manitoba	
Province	43.4 (74.2)	25.0
Specialty		
General practitioner	45.1 (45.7)	35.0
Specialist	37.6 (132.8)	3.0
Sex		
Male	47.7 (76.0)	33.0
Female	30.5 (67.2)	17.0
Age group		
<40 years	25.5 (34.2)	14.5
40-64 years	52.1 (90.4)	34.0
65+ years	47.1 (49.8)	34.5
Health region of practice		
Interlake-Eastern	45.9 (37.8)	48.0
Northern	49.4 (59.4)	20.0
Prairie Mountain	42.4 (39.7)	35.0
Southern	35.1 (29.1)	28.0
Winnipeg	44.3 (86.7)	20.0

The regression analyses produced similar results in the MB external validation data; the ratio of model deviance to degrees of freedom was close to 1.0 for the main effects model. The

model with main and two-way interaction effects resulted in a negligible decrease in the AIC. The main effects of sex ($p < .0001$), speciality ($p = .0021$), and years since specialty licensure ($p < .0001$) were statistically significant.

With respect to the internal cross-validation, for the NL model absolute bias estimates ranged from 0.2% to 12.9% across the ten data folds, while for the MB model the estimates ranged from 0.6% to 13.8%. The MAE ranged from 40.1 to 67.5 for the NL model and from 26.7 to 43.2 for the MB model. Finally, the RMSE ranged from 56.5 to 131.2 for the NL model and from 33.8 to 151.0 for the MB model.

Table 4. Observed and predicted average number of diabetes cases per fee-for-service (FFS) and non-fee-for-service (NFFS) physician in Manitoba's physician cohort

	FFS		NFFS	
	Observed	**Predicted**	**Observed**	**Predicted**
Entire province	43.4	43.8	21.7	32.7
Health region of practice				
Interlake-Eastern	45.9	49.7	20.7	31.9
Northern	49.4	43.3	15.1	30.6
Prairie Mountain	42.4	44.0	16.0	39.4
Southern	35.1	36.0	17.1	21.8
Winnipeg	44.3	44.3	39.5	37.4

Using the MB model results, we compared the observed and expected number of prevalent diabetes cases per FFS and NFFS physician (Table 4) for the entire province and by health region of practice. The provincial and regional figures were similar for FFS physicians, supporting the internal validity of the model. For NFFS physicians, the expected number of cases was 51.0% higher than the observed number for the entire province. When we examined these values by health region, we found that the expected value was 8.2% lower than the observed value for the Winnipeg health region. However, for the remaining health regions the expected values were much higher than the observed values.

Figure 1 shows the percentage of diabetes cases ascertained from each data source in both provinces. In NL, the prediction model resulted in a 37.2% increase in the number of diabetes cases ascertained from the administrative databases, while in MB it resulted in a 16.3% increase. In NL, crude diabetes prevalence based on cases ascertained only from hospital data and FFS physician claims was 8.1%, while the estimate based on observed and expected cases was 13.0% (95% CI: 12.9, 13.0). In MB, the crude diabetes prevalence estimate based on cases ascertained from hospital data and FFS physician claims was 5.6%, while the estimate based on both observed and expected cases was 6.7% (95% CI: 6.7%, 6.8%).

DISCUSSION

This study developed a prediction model for linked administrative health databases to estimate the completeness of electronic physician claims data; the model was applied to estimate under-ascertainment of prevalent diabetes cases but could be applied to other chronic or acute conditions that are primarily managed or treated in non-acute care settings. When the model was applied to data from the Canadian province of NL, the results revealed that close to 40% of diabetes cases were missed because NFFS physicians do not report contacts with patients in

claims data. When the model was externally validated in MB, a province in which some NFFS physicians submit some claims, the modeling results indicated than less than 20% of diabetes cases were missed, but this percentage varied substantially by region; there was less bias in the Winnipeg health region, which contains the largest city in Manitoba, and more substantial bias in non-Winnipeg health regions where there is a higher proportion of NFFS physicians.

Data from the 2005 Canadian Community Health Survey,[28] a national survey used for regional chronic disease surveillance, revealed a crude diabetes prevalence of 6.8% for NL and 4.4% for MB for the population 12+ years, a difference of more than 50%. When we compared crude prevalence estimates for the two provinces using only FFS claims and hospital records, rates in NL were just 8.9% higher than those in MB. However, after adjustment for potential missed cases using our prediction model, crude prevalence was 45.1% higher in NL than in MB, producing a similar difference in estimates to those observed in survey data.

Incomplete capture of claims for NFFS physicians is similar to unit non-response in survey data, both of which can bias parameter estimates and increase variance estimates. Unit non-response in surveys is often difficult to adjust for, because information about non-responders is rarely available to the researcher. In fact, administrative data have been used in previous research to estimate the effect of survey non-response bias in estimates of health care use.[29] However, our study suggests that the use of administrative data for evaluating survey non-response should be adopted with caution, as administrative databases may themselves be incomplete.

While the proposed prediction model provides a useful tool to estimate bias in disease prevalence due to incomplete claims data, it is equally important to consider how other databases can be used to address gaps in these data. Electronic medical records are increasingly being

adopted in population-based chronic disease research and surveillance studies,[30] and could represent an important additional source of data for case ascertainment. Pharmacy databases have also been used for case ascertainment [31] when the medications used for disease treatment have high specificity for case capture.

Limitations of the study include the restricted set of explanatory variables available to develop the prediction model. Residual confounding due to factors such as physician productivity,[10] type of practice, and even characteristics of the patients seen by a physician may affect prediction accuracy.[32] Strengths of the study include the use of a validated case definition to ascertain diabetes cases and the internal and external validation process.

Further research could examine the validity of the prediction model by applying it to other chronic diseases and in other jurisdictions; [33] the utility of the model is not limited to Canadian administrative data, as a similar approach has been proposed to evaluate the completeness of cancer registry data.[16] Simulation could also be used to assess the impact of patient, physician, and health system characteristics on estimates of completeness.[34] For example, the model assumes that physician characteristics will have the same distribution and association with the number of prevalent diabetes cases in FFS and NFFS populations, which may not be a valid assumption.[35]

In summary, this study revealed that completeness of physician claims data are associated with method of physician remuneration and that a predictive model can be used to estimate the magnitude of data incompleteness for disease surveillance. This predictive model makes use of routinely collected linked data, and therefore is feasible to implement over time and across jurisdictions.

ACKNOWLEDGEMENTS

The authors are indebted to Manitoba Health, Healthy Living, and Seniors (HIPC 2012/2013-04) and the Newfoundland and Labrador Centre for Health Information for the provision of data. The results and conclusions are those of the authors, and no official endorsement by Manitoba Health, Healthy Living, and Seniors is intended or should be inferred.

FUNDING

This research was funded by the Canadian Institutes of Health Research (Funding Reference Number 123357). The first author is supported by a Research Chair from the Manitoba Health Research Council. The funders had no involvement in the conduct of the research or in manuscript preparation.

STATEMENT OF CONTRIBUTIONS

LML, GK, HQ, MS, and KS designed the analysis and acquired the study data. JPK, XY, NM, and WK conducted the analyses. LML, JPK, and NM drafted the manuscript and all remaining authors read and revised it substantially. All authors approved the final version of the manuscript before submission.

DATA SHARING STATEMENT

No data are available.

REFERENCES

(1) Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Annu Rev Public Health* 2001;22:213-230.

(2) Dombkowski KJ, Wasilevich EA, Lyon-Callo S, et al. Asthma surveillance using Medicaid administrative data: a call for a national framework. *J Public Health Manag Pract* 2009;15:485-493.

(3) Potter BK, Manuel D, Speechley KN, et al. Is there value in using physician billing claims along with other administrative health care data to document the burden of adolescent injury? An exploratory investigation with comparison to self-reports in Ontario, Canada. *BMC Health Serv Res* 2005;5:15.

(4) Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Med Care* 2006;44:1011-1019.

(5) Hux JE, Ivis F, Flintoft V, et al. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* 2002;25:512-516.

(6) Saydah SH, Geiss LS, Tierney E, et al. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Annals of Epidemiology* 2004;14:507-516.

(7) Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension* 2009;54:1423-1428.

(8) Tu K, Campbell NRC, Chen Z-L, et al. Accuracy of administrative databases in identifying patients with hypertension. *Open Medicine* 2007;1:E3-E5.

(9) Saez M, Barcelo MA, Coll de TG. A selection-bias free method to estimate the prevalence of hypertension from an administrative primary health care database in the Girona Health Region, Spain. *Comput Methods Programs Biomed* 2009;93:228-240.

(10) Wranik DW, Durier-Copp M. Physician remuneration methods for family physicians in Canada: Expected outcomes and lessons learned. *Health Care Anal* 2009;18:35-59.

(11) Alshammari AM, Hux JE. The impact of non-fee-for-service reimbursement on chronic disease surveillance using administrative data. *Can J Public Health* 2009;100:472-474.

(12) Crocetti E, Miccinesi G, Paci E, et al. An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy. *Eur J Cancer Prev* 2001;10:417-423.

(13) Dockerty JD, Becroft DM, Lewis ME, et al. The accuracy and completeness of childhood cancer registration in New Zealand. *Cancer Causes Control* 1997;8:857-864.

(14) Schouten LJ, Straatman H, Kiemeney LA, et al. The capture-recapture method for estimation of cancer registry completeness: a useful tool? *Int J Epidemiol* 1994;23:1111-1116.

(15) Brenner H, Stegmaier C, Ziegler H. Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany. *J Epidemiol Community Health* 1995;49:426-430.

(16) Das B, Clegg LX, Feuer EJ, et al. A new method to evaluate the completeness of case ascertainment by a cancer registry. *Cancer Causes Control* 2008;19:515-525.

(17) Newfoundland and Labrador Centre for Health Information. Enhancing chronic disease surveillance in Newfoundland and Labrador: adjustment of rates based on physician payment methods. Newfoundland and Labrador Centre for Health Information. 2010. St. John's, NL, Newfoundland and Labrador Centre for Health Information.

(18) Canadian Institute for Health Information. National physician database, 2008-2009. 2010. Ottawa, Canadian Institute for Health Information.

(19) Clottey C, Mo F, LeBrun B, et al. The development of the National Diabetes Surveillance System (NDSS) in Canada. *Chronic Dis Can* 2001;22:67-69.

(20) McCulloch CE, Searle SR. Generalized, Linear, and Mixed Models. New York: Wiley; 2001.

(21) Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 1987;52:345-370.

(22) Watson DE, Katz A, Reid RJ, et al. Family physician workloads and access to care in Winnipeg: 1991 to 2001. *CMAJ* 2004;171:339-342.

(23) Lix LM, Walker R, Quan H, et al. Features of physician billing claims databases in Canada. *Chron Dis Can* 2012;32:186-193.

(24) Dunn G, Mirandola M, Amaddeo F, et al. Describing, explaining or predicting mental health care costs: A guide to regression models - Methodological review. *Br J Psychiatry* 2003;183:398-404.

(25) Austin PC, Rothwell DM, Tu JV. A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. *Health Serv Outcomes Res Methodol* 2002;3:107-133.

(26) Kuwornu JP, Lix LM, Quail J, et al. A comparison of statistical models for analyzing episode-of-care costs for chronic obstructive pulmonary disease. *Health Serv Outcomes Res Methodol* 2013;13:203-208.

(27) Canadian Institute for Health Information. The status of alternate payment programs for physicians in Canada: 2002-2003 and preliminary information for 2003-2004. 2005. Ottawa, ON, Canadian Institute for Health Information.

(28) Sanmartin C, Gilmore J. Diabetes prevalence and care practices. *Health Rep* 2008;19:59-63.

(29) Gundgaard J, Ekholm O, Hansen EH, et al. The effect of non-response on estimates of health care utilisation: linking health surveys and registers. *Eur J Public Health* 2008;18:189-194.

(30) Desai JR, Wu P, Nichols GA, et al. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care* 2012;50 Suppl:S30-S35.

(31) Maio V, Yuen E, Rabinowitz C, et al. Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. *J Health Serv Res Policy* 2005;10:232-238.

(32) Hanley JA, Dendukuri N. Efficient sampling approaches to address confounding in database studies. *Stat Methods Med Res* 2009;18:81-105.

(33) Kleinberg S, Elhadad N. Lessons learned in replicating data-driven experiments in multiple medical systems and patient populations. *AMIA Annu Symp Proc* 2013; 2013:786-795.

(34) Silcocks PB, Robinson D. Simulation modelling to validate the flow method for estimating completeness of case ascertainment by cancer registries. *J Public Health (Oxf)* 2007;29:455-462.

(35) Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-483.

LIST OF FIGURES

STROBE Statement—checklist of items that should be included in reports of observational studies

NOTE: ALL ITEMS THAT HAVE BEEN ACHIEVED ARE HIGHLIGHTED IN YELLOW.

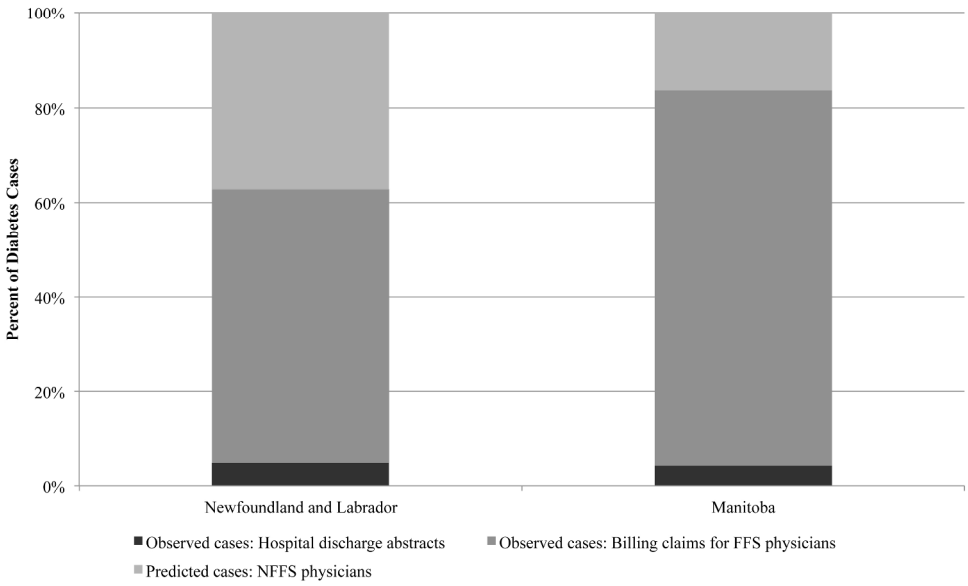
	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy NOT APPLICABLE
		(*e*) Describe any sensitivity analyses

Continued on next page

Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time
		Case-control study—Report numbers in each exposure category, or summary measures of exposure
		Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results	18	Summarise key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.



254x190mm (300 x 300 DPI)