Importance of accurately identifying chronic disease in studies using electronic health records

Douglas G Manuel,[1 2 3 4 5] Laura C Rosella,[4 5 6] Thérèse A Stukel[4 6]

[1]Ottawa Hospital Research Institute, 1053 Carling Ave, Ottawa, Ontario, Canada K1Y 4E9
[2]Statistics Canada, Ottawa
[3]Departments of Family Medicine and Epidemiology and Community Medicine, University of Ottawa
[4]Institute for Clinical Evaluative Sciences, Toronto, Ontario
[5]Dalla Lana School of Public Health, University of Toronto, Toronto
[6]Ontario Agency for Health Protection and Promotion, Toronto, Ontario
Correspondence to: D G Manuel dmanuel@ohri.ca
Accepted: 7 June 2010

Cite this as: *BMJ* 2010;341:c4226
doi: 10.1136/bmj.c4226

Use of routinely collected electronic health data to identify people for epidemiology studies and performance reports can lead to serious bias

Disease registries and similar databases have facilitated epidemiological studies that contribute to our understanding of the natural course of disease and the value of medical and surgical interventions.[1] These data have also allowed us to study the performance of health care, including patient safety and quality of care.[2 3] However, there is an increasing possibility of inaccurate results arising from a shift in the type of data used to identify people with chronic diseases. In the past, registries for cancer and other diseases were laboriously created using active reporting from individual clinical records. But increasingly, disease databases are now generated from routinely collected electronic data and applying a set of disease identification criteria en masse. For uncommon diseases, small errors in classifying people can result in a large number of incorrect entries in a database, leading to biased results and classification errors that propagate through calculations in ways that are difficult to intuitively appreciate.

How disease classification errors affect study conclusions

Routinely collected electronic data are increasingly used to identify patients with chronic diseases such as diabetes, heart disease, cancer, and arthritis, for research.[1 4] Databases that contain information on patients with a wide range of diseases are even more widely used. The United Kingdom's General Practice Research Database, for example, has been used for more than 700 studies of over 150 conditions (see table 1 on bmj.com),[5 6] and hospital discharge databases are widely used in many countries for research and performance studies.

Table 2 | Validation of sensitivity and specificity of identifying diabetes in the Ontario Diabetes Database against primary care records[10]

Ontario Diabetes Database	Primary care records		
	Diabetes	No diabetes	Total
Diabetes	335	85	420
No diabetes	54	2843	2897
Total	389	2928	3317

Specificity=97.1%
Sensitivity=86.1%
Positive predictive value=79.8%
Negative predictive value=98.1%
Diabetes prevalence from primary care records=11.7%
Diabetes prevalence from database=12.7%

SUMMARY POINTS

Routinely collected electronic health data are increasingly used to identify people with chronic conditions for research
Classification error can occur during the disease identification process
Even when the identification process has very good sensitivity and specificity, misclassification can considerably bias study findings
Studies using routinely collected data should assess the potential for classification error and adjust for bias

However, few of these studies assess whether their findings may be biased by misclassification of patients in the database. We believe that the conclusions of many studies may change if their results were adjusted for bias or if there were no misclassification errors.

To illustrate our case, we estimated the potential bias in two published studies that use the Ontario Diabetes Database.[7 8] The concerns about misclassification error are described in other areas of health care, such as diagnostic accuracy studies, where methods to reduce error and reporting guidelines to disclose potential bias have been developed.[9] We applied the same principles and methods to examine bias in the use of routinely collected data to identify disease.

Estimating bias

The Ontario Diabetes Database is a well developed database generated using only routinely collected administrative data. Both studies that we examined generated study populations directly from this database, and both studies quoted a separate development study as validation that disease identification in the database was high quality (table 2).[10]

We calculated the potential percentage of misclassified people in the two study samples using a straightforward correction method described in different epidemiology settings (see bmj.com).[11-13] Like several other approaches to assess misclassification and bias, this method centres around estimating the predictive accuracy of disease identification in terms of "false positives"—people who do not have a disease and are incorrectly enrolled in the disease database or study—and "false negatives"—people who do have the disease but are missing from the disease database.[14-16] The amount that a study is biased can be estimated after misclassification is described. For example, performance studies for diabetes care will report the proportion of patients who receive care as recommended in clinical

Table 3 | Estimates of misclassified respondents in two published studies using sensitivity and specificity from development study

	Reported prevalence of diabetes (%)	Study base	True positives	False positives	True negatives	False negatives	Positive predictive value (%)	Negative predictive value (%)
Development study[10]	11.7	3 317	335	85	2 843	54	79.8	98.1
Study of haemoglobin A_{1c} testing coverage[7]	6.9	923 174	38 186	25 513	853 319	6 155	59.9	99.2
Study of trend in diabetes prevalence[8]	8.9	9 276 945	577 579	249 840	8 356 424	93 102	69.8	98.9

The sensitivity and specificity from the development study were used to estimate the true and false positives and true and false negatives in the example studies, using the prevalence of diabetes in the example studies. See bmj.com for detailed calculations.

practice guidelines (such as regular haemoglobin A_{1c} testing). Incorrectly including people without diabetes in a study of diabetes care will bias performance towards poor care because people who do not have diabetes do not need to have regular testing. In this way, false positives and classification error will almost always bias performance reporting towards poor care.

Table 3 shows the findings from the validation study[10] for the Ontario Diabetes Database and our estimate of false positives and false negatives in the study populations for each of two examples.

The first study reported an annual rate of haemoglobin A_{1c} tests and concluded that the level of testing was unacceptably low in 2005.[7] The study reported that 58% of 36 945 patients with physician diagnosed diabetes received a haemoglobin A_{1c} test. These results have been widely cited. The Health Council of Canada, for one, used these and other findings to conclude that care for people with diabetes in Canada is possibly the worst of any country in the Organisation for Economic Cooperation and Development.[17] Applying a sensitivity of 86.1% and specificity of 97.1% from the database validation study, we estimate that 38 186 of 63 699 participants were correctly classified as having diabetes (positive predictive value 59.9%). The remaining 25 513 patients were false positives, misclassified as having diabetes and not in need of regular haemoglobin A_{1c} testing. Using this information (see bmj.com), we calculated an unbiased estimate of haemoglobin A_{1c} testing among diabetes patients of 97% (36 945/38 186).

The second study reported trends in the incidence and prevalence of diabetes[8] and concluded that more adult Ontarians were diagnosed with diabetes in 2005 (8.9% or 827 419 people) than the global rate predicted for 2030.[18] The Ontario government extrapolated findings from the study to state that diabetes prevalence will increase by an additional 30% by 2010.[19] This prevalence calculation is widely quoted and is being used to support a considerably expanded diabetes strategy.[19] However, applying the database validation study, we estimate that the unbiased prevalence of diabetes in 2005 was 19% lower than the original study found (7.2% versus 8.9%). Of the 827 419 people enrolled in the Ontario Diabetes Database, we calculated that 249 840 were wrongly classified as having been diagnosed with diabetes (positive predictive value 69.8%), and that 93 102 people had diabetes diagnosed by their physician but were not enrolled in the database (false negatives).

Steps for creating and using a disease database when disease status is not individually verified

- *Development studies*—Develop disease identification criteria by assessing the identification (or diagnostic) accuracy of different ascertainment approaches or algorithms against a reference standard of people with individually verified disease status
- *Create disease database*—Systematically apply the case identification criteria to an entire population's health data. Enrol people in the database if they satisfy case identification criteria. Regularly update the process when new data become available. Assign an enrolment (incident) date. Studies may not formally create a disease database; instead, the disease identification criteria are applied to the health data of all study participants
- *Assess for bias due to classification error*—For each use of the disease databases or disease identification algorithm, assess the potential for misclassification to bias the study results. Estimate the number of people who may be false positives and false negatives, and examine how this affects the study results
- *Validation studies*—(Re)validate the disease identification criteria in new study populations with a reference standard

Why does this problem happen?

It is important to recognise a subtle but critical distinction between disease databases that individually verify diagnoses from those that do not. It is one matter to identify patients with a positive confirmation test such as a cancer pathology report, manually verify the report, and then use this information to create a disease registry. It is another matter to access an entire population's electronic records and apply identification criteria to automatically classify people who have a disease and exclude those who do not. Routinely collected electronic data offer the advantage of identifying many diseases in large populations at low cost. However, mass application of identification criteria is more prone to error than the traditional, more expensive, approach of individually or manually verifying disease diagnoses for each person.

When individual verification is not done, disease databases should at least attempt to gauge the accuracy of the identification process in a representative sample. Unfortunately, this step is commonly omitted. Instead, diagnoses are identified using the corresponding codes within health services data such as international classification of disease (ICD) codes from hospital admission discharge summaries or Read codes from primary care data.[20 21] This approach assumes that the diseases are accurately and completely recorded in the databases, which in turn assumes that well implemented quality control procedures are in place at the point of data entry.[21]

The purpose of development and validation studies is to test these assumptions. These studies run different identification algorithms against a reference population whose disease status has been individually validated (box). Identification algorithms are constructed and tested using various diagnosis codes along with procedures and services in different combinations and intensities. Identification algorithms are then compared using tests of discrimination (sensitivity, specificity, likelihood ratios) and predictive accuracy (positive and negative predictive values).[22] Other approaches for developing and validating identification methods are available.[23-25] Because studies of identification accuracy (assessing the accuracy of tests to identify people already diagnosed with a disease) are similar to those studying diagnostic accuracy

(assessing the accuracy of tests to diagnose people who may have a disease), the approaches to development, validation, and reporting are largely applicable to both types of studies.[9]

Errors can also occur when information is abstracted from a database for a study. With increased computing power and wider availability of health data it is straightforward to apply identification criteria to an entire population, including populations beyond those represented in a development study (if one exists). Rather than formally creating a database for a specific study, it is common simply to apply the identification criteria to create a study population. For example, a hospital may assess its performance by examining the quality of care for people with an acute myocardial infarction in terms of time to thrombolytic therapy or 30 day survival by identifying people with a discharge diagnosis coded for acute myocardial infarction.[26] However, if the ICD-9 code for myocardial infarction is incorrectly used, the quality measure may be biased.

Studies using electronically collected data can be grouped into three types:

- Study denominator is drawn from the database—for example, examining healthcare performance for people with a particular condition
- Study base is entire population and the numerator is people with a disease—for example, examining the incidence and prevalence of a disease in a population
- Outcome of interest is people who develop a condition—for example, study of drug side effects such as admission for hyperkalaemia (identified from hospital discharge data) in patients prescribed spironolactone.[27]

Classification errors will potentially bias different study types in different ways. Performance reports are biased only from false positive entries, whereas estimates of disease incidence are affected by both false positive and false negative entries.

Assessment of bias

Bias from misclassification should be assessed for each use of data from electronic health record systems. Unless a diagnosis is individually verified there will inevitably be some classification error, and the resulting bias is difficult to intuitively gauge because both the amount and direction of bias are affected by the study design and by various properties of disease identification including prevalence, sensitivity, and specificity. The amount of bias may be large, even when the disease identification criteria seem to be accurate or there are well instituted data quality control procedures.

There are two general approaches that are used to estimate bias. The first approach applies the level of identification accuracy from development studies to a new study. We used this approach when we estimated bias in the two published diabetes studies. The second approach validates the identification in a new study, correcting for bias as needed.

Calculating bias is not always straightforward. First, development or validation studies are required, and they should report sensitivity and specificity or similar measures of disease identification accuracy. Many studies have not validated their method of disease identification. For example, more than two thirds of peer reviewed studies using the General Practice Research Database did not perform a validation study and most of those that did calculated only specificity and positive predictive values.[5] Even well performed validation studies carry generalisability concerns. In our examples, the validation study used a diagnosis of diabetes in general practice records as the reference standard. This reference standard is imperfect because, among other reasons, some patients may not have had their diagnosis in their general practice recorded because their diabetes was diagnosed and cared for exclusively by specialists. Furthermore, it may be inappropriate to assume that sensitivity and specificity from a validation study hold firm for studies with different population characteristics. Methods are available to overcome these concerns, including performing sensitivity testing using different reference standards or levels of identification accuracy (calculating bias by varying sensitivity and specificity).[9] We recommend the development and use of multi-attribute identification algorithms to estimate the probability of disease diagnoses (value of 0 to 1), rather than assigning disease status to a person (value of 0 or 1).[28]

Conclusion

As our two examples show, even in well performed studies with well developed identification criteria, there is considerable opportunity for misclassification to bias results—so much so that studies can arrive at incorrect conclusions. Most of the time, it is straightforward to calculate the amount of potential bias and adjust the findings accordingly. Our findings are applicable beyond diabetes, particularly when disease prevalence is below about 10% and the specificity of identification is less than perfect (say, less than 98%).

The problem is further magnified because once a disease database is generated, many different investigators may use it for a wide range of studies or reports, propagating classification errors in their wake. However, data users cannot estimate bias when the accuracy of identification is unknown, and people who generate the databases or apply identification algorithms to routinely collected data should clearly describe the accuracy of their classification process. Researchers using such data should also publish an estimate of the percentage of false positives and negatives and the effect of misclassified people on the study's findings. Readers of reports can reasonably ask if classification error potentially challenges the studies' findings, and they should expect to see calculations that estimate the amount of bias.

It would be wrong to conclude that routinely collected data are poorly suited to study people with chronic conditions. Routinely collected data are improving and increasingly include more clinical information that can be used to individually verify disease or develop more accurate identification algorithms. Nevertheless, careful development and validation can help ensure that disease identification is accurate, bias can be measured, and results accordingly adjusted.

Contributors: All authors contributed to the development of the paper. DGM led the writing. LCR and TAS provided edits. DGM is the guarantor of the paper and analyses.

Competing interests: All authors have completed the unified competing interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare support for this article was provided by the Population Health Improvement Research Network; DGM holds a chair in applied public health from the Canadian Institute for Health Research and the Public Health Agency of Canada; no other relationships or activities that could appear to have influenced the submitted work. The opinions, results, and conclusions reported in this paper are not necessarily those of the funding or employment sources.

bmj.com archive

Previous articles in this series

- Taking healthcare interventions from trial to practice (*BMJ* 2010;341:c3852)
- Random measurement error and regression dilution bias (*BMJ* 2010;340:c2289)
- An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials (*BMJ* 2010;340:c2073)
- Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses (*BMJ* 2010;340:c117)

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Newton J, Garnes H. Disease registers in England. Institute of Health Sciences, University of Oxford, 2002.
2 Hurtado M, Swift E, Corrigan J. Envisioning the national health care quality report. National Academy Press, 2001.
3 Powell AE, Davies HT, Thomson RG. Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *Qual Saf Health Care* 2003;12:122-8.
4 Joshy G, Simmons D. Diabetes information systems: a rapidly emerging support for diabetes surveillance and care. *Diabetes Technol Ther* 2006;8:587-97.
5 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010;69:4-14.
6 General Practice Research Database. www.gprd.com.
7 Woodward G, van Walraven C, Hux JE. Utilization and outcomes of HbA1c testing: a population-based study. *CMAJ* 2006;174:327-9.
8 Lipscombe LL, Hux JE. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet* 2007;369:750-6.
9 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138:w1-12.
10 Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* 2002;25:512-6.
11 Kelsey J, Thompson W, Evans A. Methods in observational epidemiology. 2nd ed. Oxford University Press, 1996.
12 Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978;107:71-6.
13 Couris CM, Polazzi S, Olive F, Remontet L, Bossard N, Gomez F, et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. *J Clin Epidemiol* 2009;62:660-6.
14 Lix L, Yogendran M, Burchill C, Metge C, McKeen N, Moore D, et al. Defining and validating chronic diseases: an administrative data approach. Manitoba Centre for Health Policy, 2006.
15 Mullooly JP. Misclassification model for person-time analysis of automated medical care databases. *Am J Epidemiol* 1996;144:782-92.
16 White E. The effect of misclassification of disease status in follow-up-studies—implications for selecting disease classification criteria. *Am J Epidemiol* 1986;124:816-25.
17 Health Council of Canada. Why health care renewal matters: lessons from diabetes. 2007. www.healthcouncilcanada.ca/docs/rpts/2007/HCC_DiabetesRpt.pdf.
18 Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004;27:1047-53.
19 Ontario Ministry of Health and Longterm Care. Diabetes strategy—backgrounder. Ontario Ministry of Health and Longterm Care, 2008.
20 World Health Organization. International statistical classification of diseases and related health problems: 10th revision. WHO, 1992.
21 Schulz EB, Barrett JW, Price C. Read code quality assurance. *J Am Med Inform Assoc* 1998;5:337-46.
22 Habbema JDF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. BMJ Publishing, 2002:117-44.
23 Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. *J Clin Epidemiol* 2008;61:1250-60.
24 Sturmer T, Thurigen D, Spiegelman D, Blettner M, Brenner H. The performance of methods for correcting measurement error in case-control studies. *Epidemiology* 2002;13:507-16.
25 Mower WR. Evaluating bias and variability in diagnostic test reports. *Ann Emerg Med* 1999;33:85-91.
26 Tu JV, Khalid L, Donovan LR, Ko DT. Indicators of quality of care for patients with acute myocardial infarction. *CMAJ* 2008;179:909-15.
27 Juurlink DN, Mamdani MM, Lee DS, Kopp A, Austin PC, Laupacis A, et al. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. *N Engl J Med* 2004;351:543-51.
28 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;63:1000-10.

CORRECTIONS AND CLARIFICATIONS

Election views

In this Feature article on contributors' views about key election issues for the NHS (*BMJ* 2010;340:c2095, print publication 24 April, pp 894-7), the biography of Angela Coulter stated that she was the chief executive of Picker Institute Europe. This was true from 2000 to 2008, but she is now an independent healthcare analyst/consultant.

Is underdiagnosis the main pitfall when diagnosing bipolar disorder? Yes

One of the authors of this Head to Head article by Daniel J Smith and Nassir Ghaemi supporting the proposal that bipolar disorder is underdiagnosed (*BMJ* 2010;340:c854, print publication 27 March pp 686-7) has told us that he should have declared a competing interest. Nassir Ghaemi currently has a research grant from Pfizer.

Spanish doctors carry out first transplantation of a full face

Some errors occurred in this News article by Lynn Eaton (*BMJ* 2010;340:c2303, print publication). Firstly, we wrongly referred to the mandible in the plural. We did not clarify that in the first facial transplantation (in 2005), Jean-Michel Dubernard, the head of the transplant team, assisted Professor Bernard Devauchelle. We also should have referred to the infraorbital nerve (not the intraorbital optic nerve) in the discussion of nerve repair.

Using the new UK-WHO growth charts

The authorship of this Practice pointer by Charlotte Wright and colleagues (*BMJ* 2010;340:c1140, print publication 20 March, pp 647-50) was incomplete. The author list should have concluded: "on behalf of a Working Group of the Science and Research Department of the Royal College of Paediatrics and Child Health.

WHO and the pandemic flu "conspiracies"

In this feature article by Deborah Cohen and Philip Carter (*BMJ* 2010;340:c2912, print publication 12 June, pp 1274-9), we misspelt Barbara Mintzes' first name (p 1275).

Effect of influenza vaccination on excess deaths occurring during periods of high circulation of influenza: cohort study in elderly people

The authors of this 2004 paper (Ben G Armstrong and colleagues) have alerted us to an error in their paper (*BMJ* 2004;329:660, print publication 18 September 2004). They say that the expression they used for estimating vaccine efficacy (VE) was only approximately consistent and that a consistent estimator is $VE = (RR_u - RR_v)/(RR_u - 1)$ where RR_u and RR_v are the ratios of outcome rates in the flu period versus the non-flu period in vaccinated and unvaccinated people respectively. The authors state that re-estimating the vaccine efficacy for all cause mortality reported in the published paper changes the estimate little: from 83% (95% confidence interval 9% to 100%) to 85% (13% to 100%) for all cause mortality; from 80% to 83% for death from cardiovascular disease; and from 79% to 83% for deaths from respiratory disease. The latter two sets of 95% confidence intervals were unchanged, spanning the entire meaningful range (0 to 100).

The authors say that, although the revision made only a small difference in their data, it could be more important in other data. Specifically, the originally published estimator ($VE_{2010} = RR_v \times VE_{2004}$) is biased by a factor of RR_v^{-1}.

A derivation of the new estimator and its relation to the old one is available from ben.armstrong@lshtm.ac.uk.