Published in final edited form as:
J Clin Epidemiol. 2012 March ; 65(3): 343–349.e2. doi:10.1016/j.jclinepi.2011.09.002.

Trade-offs between accuracy measures for electronic healthcare data algorithms

Jessica Chubak[1,2], **Gaia Pocobelli**[1,2], and **Noel S. Weiss**[2,3]

[1]Group Health Research Institute, Group Health, Seattle, WA

[2]Department of Epidemiology, University of Washington, Seattle, WA

[3]Fred Hutchinson Cancer Research Center, Public Health Sciences Division, Seattle, WA

Abstract

Objective—We review uses of electronic healthcare data algorithms, measures of their accuracy, and reasons for prioritizing one measure of accuracy over another.

Study design and setting—We use real studies to illustrate the variety of uses of automated healthcare data in epidemiologic and health services research. Hypothetical examples show the impact of different types of misclassification when algorithms are used to ascertain exposure and outcome.

Results—High algorithm sensitivity is important for reducing the costs and burdens associated with the use of a more accurate measurement tool, for enhancing study inclusiveness, and for ascertaining common exposures. High specificity is important for classifying outcomes. High positive predictive value is important for identifying a cohort of persons with a condition of interest but that need not be representative of or include everyone with that condition. Finally, a high negative predictive value is important for reducing the likelihood that study subjects have an exclusionary condition.

Conclusion—Epidemiologists must often prioritize one measure of accuracy over another when generating an algorithm for use in their study. We recommend researchers publish all tested algorithms—including those without acceptable accuracy levels—to help future studies refine and apply algorithms that are well-suited to their objectives.

Keywords

algorithms; bias; databases; factual; epidemiology; medical records systems; computerized; misclassification

INTRODUCTION

Electronic healthcare data (e.g., Medicare claims, automated data from health plans) can be used to address epidemiologic questions in large populations in real-world settings. Algorithms based on these data allow epidemiologists to classify persons according to an

CORRESPONDING AUTHOR: Jessica Chubak, PhD, Group Health Research Institute, 1730 Minor Avenue, Ste. 1600, Seattle, WA 98101, PHONE: 206-287-2556, FAX: 206-287-2871, chubak.j@ghc.org.

exposure (e.g., preexisting dementia), outcome (e.g., disease-free survival), eligibility factor (e.g., absence of immunosuppresion), or covariate (e.g., a comorbidity) (Table 1). Electronic healthcare data have been used in studies on a wide variety of health conditions (e.g., infectious disease, cancer, diabetes), and for a variety of study designs (i.e., retrospective studies, prospective studies, and surveillance).

However, because electronic healthcare data can be incomplete or inaccurate, misclassification of the variable defined by the algorithm may occur.(1) The circumstances under which electronic data could be incomplete or inaccurate, include: 1) when patients do not seek care for a condition,(2, 3) or when they are treated outside of an integrated healthcare delivery system or insurance plan;(2–6) 2) when physicians do not accurately or consistently code procedures or diagnoses;(3–8) 3) when available codes do not adequately describe the procedure or condition,(2, 6, 7) or when too many diagnoses are present for all to be coded;(6) 4) when a health plan (e.g., Medicare) does not cover a particular procedure; (4, 9) and 5) when the variable of interest is not measured well by automated data (e.g. functional status),(2–5, 10–12) tends to be missing altogether (e.g. exercise), (2, 3, 5, 10) or tends to be missing differentially by exposure or disease status (e.g. smoking).(13) Recognizing that an algorithm based on electronic healthcare data will not be completely accurate, researchers must often prioritize one measure of algorithm accuracy (sensitivity, specificity, positive predictive value, or negative predictive) over another. Herein we review uses of electronic healthcare data algorithms, measures of their accuracy, and reasons for prioritizing one measure of accuracy over another based on the goals of the analysis. Addressing the reasons for prioritizing one accuracy measure over another in subsequent algorithm development and validation studies would enhance the current effort (14) to improve reporting in such studies.

USES OF ELECTRONIC HEALTHCARE DATA ALGORITHMS

An algorithm is “a completely defined set of operations that will produce a desired outcome.”(15) The goal of using electronic data algorithms for epidemiologic and health services research is to correctly classify a characteristic or condition. At its simplest, such an algorithm is a single criterion (such as a procedure or diagnosis code) chosen by the researcher to identify a characteristic. The algorithm classifies anyone in the study population whose record contains the appropriate code as having the characteristic as of the date associated with the code; subjects with no record of the code are classified as not having the characteristic during the window of time in which the code could have been assigned to the individual. More complex algorithms may consider combinations of procedure and diagnostic codes, the timing of codes (e.g., the frequency with which one or more codes appears over a given period of time), and code order (e.g., the sequence of two or more procedure codes). A detailed discussion of methods used to develop electronic healthcare data algorithms is beyond the scope of this paper, but descriptions can be found in studies that have used such algorithms.(16–18)

Electronic healthcare data algorithms can ascertain different types of information for use in epidemiologic studies, including information on exposures, outcomes, inclusion and exclusion criteria, and covariates. Epidemiologists can use this information in different ways, ranging from identifying persons for further contact or chart review, to relying on the classification without further validation. Study designs that use algorithms include retrospective assessments (case-control or cohort), real-time surveillance, and prospective studies (including randomized trials). Table 1 presents examples of how electronic healthcare data are used in epidemiologic studies.

MEASURES OF ALGORITHM ACCURACY

Relationship among accuracy measures

Standard epidemiologic measures including sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), describe the accuracy of algorithms (Appendix for Table 2). Algorithm accuracy is usually measured relative to data sources such as patient medical charts or patient surveys that are presumed to be a gold standard. Algorithm sensitivity is computed only among study subjects with the characteristic, and specificity is computed among only those without the characteristic. Sensitivity and specificity do not depend on the prevalence of the characteristic in the study population, but they can vary across populations.(19) Both PPV and NPV depend on sensitivity, specificity, and prevalence. For conditions that are present in a minority of the study population, specificity has a greater impact than sensitivity on PPV; the reverse is true for conditions that are present in the majority of the study population. In algorithm development, there is often a tradeoff between sensitivity and specificity: increasing an algorithm's sensitivity can decrease its specificity. For example, in developing an algorithm to identify breast cancer recurrences, the sensitivity for finding recurrences can be increased by including an International Classification of Diseases (ICD)-9 diagnostic code for a primary breast cancer, such as 174.9 (malignant neoplasm of female breast, unspecified). However, this decreases the algorithm's specificity because some women with only a primary breast cancer will be falsely classified as having had a recurrence. Several studies have demonstrated that varying algorithm inputs in this way can greatly affect algorithm properties.(20–24) Thus, when developing an algorithm, epidemiologists must often weigh the relative importance of sensitivity, specificity, PPV, and NPV, and prioritize the accuracy measure that is most important to a particular study.

Prioritizing different accuracy measures

The relative importance of different measures of accuracy (i.e., specificity, specificity, PPV, and NPV) depends on the intended use of the algorithm.(20, 21, 25–27) Misclassification can lead to reduced power,(28, 29) loss of generalizability,(25) as well as increased bias,(28, 29) patient burden,(29) and study cost.(29) The relative impact of each of these depends on the study. Below we discuss several scenarios in which prioritizing sensitivity, specificity, PPV, or NPV might be important. Increasing sensitivity can compromise specificity and vice versa, and both affect PPV and NPV. Scenarios that require maximizing one accuracy measure *entirely* at the expense of the other are probably rare, but there are situations in which one measure may be more important than another.

When is algorithm sensitivity important?

Prioritizing sensitivity of an algorithm over specificity is important when the goal is identifying all persons with a given characteristic in a population. In other words, sensitivity is the primary consideration when the benefits of identifying more true positives outweigh the negative consequences of including more false positives. This may be important when the goal is: 1) reducing study costs and burdens that will be incurred from using a more accurate measurement tool; 2) enhancing the inclusiveness of an algorithm; or 3) collecting information on a common exposure.

Reducing study costs that result from using a more accurate measurement tool—In studies where additional verification or data collection with a more accurate tool is possible, prioritizing sensitivity over specificity may be preferable. For example, in a study of care processes after myocardial infarction, patients were identified based on diagnostic and procedure codes, then medical chart review was done to collect information on symptoms and other detailed clinical data (30) (Table 1). A study of breast cancer recurrence

could substantially reduce its costs by using an electronic healthcare data algorithm to identify women likely to have had a breast cancer recurrence and then use medical chart review to identify false positives (i.e., women who did not have a breast cancer recurrence but were classified by the algorithm as having had one.) An algorithm with modest specificity but high sensitivity could dramatically reduce the number of charts to be abstracted. For example, assuming recurrence in 150 subjects (15%) in a cohort of 1000 women with breast cancer, an algorithm that identified recurrences with only 60% specificity (and 100% sensitivity) would reduce the number of charts to be abstracted by about half compared to abstracting charts of all women in the cohort: abstraction would occur on only 490 women (150 true positives plus 40% [% false positives] of the 850 women without recurrence).

Another example comes from surveillance studies that monitor for adverse events. In these studies where the priority is not missing a single case and when confirmatory analysis is intended, an algorithm with high sensitivity is desirable. Nordstrom et al. developed an algorithm to identify hypersensitivity reactions to abacavir (an antiretroviral used to treat human immunodeficiency virus), to be used when monitoring claims data as part of adverse event surveillance.(31) They proposed that their algorithm could provide a timely, initial indication of an adverse event to be confirmed with supplemental information.

Similarly, high sensitivity is desirable for studies that plan to recruit patients with a particular condition who will be further screened by telephone interview or mailed questionnaire, as suggested by Warren et al. for studies that plan to survey breast cancer patients.(27) Gary et al. used this approach to identify participants for a randomized controlled trial of a management intervention for type 2 diabetes (32) (Table 1).

Enhancing study inclusiveness—Another scenario in which identifying all cases is important is a study that assesses the full range of disease outcomes rather than only the most severe. For studies relying on claims data only, Winkelmayer et al. argue that highly sensitive algorithms are important for generalizability of results, particularly if less sensitive algorithms are differentially sensitive to different disease characteristics.(25) For example, in a study of treatment effectiveness for depression, an algorithm that is more sensitive for severe depression than mild depression may fail to detect the benefit of treatment strategies that work for mild but not severe depression.

Identifying a common exposure—Using algorithms to classify exposure status—without additional data collection for verification—is common, particularly in pharmacoepidemiology studies that use electronic pharmacy data to classify subjects' medication use (for example, Chen et al.'s study of antidepressant use and risk of hemorrhagic stroke (33)) (Table 1). In this situation, lack of sensitivity in identifying a common exposure can cause bias. For example, in a cohort study where there is non-differential misclassification of the exposure but ascertainment of the outcome is perfect, bias due to low sensitivity will increase as the exposure becomes more prevalent. This occurs because the proportion of the exposed study population that is misclassified as unexposed *increases* (Appendix for Table 3). In contrast, bias due to low specificity decreases as an exposure becomes more common because the proportion of the study population without exposure who are misclassified as exposed *decreases* (Appendix for Table 3). The overall incidence of disease does not independently affect percent bias however, bias increases as the true relative risk is further from the null (Appendix for Table 3). The bias in the odds ratio is similar to the bias in the relative risk in the above examples.

When is algorithm specificity important?

Imperfect sensitivity of an algorithm that classifies outcomes will not bias the relative risk, provided that the misclassification is non-differential with respect to exposure status and specificity is perfect. The same proportion of subjects are removed from the numerator of the rate in the unexposed and exposed groups, and the denominator is unchanged, so when comparing exposed to non-exposed subjects, the *ratio* of the observed incidence of the outcome will be the same as if the sensitivity were perfect.(19) There will however, be a very small amount of bias in the odds ratio, which will increase as the outcome becomes more common.

Imperfect specificity in classifying the outcome will, however, bias the relative risk even if sensitivity is perfect (Appendix for Table 4). In their hypothetical study of medication use and risk of lymphoma, Setoguchi et al demonstrate that bias increases with decreasing specificity.(24) The proportion of subjects added to the numerator of the rates in the exposed and unexposed groups will not be the same, because a fixed proportion of each non-diseased group is added to the diseased groups resulting in a different proportional change in the diseased group (i.e., the numerator of the rates). Therefore, prioritizing specificity, even at some cost to sensitivity, is important in studies that use algorithms rather than chart review for identifying outcomes.

Once specificity is prioritized, sensitivity remains important in one respect: at a given level of specificity, bias increases as sensitivity decreases (Appendix for Table 4). As sensitivity decreases, the size of the numerator of the rates decreases, and a given addition to that numerator (due to incomplete specificity) will have a greater impact (larger percent change in the numerator). Also of note, as the outcome becomes increasingly common, imperfect specificity has less of an effect on the relative risk. Additionally, bias increases as the true relative risk becomes further from the null (Appendix for Table 4).

When is algorithm PPV important?

The primary means by which a researcher can influence algorithm PPV and NPV is by modifying sensitivity and specificity, so PPV and NPV cannot be completely disentangled from these measurements. Prevalence, which also influences PPV and NPV, cannot be modified, although the researcher may choose to apply the algorithm to a population with a high prevalence of the condition to increase the PPV of the algorithm. Conversely, selecting a population with a low prevalence of the condition increases an algorithm's NPV.

In some studies, one may want to ensure that the algorithm's PPV – and not just its specificity – is high. PPV is important when identifying a cohort defined by disease status to ensure that only persons who truly have the condition of interest are included in the study. For example, in developing an algorithm to identify persons with a relapse of acute myelogenous leukemia, Earle et al. prioritized PPV to ensure that all patients identified were receiving treatment for the relapse and not for the initial cancer.(17) Similarly, Nattinger et al. developed an algorithm with high PPV to identify women with incident breast cancer to be used when conducting patterns-of-care and survivorship studies.(16) Winkelmayer et al. developed several algorithms to identify chronic kidney disease using Medicare claims data, and recommended prioritizing PPV when the goal is to identify a cohort with this condition. (25) However, an algorithm with a high PPV may not identify all persons with a condition (i.e., there may be false negatives). Therefore, prioritizing PPV is appropriate in studies where the cohort must be limited to persons with a particular condition but need not include or be representative of all persons with the condition of interest.

In the above scenarios, high specificity is important for ensuring high PPV. However, high specificity alone is not sufficient if the overall prevalence of disease is very low because a

relatively large absolute number of persons without the condition will be misclassified as having it, even though the proportion misclassified is small.(19) This demonstrates that effective algorithm use may require selecting an appropriate population in which to apply the algorithm.

When is algorithm NPV important?

NPV is an important consideration for algorithms used to identify subjects to include in a study. Many studies seek to exclude subjects with a history of another illness. For example, a study of the relationship between medication use and risk of non-Hodgkin's lymphoma may exclude people with a history of autoimmune diseases whose inclusion would introduce confounding because they may be more likely to both take certain medications and be diagnosed with non-Hodgkin's lymphoma. Beiderbeck et al. used ICD-9 codes to identify and exclude persons with a history of cancer or human immunodeficiency virus-related illness from their case-control study of medication risk factors for non-Hodgkin's lymphoma (34) (Table 1). Similarly, a study of incident fall risk in the elderly may exclude people with a history of falls. Thus, to reduce confounding in these types of studies, persons considered to be free of the condition must truly be disease-free. In the example of a case-control study of non-Hodgkin's lymphoma, only persons with no history of autoimmune disease should be included, so an algorithm for autoimmune disease with a high NPV should be employed. This would ensure that anyone classified as having no history of autoimmune disease truly was disease-free, even if this unnecessarily excluded a few people without autoimmune disease.

CAVEATS

The above discussion provides examples of scenarios in which different types of algorithm accuracy are important. The following section has additional considerations for guiding the use of electronic healthcare data algorithms.

Relationship between misclassification and bias of estimates is complex

One of the primary reasons to prioritize one measure of algorithm accuracy over another is to reduce bias (or distortion) in the risk estimate (i.e., the magnitude of the association between the exposure and the health outcome). The relationship between misclassification and bias of risk estimates is complex, however.(35–41) We will not explore the literature in detail, although we note several factors that may make it difficult to determine how misclassification will affect the estimate of the association between the exposure and outcome:

1. Non-differential bias does not always attenuate the risk estimate toward the null, (35, 40, 42) particularly when an exposure has more than two levels,(35, 40, 42) when non-differential errors in exposure and outcome classification are not independent of one another,(38, 41, 43) or when the error in a variable is associated with its true level.(44)
2. Small departures from non-differentiality (i.e., misclassification that is approximately—but not exactly—the same in groups being compared) can lead to substantial bias away from the null.(37, 45)
3. Differential misclassification can cause bias in either direction.(19)
4. Because bias is an average, chance alone may cause results from an individual study to be in the opposite direction of the expected bias.(36, 46, 47)

5. Bias away from the null can occur when adjusting for a confounder that is non-differentially misclassified if the direction of confounding is away from the null. (48)
6. Total bias in a risk estimate depends upon factors other than misclassification.(39, 49)

Thus, when developing and using an algorithm, predicting the expected direction of the bias due to misclassification may not be possible.

Algorithm properties may vary across settings

An algorithm developed in one setting may have different sensitivities and specificities in other settings if electronic healthcare data coding practices differ or change over time. Some integrated delivery system use their own "homegrown" codes,(4) which can make applying an algorithm developed in another setting difficult, unless careful mapping of the homegrown codes to standard diagnostic and procedure codes is performed. Even when identical coding systems are used, coding practices may differ. For example, fee-for-service and health maintenance organization providers may code differently based on reimbursement structure.(4) Thus, applying an algorithm developed in one setting to a different setting requires caution and an understanding of similarities and differences in coding practices. Studies using algorithms developed in other settings may find it useful to first assess algorithm accuracy in a subset of their own study population. Lack of adequate detail in reports of validation studies (14) may make this challenging. To the extent that readers are unable to identify characteristics of the study population used for validation or the algorithm itself, they will have difficult determining whether the algorithm is appropriate for use in a subsequent study.

CONCLUSIONS

Electronic healthcare data are valuable resources for epidemiologic studies. Ideally, algorithms that identify procedures and disease states from automated healthcare data would be 100% accurate, with perfect sensitivity and specificity. In reality, however, sensitivity and specificity are a tradeoff, and depending on the goals of the study a researcher must prioritize one measure of accuracy over others. When additional data collection with a more accurate measurement tool is feasible, algorithm sensitivity should be prioritized. High sensitivity is also important for enhancing study inclusiveness and for collecting information on common exposures. High specificity is important for classifying outcomes. High positive predictive value is important for cohort identification when the cohort does not need to be representative or include everyone with the defining condition. Finally, a high negative predictive value is important for reducing the likelihood that included subjects will have an exclusionary condition. We encourage publication of all tested algorithms, in accordance with recently proposed guidelines,(14) even those with unacceptable accuracy levels, to assist future studies in refining and applying the algorithms that are the most suitable for their objectives.

Acknowledgments

SOURCES OF FINANCIAL SUPPORT:

The project described was supported by Award Number R21CA143242 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessary represent the official views of the National Cancer Institute or the National Institutes of Health.

The authors thank Drs. Diana S.M. Buist, PhD, Michael L. Jackson, PhD, Holly Janes, PhD, and Bill Barlow, PhD for comments on earlier drafts of the manuscript.

References

1. Mullooly JP. Misclassification Model for Person-Time Analysis of Automated Medical Care Databases. Am J Epidemiol. October 15; 1996 144(8):782–92. [PubMed: 8857827]
2. Ray WA, Griffin MR. Use of Medicaid data for pharmacoepidemiology. Am J Epidemiol. 1989 Apr; 129(4):837–49. [PubMed: 2646920]
3. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58(4):323–37. [PubMed: 15862718]
4. Hornbrook MC, Goodman MJ, Fishman PA, Meenan RT, O'Keeffe-Rosetti M, Bachman DJ. Building health plan databases to risk adjust outcomes and payments. Int J Qual Health Care. 1998 Dec; 10(6):531–8. [PubMed: 9928592]
5. Brookhart MAP, Sturmer TMDMPH, Glynn RJPS, Rassen JS, Schneeweiss SMDS. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. Medical Care. 2010; 48(6 Supplement 1):S114–S20. [PubMed: 20473199]
6. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med. 1997 Oct 15; 127(8 Pt 2): 666–74. [PubMed: 9382378]
7. Roos LL, Mustard CA, Nicol JP, McLerran DF, Malenka DJ, Young TK, et al. Registries and administrative data: organization and accuracy. Med Care. 1993 Mar; 31(3):201–12. [PubMed: 8450678]
8. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Med Care. 2004 Nov; 42(11):1066–72. [PubMed: 15586833]
9. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002 Aug; 40(8 Suppl):IV-3–18.
10. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects--advantages and disadvantages. Nat Clin Pract Rheumatol. 2007 Dec; 3(12):725–32. [PubMed: 18037932]
11. Nelson JC, Jackson ML, Weiss NS, Jackson LA. New strategies are needed to improve the accuracy of influenza vaccine effectiveness estimates among seniors. J Clin Epidemiol. 2009; 62(7):687–94. [PubMed: 19124221]
12. Jackson LA, Nelson JC, Benson P, Neuzil KM, Reid RJ, Psaty BM, et al. Functional status is a confounder of the association of influenza vaccine and risk of all cause mortality in seniors. Int J Epidemiol. 2006 Apr; 35(2):345–52. [PubMed: 16368724]
13. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010 Jun; 48(6 Suppl):S114–20. [PubMed: 20473199]
14. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol. 2010 Dec 30.
15. A Dictionary of Statistics. New York: Oxford University Press Inc; 2002.
16. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. Health Serv Res. 2004 Dec; 39(6 Pt 1):1733–49. [PubMed: 15533184]
17. Earle CC, Nattinger AB, Potosky AL, Lang K, Mallick R, Berger M, et al. Identifying cancer relapse using SEER-Medicare data. Med Care. 2002 Aug; 40(8 Suppl):IV-75–81.
18. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A. The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. J Clin Epidemiol. 2010; 63(12):1332–41. [PubMed: 20457509]
19. Koepsell, TD.; Weiss, NS. Epidemiologic Methods: studying the occurrence of illness. New York: Oxford University Press; 2003.
20. Frayne SM, Miller DR, Sharkansky EJ, Jackson VW, Wang F, Halanych JH, et al. Using administrative data to identify mental illness: what approach is best? Am J Med Qual. 2010 Jan-Feb; 25(1):42–50. [PubMed: 19855046]

21. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? Am J Med Qual. 2004 Sep-Oct; 19(5):201–6. [PubMed: 15532912]
22. Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. Am J Med Qual. 2005 Nov-Dec; 20(6):319–28. [PubMed: 16280395]
23. Reker DM, Hamilton BB, Duncan PW, Yeh SC, Rosen A. Stroke: who's counting what? J Rehabil Res Dev. 2001 Mar-Apr; 38(2):281–9. [PubMed: 11392661]
24. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. Cancer Causes Control. 2007 Jun; 18(5):561–9. [PubMed: 17447148]
25. Winkelmayer WC, Schneeweiss S, Mogun H, Patrick AR, Avorn J, Solomon DH. Identification of individuals with CKD from Medicare claims data: a validation study. Am J Kidney Dis. 2005 Aug; 46(2):225–32. [PubMed: 16112040]
26. Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol. 2000 Jun; 53(6):605–14. [PubMed: 10880779]
27. Warren JL, Feuer E, Potosky AL, Riley GF, Lynch CF. Use of Medicare hospital and physician data to assess breast cancer incidence. Med Care. 1999 May; 37(5):445–56. [PubMed: 10335747]
28. Fleiss, JL. The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons; 1986.
29. White, E.; Armstrong, BK.; Saracci, R. Collecting, Evaluating, and Improving Measures of Disease Risk Factors. 2. New York: Oxford University Press; 2008. Principles of Exposure Measurement in Epidemiology.
30. Petersen LA, Normand SL, Druss BG, Rosenheck RA. Process of care and outcome after acute myocardial infarction for patients with mental illness in the VA health care system: are there disparities? Health Serv Res. 2003 Feb; 38(1 Pt 1):41–63. [PubMed: 12650380]
31. Nordstrom BL, Norman HS, Dube TJ, Wilcox MA, Walker AM. Identification of abacavir hypersensitivity reaction in health care claims data. Pharmacoepidemiol Drug Saf. 2007 Mar; 16(3):289–96. [PubMed: 17245797]
32. Gary TL, Batts-Turner M, Bone LR, Yeh H-c, Wang N-Y, Hill-Briggs F, et al. A randomized controlled trial of the effects of nurse case manager and community health worker team interventions in urban African-Americans with type 2 diabetes. Control Clin Trials. 2004; 25(1): 53–66. [PubMed: 14980748]
33. Chen Y, Guo JJ, Patel NC. Hemorrhagic stroke associated with antidepressant use in patients with depression: does degree of serotonin reuptake inhibition matter? Pharmacoepidemiol Drug Saf. 2009 Mar; 18(3):196–202. [PubMed: 19115419]
34. Beiderbeck AB, Holly EA, Sturkenboom MC, Coebergh JW, Stricker BH, Leufkens HG. Prescription medications associated with a decreased risk of non-Hodgkin's lymphoma. Am J Epidemiol. 2003 Mar 15; 157(6):510–6. [PubMed: 12631540]
35. Weinberg CR, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? Am J Epidemiol. 1994 Sep 15; 140(6):565–71. [PubMed: 8067350]
36. Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. Int J Epidemiol. 2005 Jun; 34(3):680–7. [PubMed: 15802377]
37. Jurek AM, Greenland S, Maldonado G. How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? Int J Epidemiol. 2008 Apr; 37(2):382–5. [PubMed: 18184671]
38. Brenner H, Savitz DA, Gefeller O. The effects of joint misclassification of exposure and disease on epidemiologic measures of association. J Clin Epidemiol. 1993 Oct; 46(10):1195–202. [PubMed: 8410104]
39. Greenland S, Gustafson P. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. Am J Epidemiol. 2006 Jul 1; 164(1):63–8. [PubMed: 16641307]

40. Correa-Villasenor A, Stewart WF, Franco-Marina F, Seacat H. Bias from nondifferential misclassification in case-control studies with three exposure levels. Epidemiology. 1995 May; 6(3):276–81. [PubMed: 7619936]
41. Chavance M, Dellatolas G, Lellouch J. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. Int J Epidemiol. 1992 Jun; 21(3):537–46. [PubMed: 1634317]
42. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol. 1990 Oct; 132(4):746–8. [PubMed: 2403115]
43. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. Epidemiology. 1992 May; 3(3):210–5. [PubMed: 1591319]
44. Wacholder S. When measurement errors correlate with truth: surprising effects of nondifferential misclassification. Epidemiology. 1995 Mar; 6(2):157–61. [PubMed: 7742402]
45. Brenner H. Inferences on the potential effects of presumed nondifferential exposure misclassification. Ann Epidemiol. 1993 May; 3(3):289–94. [PubMed: 8275202]
46. Sorahan T, Gilthorpe MS. Non-differential misclassification of exposure always leads to an underestimate of risk: an incorrect conclusion. Occup Environ Med. 1994 Dec; 51(12):839–40. [PubMed: 7849869]
47. Wacholder S, Hartge P, Lubin JH, Dosemeci M. Non-differential misclassification and bias towards the null: a clarification. Occup Environ Med. 1995 Aug; 52(8):557–8. [PubMed: 7663646]
48. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol. 1980 Oct; 112(4):564–9. [PubMed: 7424903]
49. Greenland, S.; Rothman, KJ. Modern Epidemiology. 2. Lippincott-Raven; 1998.
50. Raji MA, Kuo Y-F, Freeman JL, Goodwin JS. Effect of a Dementia Diagnosis on Survival of Older Patients After a Diagnosis of Breast, Colon, or Prostate Cancer: Implications for Cancer Care. Arch Intern Med. October 13; 2008 168(18):2033–40. [PubMed: 18852406]
51. Lamont EB, Herndon JE II, Weeks JC, Henderson IC, Earle CC, Schilsky RL, et al. Measuring Disease-Free Survival and Cancer Relapse Using Medicare Claims From CALGB Breast Cancer Trial Participants (Companion to 9344). J Natl Cancer Inst. September 20; 2006 98(18):1335–8. [PubMed: 16985253]
52. Dublin S, Jackson ML, Nelson JC, Weiss NS, Larson EB, Jackson LA. Statin use and risk of community acquired pneumonia in older people: population based case-control study. BMJ. 2009; 338:b2137. [PubMed: 19531550]
53. Nichol KL, Nordin J, Mullooly J, Lask R, Fillbrandt K, Iwane M. Influenza Vaccination and Reduction in Hospitalizations for Cardiac Disease and Stroke among the Elderly. N Engl J Med. April 3; 2003 348(14):1322–32. [PubMed: 12672859]
54. McClish D, Penberthy L. Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. Med Care. 2004 Nov; 42(11):1111–6. [PubMed: 15586838]
55. Penberthy L, McClish D, Manning C, Retchin S, Smith T. The added value of claims for cancer surveillance: results of varying case definitions. Med Care. 2005 Jul; 43(7):705–12. [PubMed: 15970786]
56. Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S. Using hospital discharge files to enhance cancer surveillance. Am J Epidemiol. 2003 Jul 1; 158(1):27–34. [PubMed: 12835284]
57. Gage BF, Birman-Deych E, Radford MJ, Nilasena DS, Binder EF. Risk of Osteoporotic Fracture in Elderly Patients Taking Warfarin: Results From the National Registry of Atrial Fibrillation 2. Arch Intern Med. January 23; 2006 166(2):241–6. [PubMed: 16432096]
58. Katon WJ, Lin EH, Williams LH, Ciechanowski P, Heckbert SR, Ludman E, et al. Comorbid depression is associated with an increased risk of dementia diagnosis in patients with diabetes: a prospective cohort study. J Gen Intern Med. 2010 May; 25(5):423–9. [PubMed: 20108126]

WHAT'S NEW

- Researchers developing algorithms based on electronic healthcare data should prioritize different measures of accuracy based on the intended use of the algorithm.
- Sensitivity is important for reducing the costs of data collection and ascertaining common exposures, whereas specificity is important for classifying outcomes.
- Researchers should publish all tested algorithms and their properties.

Table 1

Using Information From Administrative Healthcare Data Algorithms in Epidemiologic Studies

		Example
Types of information	Exposures	Preexisting dementia in a study of the association between dementia and cancer survival(50)
	Outcomes	Disease-free survival in breast cancer patients(51)
	Exclusion criteria	Immunocompromised persons in a study of statin use and pneumonia risk(52)
	Covariates	Comorbidities in a study of influenza vaccination and hospitalization for cardiac disease and stroke(53)
Information uses	Define a cohort	Identification of a cohort of breast cancer patients in which to study outcomes(16)
	Select patients for additional chart abstraction	Identification of persons with a primary diagnosis of myocardial infarction followed by clinical verification via chart abstraction(30)
	Exclude participants	Identification of persons with a history of cancer or human immunodeficiency virus-related illness to exclude from a case-control study of medication risk factors for non-Hodgkin's lymphoma(34)
	Select patients for direct contact	Identification of potential study participants for a trial of nurse case management in urban African-Americans with diabetes, followed by eligibility assessment by telephone(32)
	Supplement data from another source	Identification of additional cancer cases missed by cancer registry(54–56)
	Analytic variable without validation	Antidepressant use and the risk of hemorrhagic stroke(33)
Study designs	Retrospective studies	Case-control study of warfarin use and risk of osteoporotic fractures(57)
	Surveillance	Surveillance of claims data for likely abacavir hypersensitivity reaction, to be followed by medical chart review(31)
	Prospective studies	Prospective study of depression and risk of dementia among diabetic patients(58)

J Clin Epidemiol. Author manuscript; available in PMC 2013 March 1.

Table 2

Measures of Algorithm Accuracy

	Truth	
Algorithm classification	**Condition present**	**Condition not present**
Condition identified	A (true positives)	B (false positives)
Condition not identified	C (false negatives)	D (true negatives)
Sensitivity: proportion of those with the condition that are identified as having the condition.	$=\frac{A}{A+C}$	
Specificity: proportion of those without the condition that are identified as not having the condition.	$=\frac{D}{B+D}$	
Positive predictive value: proportion of those identified as having the condition who truly have it.	$=\frac{A}{A+B}=\frac{prevalence \times sensitivity}{(prevalence \times sensitivity)+(1-prevalence) \times (1-specificity)}$	
Negative predictive value: proportion of those identified as not having the condition who truly do not have it.	$=\frac{D}{C+D}=\frac{specificity \times (1-prevalence)}{specificity \times (1-prevalence)+prevalence \times (1-sensitivity)}$	

Table 3

Example of Bias in Relative Risk Due to Non-Differential Misclassification of an Exposure (X)[a]

Scenario		Proportion exposed to X	Incidence of outcome (Y) in unexposed	Incidence of Y in exposed	Incidence of Y overall	True relative risk	Algorithm sensitivity for X	Algorithm specificity for X	Observed relative risk[b]	% Bias in RR[c]
		P	I_U	I_E		RR_T	sens	spec	RR_o	
A	1	0.1	0.01	0.02	0.011	2.00	0.9	1.0	1.98	−1.1%
	2	0.1	0.01	0.02	0.011	2.00	0.8	1.0	1.96	−2.1%
	3	0.1	0.01	0.02	0.011	2.00	1.0	0.9	1.53	−23.7%
	4	0.1	0.01	0.02	0.011	2.00	1.0	0.8	1.36	−32.1%
	5	0.1	0.01	0.02	0.011	2.00	0.9	0.9	1.48	−25.9%
	6	0.1	0.01	0.02	0.011	2.00	0.9	0.8	1.32	−34.2%
B	1	0.5	0.01	0.02	0.015	2.00	0.9	1.0	1.83	−8.3%
	2	0.5	0.01	0.02	0.015	2.00	0.8	1.0	1.71	−14.3%
	3	0.5	0.01	0.02	0.015	2.00	1.0	0.9	1.91	−4.5%
	4	0.5	0.01	0.02	0.015	2.00	1.0	0.8	1.83	−8.3%
	5	0.5	0.01	0.02	0.015	2.00	0.9	0.9	1.73	−13.6%
	6	0.5	0.01	0.02	0.015	2.00	0.9	0.8	1.64	−18.2%
C	1	0.1	0.04	0.08	0.044	2.00	0.9	1.0	1.98	−1.1%
	2	0.1	0.04	0.08	0.044	2.00	0.8	1.0	1.96	−2.1%
	3	0.1	0.04	0.08	0.044	2.00	1.0	0.9	1.53	−23.7%
	4	0.1	0.04	0.08	0.044	2.00	1.0	0.8	1.36	−32.1%
	5	0.1	0.04	0.08	0.044	2.00	0.9	0.9	1.48	−25.9%
	6	0.1	0.04	0.08	0.044	2.00	0.9	0.8	1.32	−34.2%
D	1	0.5	0.04	0.08	0.060	2.00	0.9	1.0	1.83	−8.3%
	2	0.5	0.04	0.08	0.060	2.00	0.8	1.0	1.71	−14.3%
	3	0.5	0.04	0.08	0.060	2.00	1.0	0.9	1.91	−4.5%
	4	0.5	0.04	0.08	0.060	2.00	1.0	0.8	1.83	−8.3%
	5	0.5	0.04	0.08	0.060	2.00	0.9	0.9	1.73	−13.6%
	6	0.5	0.04	0.08	0.060	2.00	0.9	0.8	1.64	−18.2%

Scenario		Proportion exposed to X	Incidence of outcome (Y) in unexposed	Incidence of Y in exposed	Incidence of Y overall	True relative risk	Algorithm sensitivity for X	Algorithm specificity for X	Observed relative risk[b]	% Bias in RR[c]
		P	I_U	I_E		RR_T	sens	spec	RR_o	
E	1	0.1	0.01	0.03	0.012	3.00	0.9	1.0	2.94	−2.2%
	2	0.1	0.01	0.03	0.012	3.00	0.8	1.0	2.88	−4.2%
	3	0.1	0.01	0.03	0.012	3.00	1.0	0.9	2.05	−31.6%
	4	0.1	0.01	0.03	0.012	3.00	1.0	0.8	1.71	−42.9%
	5	0.1	0.01	0.03	0.012	3.00	0.9	0.9	1.95	−34.9%
	6	0.1	0.01	0.03	0.012	3.00	0.9	0.8	1.62	−45.9%

Abbreviations: RR, relative risk

[a] Assumes no misclassification of outcome (Y)

[b] $RR_o = \frac{a/(a+b)}{c/(c+d)}$ where

$a = sens \times P \times I_E + (1 - spec) \times (1 - P) \times I_U$

$b = sens \times P \times (1 - I_E) + (1 - spec) \times (1 - P) \times (1 - I_U)$

$c = (1 - sens) \times P \times I_E + spec \times (1 - P) \times I_U$

$d = (1 - sens) \times P \times (1 - I_E) + spec \times (1 - P) \times (1 - I_U)$

[c] $\text{Percent bias} = 100 \times \frac{RR_O - RR_T}{RR_T}$

Table 4

Example of Bias in Relative Risk Due to Non-differential Misclassification of an Outcome (Y)[a]

Scenario		Proportion exposed to X	Incidence of Y in unexposed	Incidence of Y in exposed	Incidence of Y overall	True relative risk	Algorithm sensitivity for Y	Algorithm specificity for Y	Observed relative risk[b]	% Bias in RR[c]
		P	I_U	I_E		RR_T	sens	spec	RR_o	
A	1	0.1	0.01	0.02	0.011	2.00	0.9	1.0	2.00	0.0%
	2	0.1	0.01	0.02	0.011	2.00	0.8	1.0	2.00	0.0%
	3	0.1	0.01	0.02	0.011	2.00	1.0	0.9	1.08	−45.9%
	4	0.1	0.01	0.02	0.011	2.00	1.0	0.8	1.04	−48.1%
	5	0.1	0.01	0.02	0.011	2.00	0.9	0.9	1.07	−46.3%
	6	0.1	0.01	0.02	0.011	2.00	0.9	0.8	1.03	−48.3%
B	1	0.5	0.01	0.02	0.015	2.00	0.9	1.0	2.00	0.0%
	2	0.5	0.01	0.02	0.015	2.00	0.8	1.0	2.00	0.0%
	3	0.5	0.01	0.02	0.015	2.00	1.0	0.9	1.08	−45.9%
	4	0.5	0.01	0.02	0.015	2.00	1.0	0.8	1.04	−48.1%
	5	0.5	0.01	0.02	0.015	2.00	0.9	0.9	1.07	−46.3%
	6	0.5	0.01	0.02	0.015	2.00	0.9	0.8	1.03	−48.3%
C	1	0.1	0.04	0.08	0.044	2.00	0.9	1.0	2.00	0.0%
	2	0.1	0.04	0.08	0.044	2.00	0.8	1.0	2.00	0.0%
	3	0.1	0.04	0.08	0.044	2.00	1.0	0.9	1.26	−36.8%
	4	0.1	0.04	0.08	0.044	2.00	1.0	0.8	1.14	−43.1%
	5	0.1	0.04	0.08	0.044	2.00	0.9	0.9	1.24	−37.9%
	6	0.1	0.04	0.08	0.044	2.00	0.9	0.8	1.12	−43.9%
D	1	0.5	0.04	0.08	0.060	2.00	0.9	1.0	2.00	0.0%
	2	0.5	0.04	0.08	0.060	2.00	0.8	1.0	2.00	0.0%
	3	0.5	0.04	0.08	0.060	2.00	1.0	0.9	1.26	−36.8%
	4	0.5	0.04	0.08	0.060	2.00	1.0	0.8	1.14	−43.1%
	5	0.5	0.04	0.08	0.060	2.00	0.9	0.9	1.24	−37.9%
	6	0.5	0.04	0.08	0.060	2.00	0.9	0.8	1.12	−43.9%
E	1	0.25	0.01	0.03	0.012	3.00	0.9	1.0	3.00	0.0%

Scenario	Proportion exposed to X	Incidence of Y in unexposed	Incidence of Y in exposed	Incidence of Y overall	True relative risk	Algorithm sensitivity for Y	Algorithm specificity for Y	Observed relative risk[b]	% Bias in RR[c]
	P	I_U	I_E		RR_T	sens	spec	RR_o	
2	0.25	0.01	0.03	0.012	3.00	0.8	1.0	3.00	0.0%
3	0.25	0.01	0.03	0.012	3.00	1.0	0.9	1.17	−61.2%
4	0.25	0.01	0.03	0.012	3.00	1.0	0.8	1.08	−64.1%
5	0.25	0.01	0.03	0.012	3.00	0.9	0.9	1.15	−61.7%
6	0.25	0.01	0.03	0.012	3.00	0.9	0.8	1.07	−64.4%

Abbreviations: RR, relative risk

[a] Assumes no misclassification of exposure (X)

[b] $RR_o = \frac{a/(a+b)}{c/(c+d)}$ where

$a = sens \times P \times I_E + (1 - spec) \times P \times (1 - I_E)$

$b = (1 - sens) \times P \times I_E + spec \times P \times (1 - I_E)$

$c = sens \times (1 - P) \times I_U + (1 - spec) \times (1 - P) \times (1 - I_U)$

$d = (1 - sens) \times (1 - P) \times I_U + spec \times (1 - P) \times (1 - I_U)$

[c] $\text{Percent bias} = 100 \times \frac{RR_O - RR_T}{RR_T}$