Journal of Clinical Epidemiology 64 (2011) 821–829

Journal of Clinical Epidemiology

REVIEW ARTICLES

Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data

Eric I. Benchimol[a,b,c,d,e,f,g,*], Douglas G. Manuel[a,h,i,j,k], Teresa To[a,c,d], Anne M. Griffiths[b,e], Linda Rabeneck[a,d,l], Astrid Guttmann[a,d,e,m]

[a] *The Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[b] *Division of Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, Toronto, Ontario, Canada*
[c] *Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Ontario, Canada*
[d] *Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*
[e] *Department of Paediatrics, University of Toronto, Toronto, Ontario, Canada*
[f] *Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada*
[g] *Department of Pediatrics, University of Ottawa, Ottawa, Ontario, Canada*
[h] *Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada*
[i] *Department of Epidemiology and Community Medicine, University of Ottawa, Ontario, Ottawa, Canada*
[j] *Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*
[k] *Statistics Canada, Ottawa, Ontario, Canada*
[l] *Department of Medicine, University of Toronto, Toronto, Ontario, Canada*
[m] *Division of Paediatric Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada*

Accepted 7 October 2010

Abstract

Background and Objectives: Validation of health administrative data for identifying patients with different health states (diseases and conditions) is a research priority, but no guidelines exist for ensuring quality. We created reporting guidelines for studies validating administrative data identification algorithms and used them to assess the quality of reporting of validation studies in the literature.

Methods: Using Standards for Reporting of Diagnostic accuracy (STARD) criteria as a guide, we created a 40-item checklist of items with which identification accuracy studies should be reported. A systematic review identified studies that validated identification algorithms using administrative data. We used the checklist to assess the quality of reporting.

Results: In 271 included articles, goals and data sources were well reported but few reported four or more statistical estimates of accuracy (36.9%). In 65.9% of studies reporting positive predictive value (PPV)/negative predictive value (NPV), the prevalence of disease in the validation cohort was higher than in the administrative data, potentially falsely elevating predictive values. Subgroup accuracy (53.1%) and 95% confidence intervals for accuracy measures (35.8%) were also underreported.

Conclusions: The quality of studies validating health states in the administrative data varies, with significant deficits in reporting of markers of diagnostic accuracy, including the appropriate estimation of PPV and NPV. These omissions could lead to misclassification bias and incorrect estimation of incidence and health services utilization rates. Use of a reporting checklist, such as the one created for this study by modifying the STARD criteria, could improve the quality of reporting of validation studies, allowing for accurate application of algorithms, and interpretation of research using health administrative data. © 2011 Elsevier Inc. All rights reserved.

Keywords: Health administrative data; Misclassification bias; Diagnostic accuracy; Sensitivity and specificity; Predictive values; Health services research; Epidemiology

1. Introduction

Health services and epidemiologic research are best conducted with population-level data. This helps ensure the appropriate estimation of incidence and prevalence rates, the minimization of referral bias, and the overall generalizability of the study conclusions to the population of interest. Because prospective clinical registries and retrospective

This research was conducted with the support of a Clinical Research Award from the American College of Gastroenterology.

* Corresponding author. Division of Gastroenterology, Hepatology and Nutrition, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, Ontario K1H 8L1, Canada. Tel.: +1-613-737-7600; fax: +1-613-738-4854.

E-mail address: ebenchimol@cheo.on.ca (E.I. Benchimol).

0895-4356/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2010.10.006

What is new?

- Significant deficits exist in the validation and reporting of algorithms used to identify patients within health administrative data.
- Misclassification error represents an important form of bias in research using health administrative databases.
- The modified Standards for Reporting of Diagnostic accuracy criteria reported here can be used to improve the quality of reporting in studies of health state (disease or conditions) identification validation.
- Future efforts should address criteria for conduct and reporting of research using health administrative data.

chart review comprising a representative sample or all residents of a jurisdiction are impractical, health administrative data are an alternative for population-based chronic disease surveillance, outcomes research, and health services research. Health administrative data are defined as information passively collected, often by government and health care providers, for the purpose of managing the health care of patients [1], and are a subtype of automated health care data [2]. Examples include physician billing databases (such as those managed by government in single-payer health systems or by health maintenance organizations [HMOs]), and hospital discharge record databases. Accuracy of the diagnostic codes used to identify patients within these data depends on multiple factors including database quality, the specific condition being identified, and the validity of the codes in the patient group. A large gradient in data quality exists, with some databases being of higher quality than others [3]. Isolated diagnostic codes associated with physician billing records have been shown to be accurate to identify patients with some chronic diseases [4,5] but not others [6–9]. Since chronic diseases usually require multiple contacts with the health system to diagnose, a single-visit diagnostic code is often insufficient to accurately identify patients with the disease. The validity of codes is also dependent on the patient group being studied. For instance, the accuracy of diagnostic codes or combinations of codes (algorithms) varies across age groups because of variable use of the health system [6,7,10]. As such, validation of algorithms used to identify patients with different health states (including acute conditions, chronic diseases, and other health outcomes) is essential to avoid misclassification bias [11], which may threaten the internal validity and interpretation of study conclusions. For example, assessment of health services utilization in a cohort of patients with a chronic disease contaminated by large number of healthy residents falsely labeled as having a chronic disease would underestimate the burden of the disease on the health system or the quality and performance of the health system. Similarly, assessment of incidence of the disease in the cohort would overestimate risk to the population. Although the validation of administrative data coding has been identified as a priority in the health services research by an international consortium [3], the complete and accurate reporting of algorithm validation research is equally important to appropriate application. The growing availability of administrative data for research coupled with the expense, privacy concerns, and complex methodologies required to validate identification algorithms have resulted in algorithms being applied to these databases by researchers not involved in their initial validation. As such, minimum quality criteria for the conduct and reporting of algorithm validation studies would benefit scientists using these algorithms and consumers of the research on which these algorithms rely.

The purpose of this study was to appraise all studies that validated algorithms to identify patients with different health states within the administrative data with newly developed consensus criteria for the reporting of studies that validate health administrative data algorithms, based on the Standards for Reporting of Diagnostic accuracy (STARD) initiative [12]. In so doing, we aimed to identify strengths and weaknesses in the methods of such validation studies to improve the future reporting of research using health administrative data.

2. Methods

2.1. Development of validation study quality checklist

Algorithms to identify patients with different health states may be considered a type of diagnostic test applied to health administrative data, and markers of diagnostic accuracy are often reported in studies validating algorithms against reference standards. As such, we modified the criteria published by the STARD initiative for the accurate reporting of studies using diagnostic tests [12] to evaluate included studies. Four experts (E.I.B., D.M., T.T., and A.G.) in research using these data modified the STARD criteria using an iterative approach to create a 40-point data collection tool for evaluation of studies (Table 1). A fillable version of the data collection tools is available in Supplemental Data 1 and can be used to assess the algorithm validation literature for completeness of reporting.

2.2. Systematic review: search strategy and selection criteria

An online database literature search was performed for human studies, without language restrictions, using the following databases: MEDLINE (National Library of Medicine [NLM], Bethesda; January 1950 to June 2009) and

Table 1
Data collection tool with extraction results reported in the appropriate columns as percentages

Checklist criteria	Yes (%)	No (%)	Uncertain	Not applicable (%)
Title, keywords, abstract				
1. Identifies article as study of assessing diagnostic accuracy?	94.1	5.9		
2. Identifies article as study of administrative data?	97.4	2.6		
Introduction				
3. States disease identification and validation as one of the goals of study?	93.4	6.6		
Methods				
Participants in validation cohort				
4. Describes validation cohort? (cohort of patients to which reference standard was applied)	98.9	1.1		
4a. Age?	49.1	50.6		
4b. Disease?	95.2	2.2		
4c. Severity?	17.3	48.0		34.7
4d. Location/jurisdiction?	90.8	9.2		
5. Describes recruitment procedure of validation cohort?	98.2	0.7		
5a. Inclusion criteria?	94.8	3.3		
5b. Exclusion criteria?	45.4	52.4		
6. Describes patient sampling? (random, consecutive, all, etc.)	91.5	7.4		
7. Describes data collection?	88.9	7.0		
7a. Who identified patients and ensured selection adhered to patient recruitment criteria?	74.2	14.0		10.7
7b. Who collected data?	64.6	22.5		12.2
7c. A priori data collection form?	59.0	5.2		14.4
7d. How was disease classified?	78.6	14.8		
8. Was there a split sample (i.e., revalidation using a separate cohort)?	11.4	88.2		
Test methods				
9. Describe number, training and expertise of persons reading reference standard?	46.1	23.6		29.5
10. If >1 person reading reference standard, is kappa quoted?	11.4	30.6	13.3	44.7
11. Were the readers of the reference (validation) test blinded to the results of the classification by administrative data for that patient? (e.g., Was the reviewer of the charts blinded to how that chart was billed?)	19.2	9.2	42.4	29.2
Statistical methods				
12. Describes methods of calculating/comparing diagnostic accuracy?	83.4	16.6		
Results				
Participants				
13. Report when study done, start/end dates of enrollment	80.8	17.3		
14. Describe number of people who satisfied inclusion/exclusion criteria?	83.4	14.0		
15. Study flow diagram?	17.0	83.0		
Test results				
16. Reports distribution of disease severity?	19.2	46.5		34.3
17. Report cross-tabulation of index tests by results of reference standard	80.4	19.2		
Estimates				
18. Reports at least 4 estimates of diagnostic accuracy? (Estimates reported in included studies)	36.9	63.1		
18a. Sensitivity	67.2	32.8		
18b. Specificity	49.8	50.2		
18c. PPV	63.8	36.2		
18d. NPV	32.1	67.9		
18e. Likelihood ratios	3.3	96.7		
18f. Kappa	29.2	70.8		
18g. Area under the ROC curve/c-statistic	7.0	93.0		
18h. Accuracy/agreement	26.6	73.4		
19. Was the accuracy reported for any subgroup? (e.g., age, geography, different sexes, and so on)	53.1	46.9		
20. If PPV/NPV reported, does ratio of cases/controls of validation cohort approximate prevalence of condition in the population?	21.8	42.1		36.2
21. Reports 95 CIs for each of above?	35.8	63.8	0.4	
Discussion				
22. Discusses the applicability of the findings?	96.3	3.7		

Abbreviations: PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristic; CI, confidence interval.

EMBASE (Elsevier, NY; January 1980 to June 2009). Search strategy used for MEDLINE is described in Supplemental Data 2 and was modified for EMBASE terms. The search strategy was designed to capture studies using health administrative data, which described the diagnostic accuracy in validation studies of algorithms to identify patients. The reference lists of review articles were also searched for relevant publications. Included studies used health administrative data, which were defined as data (containing little or no nonadministrative clinical information) routinely and passively collected for administrative purposes without an *a priori* research question [1]. Other inclusion criteria included the identification of a health state (disease, health outcome, medical procedure, or investigation) and examination of patient-level data. Included studies validated patient identification algorithms from within the administrative data. We defined an identification algorithm as any single record or combination of records of a health services contact (e.g., physician visit, hospitalization, procedure, laboratory test, and so on) used to identify patients with a health state from within the administrative data. Studies in all languages were included; however, two studies could not be adequately translated and were therefore excluded. Exclusion criteria were the use of databases with significant clinical information (e.g., pathology/histology results, extensive laboratory or microbiological results, chart notes, and so on) and birth and death registries (unless these data were used as the reference standard for validation of the administrative data). Studies validating the performance of comorbidity measures were excluded unless they also validated the identification of disease components of the measure. Finally, studies using databases with extensive pre- and post-collection data quality improvement programs were excluded (e.g., General Practice Research Database, Surveillance Epidemiology and End Results and other cancer registries, and so on), as these were not considered truly health administrative data and were often supplemented by chart abstraction or clinical information.

Two investigators (E.I.B. and A.G.) reviewed the articles for inclusion. After assurance of consistency of the two raters, the articles were distributed randomly and the raters used the data collection tool for assessment of study quality. Ten randomly selected articles were reviewed independently by both raters to clarify the data collection tool. With discussion, they decided on 10 key points of the 40-point checklist to be used to calculate unweighted Cohen's kappa coefficients (with standard deviations [SD]) for consistency. The subsequent 40 articles were reviewed independently, and kappa statistics were calculated on each of the 10 key points, with the *a priori* determined goal of achieving kappa >0.4 (moderate agreement) before distribution of the articles. An additional 10 articles were reviewed independently after clarification and revision of the wording of the data collection tool, and kappa coefficients were calculated using SPSS, version 15, (SPSS Inc., Chicago, IL). Kappa coefficients are reported in Table 2 demonstrating moderate or better agreement in all the 10 priority points after the second iteration of independent review by the two raters.

After data collection from all included studies, description of the study characteristics and raters' evaluations of the studies were computerized and proportions were calculated using Microsoft Excel 2007 (Microsoft Corporation, Redmond, WA).

3. Results

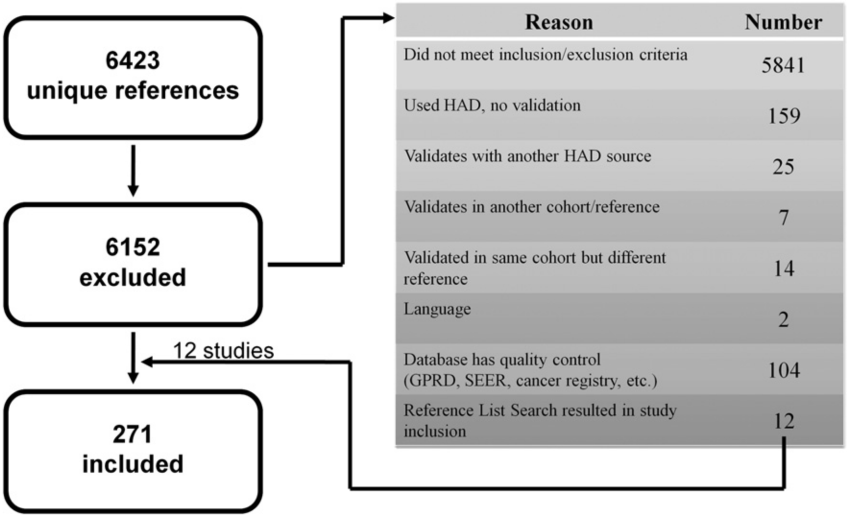
A total of 6,423 references were reviewed, resulting in 271 included studies from 16 countries (Fig. 1). A list of included studies is provided in Supplemental Data 3. Of the 271 included studies, 160 were from the United States; 50 from Canada; 12 from Australia; 7 each from Denmark, Italy, and the United Kingdom; 6 from France; 5 each from Brazil,

Table 2
Kappa coefficients demonstrating interrater consistency of the two data extractors in 10 priority areas

	Articles no. 11–50	Articles no. 51–60	Combined articles no. 11–60
Question	Kappa ± SD	Kappa ± SD	Kappa ± SD
1. Introduction	0.894 ± 0.073	0.615 ± 0.337	0.645 ± 0.233
7d. Methods—disease classification	0.479 ± 0.129	1.000 ± 0	0.505 ± 0.125
9. Methods—number and training	0.595 ± 0.122	1.000 ± 0	0.727 ± 0.096
11. Methods—blinding	0.431 ± 0.111	1.000 ± 0	0.747 ± 0.075
15. Results—study flow diagram	0.608 ± 0.176	0.615 ± 0.337	0.611 ± 0.156
18. Results—4 estimates of diagnostic accuracy	0.843 ± 0.086	1.000 ± 0	0.880 ± 0.066
19. Results—reported for subgroups	−0.291 ± 0.147	0.444 ± 0.223	0.180 ± 0.081
20. Results—PPV/NPV	0.830 ± 0.081	0.524 ± 0.208	0.773 ± 0.077
21. Results—95% CI	0.894 ± 0.073	1.000 ± 0	0.919 ± 0.056
22. Discussion	1.000 ± 0	0.615 ± 0.337	0.790 ± 0.203
Overall	0.721 ± 0.03	0.855 ± 0.047	0.750 ± 0.026

Abbreviations: SD, standard deviation; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.

Note that articles no. 1–10 were assessed to pilot the collection tool. Articles no. 11–50 were independently assessed using the collection tool, kappa calculated, and wording of the collection tool revised. Articles no. 51–60 were assessed independently, and kappa was calculated.



Reason	Number
Did not meet inclusion/exclusion criteria	5841
Used HAD, no validation	159
Validates with another HAD source	25
Validates in another cohort/reference	7
Validated in same cohort but different reference	14
Language	2
Database has quality control (GPRD, SEER, cancer registry, etc.)	104
Reference List Search resulted in study inclusion	12

Fig. 1. Flow diagram of included and excluded studies. Abbreviations: HAD, health administrative data; GPRD, General Practice Research Database; SEER, Surveillance Epidemiology and End Results.

the Netherlands, and Sweden; 2 from Israel; and one each from Finland, Germany, Norway, Spain, and Switzerland. Approximately half of the studies used identification algorithms to develop a cohort of patients for epidemiologic or health services research, whereas the goal of the other half was to validate single International Classification of Disease (ICD) diagnostic codes associated with multiple conditions without explicit cohort development. Both types of studies were included in the analyses. Included studies used single administrative databases or linked combinations of databases (e.g., physician billing, hospital discharge, and pharmacy databases). In 15 studies (5.5%), the administrative database used was not reported in the publication. Despite a search strategy which preferentially obtained studies about validation and diagnostic accuracy of algorithms, 159 studies were excluded despite using health administrative data because they did not validate or use previously validated algorithms to identify patients.

Included studies used a number of different forms of reference standards against which identification algorithms were validated. Reference standards included medical record review ($n = 153$ studies), surveys of patients or practitioners ($n = 35$), clinical registries ($n = 29$), cancer registries ($n = 10$), laboratory or radiology results ($n = 4$), death registries ($n = 3$), pharmacy records ($n = 2$), and neonatal screening programs ($n = 1$). Thirty-two studies used multiple linked sources as their reference standard. The most common statistics used to estimate diagnostic accuracy of identification algorithms included sensitivity ($n = 182$), positive predictive value (PPV) ($n = 173$), specificity ($n = 135$), negative predictive value (NPV) ($n = 87$), kappa coefficient ($n = 79$), and agreement/accuracy ($n = 72$). Quality of reporting within the studies are reported as proportions within Table 1 using the reporting criteria developed based on the STARD guidelines.

Table 3 displays the results of data extraction from five studies chosen by the raters as examples of those whose methods and reporting were of high quality, satisfying most of the modified STARD criteria. These studies can be used as examples of well-designed and reported research validating algorithms using health administrative data.

4. Discussion

The translation of research from the literature to medical practice or health policy requires the research to be appropriately designed, reported, and interpreted. As such, consortia have created criteria for the reporting of clinical trials [18], observational studies [19], and studies of diagnostic accuracy [12]. These criteria are guidelines for researchers involved in study design and for consumers of the literature to assess the quality of the research. Unfortunately, no such criteria exist for the creation or reporting of studies using health administrative data. An international symposium assessed priorities of methodological research using administrative data associated with ICD-9 and ICD-10 [3]. Five of 13 potential areas of research identified were related to reliability and validity of these data. These included assessment of internal consistency of identification algorithms, identification of reliable reference standards against which to validate data, the creation of training standards for coders, development of chart–database comparison studies, and international cross-validation of ICD-10. Using expert consensus, we developed guidelines for one component of research using administrative data, the reporting of validation studies, which enable identification of groups of patients in the administrative data. We used the criteria developed to assess the quality of the methods and reporting of research, which validated algorithms used to identify patients with different health states within the

Table 3
Examples of high-quality studies validating identification algorithms

Reference	Health state identified	Reference standard	1	2	3	4	4a	4b	4c	4d	5	5a	5b	6	7	7a	7b	7c	7d	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Armstrong et al. [13]	Mammography	Self-report and center facility report	X	X	X	X	X	X	N/A	X	X	X	X	X	X	X	X	X	X		N/A	N/A	N/A	X	X	X	X	N/A	X	X	X	X	X	X
Nattinger et al. [14]	Breast cancer	SEER Registry	X	X	X	X	X	X	X	X	X	X	X	X	N/A	N/A	N/A	N/A	X	X	N/A	N/A	N/A	X	X		X	N/A	X		X	X	X	X
Payne et al. [15]	Anthrax vaccination	Medical records	X	X	X	X		X	N/A	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X		N/A	X	X	X	X	X	X
Roberts et al. [16]	Hypertension in pregnancy	Registry and medical records from cohort study	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X		X		UC	X	X	X			X	X	X	X	X	X
Robinson et al. [17]	Diabetes, hypertension, ischemic heart disease, stroke, hypercholesterolemia	Survey of patients	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X		N/A	N/A	N/A	X	X	X			X	X	X	X		X

Abbreviations: N/A, not applicable; UC, uncertain; SEER, Surveillance Epidemiology and End Results.

These studies had the highest proportion of reported criteria to applicable criteria listed in Table 1. Question number refers to numbering available in Table 1. X mark denotes that item was reported. Blank space denotes item was not reported.

administrative data. In so doing, we identified areas of strength and weakness in the validation literature, and we hope to encourage the improvement of future validation research.

Validation is increasingly recognized as an important component of research with health administrative data. Of included studies, 246 of 271 (90.8%) were published after the 1996 report by Huston and Naylor [20], which emphasized the importance of data accuracy and validation in reporting studies using secondary data sources. Nevertheless, data accuracy continues to be a significant issue [3]. Using a group of experts, we identified areas of priority when designing and assessing the quality of validation studies using these data. We included these priority areas in a checklist of criteria, which can be used to ensure accurate and complete reporting of validation studies. In broad terms, our criteria included the clear labeling of studies as using health administrative data, description of the validation cohort, quality of methods to extract data from the validation cohorts, statistical methods accuracy assessment, and reporting the implications of the validation research.

Most authors adequately identified their research as using and validating health administrative data to identify patients. Despite this, most articles did not use the term "health administrative data" as a subject heading, and the term is not recognized as a Medical Subject Heading (MeSH) by the NLM [21] or as an EMBASE subject heading [22]. This makes the creation of systematic searches of the administrative data literature difficult, impeding the sensitivity of a search using MeSH headings. One study attempted to address this by validating an effective MEDLINE search strategy for all research using administrative data [23]. Rather than searching all administrative data research, we created a search strategy for validation studies only, which was very sensitive with low specificity, resulting in a large number of articles obtained for review. We therefore believe that we have obtained most or all of the literature pertaining to validation of administrative data algorithms. However, we recommend that the term "health administrative data" be added as a MeSH term to facilitate future systematic reviews of the literature. Our search strategy also contained terms related to validation including "verification," "identification," and multiple terms related to diagnostic accuracy. This likely resulted in lower sensitivity for identification of general articles using (but not validating) these data, with greater specificity for validation studies. Despite this, 159 studies were identified, which used administrative data without any validation of their identification algorithms, and additional studies validated with another administrative data source or used an algorithm validated in another cohort. These strategies reduce the reliability of identification methods and could alter the interpretation of the research. For example, in studies of patients with diabetes, one would expect regular screening for retinopathy. A lack of recorded physician visits for screening may reflect poor health services provision or an algorithm that was inaccurate and identified large numbers of nondiabetic patients. Without validation of identification methods, and without adequate reporting of the diagnostic accuracy of the identification algorithm, the reader cannot differentiate the two scenarios.

In describing the validation cohort, most studies described the health state identified and jurisdiction of the patients. Age was inadequately reported often because it was assumed that patients were adults. Algorithms validated in adults have been demonstrated inaccurate for the identification of children with the same disease [6,7], and some algorithms are more sensitive for older adults compared with younger adults [10]. Therefore, the age range of the validation cohort should be specified. Only 26.5% of applicable studies reported the range of disease severity in the validation cohort. This may be important to report as algorithms sensitive to patients with severe chronic disease may be less sensitive to identify those with mild disease, owing to lower health services utilization of mild patients. Although most studies reported on validation cohort recruitment (and in particular, inclusion criteria), we did not assess for the quality and randomness of recruitment strategies. For example, many studies used single ICD codes to identify potential patients with diseases, resulting in a nonrandom validation cohort and an algorithm, which may have been tainted by selection bias. Additionally, few studies addressed exclusion criteria, which might be fundamental to administrative data research. For example, insurance qualification status may have led to follow-up if authors did not exclude patients with consistent insurance coverage. Few studies revalidated algorithms using a separate cohort within the same article, and it should be stressed that algorithms validated in one jurisdiction or age group cannot necessarily be applied to other cohorts with assurances of accuracy. For example, three distinct algorithms to identify patients with inflammatory bowel disease were deemed accurate, each validated in different jurisdictions with different health administrative databases [6,24,25]. Characteristics of those interpreting the reference standards (e.g., chart reviewers) were also inadequately reported. Of applicable studies using interpreters of the reference standard, 66.1% described the number and training of the personnel. However, only 20.6% of studies using two or more personnel discussed consistency or quoted kappa coefficients, and only 27.1% reported that personnel were blinded to the codes from the administrative database.

Certain factors in the results sections of included studies were often poorly reported. Some studies included validation of coding as an issue reserved for the methods section (with the main aim of the studies to report on epidemiologic trends or health services utilization). These studies often poorly reported methodological details of the validation studies. For example, 7.0% of included studies did not report any information on data collection, in even the most general terms. If an article discussed diagnostic accuracy statistics or other properties labeled by the data extraction tool as applicable to the results section, but did so in the methods section, quality raters still counted these as reported. Nevertheless, when validation of algorithms was only reported in the methods, the expected results were often inadequately reported (likely because of space restrictions). However, most articles structured their reporting as one would expect from the data collection tool.

We reviewed the diagnostic accuracy statistics used in included studies and generally did not specifically judge their appropriateness for validation of algorithms. Only 36.9% of studies reported four or more measures of diagnostic accuracy. The most commonly reported were sensitivity and PPV, often because validation cohorts did not include patients without disease to act as reference standard true-negatives. PPV is defined as the proportion of the population with positive test (algorithm) results who are correctly identified as having the disease, whereas NPV is conversely the proportion of those with negative test results who are correctly identified as not having the disease (Fig. 2 for the full equations) [26]. Both markers of accuracy depend on the prevalence of the disease within the population of interest. In rare diseases, PPV can often be low even when sensitivity and specificity are high [26]. In some cases, even sensitivity, specificity, and likelihood ratios can depend on prevalence [27]. To understand the predictive values of algorithms to identify patients, it is important for the prevalence of the disease/condition to be equivalent in the validation cohort as in the administrative data. Investigators and readers may otherwise be falsely reassured by high predictive values in the high-prevalence validation cohort when in fact predictive values are quite low when applied to patients in the administrative databases with low prevalence of disease. Predictive values have been identified as the most important diagnostic accuracy markers in epidemiologic studies [28–31]. They define the likelihood of false-positive test results (resulting in overestimation of incidence and prevalence rates by misclassification of unaffected subjects as diseased) and false-

$$\text{PPV} = \frac{\text{number of TP}}{\text{number of TP + number of FP}} = \frac{\text{sensitivity x prevalence}}{\text{sensitivity x prevalence + (1-specificity) x (1-prevalence)}}$$

$$\text{NPV} = \frac{\text{number of TN}}{\text{number of TN + number of FN}} = \frac{\text{specificity x prevalence}}{\text{(1-sensitivity) x prevalence + specificity x (1-prevalence)}}$$

Fig. 2. Calculation of predictive values. Abbreviations: PPV = positive predictive value, NPV = negative predictive value, TP = true positives, TN = true negatives.

Table 4
Summary of recommendations based on results of systematic review and assessment of study quality

A. The term ‘‘health administrative data’’ should be added as MeSH and EMBASE subject headings and should be included as a key term in all studies using health administrative data.

B. Complete description of the validation cohort should include age, a description of the disease or health condition being studied, the distribution of disease severity (if applicable), and the geographic location or jurisdiction in which the validation cohort is located.

C. Where possible, revalidation of identification algorithms should take place in other jurisdictions before application in those jurisdictions' administrative data to ensure accuracy.

D. The training and job description of personnel interpreting the reference standard in a validation study and those personnel should be blinded to elements of administrative data when interpreting the reference standard. If two or more personnel are involved, statistics of consistency of reference standard interpretation should be reported (e.g., kappa coefficient).

E. Cross-tabulation of results should be included in the results section of articles, allowing for readers to assess the power and confidence intervals of the results.

F. Statistics describing diagnostic accuracy of algorithms should be described in the methods section, and at least four markers of diagnostic accuracy (with 95% CIs) should be reported.

G. Where PPV and NPV are reported, the prevalence of disease in the validation cohort should equal the prevalence of disease in patients contained within health administrative databases.

negative test results (resulting in an underestimation of burden of illness because of misclassification of diseased patients as unaffected). Despite this, the prevalence of disease was similar in the validation cohort compared with the administrative data in only 34.1% of studies describing predictive values. Researchers conducting future studies should make every effort to ensure adequate representation of nondiseased patients in validation cohorts to assure the reader of an accurate predictive values.

To our knowledge, this is the first study to create reporting guidelines and systematically review the validation literature. In so doing, we were able to provide general recommendations for future research and act as a first step toward development of guidelines on the use and reporting of research using these data (Table 4). Our guidelines are limited to validation studies only and do not evaluate the quality of the databases to which these algorithms are applied, nor did we address the reporting of other components of research using administrative data. By modifying the STARD criteria for use in our study and by assessing only quality of reporting, we omitted several important aspects of health administrative database research. These include the overall quality of the data, completeness of follow-up, the representative nature of the data to the general population, the quality of reference standards used to validate algorithms, the appropriateness of statistical measures of accuracy, the generalizability of algorithms to other jurisdictions, and other important issues that should be addressed in future consensus guidelines. We also did not address the collection processes for administrative data, the features of coding systems used, or the use of major or minor coding fields within the data.

The systematic review may be limited by its search strategy. Identification algorithms may not be reported in the standard scientific literature and may have been distributed by organizations or governments as internal reports. Nevertheless, we believe that the dissemination of validation studies is vital to advance the field of research using health administrative data and should therefore be reported in peer-reviewed publications. We have therefore focused on the scientific literature. Our search strategy used common terms used by studies of diagnostic accuracy to identify validation studies, and we therefore assumed that most or all validation studies used methods similar to those of studies of diagnostic accuracy (e.g., used a reference standard, compared algorithms or tests, reported statistics or diagnostic accuracy). We may have omitted novel strategies for algorithm development and validation. For example, two studies used Bayesian latent class modeling in place of a reference standard to estimate accuracy of their identification algorithms [32,33]. These studies did not meet our inclusion criteria, but this strategy may represent a valid alternative to validation with reference standards such as medical chart review.

In summary, we have identified the strengths and weaknesses of study design and reporting of results in the literature surrounding validation of algorithms and codes used to identify patients with different health states within health administrative databases. The criteria developed in this study were established by a group of experts, and the checklist detailed in Supplemental Data 1 can be used to improve study design and reporting of future research validating algorithms to identify health states within administrative data. Our checklist does not represent a final document of all criteria that should be expected in the conduct of health administrative data research. The weaknesses identified in the validation study literature suggest that a more comprehensive reporting guideline document is required. A consensus should include world expert opinion (as health administrative data quality is jurisdiction specific), develop a detailed list of criteria for all types of administrative data research (including validation, epidemiologic, and health services research), and involve a rigorous evaluation of all issues faced by researchers using these data. This study is the first stage in an ongoing process to improve the quality of epidemiologic and health services research conducted using health administrative data to ensure that interpretation and implementation of such research is accurate and reliable.

Acknowledgments

The authors wish to thank Ms Elizabeth Uleryk (Director, Hospital Library, The Hospital for Sick Children) for aiding in the search strategy used in this review, Ms Danielle Benchimol for data entry, and Drs Yaron Avitzur and Tanja Gonska for provision of translation services. Eric Benchimol is a Canadian Institutes of Health Research (CIHR) training fellow in the Canadian Child Health Clinician Scientist Program, in partnership with SickKids Foundation and the Child and Family Research Institute of British Columbia, and was also supported by a fellowship from the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition-Children's Digestive Health and Nutrition Foundation. Douglas Manuel holds a Chair in Applied Public Health from CIHR and the Public Health Agency of Canada. Teresa To was supported by the University of Toronto Dales Award. Astrid Guttmann was supported by a CIHR New Investigator Award.

Appendix

Supplementary data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jclinepi.2010.10.006

References

[1] Spasoff RA. Epidemiologic methods for health policy. New York, NY: Oxford University Press, Inc.; 1999.

[2] FDA's Sentinel Initiative. Silver Spring, MD: U.S. Food and Drug Administration; 2010. Available at. http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm. Accessed July 2010.

[3] De Coster C, Quan H, Finlayson A, Gao M, Halfon P, Humphries KH, et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. BMC Health Serv Res 2006;6:77.

[4] Lix L, Yogendran M, Burchill C, Metge C, McKeen N, Moore D, et al. Defining and validating chronic diseases: an administrative data approach. Winnipeg, Manitoba: Manitoba Centre for Health Policy; 2006.

[5] Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Med Res Methodol 2009;9:5.

[6] Benchimol EI, Guttmann A, Griffiths AM, Rabeneck L, Mack DR, Brill H, et al. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. Gut 2009;58:1490–7.

[7] Guttmann A, Nakhla M, Henderson M, To T, Daneman D, Cauch-Dudek K, et al. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. Pediatr Diabetes 2010;11:122–8.

[8] 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.

[9] To T, Dell S, Dick PT, Cicutto L, Harris JK, MacLusky IB, et al. Case verification of children with asthma in Ontario. Pediatr Allergy Immunol 2006;17:69–76.

[10] Ahmed F, Janes GR, Baron R, Latts LM. Preferred provider organization claims showed high predictive value but missed substantial proportion of adults with high-risk conditions. J Clin Epidemiol 2005;58:624–8.

[11] Manuel DG, Lim JJ, Tanuseputro P, Stukel TA. How many people have had a myocardial infarction? Prevalence estimated using historical hospital data. BMC Public Health 2007;7:174.

[12] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41–4.

[13] Armstrong K, Long JA, Shea JA. Measuring adherence to mammography screening recommendations among low-income women. Prev Med 2004;38:754–60.

[14] 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;39:1733–49.

[15] Payne DC, Rose CE Jr, Aranas A, Zhang Y, Tolentino H, Weston E, et al. Assessment of anthrax vaccination data in the Defense Medical Surveillance System, 1998-2004. Pharmacoepidemiol Drug Saf 2007;16:605–11.

[16] Roberts CL, Bell JC, Ford JB, Hadfield RM, Algert CS, Morris JM. The accuracy of reporting of the hypertensive disorders of pregnancy in population health data. Hypertens Pregnancy 2008;27:285–97.

[17] Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease. Comparing administrative data and self-reports. Med Care 1997;35:932–47.

[18] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–9.

[19] von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4(10):e296.

[20] Huston P, Naylor CD. Health services research: reporting on studies using secondary data sources. CMAJ 1996;155(12):1697–709.

[21] National Library of Medicine. Medical Subject Headings. Bethesda, MD: National Institutes of Health; 1999. Available at. http://www.nlm.nih.gov/mesh/. Accessed January 2010.

[22] EMBASE. New York, NY: Elsevier Inc.; 2010. Available at. http://www.info.embase.com/. Accessed January 2010.

[23] van Walraven C, Bennett C, Forster AJ. Derivation and validation of a MEDLINE search strategy for research studies that use administrative data. Health Serv Res 2010;45:1836–45.

[24] Bernstein CN, Blanchard JF, Rawsthorne P, Wajda A. Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. Am J Epidemiol 1999;149:916–24.

[25] Liu L, Allison JE, Herrinton LJ. Validity of computerized diagnoses, procedures, and drugs for inflammatory bowel disease in a northern California managed care organization. Pharmacoepidemiol Drug Saf 2009;18:1086–93.

[26] Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309:102.

[27] Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 1997;16:981–91.

[28] Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. Am J Epidemiol 1993;138:1007–15.

[29] Green MS. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. Am J Epidemiol 1983;117:98–105.

[30] Mullooly JP, Donahue JG, DeStefano F, Baggs J, Eriksen E. Predictive value of ICD-9-CM codes used in vaccine safety research. Methods Inf Med 2008;47(4):328–35.

[31] Pekkanen J, Pearce N. Defining asthma in epidemiological studies. Eur Respir J 1999;14(4):951–7.

[32] Prosser RJ, Carleton BC, Smith MA. Identifying persons with treated asthma using administrative data via latent class modelling. Health Serv Res 2008;43:733–54.

[33] Bernatsky S, Joseph L, Pineau CA, Belisle P, Hudson M, Clarke AE. Scleroderma prevalence: Demographic variations in a population-based sample. Arthritis Care Res 2009;61:400–4.