TITLE PAGE

Title: A Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases

Authors: Jessica Widdifield[1] BSc, PhD(c); Jeremy Labrecque[2] MSc; Lisa Lix[3] PhD; J. Michael Paterson[1,4,5] MSc; Sasha Bernatsky[2] MD, FRCPC, PhD; Karen Tu[1,4] MD, MSc, Noah Ivers[1] MD, PhD(c); Claire Bombardier[1] MD, FRCPC.

Author Affiliations: 1University of Toronto, Toronto, ON; 2McGill University, Montreal, PQ; 3University of Manitoba, Winnipeg, MB; 4Institute for Clinical Evaluative Sciences, Toronto, ON; 5McMaster University, Hamilton, ON;

Financial Support: Canadian Arthritis Network Rapid Impact Platform Program: Administrative Data in Rheumatic Disease Research and Surveillance

Acknowledgements: We wish to thank: the Canadian Rheumatology Administrative Data (CANRAD) Network members, especially Dr. Diane Lacaille (University of British Columbia, Vancouver, BC); Information Specialists: Amy Faulkner, Rouhi Fazelzad, Marina Englesakis, University Health Network; We also wish to thank Dr. Debra Butt (University of Toronto, ON) for her review of the manuscript.

Dr. Lix holds a Manitoba Research Chair; Dr. Tu holds a Canadian Institutes of Health Research Primary Care Fellowship Award (2011-2013). Dr. Ivers holds a CIHR Fellowship Award in Clinical Research and a Fellowship Award from the Department of Family and Community Medicine, University of Toronto; Dr. Bombardier holds a Canada Research Chair in Knowledge Transfer for Musculoskeletal Care (2002-2016) and a Pfizer Research Chair in Rheumatology.

Manuscript word count: N= 3562/3800

Corresponding author:
J Widdifield
200 Elizabeth St 13EN 224, Toronto, ON, M5G2C4
Telephone: 416-316-1697; Fax: 416-340-4814; E-mail: Jessica.widdifield@utoronto.ca

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/acr.21993

Received: Aug 23, 2012; Revised: Jan 31, 2013; Accepted: Feb 13, 2013

ABSTRACT WORD COUNT: N=250/250

Objective: To evaluate the quality of the methods and reporting of published studies that validate administrative database algorithms for rheumatic disease case ascertainment.

Methods: We systematically searched MEDLINE, Embase and the reference lists of articles published from 1980 to 2011. We included studies that validated administrative data algorithms for rheumatic disease case ascertainment using medical record or patient-reported diagnoses as the reference standard. Each study was evaluated using published standards for the reporting and quality assessment of diagnostic accuracy, which informed the development of a methodological framework to help critically appraise and guide research in this area.

Results: Twenty-three studies met the inclusion criteria. Administrative database algorithms to identify cases were most frequently validated against diagnoses in medical records (83%). Almost two-thirds of the studies (61%) used diagnosis codes in administrative data to identify potential cases, and then reviewed medical records to confirm the diagnoses. The remaining studies did the reverse, identifying patients using a reference standard, and then testing algorithms to identify cases in administrative data. Many authors (61%) described the patient population, but few (26%) reported key measures of diagnostic accuracy (sensitivity, specificity, positive and negative predictive values). Only one-third of studies reported disease prevalence in the validation study sample.

Conclusion: Methods used in administrative data validation studies of rheumatic diseases are highly variable. Few studies report key measures of diagnostic accuracy, despite their importance for drawing conclusions about the validity of administrative database algorithms. We developed a methodological framework and recommendations for validation study conduct and reporting.

Key words:
accuracy
health administrative data
rheumatic diseases
validation
systematic review

Significance and Innovations.

- Few studies have validated administrative data algorithms for accurate identification of rheumatic diseases.

- Validation studies of administrative data algorithms often lack consistent methodology and many underreport key measures to evaluate their accuracy. This may bias results and limit their generalizability.

- Improvements to validation study methods are *essential* to fully leverage administrative data for rheumatology research. Using basic epidemiologic principles and the consensus criteria for the reporting of diagnostic accuracy studies, we present a methodological framework and suggest standards for best practice for future validation studies of rheumatic disease algorithms in administrative data.

- Higher quality studies, employing more rigorous methodology, are needed.

INTRODUCTION

Health administrative databases are an efficient source of data for population-based rheumatology research and are increasingly being used to study disease burden, disease and treatment outcomes[1] and quality of care.[2,3] The value of studies that use administrative databases for secondary research rests heavily upon the accuracy of data for ascertaining disease cases. To reduce misclassification error in case ascertainment, researchers often make use of case definitions (usually in the form of algorithms based on diagnosis codes and/or other information such as pharmacy dispensations). However, estimates of disease prevalence using different algorithms may vary by as much as 50%.[4,5] For example, an algorithm with 100% sensitivity will capture all individuals with the disease, however, an algorithm with a sensitivity of 50% will identify fewer individuals and this will reduce the disease prevalence estimate ascertained from administrative data. Therefore, confirming the accuracy of case ascertainment algorithms through a validation study (see Box 1) is an important step to improving rheumatology surveillance and research using administrative databases.

Box 1. Steps in performing an administrative database validation study[6]

PARTICIPANT SAMPLING:
- Sample potential patients to comprise a validation cohort

PARTICIPANT SELECTION (TO CLASSIFY PATIENTS AS CASES AND NON-CASES):
- Develop or define a reference standard to classify patients with and without the disease within the validation cohort.

METHODS:
- Develop one or more case ascertainment algorithms to apply to the administrative database.
- Test each administrative data algorithm against the reference standard for ability to accurately identify patients with the disease (similar to testing the accuracy of a diagnostic test).

RESULTS:
- Report measures of diagnostic accuracy: sensitivity, specificity and predictive values.
- Interpret results, recognizing tradeoffs between these measures.

Complete and accurate reporting of the methods used in validation studies is important to assess the potential biases and generalizability of results. Benchimol and colleagues[7] recently developed consensus criteria for the reporting of studies that validate administrative database algorithms, but a methodological framework to guide the conduct of such studies was not established. We performed a systematic review to identify studies that validate administrative database algorithms for rheumatic diseases and evaluate the quality of the methods and reporting of these studies. Here we summarize the various approaches to performing administrative data validation studies, we illustrate the outcome measures associated with each approach, and provide practical advice for how to achieve reliable and meaningful results.

MATERIALS AND METHODS

Our systematic review used the Consort Group's Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and followed a protocol that pre-specified study selection, eligibility criteria, quality assessment, and data abstraction.[8]

Search Strategy. A systematic literature search was conducted of Ovid MEDLINE and Embase covering the period of January 1980 to May 2011 to identify all validation studies using administrative data for rheumatology diagnoses. As the term ''health administrative data'' is not recognized as a Medical Subject Heading (MeSH) by the National Library of Medicine[9] or as an Embase subject heading[10], we developed a sensitive search strategy with the assistance of a health librarian and adapted it to each database. A complete list of the search terms is available in Supplementary Materials 1. We additionally hand-searched reference lists and performed a “grey literature” review, which included the websites of health policy units for relevant articles not captured by the electronic searches.

Study Selection. Two reviewers (JL and JW) independently screened the titles and abstracts of all studies for eligibility. The inclusion criteria were: (a) studies that addressed the validation of health administrative databases or health information systems for case ascertainment of rheumatology diagnoses using medical records and/or patient-reported diagnoses as the reference standard, and (b) the article was written in English. ‘Health administrative data’ was defined as information passively collected, often by government and health care providers, for the purpose of managing the health care of patients[11] and ‘health information system’ was defined as administrative data supplemented with detailed clinical information.[12] Rheumatology diagnoses included all diagnoses according to Medical Subject Headings.[9] There was no geographic restriction on included studied. Studies evaluating the agreement between two or more administrative data sources were excluded.

Data Abstraction for Reporting and Quality Assessment. For data abstraction, we used the STAtement for Reporting of Diagnostic accuracy (STARD)[13] and the QUality Assessment of Diagnostic Accuracy Studies (QUADAS)[14] tools. The purpose of the STARD criteria is to evaluate the *reporting* of diagnostic accuracy studies whereas the purpose of the QUADAS tool is to assess the *quality* of diagnostic accuracy studies. Both criteria were harmonized and modified to be applicable to the administrative database setting. Each individual item was adapted for this review by consensus of three authors (JL, JW and LL) and pilot tested. Items were re-phrased to increase their clarity and to be action oriented with the goal of improving validation protocol development for future research. Consensus on all issues was established prior to commencing quality assessment. Data were abstracted by two of the authors (JL and JW) and any disagreement between the two reviewers was resolved by consensus, or if necessary, by a third party. In addition, we abstracted details of the data sources [country origin, type of administrative data (e.g., inpatient, outpatient)], the specific rheumatic disease that was studied, the choice of reference standard, sample sizes, and measures of diagnostic accuracy for the algorithms tested. The data were descriptively analyzed.

Methodological Framework Development. The results of the data abstraction for reporting and quality assessment were used to develop a framework to help critically appraise and guide research in this area. Using basic epidemiologic principles and the consensus criteria for the reporting of diagnostic accuracy studies, several factors that threaten the internal and external validity were identified. We assessed the methodological merit (internal validity) by classifying the studies according to the method of patient sampling and presence or absence of a comparator group without the disease, and identified measures of diagnostic accuracy that could be computed with each approach. Second, we report the strengths and weaknesses of the various approaches to ensure the results are generalizable to the target population (i.e., external validity).

RESULTS

Studies Included. Our search identified 486 and 1063 references in MEDLINE and Embase, respectively. The number of articles assessed for inclusion and the reasons for exclusion are detailed in Figure 1. Sixteen studies were identified in the bibliographic databases and seven studies were further identified from reference lists and health policy research unit websites.

For the 23 studies identified in the published literature, Table 1 summarizes the details of the administrative data sources, diseases and reference standards. Most studies were conducted in the United States (n=15; 65%) for rheumatoid arthritis (RA) (n=13; 57%) using a combination of medical records sampled from hospitalized, ambulatory and rheumatology clinics (n=14; 61%). Most authors (n=18; 78%) evaluated algorithms that were derived from various linked data sources (inpatient, outpatient and/or prescription data). Reference standard definitions to classify individuals as true cases and non-cases came from various sources: (a) strict clinical classification criteria (e.g., 1987 RA classification criteria[15]); (n=9; 39%), (b) clinical case definitions involving diagnoses documented in medical records (n=7; 30%); (c) both clinical classification criteria and a clinical case definition (n=3; 13%); and (d) patient-reported data from surveys (n=4; 17%).

Table 2 describes the characteristics of the included studies. There is important heterogeneity with respect to the diseases evaluated, administrative data sources, types of reference standard definitions (previously described), and sample sizes. For example, sources of data included health maintenance organizations (HMOs), Medicare, Medicaid, Veteran's Affairs databases, the clinical information system of Rochester, Minnesota (Mayo Clinic) in the United States, Canadian administrative claims databases, Scandinavian population registers, and the comprehensive record linkages of the General Practice Research Database in the United Kingdom. Sample sizes ranged from 151 to 18,464 patients. The tested algorithms differed in the number (and timing) of diagnosis codes, the source of diagnoses (e.g., specialist versus general practice physician), and the use of prescription drug and procedure codes. Also, the results of diagnostic accuracy that were used to evaluate the algorithms varied considerably and appear to depend on methodology (both study design and study population). For example, studies that produced high estimates of sensitivity and PPV selected their subjects from rheumatology specialty clinics[16-19] (highlighted in table 2), which may imply that these estimates may not be representative across all populations. In general, increasing the number of diagnosis codes improved algorithm specificity; the addition of pharmacy information to diagnosis codes also improved specificity slightly, but at the cost of a dramatic reduction in sensitivity.

Quality Assessment for Reporting and Methodological Conduct. Table 3 lists the number of studies that met each of the data quality and reporting criteria (modified STARD/QUADAS criteria). Most authors (n=21; 91%) identified their research as validating administrative data to identify rheumatic diseases, and all described the data source, the setting and locations where the data were collected, and the data abstraction method. All studies described participant selection methods and just over half of the studies (n=14; 61%) reported patient clinical and/or demographic characteristics with the most common being age and sex (n=12; 52%). Very few studies reported patients' duration of disease (n=3; 13%) or co-morbid conditions (n=2; 9%).

Few studies provided study flow diagrams (n=3; 13%), statistical justification for the sample size (n=1; 4%), or confirmed that abstractors were blind to the diagnosis codes of patients (n=5; 26%). The most common statistics used to estimate diagnostic accuracy were positive predictive value (PPV) (n=14; 61%), sensitivity (n=11; 48%), specificity (n=9; 39%), and negative predictive value (NPV) (n=7; 30%). Most authors (n=16; 70%) reported results of multiple algorithms tested but only one-quarter of studies (n=6; 26%) reported at least four measures of diagnostic accuracy, and only a third (n=8; 35%) reported disease prevalence within their samples (pre-test prevalence).

Methodological Framework. Testing the accuracy of administrative database algorithms is measured on a binary scale and the results can be classified as a true positive (TP), a true negative (TN), a false positive (FP) or a false negative (FN). In order to properly evaluate a diagnostic test, both cases and non-cases are needed to populate all four cells of a 2 x 2 contingency table.

Our review identified in the published studies two main approaches to conducting administrative data validation studies (Figure 2). The defining characteristic of each approach is the manner in which patients are sampled (either by the reference standard or by diagnosis codes in administrative data) and the corresponding absence or presence of a comparator group (non-cases).

Nine studies (39%) sampled patients using the reference standard prior to testing administrative data algorithms (Figure 2: Diagram 1 A-B). Of these studies, seven applied diagnostic criteria to a random sample of patients to develop a reference standard that included cases and non-cases prior to analysis (Diagram 1A).[20,21,22,23,17,24,25] Only four studies reported the four key measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV that can be computed using this approach. Authors commonly reported kappa in place of key measures of diagnostic accuracy and one study performed multivariable logistic regression analyses to identify predictors of discordance between the reference standard and administrative database diagnosis. Two of the nine studies that sampled from a reference standard (Diagram 1B) tested their administrative data algorithms using only cases (e.g., a sample of patients with known disease status) and no comparator group without the disease[26,18]. With this approach only sensitivity can be computed.

In contrast, 16 studies (61%) initially identified patients using administrative data algorithms prior to confirming the diagnoses within the reference standard (Diagram 2A-B). Of these studies, six (26%) sampled patients with positive *and* negative test results in the administrative data (e.g., patients with *and* without specific diagnosis codes who fulfill the initial administrative data case definition) and then diagnostic criteria were applied to this sample to develop a reference standard of true cases and non-cases (Diagram 2A).[27,28,29,30,31,19] Sensitivity, specificity, and predictive values can be computed using this approach. The remaining ten studies sampled only patients with a positive test result in the administrative data source (those who fulfill the initial administrative data case definition) and then patients were subsequently classified as true cases and non-cases by the reference standard (Diagram 2B).[26,32,33,18,34,35,36,37,38,39] With this approach only the false positive fraction can be computed.

Discussion

Despite the widespread use of health administrative databases for epidemiological research in rheumatology, few studies have rigorously evaluated the accuracy of administrative data algorithms for rheumatic disease case ascertainment. We conducted a systematic review, finding 23 studies and used a modified version of the STARD/QUADAS criteria to assess the quality of the methods and reporting. Based on the variable methods to conducting validation studies, we developed a methodological framework to guide the conduct of such studies.

Thorough assessment of the internal and external validity of individual validation studies is important for assessing risk of bias. However, our quality assessment identified important heterogeneity with regards to patient sampling, reference standards to classify patients, and the measures of diagnostic accuracy that were reported. Our methodological framework also identified important heterogeneity with regards to study conduct, including the direction of patient sampling (patients are either initially sampled from the reference standard or, alternatively, from diagnosis codes in the administrative database) and the inclusion or exclusion of a comparator group without the disease.

The usefulness of validation studies depends greatly upon how potential patients are initially identified as this can impact disease prevalence, the generalizability of patient characteristics, and the measures of diagnostic accuracy that can be computed; All of which impact the outcomes of algorithms tested. When an appropriate reference standard is applied (to accurately classify cases by the reference standard) and patients are randomly sampled (ideally from a general or generalizable population), the disease prevalence approximates the population prevalence and provides unbiased estimates of sensitivity, specificity, PPV and NPV (Diagram 1A). Unfortunately, our review did not find any studies that randomly sampled patients from the general population and reported all four key measures of diagnostic accuracy. Rather, several studies randomly selected patients from specialty clinics, which can generate falsely elevated PPVs due to the high prevalence of case patients. Recognizing that the study of randomly sampled patients from the general population is not always feasible (especially for diseases of low prevalence), it remains critical for authors to report the pre-test disease prevalence ascertained from their study population to avoid errors in interpretation. As previously stated, in order to properly evaluate the characteristics of a diagnostic test and be able to report the pre-test disease prevalence, both cases and non-cases are needed to populate all four cells of a 2 x 2 contingency table. Thus, for diseases of low prevalence, strategically sampling from a source population that has a high concentration of case patients may be the only viable option. Even if the disease prevalence is falsely elevated in the validation cohort, the pre-test prevalence should approximate the post-test prevalence for the administrative data algorithm to perform well.

While the alternative approach to sampling patients by the presence or absence of diagnosis codes in administrative data (Diagram 2A) also enables computation of important parameters (true and false positives, true and false negatives), unbiased estimates of accuracy can not be generated because estimates of underlying prevalence are unknown. Furthermore, very few studies randomly sampled patients, which may have introduced verification bias and reduced external validity by impacting the spectrum of disease in the validation cohort. Sensitivity and specificity estimates are dependent on the spectrum of patients in the study sample and may vary among subpopulations defined by patient age, sex, disease duration and severity, co-morbidity or drug exposures. Unfortunately, such characteristics were not consistently reported; this is one consideration for authors wishing to optimize the usefulness of future studies. Therefore, it may not be suitable to generalize findings about sensitivity and specificity without accurate reporting of the characteristics of both cases and non-cases. In addition, because predictive values are dependent on the disease prevalence[40], future studies that wish to generalize findings regarding PPV and NPV estimates should provide accurate information on disease prevalence in the study cohort. In sum, future validation studies should follow the modified STARD recommendations[7] and provide a complete description of the patients under study (spectrum of disease). This would allow investigators to assess the effect of specific patient characteristics and disease prevalence on their results.

Different reference standards were used to classify rheumatic diseases, and this influenced the study results. In our review, medical records were the most frequently used reference source. However, their use assumes that the records contain complete information to determine a patient's disease status.[41] A related challenge in studies of rheumatic disease is that diagnoses may evolve over time: for example, a patient who initially fulfills RA criteria may later meet clinical criteria for systemic lupus. A separate problem is the use of patient-reported diagnoses (such as patient surveys) as a reference standard. However, studies that tested algorithms against patient-reported diagnoses had poor estimates of sensitivity (<50%) and substantially higher pre-test prevalence estimates. Thus, patients may not be aware of their specific underlying diagnosis or arthritis subtype.[42] Sensitivity of self-report is generally highest for medical conditions that are well-defined (from both the perspective of the layperson and the physician), and relatively easily diagnosed.[43] Therefore, clinical classification criteria and clinical case definitions derived from medical records should be encouraged as a reference standard (as opposed to using patient-reported diagnoses).

Our review identified a lack of explicit reporting of statistical methods and all but one study failed to provide statistical justification for their sample size. As there is no single statistic for the measure of diagnostic accuracy, ideally, researchers should report all relevant measures.[44] Only a quarter of the studies reported four or more measures of diagnostic accuracy with the most commonly reported being PPV and sensitivity as studies commonly sampled patients by diagnosis codes or did not include patients without disease to act as true-negatives.

The majority of authors are testing and reporting results of multiple algorithms. As the selection of algorithms for future research will vary according to their application,[45] authors of administrative data validation studies should continue to test and report results for multiple algorithms.[46] Depending on the research question, algorithms can be selected based on high sensitivity to optimize detection of cases (e.g., studying population-level burden of disease), or on high specificity and/or PPV to create a more homogeneous sample and to avoid detecting false disease cases (e.g., evaluating quality of care and/or outcomes), or the maximum combination of sensitivity and specificity. Generally, additional criteria in algorithms are expected to increase specificity at the expense of sensitivity. For example, in our review, the addition of pharmacy claims data or specialists diagnosis codes improved algorithm performance, but at the cost of dramatic reductions in sensitivity.

Limitations to this review include the inclusion of English only studies and the lack of a standardized approach to identify administrative data validation studies in the scientific literature. We did not include abstracts presented at scientific meetings because they do not contain sufficient information to properly assess study quality. Finally, we did not address the ethical issues associated with each study as ethical considerations vary by jurisdiction; however these principles may be guiding the conduct of research using administrative data.[47,48] Feasibility, practicality, or ethical considerations may have played a role in the different methodological approaches that we identified and future work is required to fully understand and address solutions to these real-world problems that may impede optimal administrative data validation methodology.

This review highlights important gaps with respect to the methodology and reporting of administrative data validation studies. Due to these gaps, each study published to date has to be interpreted individually in light of its potential for bias and generalizability. We identified strengths and weaknesses in the published literature and provide a framework to guide future study conduct in this field. Box 2 lists several recommendations for improving the design and reporting of administrative data validation studies. Our best practice statements can be used by investigators in the planning and reporting of administrative data validation studies, and by reviewers, editors, and readers to evaluate the studies to avoid errors in interpretation. Additional high quality studies, employing more rigorous methodology, would be an essential step towards improving rheumatic disease surveillance and research using administrative data.

BOX 2: SUMMARY OF RECOMMENDATIONS

1. The optimal approach to patient selection includes developing a reference standard among a random sample of patients to classify patients as cases and non-cases.
2. Authors should provide a complete description of the validation cohort, including age, sex, a description of the disease or health condition under study, distribution of disease severity, co-morbidities (if applicable) and the setting from which patients are sampled. Ideally, patients should be drawn from a sampling frame that is otherwise similar (the source population), such that the disease prevalence in the sample can approximate the disease prevalence in the administrative data source. This will enable the pre-test prevalence (disease prevalence ascertained from the reference standard) to closely approximate a post-test prevalence (disease prevalence ascertained from administrative data).
3. Clinical classification criteria and clinical case definitions derived from medical records should be encouraged as a reference standard and readers of the reference standard should be blinded to the results of the classification by administrative data for that patient.
4. Authors should test and report multiple measures of diagnostic accuracy (sensitivity, specificity and predictive values) for multiple administrative algorithms and report information on study prevalence in order to provide comprehensive information about their study.

REFERENCES

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323-37.
2. Goldfield N, Villani J. The use of administrative data as the first step in the continuous quality improvement process. Am J Med Qual 1996;11:S35-8.
3. Schwartz RM, Gagnon DE, Muri JH, Zhao QR, Kellogg R. Administrative data for quality improvement. Pediatrics 1999;103:291-301.
4. Ladouceur M, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics 2007;63:272-9.
5. Lin KJ, Garcia Rodriguez LA, Hernandez-Diaz S. Systematic review of peptic ulcer disease incidence rates: do studies without validation provide reliable estimates? Pharmacoepidemiol Drug Saf 2011;20:718-28.
6. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 1996;25:435-42.
7. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol 2011;64:821-9.
8. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol 2009;62:1006-12.
9. Medical Subject Headings. 1999. (Accessed at http://www.nlm.nih.gov/mesh/.)
10. EMBASE. Elsevier Inc., 2010. (Accessed at Available at. http://www.info.embase.com/. .)
11. Spasoff R. Epidemiologic methods for health policy. New York, NY; 1999.
12. Shortliffe E, Cimino J. Biomedical Informatics: Computer Applications in Health Care and Biomedicine (3rd edition) New York Springer; 2006.
13. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-12.
14. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9.
15. Arnett F, Edworthy S, Bloch D, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
16. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. . Arthritis Rheum 2004;51:952-7.
17. Singh JA, Holmgren AR, Krug H, Noorbaloochi S. Accuracy of the diagnoses of spondylarthritides in veterans affairs medical center databases. Arthritis Rheum 2007;57:648-55.
18. Katz J, Barrett J, Liang M, al e. Sensitivity and positive predictive value of medicare part B physician claims for rheumatologic diagnoses and procedures. Arthritis Rheum 1997;40:1594-600.
19. Bernatsky S, Linehan T, Hanly JG. The accuracy of administrative data diagnoses of systemic autoimmune rheumatic diseases. J Rheumatol 2011;38:1612-6.
20. Fowles JB, Lawthers AG, Weiner JP, Garnick DW, Petrie DS, Palmer RH. Agreement between physicians' office records and Medicare Part B claims data. Health Care Financ Rev 1995;16:189-99.
21. Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. Health Serv Res 2004;39:1839-57.
22. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Rheum 2004;51:952-7.
23. Lix L, Yogendran M, Burchill C, et al. Defining and Validating Chronic Diseases: An Administrative Data Approach. Winnipeg, MB: Manitoba Centre for Health Policy; 2006.

24. Lix L, Yogendran M, Mann J. Defining and Validating Chronic Diseases: An Administrative data Approach: An Update with ICD-10-CA. Winnipeg: Manitoba Centre for Health Policy, University of Manitoba; 2008.
25. Singh J. Discordance Between Self-report of Physician Diagnosis and Administrative Database Diagnosis of Arthritis and Its Predictors J Rheum 2009;36:1858-60.
26. Allebeck P, Ljungstrom K, Allander E. Rheumatoid arthritis in a medical information system: how valid is the diagnosis? Scand J Soc Med 1983;11:27-32.
27. Tennis P, Bombardier C, Malcolm E, Downey W. Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan Hospital Separations Database. J Clin Epidemiol 1993;46:675-83.
28. Gabriel S. The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. Arthritis Rheum 1994;37:821-3.
29. Harrold LR, Yood RA, Andrade SE, et al. Evaluating the predictive value of osteoarthritis diagnoses in an administrative database. Arthritis Rheum 2000;43:1881-5.
30. Losina E, Barrett J, Baron JA, Katz JN. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. J Clin Epidemiol 2003;56:515-9.
31. Thomas SL, Edwards CJ, Smeeth L, Cooper C, Hall AJ. How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? Arthritis Rheum 2008;59:1314-21.
32. Hakala M, Pollanen R, Nieminen P. The ARA 1987 revised criteria select patients with clinical rheumatoid arthritis from a population based cohort of subjects with chronic rheumatic diseases registered for drug reimbursement. J Rheumatol 1993;20:1674-8.
33. Gabriel SE, Crowson CS, O'Fallon WM. A mathematical model that improves the validity of osteoarthritis diagnoses obtained from a computerized diagnostic database. J Clin Epidemiol 1996;49:1025-9.
34. Pedersen M, Klarlund M, Jacobsen S, Svendsen A, Frisch M. Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. . Eur J Epidemiol 2004;19:1097-103.
35. Harrold LR, Saag KG, Yood RA, et al. Validity of gout diagnoses in administrative data. Arthritis Rheum 2007;57:103-8.
36. Icen M, Crowson CS, McEvoy MT, Gabriel SE, Maradit Kremers H. Potential misclassification of patients with psoriasis in electronic databases. J Am Acad Dermatol 2008;59:981-5.
37. Malik A, Dinnella JE, Kwoh CK, Schumacher HR. Poor validation of medical record ICD-9 diagnoses of gout in a veterans affairs database. J Rheumatol 2009;36:1283-6.
38. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. Lupus 2010;19:741-3.
39. Kim SY, Servi A, Polinski JM, et al. Validation of rheumatoid arthritis diagnoses in health care utilization data. Arthritis Res Ther 2011;13:R32.
40. Pepe M. The Statistical Evaluation of Medical Tests for Classification and Prediction: Oxford University Press; 2003.
41. Worster A, Haines T. Medical Record Review Studies: an Overview Israel Journal of Trauma, Intensive Care and Emergency Medicine 2002;2:18-23.
42. Gariepy G, Rossignol M, Lippman A. Characteristics of subjects self-reporting arthritis in a population health survey: distinguishing between types of arthritis Can J Public Health 2009 100:467-71.
43. Kehoe R, Wu SY, Leske MC, Chylack LT, Jr. Comparing self-reported and physician-reported medical history. Am J Epidemiol 1994;139:813-8.
44. Chen G, Faris P, Hemmelgarn B, Walker R, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Med Res Methodol 2009;9:5.
45. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343-9 e2.

46. Carnahan R. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. Pharmacoepidemiology and Drug Safety 2012;21:90–9.
47. Stiles P, Boothroyd R, Robst J, Ray J. Ethically using administrative data in research: Medicaid administrators current practices and best practices recommendations. Administration & Society 2011;43:171-92.
48. Stiles P, Boothroyd R. Ethical Use of Administrative Data for Research Purposes; 2012.

Figure 1. Flow chart for studies evaluated for inclusion in the systematic review.

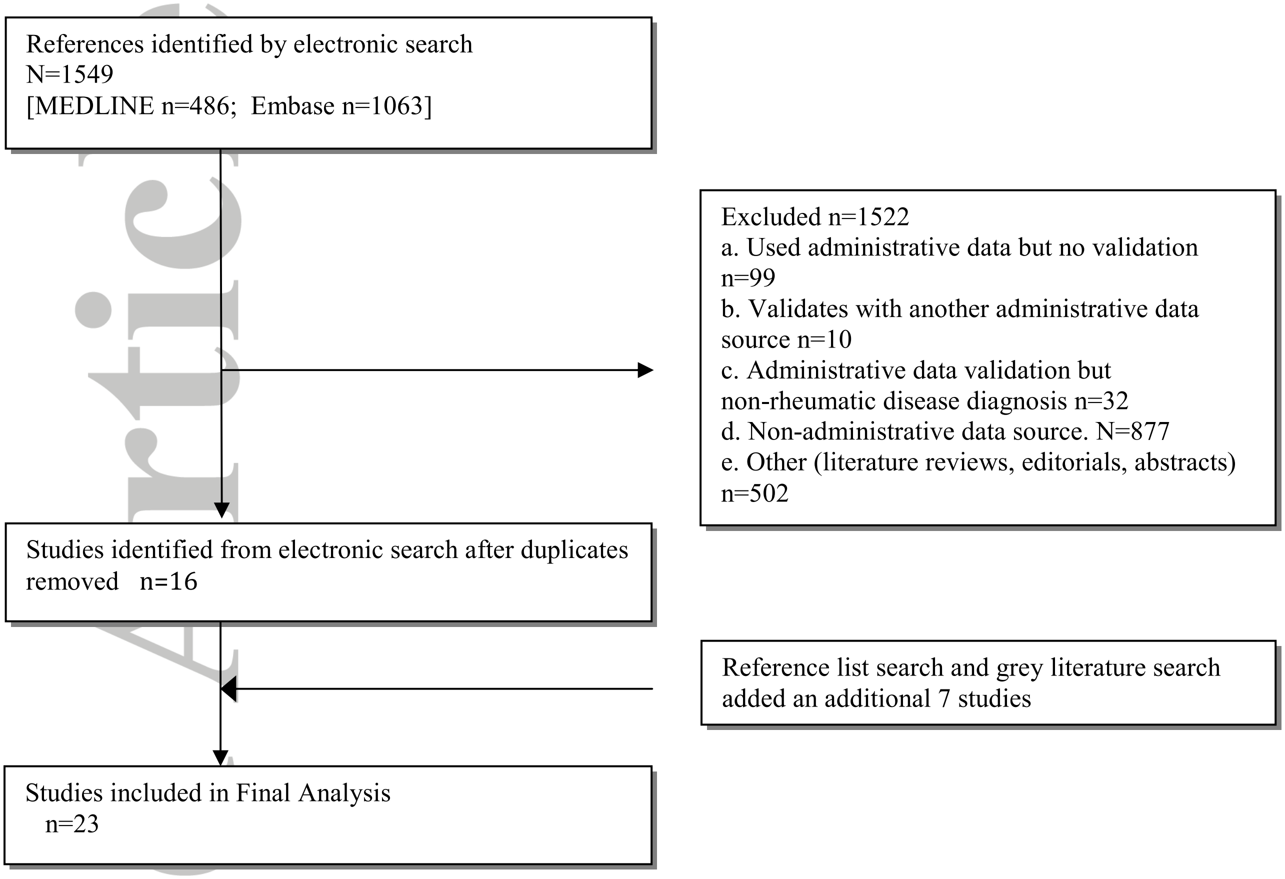

Table 1: Number of studies by Country of Data Source, Type of Secondary Data Source, Sampling source, type of records in which diagnoses for case definitions were selected from, types of specific rheumatic diseases that were evaluated and the choice of reference standard.

Characteristics	Frequency (N=23)
Studies by Country of Data Source	
USA	15 (65%)
Canada	4 (17%)
UK/Europe	4 (17%)
Type of Secondary Automated Data Source	
Health administrative *claims* database*	11 (48%)
Clinical information systems**	12 (52%)
Sampling Population Source	
Hospitalized patients only	3 (13%)
Rheumatology clinic patients only	4 (17%)
Ambulatory patients only	2 (9%)
Across all 3 domains (above)	14 (61%)
Source of Data for case definition	
Inpatient diagnosis only	3 (13%)
Outpatient diagnosis only	2 (9%)
Linked records (inpatient, outpatient +/- pharmacy/laboratory)	18 (78%)
Diagnoses***	
Rheumatoid Arthritis	13 (57%)
Osteoarthritis	6 (26%)
Connective Tissue Diseases	3 (13%)
Gout	2 (9%)
Spondylarthropathies	2 (9%)
Fibromyalgia	1 (4%)
Unspecified arthritis	2 (9%)
Reference Standard Definitions	
Clinical Classification Criteria	9 (39%)
Medical record diagnoses	7 (30%)
Both classification and medical record diagnoses	3 (13%)
Patient-Reported diagnoses	4 (17%)

*defined as information passively collected, often by government and health care providers, for the purpose of managing the health care of patients e.g., claims data);
**Clinical or health information systems (administrative data incorporating electronic health records)
***Total do not add up to 23 as several studies evaluated >1 diagnoses

Table 2: List of included articles and their characteristics.

Citation	Diagnosis	Data Source	Record Type	Reference Standard Definition(s)	Sample Size	Administrative Data Algorithm (Case Definition)	Measures of Diagnostic Accuracy
Allebeck1983 [26]	RA	Stockholm County Medical Information System; Sweden	Inpatient	Clinical classification criteria	276	≥ 1 inpatient diagnosis	**SENS:** 65 – 91%
Hakala 1993 [32]	RA	Sickness Insurance Register, Finland	Insurance + inpatient	Clinical classification criteria	151	≥ 1 diagnosis	**SENS:** 56%
Tennis 1993 [27]	RA	Saskatchewan Health, Canada	Inpatient + outpatient	Clinical classification criteria	432	≥ 1 inpatient diagnosis	**SENS:** 84%
Gabriel 1994 [28]	RA	REP, USA (health information system)	Inpatient + outpatient	Clinical classification criteria	1602	≥ 1 diagnosis	**SENS:** 89%; **SPEC**: 74%; **PPV:** 57%; **NPV**: 94%; ***κ*** = 0.54
Fowles 1995 [20]	RA	Medicare Part B, USA	Outpatient (ambulatory clinics only)	Case definition†	1596	≥ 1 outpatient diagnosis	**κ:** 0.44
Gabriel 1996 [33]	OA	REP, USA (health information system)	Inpatient + outpatient	Case definition†	387	≥ 1 diagnosis	**PPV:** 60%
						Classification tree	**SENS:** 75%; **SPEC:** 86%; **PPV:** 89%; **NPV:** 70%
Katz 1997 [18]	RA, OA, FM, SLE	Medicare Part B, USA	Outpatient (Rheumatology clinics only)	Clinical classification criteria	378	≥ 1 diagnosis (rheumatologist only)	**SENS:** 90% **PPV:** 95%
Harrold 2000 [29]	OA	HMO, USA (health information system)	Inpatient + outpatient	Clinical classification criteria	599	≥ 1 diagnosis	**PPV:** 62%
						≥ 2 diagnosis	**PPV:** 67%
						≥ 1 diagnosis AND ≥ 1 rheumatology/ orthopedic surgeon visit	**PPV:** 83%
Losina 2003 [30]	RA, OA, AVN	Medicare Part A and Part B, USA	Inpatient (recipients of total hip replacement)	Case definition†	922	≥ 1 inpatient diagnosis	**SENS:** 65% (RA), 54% (AVN), 96% (OA); **PPV:** 86-89%; **κ:** 0.73
Pedersen 2004 [34]	RA	National Patient Registry, Denmark	Inpatient + outpatient	Clinical classification criteria and case definition†	217	≥ 1 diagnosis	**SENS:** 59% (clinical case definition), 46% (ACR criteria)
Rector 2004 [21]	Arthritis[Ω]	Medicare + HMO, USA	Inpatient + outpatient + pharmacy	Patient-reported diagnoses	3633	≥ 1 diagnosis	**SENS:** 43%; **SPEC:** 87%
						≥ 2 diagnosis	**SENS:** 28%; **SPEC:** 94%
						≥ 1 outpatient diagnosis	**SENS:** 29%; **SPEC:** 93%
						≥ 1 outpatient diagnosis (primary only)	**SENS:** 23%; **SPEC:** 95%
						≥ 1 Rx	**SENS:** 32%; **SPEC:** 87%
						≥ 1 diagnosis OR ≥ 1 Rx	**SENS:** 55%; **SPEC:** 77%
						≥ 1 diagnosis AND ≥ 1 Rx	**SENS:** 20%; **SPEC:** 96%
Singh 2004 [22]	RA	VHA, USA (health information	Outpatient + pharmacy +	Case definition†	184	≥1 outpatient diagnosis	**SENS:** 100%; **SPEC:** 55.1%; **PPV**: 66.2%; **NPV**: 100%

		system)	laboratory (Rheumatology clinics only)			≥1 outpatient diagnosis AND ≥ 1 Rx (≥ 3 month duration)	**SENS:** 84.9%; **SPEC:** 82.7%; **PPV**: 81.1%; **NPV:** 86.2%
						≥1 outpatient diagnosis AND RF positive	**SENS**: 88.2%; **SPEC**: 91.4%; **PPV**: 92.6%; **NPV**:86.5%
						≥ 1 Rx (≥ 3 month duration) AND RF positive	**SENS:** 76.5%; **SPEC:** 95.7%; **PPV**: 95.6%; **NPV**: 77%
						≥1 outpatient diagnosis AND ≥ 1 Rx AND RF positive	**SENS:** 76.5%; **SPEC:** 97.1%; **PPV:** 97%; **NPV:** 77.3%
Lix 2006◊ [23,24]	RA, OA	Manitoba Population Health Research Data Repository, Canada	Inpatient + outpatient + pharmacy	Patient-reported diagnosis	5589	≥ 1 outpatient diagnosis ≤ 5 years	**SENS:** 78.1% (OA), 11.3% (RA); **SPEC:** 58.6% (OA), 99.2% (RA); **PPV:** 37.4% (OA), 55.9% (RA); **NPV:** 89.4%(OA), 92.6% (RA); **κ:** 0.27 (OA). 0.17 (RA) ; **Youden:** 0.37 (OA), 0.11(RA)
						≥ 2 outpatient diagnosis ≤ 5 years	**SENS:** 63.1% (OA), 8.3% (RA); **SPEC:** 76.2% (OA), 99.7% (RA); **PPV:** 45.7% (OA), 69.1% (RA); **NPV:** 86.7% (OA), 92.4% (RA); **κ:** 0.35 (OA), 0.13 (RA); **Youden:** 0.39 (OA), 0.08 (RA)
						≥ 1 inpatient diagnosis or ≥ 2 outpatient diagnosis ≤ 5 years	**SENS:** 63.7% (OA), 8.9% (RA); **SPEC:** 75.9% (OA), 99.7% (RA); **PPV:** 45.6% (OA), 70.7% (RA); **NPV:** 88.4% (OA), 92.4% (RA); **κ:** 0.35 (OA), 0.14 (RA); **Youden:** 0.40 (OA), 0.09 (RA)
						≥ 1 inpatient diagnosis or ≥ 2 outpatient diagnosis; OR ≥1 outpatient diagnosis and ≥ 2 Rx ≤ 5 years	**SENS**: 71.1% (OA), 9.4% (RA); **SPEC** : 70.1% (OA), 99.6% (RA); **PPV**: 42.9% (OA), 68.3% (RA); **NPV:** 88.4% (OA), 92.5% (RA); **κ** : 0.34 (OA), 0.17 (RA); **Youden:** 0.41 (OA), 0.11 (RA)
Harrold 2007 [35]	Gout	HMO database, USA (health information system)	Outpatient + pharmacy	Clinical classification criteria and case definition†	200	≥ 2 outpatient diagnosis ≥ 30 days apart	**PPV:** 61% (case definition)
						≥ 3 outpatient diagnosis	**PPV:** 64% (case definition)
						≥ 4 outpatient diagnosis	**PPV:** 67% (case definition)
						≥ 1 Rx	**PPV:** 39% (case definition)
						Seen by a rheumatologist	**PPV:** 92% (case definition)
Singh 2007 [17]	SpA	VHA, USA (health information system)	Inpatient + outpatient + pharmacy (rheumatology clinics only)	Case definition†	184	≥ 1 Diagnosis	**SENS:** 91% (AS), 100% (PsA), 71% (ReA) ; **SPEC:** 99% (AS), 100% (PsA), 100% (ReA); **PPV:** 83% (AS), 100% (PsA), 100% (ReA); **NPV:** 99% (AS), 100% (PsA), 99% (ReA); **κ:** 0.82 (AS), 1.0 (PsA), 0.83 (ReA)
						≥ 1 Diagnosis AND ≥ 1 Rx	**SENS**: 27% (AS), 65% (PsA), 57% (ReA); **SPEC:** 99% (AS), 100% (PsA), 100% (ReA); **PPV:** 75% (AS), 100% (PsA), 100% (ReA); **NPV:** 96% (AS), 97% (PsA), 98% (ReA); **κ**: 0.34 (AS), 0.77 (PsA), 0.72 (ReA)
						≥ 2 Diagnosis	**SENS:** 82% (AS), 94% (PsA), 57% (ReA); **SPEC:** 100% (AS), 100% (PsA), 100% (ReA); **PPV:** 100% (AS), 100% (PsA), 100% (ReA); **NPV:** 99% (AS), 99% (PsA), 98% (ReA); **κ**: 0.89 (AS), 0.97 (PsA), 0.72 (ReA)
						≥ 2 Diagnosis AND ≥ 1 Rx	**SENS:** 27% (AS), 59% (PsA), 57% (ReA); **SPEC:** 100% (AS), 100% (PsA), 100% (ReA); **PPV:** 100% (AS), 100% (PsA), 100% (ReA); **NPV:** 96% (AS), 96% (PsA), 98% (ReA); **κ**: 0.41

							(AS), 0.72 (PsA), 0.72 (ReA)
Icen 2008 [36]	PsO	REP, USA (health information system)	Inpatient + outpatient	Case definition†	2556	≥ 1 Diagnosis	**PPV:** 68.7% (PsO), 94.0% (PsO, vulgaris), 18.7% (PsO, dermatitis), 77.8% (PsO, guttate), 90.2% (PsO, pustular), 84.2% (seborrhiasis/sebopsoriasis), 11.8% (pityriasis and other PsO)
Thomas 2008 [31]	RA	GPRD, UK (health information system)	Inpatient + outpatient + pharmacy (ambulatory records)	Clinical classification criteria	224	≥3 Diagnosis	**SENS:** 80%; **SPEC:** 81%
						≥ 1 Diagnosis AND ≥ 2 Rx (NSAID) ≤ 6 months	**SENS:** 93%; **SPEC:** 27%
						≥ 1 Diagnosis AND ≥ 1 Rx (DMARD)	**SENS:** 78%; **SPEC:** 96%
						≥ 1 Diagnosis AND ≥ 1 Rx (oral steroid)	**SENS:** 37%; **SPEC:** 82%
						≥ 1 Diagnosis AND ≥ 1 steroid injection	**SENS:** 21%; **SPEC:** 86%
Malik 2009 [37]	Gout	VHA, USA (health information system)	Inpatient + outpatient	Clinical classification criteria	289	≥ 2 outpatient diagnosis OR ≥ 1 inpatient diagnosis AND ≥ 1 outpatient diagnosis	**PPV:** 36% (ACR criteria), 30% (Rome criteria), 33% (New York criteria)
Singh 2009 [25]	Arthritis	VHA, USA (health information system)	Inpatient + outpatient + pharmacy	Patient-reported diagnosis	18464	≥ 1 diagnosis one year prior to survey	***κ*:** 0.25
						≥ 1 diagnosis one year after survey	***κ*:** 0.23
						≥ 1 diagnosis ≤ 2 years	***κ*:** 0.28
						≥ 1 diagnosis OR ≥ 1 Rx ≤ 2 years	***κ*:** 0.32
						≥ 1 inpatient diagnosis AND ≥ 1 outpatient diagnosis AND Rx ≤ 2 years	***κ*:** 0.19
Chibnik 2010 [38]	SLE	Medicaid, USA	Inpatient + outpatient	Clinical classification criteria	234	>2 diagnosis AND > 2 nephrologist visits	**PPV** 92% (SLE), 86% (nephritis)
						>2 diagnosis AND > 2 renal diagnosis	**PPV**: 89% (SLE), 80% (nephritis)
						>2 diagnosis AND > 2 nephrologist visits OR > 2 renal Diagnosis	**PPV**: 91% (SLE), 88% (nephritis)
						>2 diagnosis AND > 2 nephrologist visits AND > 2 renal Diagnosis	**PPV**: 89% (SLE), 79% (nephritis)
Kim 2011 [39]	RA	Medicare, USA	Inpatient + outpatient + pharmacy	Case definition† + clinical classification criteria	325	≥ 2 outpatient diagnosis, ≥7 days apart	**PPV:** 55.7% (clinical case definition), 33.6% (≥ 4 ACR criteria), 42.8% (≥ 3 ACR criteria)
						≥ 2 outpatient diagnosis, ≥7 days apart AND ≥ 1 Rx (DMARD)	**PPV:** 86.2% (clinical case definition), 58.6% (≥ 4 ACR criteria), 72.4% (≥ 3 ACR criteria)
						≥ 3 outpatient diagnosis, ≥7 days apart	**PPV:** 65.5% (clinical case definition), 40.0% (≥ 4 ACR criteria), 50.9% (≥ 3 ACR criteria)
						≥ 3 outpatient diagnosis, ≥7 days apart AND ≥ 1 Rx (DMARD)	**PPV:** 87.5% (clinical case definition), 60.7% (≥ 4 ACR criteria), 75.0% (≥ 3 ACR criteria)

						≥ 2 outpatient diagnosis from a rheumatologist, ≥7 days apart	**PPV:** 66.7% (clinical case definition), 39.3% (≥ 4 ACR criteria), 50.0% (≥ 3 ACR criteria)
						≥ 2 outpatient diagnosis from a rheumatologist, ≥7 days apart AND ≥ 1 Rx (DMARD)	**PPV:** 88.9% (clinical case definition), 55.6% (≥ 4 ACR criteria), 73.3% (≥ 3 ACR criteria)
Bernatsky 2011 [19]	SLE, SSc, myositis, SS, vasculitis, and PMR	Nova Scotia Population Health Research Unit, Canada	Inpatient + outpatient (rheumatology clinics only)	Case definition†	824	≥ 1 inpatient diagnosis or ≥ 2 outpatient diagnosis ≥8 weeks apart but ≤2 years, or ≥1 outpatient diagnosis by a rheumatologist	**SENS:** 98.2% (SLE), 80.5% (SSc) 88.4% (myositis), 95.5% (SS), 93.5% (vasculitis), 99.5% (PMR); **SPEC:** 72.5% (SLE), 94.9% (SSc), 96.4% (myositis), 95.8% (SS), 95.4% (vasculitis), 92.2% (PMR)

***Abbreviations* – AS:** ankylosing spondylitis; **AVN:** avascular necrosis; **DMARD:** Disease-modifying anti-rheumatic drug; **Dx:** Diagnosis **FM:** fibromyalgia, **GPRD:** General Practice Research Database; **κ:** kappa; **NPV**: Negative Predictive Value; **NSAID:** Non-steroidal anti-inflammatory drug; **OA**: osteoarthritis; **PMR:** polymyalgia rheumatica; **PPV**: Positive Predictive Value; **PsA:** Psoriatic Arthritis; **PsO:** Psoriasis; **RA:** Rheumatoid Arthritis; **ReA:** Reactive Arthritis; **REP:** Rochester Epidemiology Project; **RF:** Rheumatoid Factor; **Rx:** pharmacy claim; **SENS**: Sensitivity; **SLE:** systemic lupus erythematosus; **SpA**: spondylarthritides; **SPEC**: Specificity; **SS:** Sjögren's syndrome; **SSc:** systemic sclerosis; **VHA:** Veterans Health Administration

Case Definitions –

Prescription (Rx) classes are not defined

† Clinical case definition based on medical record review and not as stringent as clinical classification criteria

Diagnosis codes: 714.x-720.x; except 720.1

Lix and colleagues re-ran analysis in 2008 and only results of the initial study are presented.

TABLE 3: NUMBER OF STUDIES MEETING INDIVIDUAL DATA QUALITY AND REPORTING ITEMS N=23 UNLESS OTHERWISE STATED

Section/Topic	#	Item	Frequency
Title/Abstract/ keywords	**1**	Identifies article as a study of diagnostic accuracy	23 (100%)
	2	Identifies article as a study of health administrative data	23 (100%)
Introduction	**3**	States disease ascertainment or estimating diagnostic accuracy from administrative data as a study aim?	21 (91%)
METHODS ***Participants***	**4**	Describes the data source	23 (100%)
		Describes type of records (inpatient, outpatient, linked records)	23 (100%)
		Describes setting and locations where the data were collected	23 (100%)
	5	Reports a priori sample size	4 (17%)
		Provides statistical justification for the sample size	1 (4%)
	6	PARTICIPANT SAMPLING of how patients were identified for data collection	23 (100%)
		a) Patients were first identified by diagnosis codes within the administrative data	14 (61%)
		b) Patients were first identified by clinical records or self-reported diagnosis irrespective of diagnosis codes	9 (39%)
		c) Describes a systematic sampling method?	10 (44%)
		d) Describes a non-systematic sampling method?	2 (9%)
		e) All patients within the study population were sampled	11 (48%)
	7	PARTICIPANT SELECTION: How patients were chosen for data collection *and* analysis	23 (100%)
		Describes Inclusion/Exclusion criteria	23 (100%)
		Describes who identified patients (for patients identified from medical records n=4)	4 (100%)
	8	Describes data collection	23 (100%)
		Describes use of a priori data collection form	17 (74%)
	9	Reports use of a split sample or an independent sample (re-validation using a separate cohort)	7 (30%)
Test methods	**10**	Describes the reference standard	23 (100%)
	11	Reports the number of persons reading the reference standard n=19	18 (95%)
		Describes training or expertise of persons reading reference standard (medical records) n=19	17 (90%)
	12	Reports a measure of concordance if >1 persons reading the reference standards n=11	6 (55%)
	13	Readers of the reference standard were blinded to the results of the classification by administrative data for that patient (reference standard: medical records) n=19	5 (26%)
Statistical methods	**14**	Describes explicit methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty.	15 (65%)
RESULTS ***Participants***	**15**	Reports the number of participants satisfying the inclusion/exclusion criteria	23 (100%)
	16	Provides study flow diagram	3 (13%)
	17	If patients are sampled by reference standard, reports the number of records unable to link n=9	3 (33%)
		Reports missing medical records or reports the number of patients unwilling to participate	12 (52%)
		Reports incomplete records	7 (30%)
	18	Reports clinical and demographic characteristics of the study population	14 (61%)
		a) Reports age	12 (52%)
		b) Reports sex	12 (52%)
		c) Reports disease duration	3 (13%)
		d) Reports a measure of disease severity	0 (0%)
		e) Reports co-morbid conditions	2 (9%)
Test results	**19**	Describes the characteristics of misclassified patients (false positives and/or false negatives)	7 (30%)
	20	Presents a cross tabulation of the results of the index tests by the results of the reference standard	7 (30%)
	21	Reports the pre-test prevalence in the study sample	8 (35%)
	22	Tests and Reports results of multiple algorithms	16 (70%)
Measures of Diagnostic Accuracy	**23**	Reports estimates of diagnostic accuracy	18 (83%)
		a) Reports sensitivity	11 (48%)
		b) Reports specificity	9 (39%)
		c) Reports PPV	14 (61%)
		d) Reports NPV	7 (30%)
		e) Reports $\geq$ 4 measures of diagnostic accuracy	6 (26%)
		f) Reports Youden's Index	2 (9%)
		g) Reports Kappa	7 (30%)
		h) Reports likelihood ratio(s)	1 (4%)
		i) Reports area under the receiver operating characteristic (ROC) curve	2 (9%)
		Reports 95% confidence intervals	13 (57%)
	24	Report estimates of test reproducibility of the split or independent sample(s), if done n=7	4 (57%)
DISCUSSION	**25**	Discusses the applicability of the study findings	23 (100%)

Figure 2: Various approaches to performing administrative data validation studies and the measures of diagnostic accuracy associated with each approach

Approach 1A		Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev* = $\frac{TP + FN}{TP+FP+FN+TN}$
Administrative data	Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Administrative data	Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
		Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the reference standard.
2. Classify patients as cases and non-cases.
3. Test administrative data algorithms.
4. All measures of diagnostic accuracy can be computed.

*Seven studies used this approach; *Prev = Prevalence.*

Approach 1B		Reference Standard: Cases	Reference Standard: ~~Non-Cases~~	~~Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$~~
Administrative data	Positive	True Positive	~~False Positive~~	~~PPV = $\frac{TP}{TP + FP}$~~
Administrative data	Negative	False Negative	~~True Negative~~	~~NPV = $\frac{TN}{FN + TN}$~~
		Sensitivity = $\frac{TP}{TP + FN}$	~~Specificity = $\frac{TN}{FP + TN}$~~	~~Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$~~

1. Sample patients by the reference standard.
2. All patients are cases.
3. Test administrative data algorithms.
4. Only sensitivity can be computed.

Two studies used this approach.

Approach 2A		Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$
Administrative data	Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Administrative data	Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
		Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the presence and absence of diagnosis codes.
2. Classify patients as cases and non-cases by the reference standard.
3. Test administrative data algorithms.
4. All measures of diagnostic accuracy can be computed (with limitations). *Six studies used this approach.*

Approach 2B		Reference Standard: Cases	Reference Standard: Non-Cases	~~Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$~~
Administrative data	Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Administrative data	~~Negative~~	~~False Negative~~	~~True Negative~~	~~NPV = $\frac{TN}{FN + TN}$~~
		~~Sensitivity = $\frac{TP}{TP + FN}$~~	~~Specificity = $\frac{TN}{FP + TN}$~~	~~Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$~~

1. Sample patients by the presence of diagnosis codes.
2. Classify patients as cases and non-cases by the reference standard.
3. Test administrative data algorithms.
4. Only PPV can be computed.

Ten studies used this approach.

*Citations for Approach 1A: 16-22. **Citations for Approach 1B: 23-24. ***Citations for Approach 2A: 25-30. ****Citations for Approach 2B: 23-24, 31-38.

Figure 2: Various approaches to performing administrative data validation studies and the measures of diagnostic accuracy associated with each approach

Approach IA

Administrative data	Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev* = $\frac{TP + FN}{TP+FP+FN+TN}$
Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
	Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the reference standard.
2. Classify patients as cases and non-cases.
3. Test administrative data algorithms.
4. All measures of diagnostic accuracy can be computed.

*Seven studies used this approach; *Prev = Prevalence.*

Approach IB

Administrative data	Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$
Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
	Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the reference standard.
2. All patients are cases.
3. Test administrative data algorithms.
4. Only sensitivity can be computed.

Two studies used this approach.

Approach 2A

Administrative data	Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$
Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
	Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the presence and absence of diagnosis codes.
2. Classify patients as cases and non-cases by the reference standard.
3. Test administrative data algorithms.
4. All measures of diagnostic accuracy can be computed (with limitations). *Six studies used this approach.*

Approach 2B

Administrative data	Reference Standard: Cases	Reference Standard: Non-Cases	Pre-test Prev = $\frac{TP + FN}{TP+FP+FN+TN}$
Positive	True Positive	False Positive	PPV = $\frac{TP}{TP + FP}$
Negative	False Negative	True Negative	NPV = $\frac{TN}{FN + TN}$
	Sensitivity = $\frac{TP}{TP + FN}$	Specificity = $\frac{TN}{FP + TN}$	Post-test Prev = $\frac{TP + FP}{TP+FP+FN+TN}$

1. Sample patients by the presence of diagnosis codes.
2. Classify patients as cases and non-cases by the reference standard.
3. Test administrative data algorithms.
4. Only PPV can be computed.

Ten studies used this approach.

*Citations for Approach 1A: 16-22. **Citations for Approach 1B: 23-24. ***Citations for Approach 2A: 25-30. ****Citations for Approach 2B: 23-24, 31-38.