BMJ Open

Development and full validation of classification algorithms to identify rheumatoid arthritis at population level using administrative health databases - Results from the RECord linkage On Rheumatic Diseases (RECORD) study of the Italian Society for Rheumatology

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006029
Article Type:	Research
Date Submitted by the Author:	02-Jul-2014
Complete List of Authors:	Carrara, Greta; Italian Society for Rheumatology (SIR), Epidemiology Unit Scirè, Carlo; Italian Society for Rheumatology (SIR), Epidemiology Unit; IRCCS Policlinico San Matteo Foundation, Rheumatology Unit Zambon, Antonella; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Cimmino, Marco; Clinica Reumatologica, Dipartimento di Medicina Interna Cerra, Carlo; Local Health Authority, Information system and management control Caprioli, Marta; Istituto Clinico Beato Matteo, Department of Medicine Cagnotto, Giovanni; IRCCS Policlinico San Matteo Foundation, Rheumatology Unit Nicotra, Federica; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Arfè, Andrea; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Migliazza, Simona; Local Health Authority, Information system and management control Corrao, Giovanni; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Minisola, Giovanni; Ospedale S. Camillo, Clinica Reumatologica Montecucco, carlomaurizio; IRCCS Policlinico San Matteo Foundation, Rheumatology Unit
Primary Subject Heading:	Rheumatology
Secondary Subject Heading:	Health services research
Keywords:	Rheumatology < INTERNAL MEDICINE, EPIDEMIOLOGY, PUBLIC HEALTH

Development and full validation of classification algorithms to identify rheumatoid arthritis at population level using administrative health databases

Results from the RECord linkage On Rheumatic Diseases study of the Italian Society for Rheumatology

Greta Carrara[1], Carlo A. Scirè[1], Antonella Zambon[2], Marco A Cimmino[3], Carlo Cerra[4], Marta Caprioli[5], Giovanni Cagnotto[6], Federica Nicotra[2], Andrea Arfè[2], Simona Migliazza[4], Giovanni Corrao[2], Giovanni Minisola[7], Carlomaurizio Montecucco[6]

Affiliations

1. Epidemiology Unit, Italian Society for Rheumatology, Milan, Italy
2. Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy
3. Research Laboratory and Academic Unit of Clinical Rheumatology, Department of Internal Medicine, University of Genova, Genoa, Italy.
4. Information system and management control, Local Health Authority (ASL), Pavia, Italy;
5. Istituto Clinico Beato Matteo, Department of Medicine, Vigevano, Italy;
6. Department of Rheumatology, IRCCS San Matteo Foundation, Pavia, Italy;
7. Division of Rheumatology - "San Camillo" Hospital - Rome - Italy

Running title: AHD algorithm validation RA

Word Count: 2977

Correspondence to:

Carlo Alberto Scirè

Via Turati 40 – 20121 Milano – Italy

c.scire@reumatologia.it

Key words: rheumatoid arthritis, health administrative databases, validation study, sensitivity and specificity, prevalence

ABSTRACT

Objectives: to develop and validate algorithms to identify rheumatoid arthritis (RA) patients in administrative health databases (AHD) of the national health system of the Lombardy region (Italy).

Design: Case-control and cohort diagnostic accuracy study.

Methods: In a randomly selected sample of 827 patients drawn from a tertiary rheumatology centre (training set), clinically validated diagnoses were linked to administrative data including diagnostic codes and drug prescriptions. An algorithm in steps of decreasing specificity was developed and its accuracy assessed calculating sensitivity/specificity, positive/negative predictive values (PPV/NPV), and confidence intervals (CI).

The algorithm was applied to two validating sets: 106 patients from a secondary rheumatology centre and 6087 subjects from the primary care. Alternative algorithms were developed to increase PPV at population level.

Crude and adjusted prevalence estimates taking into account misclassification were also obtained for the Lombardy Region (Italy).

Results: The algorithms included: RA certification by rheumatologist, certification for other autoimmune diseases by specialists, RA code in the Hospital Discharge Form, prescription of disease-modifying anti-rheumatic drugs and oral glucocorticoids. In the training set, a four-steps algorithm identified clinically diagnosed RA cases with a sensitivity of 96.3 (95%CI:93.6-98.2) and a specificity of 90.3 (87.4-92.7). Both external validations showed highly consistent results. More specific algorithms achieved >80% PPV at the population level. The crude RA prevalence in Lombardy was 0.52%, and estimates adjusted for misclassification ranged from 0.31% (95%CI:0.14-0.42) to 0.37% (0.25-0.47).

Conclusions: AHD are a valuable tool for the identification of RA cases at the population level, and allow to measure disease occurrence and select cohorts for retrospective longitudinal studies.

ARTICLE SUMMARY: STRENGHT AND LIMITATIONS

- this study provides a complete validation of classification algorithms for the identification of rheumatoid arthritis patients at the population level through healthcare administrative databases;

- a novel approach to adjust for the misclassification inherent to the classification algorithms is proposed for the estimation of the disease occurrence;

- classification of disease according to algorithms from administrative data are setting-specific and not directly transferred to other systems;

- proper classification algorithm validations are useful to develop consistent instruments to compare disease burden in different healthcare systems.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease that is associated with development of disability, increased mortality and significant costs to society.[1] Population-based studies help to monitor disease burden, to evaluate long-term consequences of the disease and its treatments, and to assess quality of care, for both research and governance purposes.[2]

The increasing diffusion and completeness of administrative health databases (AHD) - which record healthcare services dispensed to all members of a specific population - provide a straightforward way to perform such population-based studies in RA.[3] The validity of AHD studies primarily relies on the diagnostic accuracy of case definition. The huge methodological variability of validation studies of AHD-based classification algorithms in RA makes difficult to evaluate the potentialities of AHD for population studies of RA.[4–12]

The majority of the studies in RA develop classification algorithms sampling from populations with high prevalence of RA (e.g. rheumatology clinics), focusing on the positive predictive value (PPV) - the probability of being a true case if classified as a potential one by AHD-based criteria. Even if high PPV were achieved in this setting, it does not reflect the performance achievable in the general population, where the prevalence of RA is 30-50 fold lower. Thus, in order to develop a valid instrument to perform a population study, a validation study sampling from the same population where it will be applied is highly informative. Nevertheless no study has currently provided a full validation of algorithms developed for the classification of RA by AHD at population level.[4]

The RECord linkage On Rheumatic Diseases (RECORD) study - promoted by the Italian Society for Rheumatology – aims to set up a national surveillance system to monitor the health burden of rheumatic diseases in Italy using data from AHD.

In this paper we report the methodological approach and the results of the development and validation of different classification algorithms for RA in different levels of the health care system, including primary care. We linked clinically validated diagnoses of randomly selected samples of cases and controls with the

AHD of the Lombardy region (Italy). The prevalence of RA was then derived both as crude estimate and adjusting for the inherent misclassification.

PATIENTS AND METHODS

Reporting of this study compiles with the guidelines for diagnostic and validation studies of health administrative data.[5,6]

Study design and samples

Training set

A random sample of visits of 900 patients (300 cases and 600 controls on the basis of the diagnosis reported on the electronic medical records) aged over 16 assisted by a tertiary rheumatology clinic in Pavia (Rheumatology Unit, IRCCS Policlinico San Matteo Foundation) between 2007-2010 was extracted from the medical record database of this centre according to a case-control diagnostic design nested in the resident population Pavia.[7]

A sample size >700 subjects with a proportion of one third of cases in the training set was defined to precisely estimate both negative predictive value (NPV) >0.95 and PPV >0.50 setting alpha 0.05 and beta 0.8, as proposed by Steinberg for case-control diagnostic studies.[8]

Validating sets

Two different samples were drawn for validation purpose: one from a secondary rheumatology centre, and one from the primary care within the same catchment area. In these validating samples a cross-sectional 'cohort' diagnostic study design was applied.[9] The first validating set included a random sample of 138 patients from the clinical registries of the Rheumatology outpatient clinic of the Clinical Institute Beato Matteo of Vigevano, a secondary care rheumatology clinic. A second validating set included all the 6087 subjects extracted from the primary care registries of a convenience sample of 6 primary care physicians of the LHA of Pavia.

The study was approved by the Ethics Committee of the IRCCS Policlinico San Matteo Foundation of Pavia, and participants gave their consent to the processing of their personal data.

Test methods

Reference standard

The clinical diagnosis from medical records was considered the reference standard. Diagnoses were clinically validated by an external investigator (GiCa), who was unaware of the content and results of the algorithm. Only when the diagnosis was unclear or varied over time, patients were classified according to specific classification criteria[10], cumulatively applied until date of the randomly selected visit, based on a data collection form including: gender, age, disease duration, morning stiffness, joint involvement, rheumatoid factor, X-ray abnormalities.

Administrative healthcare database variables and record linkage

AHD are an automated system of databases consisting in: i) an archive of all residents receiving NHS assistance (virtually the whole resident population), reporting demographic and administrative data; ii) an archive of all hospital discharges from public or private hospitals, reporting all diagnoses related to the hospitalization; iii) an archive of all outpatient drug prescriptions reimbursable by the NHS.

The AHD variables useful for the identification of RA cases were defined *a priori* through a consensus process, informed by a literature review, held in February 2012 and involving 5 clinicians, 1 epidemiologist, 3 database owners and 2 statisticians.

Administrative data were extracted from the data warehouse of Pavia's LHA within an interval of ±1 year over the index date (e.g. date of clinical assessment ranging from 2006 to 2011) for rheumatologic samples, and 1st of September 2011 for the primary care sample.

Clinically validated diagnoses and administrative data from the LHA of Pavia were linked using deterministic record linkage through encrypted unique identifier code. A parallel extraction from the regional data warehouse from the 1st January 2009 to the 31st December 2010 only included the items for the classification according to the developed algorithm.

Statistical methods

Development and validation of the algorithm

For each variable identified in the consensus-based phase, sensitivity and specificity were evaluated in the training set. Combining a priori knowledge and empirical estimates of sensitivity and specificity of each variable, a first candidate algorithm was developed, including in the first step variables with high specificity.

The algorithm was then changed by sequentially including other variables with lower specificity but higher sensitivity. This process was stopped when a high sensitivity was reached at the expense of the least decrease in specificity.

Once the algorithm was fully defined, its overall accuracy was assessed by estimating its sensitivity, specificity, PPV and NPV - with exact 95% confidence intervals (CIs).

The robustness of these estimates was tested in the training set by bootstrap procedure, using 1000 samples extracted with replacement.

Two automated statistical procedures were also applied: a backward variable selection approach applied to a parametric penalized logistic regression model with multiple interaction terms and non-parametric classification trees.

A sensitivity analysis was performed by considering alternative algorithms stratified by age (with a cut-off at 65 years) and a narrower temporal range of ±6 months from the index date for the extraction of the selected variables.

Two external fully independent validations[9] were carried out using datasets from different levels of health care: a secondary rheumatology centre and primary care. The performance of the algorithm was tested estimating sensitivity, specificity, PPV and NPV.

Estimation of disease prevalence

To estimate the prevalence of RA in Lombardy (an Italian region of about 9 million residents) in 2010, the final algorithm was applied to the required variables extracted from the AHD of Lombardy, which have the same structure of the AHD of the Pavia LHA (i.e., an archive of: residents, hospital discharges and of outpatient drug prescriptions). The target population consisted of all 16 years aged or older residents.

The crude prevalence estimate was adjusted for the impact of misclassification due to the possible classification errors of the algorithm, quantified during the validation phase, by applying two different methods. The first method – proposed by Rogan & Gladen– is based on a direct relationship which expresses the adjusted value of the prevalence as a function of the crude prevalence and the sensitivity and specificity of the algorithm (Formula 1). Using the estimates of sensitivity and specificity derived from the

validation study in the general population sample the crude prevalence was corrected for the impact of misclassification and 95% CI was calculated.[11]

Formula 1.

$$adjusted\ prevalence = \frac{crude\ prevalence + \widehat{SP} - 1}{\widehat{SE} + \widehat{SP} - 1}$$

The second method – proposed by Joseph et al. – provides a more precise adjusted estimate by giving preference only to the most plausible range of values for the parameters of interest (prevalence, sensitivity and specificity of the algorithm). [12] Specifically, following the Bayesian framework, an initial quantification of the plausibility of each possible value of the parameters of interest was summarised in a probability distribution (*prior distribution*), based on estimates of sensitivity and specificity obtained from the validation study in the general population sample and on prevalence obtained from previous population studies.[13–16] The prior distribution was then updated in light of the observed data through their likelihood, leading to a posterior distribution, and the mean and 2.5% and 97.5% percentiles of the posterior distribution provide an estimate of the parameters and a corresponding credibility interval.

Data management and statistical analyses were conducted using SAS software (version 9.2; SAS Institute, Cary, NC), R Statistical Software (Foundation for Statistical Computing, Vienna, Austria) and WinBUGS software version 1.4.3 [17].

RESULTS

Variable selection

The following items were selected to be included in the algorithm: RA certification by rheumatologist, absence of certification for other chronic autoimmune diseases (ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases), ICD9-CM code 714.0 in the HDF, prescription of DMARDs including biologics and oral glucocorticoids (Supplementary file 1).

Other possible candidates not included in the analyses were: outpatient diagnostic procedures, outpatient visits, diagnostic procedures in the HDF, and blood tests.

Study samples

In the first rheumatologic sample (training sample), in 862/900 subjects (96%) record linkage between the clinical dataset and administrative data was successful. Complete information for diagnosis validation was available for 827/862 subjects (96%) (Figure 1). Demographic, disease and treatment characteristics are reported in Table 1.

Development of the algorithm

Combining the variables of progressively increasing sensitivity (Table 2), we developed a four-step algorithm that, at the final step, identified clinically diagnosed RA cases with a sensitivity of 96.35 (95%CI:93.56-98.16) and a specificity of 90.30 (95%IC:87.45-92.70) (Table 3).

Bootstrap procedure confirmed the robustness of the estimates in the training set (Table 4).

More flexible methods tested in sensitivity analyses confirmed similar accuracy: logistic penalized models with multiple interaction terms showed a sensitivity of 94.35 (95%IC: 91.36-96.68) and a specificity of 92.59 (95%IC: 90.11-94.68); and classification trees did not identified alternative pathways able to significantly improve accuracy for the classification of cases.

Validation of the algorithm

The first external validation was performed in 106 out of 138 patients, in which record linkage was successful and sufficient clinical data available. This sample included 32 cases (30.2%) with a median age of

59.5 years (IQR: 51-74) and a M:F ratio of 1:4; 30 (93.8%) cases are treated with at least one DMARDs. In the sample of cases, the median age was 62.5 years (IQR:53.5-73.5). The second validation set included 6087 subjects (40 cases of RA and 6047 controls), with a median age of 45 years (IQR:35-59) and M:F ratio of 1:1 in controls and median age of 70.5 years (IQR:57-78) with a M:F ratio of 1:3 in cases. 27/40 (67.5%) cases are treated with at least one DMARD.

The first external validation showed highly consistent results compared with the training set (Table 4). Accuracy measures in general population sample showed a substantial increase in specificity (99.8; 95%CI:99.6-99.9) and decrease in PPV (72.5; 95%CI: 58.3-84.1).

PPVs over 80% were achievable both in rheumatologic samples (85.04 (80.81 – 88.66) and 81.08 (64.84 – 92.04) in training and first validating set, respectively) and in general population restricting the algorithm to DMARD-users (PPV 85.7%; 95%CI: 63.7-96.9).

Estimation of disease prevalence

Applying the four-steps algorithm to the population of the Lombardy Region, a crude prevalence of 0.52% (0.30% for males and 0.73% for females) was obtained, with a M:F ratio of 1:3 and a peak of prevalence between 72-75 years for females, and between 75-78 years for males.

Adjusting for the estimated misclassification, prevalence fell to 0.31% (95%CI:0.18-0.45) using Gladen & Rogan's method and to 0.37% (95%CI:0.26-0.48), using Joseph's method with a plausible range of values included between 0.2% and 0.7%.

DISCUSSION

This study supports the overall validity of the administrative databases of the Italian NHS of the Lombardy region in the identification of patients with RA.

Previous studies showed the validity of AHD-based algorithms to identify cases of RA, with sensitivity and specificity ranging from 56% to 100% and from 55% to 97%.[4] The accuracy achieved in our study is highly consistent with those obtained by studies following similar methodology. In particular, Widdifield and co-workers recently developed a set of classification algorithms for RA using AHD in Ontario, Canada. These algorithms - derived in a randomly selected rheumatologic sample with a 33% prevalence of RA - showed optimal accuracy in identifying clinical diagnoses of RA, with sensitivity/specificity up to 97/85% and PPV/NPV up to 76/98%. Though we used different items to construct our instruments in our training rheumatologic sample (34% prevalence of RA), we obtained highly consistent accuracy (sensitivity/specificity 96/90% and PPV/NPV 85/98%).

Despite several algorithms are available for different AHD in different settings, none of these have been fully validated in the general population. This leads to high PPV, whose generalizability is limited to high prevalence study samples - such as for example rheumatologic outpatient samples -, where the prevalence of RA may be more than 50-fold higher.[18] Once developed the algorithm in a rheumatologic sample, we measured the diagnostic performance of the algorithm in a general population sample. As expected, PPV significantly decreased to 72%, while NPV increased over 99%. Only alternative algorithms restricted to DMARD users and to rheumatology samples were associated to PPV higher than 80%. Different algorithms with different operative characteristics may be suitable for studies with different purposes: high sensitivity for impact studies and high specificity for cohort studies.[18]

Beyond the usefulness of misclassification data to drive decision on the criterion to apply in selecting cohort of patients, sensitivity and specificity estimates are useful to adjust occurrence measures at population level.[19] This is the first study taking into account empirical misclassification in the adjustment of prevalence estimates of RA. In order to obtain unbiased estimates of prevalence we applied a first method that arithmetically adjusts the crude estimates taking into account the false positive and false

negative rates.[19] A more complex method that incorporates both a priori available information about the RA prevalence in Italy and empirical misclassification was also tested in order to improve the estimation based on the current knowledge.[12] Regardless of the method applied, prevalence estimates ranging from 0.31 to 0.37% are consistent with the expected on the basis of the literature for Italy[13–16], providing further validation to the developed tool. Using Joseph's method with a larger range of plausible values (0.2%-1%) we obtained an estimated prevalence of 0.36%.

In the design of this study we tried to limit mayor bias of diagnostic studies and to ensure external validity of the results.[4–6]

Study samples were randomly selected, in order to limit selection bias and to represent the entire spectrum of disease severity. To avoid observation bias due to differential misclassification, an independent investigator - who was unaware of the items included in the algorithm and of the subject classification – validated clinical diagnoses. AHD data suitable to be included in a diagnostic algorithm were identified through a literature-informed consensus process. We included this first step to avoid a completely data-driven algorithm, which could have overestimated the accuracy in the development sample. Only items from the domain of diagnostic codes and drug utilisation were deemed to be relevant, as most previous algorithms have done.

The robustness of our findings was also confirmed by the bootstrap procedure and by the exploration of other possible combinations of the candidate items using different statistical methods. These alternative methods achieved similar accuracy, though never significantly better than the multi-step algorithm, confirming the internal validity of the results.

The generalizability of the results was evaluated by different external validations, carried out using different healthcare levels, investigators, temporal ranges and study designs.[9]

This study has several limitations. Cross-sectional diagnostic 'case-control' studies tend to overestimate diagnostic accuracy.[7] However, accuracy was still satisfactory even when a cross-sectional diagnostic 'cohort' design was applied in a same prevalence sample. Beyond the higher prevalence of RA in the training and the first validation set, patients drawn from rheumatology samples may include subject with

more severe disease and different socio-demographic characteristics. However, the algorithm still performed well with a similar sensitivity in the general population, where the entire spectrum of the disease is represented. Furthermore drug prescriptions of elderly patients who are hospitalized in retirement homes are not tracked by the AHD, leading to a substantial underestimation of the prevalence of the disease. Another possible source of bias is linked to the choice of the reference standard. Despite the majority of the studies apply this type of case definition, clinical diagnosis is less reliable than classification criteria. However, classification criteria are developed to increase specificity in order to include patients in clinical trials and not for epidemiologic purpose.[20] We only adopted classification criteria[10] to validate unclear diagnoses. This might have introduced a verification bias in our study, slightly increasing the specificity of the algorithms. Differential misclassification may take place based on disease duration, since the probability to have diagnostic codes and DMARD prescription may increase with disease duration, leading to underrepresentation of incident RA.

In conclusion this study shows the accuracy of administrative data algorithms for identifying RA patients both in rheumatology clinics and general population in Italy. This study also supports the usefulness of misclassification data to adjust estimates and to drive the decision of the appropriate algorithm to adopt based on the study objectives. Beyond the content of the applied classification criterion, validation data are useful to select homogeneous cohorts of patients with RA across countries and health care systems, making feasible the implementation of surveillance systems aiming to improve care of patients with RA.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of the entire clinical staff including the rheumatologists of the clinical centres for their former data collection and for making available their registers of patients.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

The RECORD study is funded by the Italian Society for Rheumatology (SIR) as part of the Epidemiology Unit development program.

CONRIBUTORSHIP

CAS, AZ, MAC, GCo, CM planned the study. GM funded the study. MC, GCa collected the data. GrC, FN, AA performed the analyses. All the authors partipated in writing the manuscript and reviewed and approved the final version.

DATA SHARING STATEMENT

Extra data is available by emailing the corresponding author.

REFERENCE LIST

1 Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. *BullWorld Health Organ* 2003;**81**:646–56.

2 Woolf AD, Vos T, March L. How to measure the impact of musculoskeletal conditions. *BestPractResClinRheumatol* 2010;**24**:723–32.

3 Kim SY, Solomon DH. Use of administrative claims data for comparative effectiveness research of rheumatoid arthritis treatments. *Arthritis ResTher* 2011;**13**:129.

4 Widdifield J, Labrecque J, Lix L, *et al.* Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases. *Arthritis Care Res* 2013;**65**:1490–503. doi:10.1002/acr.21993

5 Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;**64**:821–9. doi:10.1016/j.jclinepi.2010.10.006

6 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. doi:10.7326/0003-4819-138-1-200301070-00012-w1

7 Biesheuvel CJ, Vergouwe Y, Oudega R, *et al.* Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;**8**:48. doi:10.1186/1471-2288-8-48

8 Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. *Biostatistics* 2009;**10**:94–105.

9 Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. New York: : Springer 2009.

10 MacGregor AJ, Bamber S, Silman AJ. A comparison of the performance of different methods of disease classification for rheumatoid arthritis. Results of an analysis from a nationwide twin study. *JRheumatol* 1994;**21**:1420–6.

11 Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978;**107**:71–6.

12 Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *AmJEpidemiol* 1995;**141**:263–72.

13 Cimmino MA, Parisi M, Moggiana G, *et al.* Prevalence of rheumatoid arthritis in Italy: the Chiavari Study. *AnnRheumDis* 1998;**57**:315–8.

14 Salaffi F, De AR, Grassi W. Prevalence of musculoskeletal conditions in an Italian population sample: results of a regional community-based study. I. The MAPPING study. *ClinExpRheumatol* 2005;**23**:819–28.

15 Marotto D, Nieddu ME, Cossu A, *et al.* [Prevalence of rheumatoid arthritis in North Sardinia: the Tempio Pausania's study]. *Reumatismo* 2005;**57**:273–6.

16 Della Rossa A, Neri R, Talarico R, *et al.* Diagnosis and referral of rheumatoid arthritis by primary care physician: results of a pilot study on the city of Pisa, Italy. *Clin Rheumatol* 2010;**29**:71–81. doi:10.1007/s10067-009-1285-x

17 Lunn DJ, Thomas A, Best N, *et al.* WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput* 2000;**10**:325–37.

18 Widdifield J, Bernatsky S, Paterson JM, *et al.* Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: A validation study using the medical records of rheumatologists. *Arthritis Care ResHoboken* Published Online First: 16 April 2013.PM:23592598

19 Gladen B, Rogan WJ. Misclassification and the design of environmental studies. *AmJEpidemiol* 1979;**109**:607–16.

20 Sakellariou G, Scire CA, Zambon A, *et al.* Performance of the 2010 classification criteria for rheumatoid arthritis: a systematic literature review and a meta-analysis. *PLoSOne* 2013;**8**:e56528.

Table 1. Characteristics of the training sample.

Characteristic		RA (n=301)	Non-RA (n=526)
Age (years) - mean (SD)		66.8 (13.1)	57.7 (15.7)
Female gender- n (%)		218 (72.4)	405 (77)
Disease duration <2years – n (%)		81* (27.6)	
Rheumatoid Factor positive – n (%)		151** (54.4)	
NSAID or COX-2 Inhibitor - n (%)		198 (65.8)	298 (56.7)
Glucocorticoids - n (%)		228 (75.8)	178 (33.8)
DMARDs			
	Methotrexate - n (%)	182 (60.5)	31 (5.9)
	Antimalarials - n (%)	153 (50.8)	67 (12.7)
	Sulfasalazine - n (%)	14 (4.7)	24 (4.6)
	Leflunomide - n (%)	12 (4)	0 (0)
	Other DMARDs - n (%)	5 (1.7)	7 (1.3)
	Any DMARD - n (%)	271 (90)	114 (21.7)
Biologic		30 (10)	7 (1.3)

*Data available on *293 and **277subjects*

Table 2. Empirical values of sensitivity and specificity of candidate items to be included in the algorithm in the first rheumatologic sample.

Variable	Cases		Controls		Sensitivity (95% CI)	Specificity (95% CI)
	+	-	+	-		
RA certification by Rheumatologist	232	69	19	507	77.08 (71.91-81.70)	96.39 (94.42-97.81)
Absence of certification for other autoimmune diseases*	294	7	449	77	97.67 (95.27-99.06)	14.64 (11.73-17.95)
ICD9-CM code 714 in HDF	57	244	2	524	18.94 (14.67-23.83)	99.62 (98.63-99.95)
Methotrexate	182	119	31	495	60.47 (54.69-66.03)	94.11 (91.74-95.96)
Antimalarials	153	148	67	459	50.83 (45.03-56.61)	87.26 (84.11-89.99)
Sulfasalazine	14	287	24	502	4.65 (2.57-7.68)	95.44 (93.29-97.06)
Leflunomide	12	289	0	526	3.99 (2.08-6.86)	100 (99.30-100)
Azathioprine	1	300	4	522	0.33 (0.01-1.84)	99.24 (98.06-99.79)
Cyclosporine	4	297	3	523	1.33 (0.36-3.37)	99.43 (98.34-99.88)
Anti-Tumor Necrosis Factor alpha	29	272	5	521	9.63 (6.55-13.54)	99.05 (97.80-99.69)
Abatacept	4	297	0	526	1.33 (0.36-3.37)	100 (99.30-100)
Rituximab	0	301	2	524	0	99.62 (98.63-99.95)
RA certification by Rheumatologist + ICD9 code 714 in HDF	41	260	1	525	13.62 (9.96-18.02)	99.81 (98.95-100)
RA certification by Rheumatologist + any DMARD	211	90	14	512	70.10 (64.58-75.22)	97.34 (95.57-98.54)
RA certification by Rheumatologist + ICD9 code 714 in HDF + any DMARD	38	263	1	525	12.62 (9.09-16.91)	99.81 (98.95-100)

*ICD: international classification of diseases; HDF: hospital discharge form; * ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases*

Table 3. Accuracy of the algorithm in the training sample by step

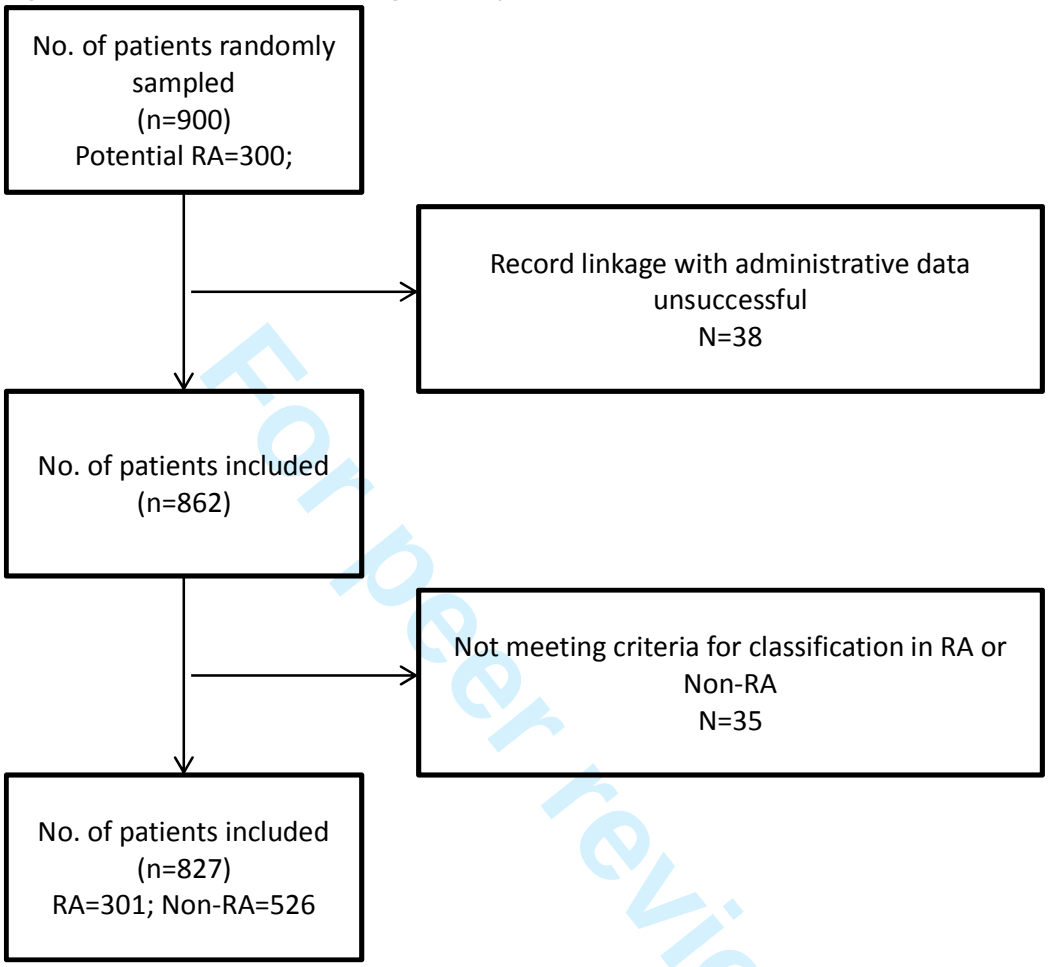
Step	Sensitivity (95% CI)	Specificity (95% CI)
Step1: RA certification by rheumatologist OR ICD9-CM code 714 in HDF OR leflunomide OR tocilizumab OR abatacept OR Gold Salts	82.39 (77.61-86.52)	96.20 (94.19-97.66)
Step2: Step 1 OR (methotrexate AND antimalarials AND no certification for other autoimmune diseases))	85.38 (80.88-89.17)	95.63 (93.51-97.21)
Step3: Step 2 OR (glucocorticoids ≥ 3 prescriptions AND antimalarials AND no certification for other autoimmune diseases))	91.36 (87.60-94.28)	92.21 (89.57-94.35)
Step4: Step 3 OR (methotrexate ≥ 3 prescriptions AND no certification for other autoimmune diseases)	96.35 (93.56-98.16)	90.30 (87.45-92.70)

Table 4. Accuracy of the algorithm in the validation samples.

	Training set- Rheumatologic sample*	Validating set - Rheumatologic sample	Validating set – General population
Sensitivity (95% CI)	96.32 (96.25-96.38)	93.75 (79.19-99.23)	92.50 (79.61-98.43)
Specificity (95% CI)	90.33 (90.24-90.41)	90.54 (81.48-96.11)	99.77 (99.61-99.87)
PPV (95% CI)	85.04 (80.81–88.66)	81.08 (64.84–92.04)	72.55 (58.26-84.11)
NPV (95% CI)	97.74 (95.99–98.86)	97.10 (89.92-99.65)	99.95 (99.85-99.99)

* bootstrap estimates

Figure 1. Flow-chart of the training set sample



STARD checklist for reporting of studies of diagnostic accuracy

(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	2
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	3
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	4
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	4
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	4
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	4
Test methods	7	The reference standard and its rationale.	5
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	5
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	NA
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	5
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	5
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	6
	13	Methods for calculating test reproducibility, if done.	NA
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	4
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	Tab1
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	Fig1
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	NA
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	Tab1
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	Tab2
	20	Any adverse events from performing the index tests or the reference standard.	NA
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	TAB2-4
	22	How indeterminate results, missing data and outliers of the index tests were handled.	8
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	9
	24	Estimates of test reproducibility, if done.	NA
DISCUSSION	25	Discuss the clinical applicability of the study findings.	12

Appendix 1 -Administrative data included:

Diagnostic codes

Life-long certification codes for chronic diseases by specialist for RA (006.714.0) and other systemic autoimmune diseases (arteritis 002.447.6; Crohn disease 009.555; ulcerative colitis 009.556; systemic lupus erythematosus 028.710.0; Sjögren syndrome 030.710.2; psoriasis or psoriatic arthritis 045.696; systemic sclerosis 047.710.1; ankylosing spondylitis 054.720.0; Behçet disease RC0210; IgA vasculitis RD0030; microscopic polyangiitis RG0020; polyarteritis nodosa RG0030; eosinophilic granulomatosis with polyangiitis RG0050; granulomatosis with polyangiitis RG0070; giant cell (temporal) arteritis RG0080; Takayasu arteritis RG0090; dermatomyositis RM0010; polymyositis RM0020; mixed connective tissue disease RM0030; undifferentiated connective tissue disease RMG010; relapsing polychondritis RM0060);

Hospital Discharge Form

ICD9-CM 714.0 code (RA)

Drug prescriptions

Anatomical Therapeutic Chemical (ATC) codes for disease modifying anti-rheumatic drugs (DMARD) (methotrexate L01BA01; antimalarials P01BA01-02; sulfasalazine A07EC01; leflunomide L04AA13; azathioprine L04AX01; cyclosporine A L04AD01; gold salts M01CB01-03; anti-tumour necrosis factor L04AB; anakinra L04AC03; tocilizumab L04AC07; abatacept L04AA24; rituximab L01XC02) and glucocorticoids (H02AB).

BMJ Open

A validation study of a new classification algorithm to identify rheumatoid arthritis using administrative health databases: case-control and cohort diagnostic accuracy studies - Results from the RECord linkage On Rheumatic Diseases study of the Italian Society for Rheumatology

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2014-006029.R1
Article Type:	Research
Date Submitted by the Author:	03-Dec-2014
Complete List of Authors:	Carrara, Greta; Italian Society for Rheumatology (SIR), Epidemiology Unit Scirè, Carlo; Italian Society for Rheumatology (SIR), Epidemiology Unit Zambon, Antonella; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Cimmino, Marco; University of Genova, Department of Internal Medicine, Research Laboratory and Academic Division of Clinical Rheumatology Cerra, Carlo; Local Health Authority, Information system and management control Caprioli, Marta; Istituto Clinico Beato Matteo, Department of Medicine Cagnotto, Giovanni; IRCCS Policlinico San Matteo Foundation, Rheumatology Unit Nicotra, Federica; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Arfè, Andrea; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Migliazza, Simona; Local Health Authority, Information system and management control Corrao, Giovanni; University of Milano-Bicocca, Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health Minisola, Giovanni; Ospedale S. Camillo, Clinica Reumatologica Montecucco, carlomaurizio; IRCCS Policlinico San Matteo Foundation, Rheumatology Unit
Primary Subject Heading:	Rheumatology
Secondary Subject Heading:	Health services research
Keywords:	Rheumatology < INTERNAL MEDICINE, EPIDEMIOLOGY, PUBLIC HEALTH

A validation study of a new classification algorithm to identify rheumatoid arthritis using administrative health databases: case-control and cohort diagnostic accuracy studies

Results from the RECord linkage On Rheumatic Diseases study of the Italian Society for Rheumatology

Greta Carrara[1], Carlo A. Scirè[1], Antonella Zambon[2], Marco A Cimmino[3], Carlo Cerra[4], Marta Caprioli[5], Giovanni Cagnotto[6], Federica Nicotra[2], Andrea Arfè[2], Simona Migliazza[4], Giovanni Corrao[2], Giovanni Minisola[7], Carlomaurizio Montecucco[6]

Affiliations

1. Epidemiology Unit, Italian Society for Rheumatology, Milan, Italy
2. Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy
3. Research Laboratory and Academic Division of Clinical Rheumatology, Department of Internal Medicine, University of Genova, Genoa, Italy.
4. Information system and management control, Local Health Authority (ASL), Pavia, Italy;
5. Istituto Clinico Beato Matteo, Department of Medicine, Vigevano, Italy;
6. Department of Rheumatology, IRCCS San Matteo Foundation, Pavia, Italy;
7. Division of Rheumatology - "San Camillo" Hospital - Rome - Italy

Running title: AHD algorithm validation RA

Word Count: 3263

Correspondence to:

Carlo Alberto Scirè

Via Turati 40 – 20121 Milano – Italy

c.scire@reumatologia.it

Key words: rheumatoid arthritis, health administrative databases, validation study, sensitivity and specificity, prevalence

ABSTRACT

Objectives: to develop and validate a new algorithm to identify rheumatoid arthritis (RA) patients and estimate disease prevalence using administrative health databases (AHD) of the Italian Lombardy region.

Design: Case-control and cohort diagnostic accuracy study.

Methods: In a randomly selected sample of 827 patients drawn from a tertiary rheumatology centre (training set), clinically validated diagnoses were linked to administrative data including diagnostic codes and drug prescriptions. An algorithm in steps of decreasing specificity was developed and its accuracy assessed calculating sensitivity/specificity, positive/negative predictive values (PPV/NPV), with corresponding confidence intervals (CI).

The algorithm was applied to two validating sets: 106 patients from a secondary rheumatology centre and 6087 subjects from the primary care. Alternative algorithms were developed to increase PPV at population level.

Crude and adjusted prevalence estimates taking into account algorithm misclassification rates were obtained for the Lombardy Region.

Results: The algorithms included: RA certification by rheumatologist, certification for other autoimmune diseases by specialists, RA code in the Hospital Discharge Form (HDF), prescription of disease-modifying anti-rheumatic drugs and oral glucocorticoids. In the training set, a four-steps algorithm identified clinically diagnosed RA cases with a sensitivity of 96.3 (95%CI:93.6-98.2) and a specificity of 90.3 (87.4-92.7). Both external validations showed highly consistent results. More specific algorithms achieved >80% PPV at the population level. The crude RA prevalence in Lombardy was 0.52%, and estimates adjusted for misclassification ranged from 0.31% (95%CI:0.14-0.42) to 0.37% (0.25-0.47).

Conclusions: AHD are a valuable tool for the identification of RA cases at the population level, and allow to estimate disease prevalence and potentially to select retrospective cohorts.

ARTICLE SUMMARY: STRENGHT AND LIMITATIONS

- this study provides a complete validation of classification algorithms for the identification of rheumatoid arthritis patients at the population level through healthcare administrative databases;

- two different approaches were applied in this study to estimate RA prevalence accounting for misclassification inherent to the classification algorithm;

- classification of disease according to algorithms from administrative data are setting-specific and not directly transferred to other systems;

- proper classification algorithm validations are useful to develop consistent instruments to compare disease burden in different healthcare systems.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease that is associated with development of disability, increased mortality and significant costs to society.[1] Population-based studies help to monitor disease burden, to evaluate long-term consequences of the disease and its treatments, and to assess quality of care, for both research and governance purposes.[2]

The increasing diffusion and completeness of administrative health databases (AHD) - which record healthcare services dispensed to all members of a specific population - provide a straightforward way to perform such population-based studies in RA.[3–5] The validity of AHD studies primarily relies on the diagnostic accuracy of case definition. The huge methodological variability of validation studies of AHD-based classification algorithms in RA makes difficult to evaluate the potentialities of AHD for population studies of RA.[6–16]

The majority of the studies in RA develops classification algorithms sampling from populations with high prevalence of RA (e.g. rheumatology clinics), focusing on the positive predictive value (PPV) - the probability of being a true case if classified as a potential one by AHD-based criteria. Even if high PPV was achieved in this setting, it does not reflect the performance achievable in the general population, where the prevalence of RA is 30-50 fold lower. Thus, in order to develop a valid instrument to perform a population study, a validation study sampling from the same population where it will be applied is highly informative. Nevertheless no study has currently provided a full validation of algorithms developed for the classification of RA by AHD at population level.[15]

The RECord linkage On Rheumatic Diseases (RECORD) study - promoted by the Italian Society for Rheumatology – aims to set up a national surveillance system to monitor the health burden of rheumatic diseases in Italy using data from AHD. The RECORD study of RA is structured in 3 phases: the first phase aims to evaluate the frequency of the disease; the second phase to evaluate the impact of the disease and its treatment on hard disease outcomes at population level; and the third phase to evaluate the quality of care delivered to RA patients.

In order to reach the objectives of the first step of the RECORD study, we report the methodological approach and the results of the development and validation of different classification algorithms for RA at different levels of the health care system, including primary care. We linked clinically validated diagnoses of randomly selected samples of cases and controls with the AHD of the Lombardy region (Italy). The prevalence of RA was then derived both as crude estimate and adjusting for the inherent misclassification.

PATIENTS AND METHODS

Reporting of this study compiles with the guidelines for diagnostic and validation studies of health administrative data.[17,18]

Study design and samples

Training set

A random sample of visits of 900 outpatients (300 cases with RA and 600 controls with rheumatic diseases other than RA, on the basis of the diagnosis reported on the electronic medical records) aged over 16 years and assisted by a tertiary rheumatology clinic (Rheumatology Unit, IRCCS Policlinico San Matteo Foundation, Pavia) between 2007-2010 was extracted from the medical record database of this centre according to a case-control diagnostic design nested in the resident population of Pavia.[19]

A sample size >700 subjects with a proportion of one third of cases in the training set was defined to precisely estimate both negative predictive value (NPV) >0.95 and PPV >0.50, setting alpha at 0.05 and beta at 0.8, as proposed by Steinberg et al. for case-control diagnostic studies.[20]

Validating sets

Two different samples were drawn for validation purpose: one from a secondary rheumatology centre and one from the primary care (general population sample) within the same catchment area. In these validating samples, a cohort diagnostic study design was applied.[21] The first validating set included a random sample of 138 patients from the electronic medical records of the Rheumatology outpatient clinic of the Clinical Institute Beato Matteo of Vigevano, a secondary care rheumatology clinic. A second validating set included all the 6087 subjects extracted from the primary care electronic medical records of a convenience sample of 6 primary care physicians of the LHA of Pavia.

The study was approved by the Ethics Committee of the IRCCS Policlinico San Matteo Foundation of Pavia, and participants gave their consent to the processing of their personal data.

Test methods

Reference standard

The clinical diagnosis from medical records was considered the reference standard. Diagnoses were clinically validated by an external investigator (GiCa), who was unaware of the content and results of the algorithm. Only when the diagnosis was unclear or varied over time, patients were classified according to specific classification criteria[22], cumulatively applied until date of the randomly selected visit, based on a data collection form including: gender, age, disease duration, morning stiffness, joint involvement, rheumatoid factor, X-ray abnormalities.

Administrative healthcare database variables and record linkage

AHD are an automated system of databases consisting in: i) an archive of all residents receiving NHS assistance (virtually the whole resident population), reporting demographic and administrative data; ii) an archive including all the certifications of chronic diseases for the exemption from co-payment; iii) an archive of all HDF from public or private hospitals, reporting all diagnoses related to the hospitalization; iv) an archive of all outpatient drug prescriptions reimbursable by the NHS.

The AHD variables useful for the identification of RA cases were defined *a priori* through a consensus process, informed by a literature review, held in February 2012 and involving 5 clinicians, 1 epidemiologist, 3 database owners and 2 statisticians. The literature was searched via Pubmed using a combination of free-text and MeSH terms regarding 'rheumatoid arthritis' and 'administrative database'. The relevant variable were selected among a list of items extracted from the retrieved literature [3,6–12,23] (see Appendix 1). These variables represented the set of potential index texts to be included in the classification algorithm: RA certification by rheumatologist and certification for other chronic autoimmune diseases (ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases), ICD9-CM code 714.0 in the HDF, prescription of DMARDs including biologics and oral glucocorticoids, outpatient diagnostic procedures, outpatient visits, diagnostic procedures in the HDF, blood tests and instrumental tests (as radiographs).

The following items were selected to be included in the algorithm: RA certification by rheumatologist, absence of certification for other chronic autoimmune diseases, ICD9-CM code 714.0 in the HDF, prescription of DMARDs including biologics and oral glucocorticoids (see Appendix 2).

Administrative data (selected items needed to create the algorithm) relative to patients included in both rheumatologic samples and general population sample were extracted from the data warehouse of Pavia's LHA within an interval of ±1 year over the index date (i.e. date of clinical assessment ranging from 2006 to 2011) for rheumatologic samples, and ±1 year over 1st of September 2011 for the primary care sample. Clinically validated diagnoses and administrative data from the LHA of Pavia were linked using deterministic record linkage through encrypted unique identifier code (only subjects successfully linked were retained for the analyses). A parallel extraction from the regional data warehouse from the 1st January 2009 to the 31st December 2010 only included the items needed for the classification according to the developed algorithm, in order to estimate the prevalence of RA.

Statistical methods

Development and validation of the algorithm

For each variable identified in the consensus-based phase, sensitivity and specificity were evaluated in the training set. Combining a priori knowledge and empirical estimates of sensitivity and specificity of each variable, a first candidate algorithm was developed, including in the first step variables with high specificity. The algorithm was then changed by sequentially including other variables with lower specificity but higher sensitivity. This process was stopped when a high sensitivity was reached at the expense of the least decrease in specificity.

For example, RA certification by rheumatologist, ICD9-CM code 714 in HDF and some drugs, like leflunomide and abatacept, showed high specificity. Knowing that other drugs, like tocilizumab and gold salts also have high specificity, we combined these items in the first step. Afterwards, items that are more sensitive and less specific, like methotrexate, antimalarials drugs and glucocorticoids, were combined in the successive steps.

Once the algorithm was fully defined, its overall accuracy was assessed by estimating sensitivity, specificity, PPV and NPV - with exact 95% confidence intervals (CIs).

The robustness of these estimates was tested in the training set by bootstrap procedure, using 1000 samples extracted with replacement.

Two automated statistical procedures were also applied: a backward variable selection approach applied to a parametric penalized logistic regression model with multiple interaction terms and non-parametric classification trees.

A sensitivity analysis was performed by considering alternative algorithms stratified by age (with a cut-off at 65 years) and a narrower temporal range of ±6 months from the index date for the extraction of the selected variables.

Two external fully independent validations[21] were carried out using datasets from different levels of health care: a secondary rheumatology centre and primary care. The performance of the algorithm was tested estimating sensitivity, specificity, PPV and NPV.

Estimation of disease prevalence

To estimate the prevalence of RA in Lombardy (an Italian region of about 9 million residents) in 2010, the final algorithm was applied to the required variables extracted from the AHD of Lombardy, which have the same structure of the AHD of the Pavia LHA (i.e., an archive of: residents, certifications of exemption, hospital discharges and of outpatient drug prescriptions). The target population consisted of all residents aged 16 years or older.

The crude prevalence estimate was adjusted for the impact of misclassification due to the possible classification errors of the algorithm, quantified during the validation phase, by applying two different methods. The first method – proposed by Rogan & Gladen– is based on a direct relationship, which expresses the adjusted value of the prevalence as a function of the crude prevalence and the sensitivity and specificity of the algorithm (Formula 1). Using the estimates of sensitivity and specificity derived from the validation study in the general population sample the crude prevalence was corrected for the impact of misclassification and 95% CI was calculated.[24]

Formula 1.

$$adjusted\ prevalence\ = \frac{crude\ prevalence + \widehat{SP} - 100}{\widehat{SE} + \widehat{SP} - 100}$$

The second method – proposed by Joseph et al. – provides a more precise adjusted estimate by giving preference only to the most plausible range of values for the parameters of interest (prevalence, sensitivity

and specificity of the algorithm). [25] Specifically, following the Bayesian framework, an initial quantification of the plausibility of each possible value of the parameters of interest was summarised in a probability distribution (*prior distribution*), based on estimates of sensitivity and specificity obtained from the validation study in the general population sample and on prevalence obtained from previous population studies.[26–29] The prior distribution was then updated in light of the observed data through their likelihood, leading to a posterior distribution, and the mean and 2.5% and 97.5% percentiles of the posterior distribution provide an estimate of the parameters and a corresponding credibility interval. Data management and statistical analyses were conducted using SAS software (version 9.2; SAS Institute, Cary, NC), R Statistical Software (Foundation for Statistical Computing, Vienna, Austria) and WinBUGS software version 1.4.3 [30].

RESULTS

Study samples

In the first rheumatologic sample (training sample), in 862/900 subjects (96%) record linkage between the clinical dataset and administrative data was successful. Complete information for diagnosis validation (criteria for classification in RA e Non-RA) was available for 827/862 subjects (96%) (Figure 1). Demographic, disease and treatment characteristics are reported in Table 1.

Development of the algorithm

Combining the variables of progressively increasing sensitivity (Table 2), we developed a final 4-step algorithm that identifies clinically diagnosed RA cases with a sensitivity of 96.35 (95%CI:93.56-98.16) and a specificity of 90.30 (95%CI:87.45-92.70) (Table 3).

Bootstrap procedure confirmed the robustness of the estimates in the training set (Table 4).

More flexible methods tested in sensitivity analyses confirmed similar accuracy: logistic penalized models with multiple interaction terms showed a sensitivity of 94.35 (95%CI: 91.36-96.68) and a specificity of 92.59 (95%CI: 90.11-94.68); classification trees did not identified alternative pathways able to significantly improve accuracy for the classification of cases.

Validation of the algorithm

The first external validation was performed in 106 out of 138 patients, in which record linkage was successful and sufficient clinical data available. This sample included 32 cases (30.2%) with a median age of 62.5 years (IQR: 53.5-73.5) and a M:F ratio of 1:2; 30 (93.8%) cases were treated with at least one DMARDs. In the sample of controls, the median age was 57 years (IQR:51-74). The second validation set included 6087 subjects (40 cases of RA and 6047 controls), with a median age of 70.5 years (IQR:57-78) with a M:F ratio of 1:3 in cases and median age of 45 years (IQR:35-59) and M:F ratio of 1:1 in controls. 27/40 (67.5%) cases were treated with at least one DMARD.

The first external validation showed highly consistent results compared with the training set (Table 4). Accuracy measures in general population sample showed a substantial increase in specificity (99.8; 95%CI:99.6-99.9) and decrease in PPV (72.5; 95%CI: 58.3-84.1).

PPVs over 80% were achievable both in rheumatologic samples (85.04 (80.81 – 88.66) and 81.08 (64.84 – 92.04) in training and first validating set, respectively) and in general population restricting the algorithm to DMARD-users (PPV 85.7%; 95%CI: 63.7-96.9).

Estimation of disease prevalence

Applying the four-steps algorithm to the population of the Lombardy Region, a crude prevalence of 0.52% (0.30% for males and 0.73% for females) was obtained, with a M:F ratio of 1:3 and a peak of prevalence between 72-75 years for females, and between 75-78 years for males.

Adjusting for the estimated misclassification, prevalence fell to 0.31% (95%CI:0.18-0.45) using Gladen & Rogan's method (in Formula 1: crude prevalence=0.52%, sensibility=92.5%, specificity=99.77%) and to 0.37% (95%CI:0.26-0.48), using Joseph's method with a plausible range of values included between 0.2% and 0.7%.

DISCUSSION

This study supports the overall validity of the administrative databases of the Italian NHS of the Lombardy region in the identification of patients with RA.

Previous studies showed the validity of AHD-based algorithms to identify cases of RA, with sensitivity and specificity ranging from 56% to 100% and from 55% to 97%.[15] The accuracy achieved in our study is highly consistent with those obtained by studies following similar methodology. In particular, Widdifield and co-workers recently developed a set of classification algorithms for RA using AHD in Ontario, Canada. These algorithms - derived in a randomly selected rheumatologic sample with a 33% prevalence of RA - showed optimal accuracy in identifying clinical diagnoses of RA, with sensitivity/specificity up to 97/85% and PPV/NPV up to 76/98%. Though we used different items to construct our instruments in our training rheumatologic sample (34% prevalence of RA), we obtained highly consistent accuracy (sensitivity/specificity 96/90% and PPV/NPV 85/98%).

Despite several algorithms are available for different AHD in different settings, none of these have been fully validated in the general population. This leads to high PPV, whose generalizability is limited to high prevalence study samples - such as for example rheumatologic outpatient samples -, where the prevalence of RA may be more than 50-fold higher.[14] Once developed the algorithm in a rheumatologic sample, we measured the diagnostic performance of the algorithm in a general population sample. As expected, PPV significantly decreased to 72%, while NPV increased over 99%. Only alternative algorithms restricted to DMARD users and to rheumatology samples were associated to PPV higher than 80%. Different algorithms with different operative characteristics may be suitable for studies with different purposes: high sensitivity for impact studies and high specificity for cohort studies.[14]

Beyond the usefulness of misclassification data to drive decision on the criterion to apply in selecting cohort of patients, sensitivity and specificity estimates are useful to adjust occurrence measures at population level.[31] This is the first study taking into account empirical misclassification in the adjustment of prevalence estimates of RA. In order to obtain unbiased estimates of prevalence we applied a first method that arithmetically adjusts the crude estimates taking into account the false positive and false

negative rates.[31] A more complex method that incorporates both a priori available information about the RA prevalence in Italy and empirical misclassification was also tested in order to improve the estimation based on the current knowledge.[25] Regardless of the method applied, prevalence estimates ranging from 0.31 to 0.37% are consistent with the expected on the basis of the literature for Italy[26–29], providing further validation to the developed tool. Using Joseph's method with a larger range of plausible values (0.2%-1%) we obtained an estimated prevalence of 0.36%.

In the design of this study we tried to limit mayor bias of diagnostic studies and to ensure external validity of the results.[15,17,18]

Study samples were randomly selected, in order to limit selection bias and to represent the entire spectrum of disease severity. To avoid observation bias due to differential misclassification, an independent investigator - who was unaware of the items included in the algorithm and of the subject classification – validated clinical diagnoses. AHD data suitable to be included in a diagnostic algorithm were identified through a literature-informed consensus process. We included this first step to avoid a completely data-driven algorithm, which could have overestimated the accuracy in the development sample. Only items from the domain of diagnostic codes and drug utilisation were deemed to be relevant, as most previous algorithms have done.

The robustness of our findings was also confirmed by the bootstrap procedure and by the exploration of other possible combinations of the candidate items using different statistical methods. These alternative methods achieved similar accuracy, though never significantly better than the multi-step algorithm, confirming the internal validity of the results.

The generalizability of the results was evaluated by different external validations, carried out using different healthcare levels, investigators, temporal ranges and study designs.[21]

This study has several limitations. Cross-sectional diagnostic 'case-control' studies tend to overestimate diagnostic accuracy.[19] However, accuracy was still satisfactory even when a cross-sectional diagnostic 'cohort' design was applied in a same prevalence sample. Beyond the higher prevalence of RA in the training and the first validation set, patients drawn from rheumatology samples may include subjects with

more severe disease and different socio-demographic characteristics. However, the algorithm still performed well with a similar sensitivity in the general population, where the entire spectrum of the disease is represented. Furthermore drug prescriptions of elderly patients who are hospitalized in retirement homes are not tracked by the AHD, leading to a substantial underestimation of the prevalence of the disease. Another possible source of bias is linked to the choice of the reference standard. Despite the majority of the studies apply this type of case definition, clinical diagnosis is less reliable than classification criteria. However, classification criteria are developed to increase specificity in order to include patients in clinical trials and not for epidemiologic purpose.[32] We only adopted classification criteria[22] to validate unclear diagnoses. This might have introduced a verification bias in our study, slightly increasing the specificity of the algorithms. Differential misclassification may take place based on disease duration, since the probability to have diagnostic codes and DMARD prescription may increase with disease duration, leading to underrepresentation of incident RA.

In conclusion this study shows the accuracy of administrative data algorithms for identifying RA patients both in rheumatology clinics and general population in Italy. This study also supports the usefulness of misclassification data to adjust estimates and to drive the decision of the appropriate algorithm to adopt based on the study objectives. Beyond the content of the applied classification criterion, validation data are useful to select homogeneous cohorts of patients with RA across countries and health care systems, making feasible the implementation of surveillance systems aiming to improve care of patients with RA.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of the entire clinical staff including the rheumatologists of the clinical centres for their former data collection and for making available their registers of patients.

CONTRIBUTORSHIP STATEMENT

CAS, AZ, MAC, GiCo, GM, CM planned the study. CC, MC, GiCa and SM collected data. GrCa, FN, AA analysed data. GrCa wrote the first draft, and all the authors critically revised and approved the final manuscript. GM was involved in obtaining of funding.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

The RECORD study is funded by the Italian Society for Rheumatology (SIR) as part of the Epidemiology Unit development program.

DATA SHARING

No additional data available.

REFERENCE LIST

1 Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. *BullWorld Health Organ* 2003;**81**:646–56.

2 Woolf AD, Vos T, March L. How to measure the impact of musculoskeletal conditions. *BestPractResClinRheumatol* 2010;**24**:723–32.

3 Kim SY, Solomon DH. Use of administrative claims data for comparative effectiveness research of rheumatoid arthritis treatments. *Arthritis ResTher* 2011;**13**:129.

4 Corrao G, Cesana G, Merlino L. Pharmacoepidemiological research and the linking of electronic healthcare databases available in the Italian region of Lombardy. *BioMed Stat Clin Epidemiol* 2008;**2**:117–25.

5 Scotti L, Arfè A, Zambon A, *et al.* Cost-effectiveness of enhancing adherence with oral bisphosphonates treatment in osteoporotic women: an empirical approach based on healthcare utilisation databases. *BMJ Open* 2014;**4**:e003758. doi:10.1136/bmjopen-2013-003758

6 Gabriel SE. The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. *Arthritis Rheum* 1994;**37**:821–3.

7 Katz JN, Barrett J, Liang MH, *et al.* Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. *Arthritis Rheum* 1997;**40**:1594–600.

8 MacLean CH, Park GS, Traina SB, *et al.* Positive predictive value (PPV) of an administrative data-based algorithm for the identification of patients with rheumatoid arthritis (RA). *Arthritis Rheum* 2001;**44**:1–520.

9 Losina E, Barrett J, Baron JA, *et al.* Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. *JClinEpidemiol* 2003;**56**:515–9.

10 Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum* 2004;**51**:952–7.

11 Pedersen M, Klarlund M, Jacobsen S, *et al.* Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. *EurJEpidemiol* 2004;**19**:1097–103.

12 Thomas SL, Edwards CJ, Smeeth L, *et al.* How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? *Arthritis Rheum* 2008;**59**:1314–21.

13 Kim SY, Servi A, Polinski JM, *et al.* Validation of rheumatoid arthritis diagnoses in health care utilization data. *Arthritis Res Ther* 2011;**13**:R32. doi:10.1186/ar3260

14 Widdifield J, Bernatsky S, Paterson JM, *et al.* Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: A validation study using the medical records of rheumatologists. *Arthritis Care ResHoboken* Published Online First: 16 April 2013.PM:23592598

15 Widdifield J, Labrecque J, Lix L, *et al.* Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases. *Arthritis Care Res* 2013;**65**:1490–503. doi:10.1002/acr.21993

16 Chung CP, Rohan P, Krishnaswami S, *et al.* A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine* 2013;**31, Supplement 10**:K41–K61. doi:10.1016/j.vaccine.2013.03.075

17 Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;**64**:821–9. doi:10.1016/j.jclinepi.2010.10.006

18 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. doi:10.7326/0003-4819-138-1-200301070-00012-w1

19 Biesheuvel CJ, Vergouwe Y, Oudega R, *et al.* Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;**8**:48. doi:10.1186/1471-2288-8-48

20 Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. *Biostatistics* 2009;**10**:94–105.

21 Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. New York: : Springer 2009.

22 MacGregor AJ, Bamber S, Silman AJ. A comparison of the performance of different methods of disease classification for rheumatoid arthritis. Results of an analysis from a nationwide twin study. *JRheumatol* 1994;**21**:1420–6.

23 Ng B, Aslam F, Petersen NJ, *et al.* Identification of rheumatoid arthritis patients using an administrative database: a Veterans Affairs study. *Arthritis Care ResHoboken* 2012;**64**:1490–6.

24 Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978;**107**:71–6.

25 Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *AmJEpidemiol* 1995;**141**:263–72.

26 Cimmino MA, Parisi M, Moggiana G, *et al.* Prevalence of rheumatoid arthritis in Italy: the Chiavari Study. *AnnRheumDis* 1998;**57**:315–8.

27 Salaffi F, De AR, Grassi W. Prevalence of musculoskeletal conditions in an Italian population sample: results of a regional community-based study. I. The MAPPING study. *ClinExpRheumatol* 2005;**23**:819–28.

28 Marotto D, Nieddu ME, Cossu A, *et al.* [Prevalence of rheumatoid arthritis in North Sardinia: the Tempio Pausania's study]. *Reumatismo* 2005;**57**:273–6.

29 Della Rossa A, Neri R, Talarico R, *et al.* Diagnosis and referral of rheumatoid arthritis by primary care physician: results of a pilot study on the city of Pisa, Italy. *Clin Rheumatol* 2010;**29**:71–81. doi:10.1007/s10067-009-1285-x

30 Lunn DJ, Thomas A, Best N, *et al.* WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput* 2000;**10**:325–37.

31 Gladen B, Rogan WJ. Misclassification and the design of environmental studies. *AmJEpidemiol* 1979;**109**:607–16.

32 Sakellariou G, Scire CA, Zambon A, *et al.* Performance of the 2010 classification criteria for rheumatoid arthritis: a systematic literature review and a meta-analysis. *PLoSOne* 2013;**8**:e56528.

Table 1. Characteristics of the training sample.

Characteristic		RA (n=301)	Non-RA (n=526)
Age (years) - mean (SD)		66.8 (13.1)	57.7 (15.7)
Female gender- n (%)		218 (72.4)	405 (77)
Disease duration <2years – n (%)		81* (27.6)	
Rheumatoid Factor positive – n (%)		151** (54.4)	
NSAID or COX-2 Inhibitor - n (%)		198 (65.8)	298 (56.7)
Glucocorticoids - n (%)		228 (75.8)	178 (33.8)
DMARDs			
	Methotrexate - n (%)	182 (60.5)	31 (5.9)
	Antimalarials - n (%)	153 (50.8)	67 (12.7)
	Sulfasalazine - n (%)	14 (4.7)	24 (4.6)
	Leflunomide - n (%)	12 (4)	0 (0)
	Other DMARDs - n (%)	5 (1.7)	7 (1.3)
	Any DMARD - n (%)	271 (90)	114 (21.7)
Biologic		30 (10)	7 (1.3)

*Data available on *293 and **277subjects*

Table 2. Empirical values of sensitivity and specificity of candidate items to be included in the algorithm in the first rheumatologic sample.

Variable	Cases (N=301)		Controls (N=526)		Sensitivity (95% CI)	Specificity (95% CI)
	+	-	+	-		
RA certification by Rheumatologist	232	69	19	507	77.08 (71.91-81.70)	96.39 (94.42-97.81)
Absence of certification for other autoimmune diseases*	294	7	449	77	97.67 (95.27-99.06)	14.64 (11.73-17.95)
ICD9-CM code 714 in HDF	57	244	2	524	18.94 (14.67-23.83)	99.62 (98.63-99.95)
Methotrexate	182	119	31	495	60.47 (54.69-66.03)	94.11 (91.74-95.96)
Antimalarials	153	148	67	459	50.83 (45.03-56.61)	87.26 (84.11-89.99)
Sulfasalazine	14	287	24	502	4.65 (2.57-7.68)	95.44 (93.29-97.06)
Leflunomide	12	289	0	526	3.99 (2.08-6.86)	100 (99.30-100)
Azathioprine	1	300	4	522	0.33 (0.01-1.84)	99.24 (98.06-99.79)
Cyclosporine	4	297	3	523	1.33 (0.36-3.37)	99.43 (98.34-99.88)
Anti-Tumor Necrosis Factor alpha	29	272	5	521	9.63 (6.55-13.54)	99.05 (97.80-99.69)
Abatacept	4	297	0	526	1.33 (0.36-3.37)	100 (99.30-100)
Rituximab	0	301	2	524	0	99.62 (98.63-99.95)
RA certification by Rheumatologist + ICD9 code 714 in HDF	41	260	1	525	13.62 (9.96-18.02)	99.81 (98.95-100)
RA certification by Rheumatologist + any DMARD	211	90	14	512	70.10 (64.58-75.22)	97.34 (95.57-98.54)
RA certification by Rheumatologist + ICD9 code 714 in HDF + any DMARD	38	263	1	525	12.62 (9.09-16.91)	99.81 (98.95-100)

*ICD: international classification of diseases; HDF: hospital discharge form; * ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases*

Table 3. Accuracy of the algorithm in the training sample by step

Step	Sensitivity (95% CI)	Specificity (95% CI)
Step 1: RA certification by rheumatologist OR ICD9-CM code 714 in HDF OR leflunomide OR tocilizumab OR abatacept OR Gold Salts	82.39 (77.61-86.52)	96.20 (94.19-97.66)
Step 2: Step 1 OR (methotrexate AND antimalarials AND no certification for other autoimmune diseases))	85.38 (80.88-89.17)	95.63 (93.51-97.21)
Step 3: Step 2 OR (glucocorticoids ≥ 3 prescriptions AND antimalarials AND no certification for other autoimmune diseases))	91.36 (87.60-94.28)	92.21 (89.57-94.35)
Step 4: Step 3 OR (methotrexate ≥ 3 prescriptions AND no certification for other autoimmune diseases)*	96.35 (93.56-98.16)	90.30 (87.45-92.70)

*the final algorithm used in the analysis

Table 4. Accuracy of the algorithm in the validation samples.

	Training set- Rheumatologic sample*	Validating set - Rheumatologic sample	Validating set – General population
Sensitivity (95% CI)	96.32 (96.25-96.38)	93.75 (79.19-99.23)	92.50 (79.61-98.43)
Specificity (95% CI)	90.33 (90.24-90.41)	90.54 (81.48-96.11)	99.77 (99.61-99.87)
PPV (95% CI)	85.04 (80.81–88.66)	81.08 (64.84–92.04)	72.55 (58.26-84.11)
NPV (95% CI)	97.74 (95.99–98.86)	97.10 (89.92-99.65)	99.95 (99.85-99.99)

* bootstrap estimates

A validation study of a new classification algorithm to identify rheumatoid arthritis using administrative health databases: case-control and cohort diagnostic accuracy studies

Results from the RECord linkage On Rheumatic Diseases study of the Italian Society for Rheumatology

Greta Carrara[1], Carlo A. Scirè[1], Antonella Zambon[2], Marco A Cimmino[3], Carlo Cerra[4], Marta Caprioli[5], Giovanni Cagnotto[6], Federica Nicotra[2], Andrea Arfè[2], Simona Migliazza[4], Giovanni Corrao[2], Giovanni Minisola[7], Carlomaurizio Montecucco[6]

Affiliations

1. Epidemiology Unit, Italian Society for Rheumatology, Milan, Italy
2. Department of Statistics and Quantitative Methods, Section of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy
3. Research Laboratory and Academic Division of Clinical Rheumatology, Department of Internal Medicine, University of Genova, Genoa, Italy.
4. Information system and management control, Local Health Authority (ASL), Pavia, Italy;
5. Istituto Clinico Beato Matteo, Department of Medicine, Vigevano, Italy;
6. Department of Rheumatology, IRCCS San Matteo Foundation, Pavia, Italy;
7. Division of Rheumatology - "San Camillo" Hospital - Rome - Italy

Running title: AHD algorithm validation RA

Word Count: 3263

Correspondence to:

Carlo Alberto Scirè

Via Turati 40 – 20121 Milano – Italy

c.scire@reumatologia.it

Key words: rheumatoid arthritis, health administrative databases, validation study, sensitivity and specificity, prevalence

ABSTRACT

Objectives: to develop and validate a new algorithm to identify rheumatoid arthritis (RA) patients and estimate disease prevalence using administrative health databases (AHD) of the Italian Lombardy region.

Design: Case-control and cohort diagnostic accuracy study.

Methods: In a randomly selected sample of 827 patients drawn from a tertiary rheumatology centre (training set), clinically validated diagnoses were linked to administrative data including diagnostic codes and drug prescriptions. An algorithm in steps of decreasing specificity was developed and its accuracy assessed calculating sensitivity/specificity, positive/negative predictive values (PPV/NPV), with corresponding confidence intervals (CI).

The algorithm was applied to two validating sets: 106 patients from a secondary rheumatology centre and 6087 subjects from the primary care. Alternative algorithms were developed to increase PPV at population level.

Crude and adjusted prevalence estimates taking into account algorithm misclassification rates were obtained for the Lombardy Region.

Results: The algorithms included: RA certification by rheumatologist, certification for other autoimmune diseases by specialists, RA code in the Hospital Discharge Form (HDF), prescription of disease-modifying anti-rheumatic drugs and oral glucocorticoids. In the training set, a four-steps algorithm identified clinically diagnosed RA cases with a sensitivity of 96.3 (95%CI:93.6-98.2) and a specificity of 90.3 (87.4-92.7). Both external validations showed highly consistent results. More specific algorithms achieved >80% PPV at the population level. The crude RA prevalence in Lombardy was 0.52%, and estimates adjusted for misclassification ranged from 0.31% (95%CI:0.14-0.42) to 0.37% (0.25-0.47).

Conclusions: AHD are a valuable tool for the identification of RA cases at the population level, and allow to estimate disease prevalence and potentially to select retrospective cohorts.

ARTICLE SUMMARY: STRENGHT AND LIMITATIONS

- this study provides a complete validation of classification algorithms for the identification of rheumatoid arthritis patients at the population level through healthcare administrative databases;

- two different approaches were applied in this study to estimate RA prevalence accounting for misclassification inherent to the classification algorithm;

- classification of disease according to algorithms from administrative data are setting-specific and not directly transferred to other systems;

- proper classification algorithm validations are useful to develop consistent instruments to compare disease burden in different healthcare systems.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease that is associated with development of disability, increased mortality and significant costs to society.[1] Population-based studies help to monitor disease burden, to evaluate long-term consequences of the disease and its treatments, and to assess quality of care, for both research and governance purposes.[2]

The increasing diffusion and completeness of administrative health databases (AHD) - which record healthcare services dispensed to all members of a specific population - provide a straightforward way to perform such population-based studies in RA.[3–5] The validity of AHD studies primarily relies on the diagnostic accuracy of case definition. The huge methodological variability of validation studies of AHD-based classification algorithms in RA makes difficult to evaluate the potentialities of AHD for population studies of RA.[6–16]

The majority of the studies in RA develops classification algorithms sampling from populations with high prevalence of RA (e.g. rheumatology clinics), focusing on the positive predictive value (PPV) - the probability of being a true case if classified as a potential one by AHD-based criteria. Even if high PPV was achieved in this setting, it does not reflect the performance achievable in the general population, where the prevalence of RA is 30-50 fold lower. Thus, in order to develop a valid instrument to perform a population study, a validation study sampling from the same population where it will be applied is highly informative. Nevertheless no study has currently provided a full validation of algorithms developed for the classification of RA by AHD at population level.[15]

The RECord linkage On Rheumatic Diseases (RECORD) study - promoted by the Italian Society for Rheumatology – aims to set up a national surveillance system to monitor the health burden of rheumatic diseases in Italy using data from AHD. The RECORD study of RA is structured in 3 phases: the first phase aims to evaluate the frequency of the disease; the second phase to evaluate the impact of the disease and its treatment on hard disease outcomes at population level; and the third phase to evaluate the quality of care delivered to RA patients.

In order to reach the objectives of the first step of the RECORD study, we report the methodological approach and the results of the development and validation of different classification algorithms for RA at different levels of the health care system, including primary care. We linked clinically validated diagnoses of randomly selected samples of cases and controls with the AHD of the Lombardy region (Italy). The prevalence of RA was then derived both as crude estimate and adjusting for the inherent misclassification.

PATIENTS AND METHODS

Reporting of this study compiles with the guidelines for diagnostic and validation studies of health administrative data.[17,18]

Study design and samples

Training set

A random sample of visits of 900 outpatients (300 cases with RA and 600 controls with rheumatic diseases other than RA, on the basis of the diagnosis reported on the electronic medical records) aged over 16 years and assisted by a tertiary rheumatology clinic (Rheumatology Unit, IRCCS Policlinico San Matteo Foundation, Pavia) between 2007-2010 was extracted from the medical record database of this centre according to a case-control diagnostic design nested in the resident population of Pavia.[19]

A sample size >700 subjects with a proportion of one third of cases in the training set was defined to precisely estimate both negative predictive value (NPV) >0.95 and PPV >0.50, setting alpha at 0.05 and beta at 0.8, as proposed by Steinberg et al. for case-control diagnostic studies.[20]

Validating sets

Two different samples were drawn for validation purpose: one from a secondary rheumatology centre and one from the primary care (general population sample) within the same catchment area. In these validating samples, a cohort diagnostic study design was applied.[21] The first validating set included a random sample of 138 patients from the electronic medical records of the Rheumatology outpatient clinic of the Clinical Institute Beato Matteo of Vigevano, a secondary care rheumatology clinic. A second validating set included all the 6087 subjects extracted from the primary care electronic medical records of a convenience sample of 6 primary care physicians of the LHA of Pavia.

The study was approved by the Ethics Committee of the IRCCS Policlinico San Matteo Foundation of Pavia, and participants gave their consent to the processing of their personal data.

Test methods

Reference standard

The clinical diagnosis from medical records was considered the reference standard. Diagnoses were clinically validated by an external investigator (GiCa), who was unaware of the content and results of the algorithm. Only when the diagnosis was unclear or varied over time, patients were classified according to specific classification criteria[22], cumulatively applied until date of the randomly selected visit, based on a data collection form including: gender, age, disease duration, morning stiffness, joint involvement, rheumatoid factor, X-ray abnormalities.

Administrative healthcare database variables and record linkage

AHD are an automated system of databases consisting in: i) an archive of all residents receiving NHS assistance (virtually the whole resident population), reporting demographic and administrative data; ii) an archive including all the certifications of chronic diseases for the exemption from co-payment; iii) an archive of all HDF from public or private hospitals, reporting all diagnoses related to the hospitalization; iv) an archive of all outpatient drug prescriptions reimbursable by the NHS.

The AHD variables useful for the identification of RA cases were defined *a priori* through a consensus process, informed by a literature review, held in February 2012 and involving 5 clinicians, 1 epidemiologist, 3 database owners and 2 statisticians. The literature was searched via Pubmed using a combination of free-text and MeSH terms regarding 'rheumatoid arthritis' and 'administrative database'. The relevant variable were selected among a list of items extracted from the retrieved literature [3,6–12,23] (see Appendix 1). These variables represented the set of potential index texts to be included in the classification algorithm: RA certification by rheumatologist and certification for other chronic autoimmune diseases (ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases), ICD9-CM code 714.0 in the HDF, prescription of DMARDs including biologics and oral glucocorticoids, outpatient diagnostic procedures, outpatient visits, diagnostic procedures in the HDF, blood tests and instrumental tests (as radiographs).

The following items were selected to be included in the algorithm: RA certification by rheumatologist, absence of certification for other chronic autoimmune diseases, ICD9-CM code 714.0 in the HDF, prescription of DMARDs including biologics and oral glucocorticoids (see Appendix 2).

Administrative data (selected items needed to create the algorithm) relative to patients included in both rheumatologic samples and general population sample were extracted from the data warehouse of Pavia's LHA within an interval of ±1 year over the index date (i.e. date of clinical assessment ranging from 2006 to 2011) for rheumatologic samples, and ±1 year over 1st of September 2011 for the primary care sample. Clinically validated diagnoses and administrative data from the LHA of Pavia were linked using deterministic record linkage through encrypted unique identifier code (only subjects successfully linked were retained for the analyses). A parallel extraction from the regional data warehouse from the 1st January 2009 to the 31st December 2010 only included the items needed for the classification according to the developed algorithm, in order to estimate the prevalence of RA.

Statistical methods

Development and validation of the algorithm

For each variable identified in the consensus-based phase, sensitivity and specificity were evaluated in the training set. Combining a priori knowledge and empirical estimates of sensitivity and specificity of each variable, a first candidate algorithm was developed, including in the first step variables with high specificity. The algorithm was then changed by sequentially including other variables with lower specificity but higher sensitivity. This process was stopped when a high sensitivity was reached at the expense of the least decrease in specificity.

For example, RA certification by rheumatologist, ICD9-CM code 714 in HDF and some drugs, like leflunomide and abatacept, showed high specificity. Knowing that other drugs, like tocilizumab and gold salts also have high specificity, we combined these items in the first step. Afterwards, items that are more sensitive and less specific, like methotrexate, antimalarials drugs and glucocorticoids, were combined in the successive steps.

Once the algorithm was fully defined, its overall accuracy was assessed by estimating sensitivity, specificity, PPV and NPV - with exact 95% confidence intervals (CIs).

The robustness of these estimates was tested in the training set by bootstrap procedure, using 1000 samples extracted with replacement.

Two automated statistical procedures were also applied: a backward variable selection approach applied to a parametric penalized logistic regression model with multiple interaction terms and non-parametric classification trees.

A sensitivity analysis was performed by considering alternative algorithms stratified by age (with a cut-off at 65 years) and a narrower temporal range of ±6 months from the index date for the extraction of the selected variables.

Two external fully independent validations[21] were carried out using datasets from different levels of health care: a secondary rheumatology centre and primary care. The performance of the algorithm was tested estimating sensitivity, specificity, PPV and NPV.

Estimation of disease prevalence

To estimate the prevalence of RA in Lombardy (an Italian region of about 9 million residents) in 2010, the final algorithm was applied to the required variables extracted from the AHD of Lombardy, which have the same structure of the AHD of the Pavia LHA (i.e., an archive of: residents, certifications of exemption, hospital discharges and of outpatient drug prescriptions). The target population consisted of all residents aged 16 years or older.

The crude prevalence estimate was adjusted for the impact of misclassification due to the possible classification errors of the algorithm, quantified during the validation phase, by applying two different methods. The first method – proposed by Rogan & Gladen– is based on a direct relationship, which expresses the adjusted value of the prevalence as a function of the crude prevalence and the sensitivity and specificity of the algorithm (Formula 1). Using the estimates of sensitivity and specificity derived from the validation study in the general population sample the crude prevalence was corrected for the impact of misclassification and 95% CI was calculated.[24]

Formula 1.

$$adjusted\ prevalence\ = \frac{crude\ prevalence + \widehat{SP} - 100}{\widehat{SE} + \widehat{SP} - 100}$$

The second method – proposed by Joseph et al. – provides a more precise adjusted estimate by giving preference only to the most plausible range of values for the parameters of interest (prevalence, sensitivity

and specificity of the algorithm). [25] Specifically, following the Bayesian framework, an initial quantification of the plausibility of each possible value of the parameters of interest was summarised in a probability distribution (*prior distribution*), based on estimates of sensitivity and specificity obtained from the validation study in the general population sample and on prevalence obtained from previous population studies.[26–29] The prior distribution was then updated in light of the observed data through their likelihood, leading to a posterior distribution, and the mean and 2.5% and 97.5% percentiles of the posterior distribution provide an estimate of the parameters and a corresponding credibility interval. Data management and statistical analyses were conducted using SAS software (version 9.2; SAS Institute, Cary, NC), R Statistical Software (Foundation for Statistical Computing, Vienna, Austria) and WinBUGS software version 1.4.3 [30].

RESULTS

Study samples

In the first rheumatologic sample (training sample), in 862/900 subjects (96%) record linkage between the clinical dataset and administrative data was successful. Complete information for diagnosis validation (criteria for classification in RA e Non-RA) was available for 827/862 subjects (96%) (Figure 1). Demographic, disease and treatment characteristics are reported in Table 1.

Development of the algorithm

Combining the variables of progressively increasing sensitivity (Table 2), we developed a final 4-step algorithm that identifies clinically diagnosed RA cases with a sensitivity of 96.35 (95%CI:93.56-98.16) and a specificity of 90.30 (95%CI:87.45-92.70) (Table 3).

Bootstrap procedure confirmed the robustness of the estimates in the training set (Table 4).

More flexible methods tested in sensitivity analyses confirmed similar accuracy: logistic penalized models with multiple interaction terms showed a sensitivity of 94.35 (95%CI: 91.36-96.68) and a specificity of 92.59 (95%CI: 90.11-94.68); classification trees did not identified alternative pathways able to significantly improve accuracy for the classification of cases.

Validation of the algorithm

The first external validation was performed in 106 out of 138 patients, in which record linkage was successful and sufficient clinical data available. This sample included 32 cases (30.2%) with a median age of 62.5 years (IQR: 53.5-73.5) and a M:F ratio of 1:2; 30 (93.8%) cases were treated with at least one DMARDs. In the sample of controls, the median age was 57 years (IQR:51-74). The second validation set included 6087 subjects (40 cases of RA and 6047 controls), with a median age of 70.5 years (IQR:57-78) with a M:F ratio of 1:3 in cases and median age of 45 years (IQR:35-59) and M:F ratio of 1:1 in controls. 27/40 (67.5%) cases were treated with at least one DMARD.

The first external validation showed highly consistent results compared with the training set (Table 4). Accuracy measures in general population sample showed a substantial increase in specificity (99.8; 95%CI:99.6-99.9) and decrease in PPV (72.5; 95%CI: 58.3-84.1).

PPVs over 80% were achievable both in rheumatologic samples (85.04 (80.81 – 88.66) and 81.08 (64.84 – 92.04) in training and first validating set, respectively) and in general population restricting the algorithm to DMARD-users (PPV 85.7%; 95%CI: 63.7-96.9).

Estimation of disease prevalence

Applying the four-steps algorithm to the population of the Lombardy Region, a crude prevalence of 0.52% (0.30% for males and 0.73% for females) was obtained, with a M:F ratio of 1:3 and a peak of prevalence between 72-75 years for females, and between 75-78 years for males.

Adjusting for the estimated misclassification, prevalence fell to 0.31% (95%CI:0.18-0.45) using Gladen & Rogan's method (in Formula 1: crude prevalence=0.52%, sensibility=92.5%, specificity=99.77%) and to 0.37% (95%CI:0.26-0.48), using Joseph's method with a plausible range of values included between 0.2% and 0.7%.

DISCUSSION

This study supports the overall validity of the administrative databases of the Italian NHS of the Lombardy region in the identification of patients with RA.

Previous studies showed the validity of AHD-based algorithms to identify cases of RA, with sensitivity and specificity ranging from 56% to 100% and from 55% to 97%.[15] The accuracy achieved in our study is highly consistent with those obtained by studies following similar methodology. In particular, Widdifield and co-workers recently developed a set of classification algorithms for RA using AHD in Ontario, Canada. These algorithms - derived in a randomly selected rheumatologic sample with a 33% prevalence of RA - showed optimal accuracy in identifying clinical diagnoses of RA, with sensitivity/specificity up to 97/85% and PPV/NPV up to 76/98%. Though we used different items to construct our instruments in our training rheumatologic sample (34% prevalence of RA), we obtained highly consistent accuracy (sensitivity/specificity 96/90% and PPV/NPV 85/98%).

Despite several algorithms are available for different AHD in different settings, none of these have been fully validated in the general population. This leads to high PPV, whose generalizability is limited to high prevalence study samples - such as for example rheumatologic outpatient samples -, where the prevalence of RA may be more than 50-fold higher.[14] Once developed the algorithm in a rheumatologic sample, we measured the diagnostic performance of the algorithm in a general population sample. As expected, PPV significantly decreased to 72%, while NPV increased over 99%. Only alternative algorithms restricted to DMARD users and to rheumatology samples were associated to PPV higher than 80%. Different algorithms with different operative characteristics may be suitable for studies with different purposes: high sensitivity for impact studies and high specificity for cohort studies.[14]

Beyond the usefulness of misclassification data to drive decision on the criterion to apply in selecting cohort of patients, sensitivity and specificity estimates are useful to adjust occurrence measures at population level.[31] This is the first study taking into account empirical misclassification in the adjustment of prevalence estimates of RA. In order to obtain unbiased estimates of prevalence we applied a first method that arithmetically adjusts the crude estimates taking into account the false positive and false

negative rates.[31] A more complex method that incorporates both a priori available information about the RA prevalence in Italy and empirical misclassification was also tested in order to improve the estimation based on the current knowledge.[25] Regardless of the method applied, prevalence estimates ranging from 0.31 to 0.37% are consistent with the expected on the basis of the literature for Italy[26–29], providing further validation to the developed tool. Using Joseph's method with a larger range of plausible values (0.2%-1%) we obtained an estimated prevalence of 0.36%.

In the design of this study we tried to limit mayor bias of diagnostic studies and to ensure external validity of the results.[15,17,18]

Study samples were randomly selected, in order to limit selection bias and to represent the entire spectrum of disease severity. To avoid observation bias due to differential misclassification, an independent investigator - who was unaware of the items included in the algorithm and of the subject classification – validated clinical diagnoses. AHD data suitable to be included in a diagnostic algorithm were identified through a literature-informed consensus process. We included this first step to avoid a completely data-driven algorithm, which could have overestimated the accuracy in the development sample. Only items from the domain of diagnostic codes and drug utilisation were deemed to be relevant, as most previous algorithms have done.

The robustness of our findings was also confirmed by the bootstrap procedure and by the exploration of other possible combinations of the candidate items using different statistical methods. These alternative methods achieved similar accuracy, though never significantly better than the multi-step algorithm, confirming the internal validity of the results.

The generalizability of the results was evaluated by different external validations, carried out using different healthcare levels, investigators, temporal ranges and study designs.[21]

This study has several limitations. Cross-sectional diagnostic 'case-control' studies tend to overestimate diagnostic accuracy.[19] However, accuracy was still satisfactory even when a cross-sectional diagnostic 'cohort' design was applied in a same prevalence sample. Beyond the higher prevalence of RA in the training and the first validation set, patients drawn from rheumatology samples may include subjects with

more severe disease and different socio-demographic characteristics. However, the algorithm still performed well with a similar sensitivity in the general population, where the entire spectrum of the disease is represented. Furthermore drug prescriptions of elderly patients who are hospitalized in retirement homes are not tracked by the AHD, leading to a substantial underestimation of the prevalence of the disease. Another possible source of bias is linked to the choice of the reference standard. Despite the majority of the studies apply this type of case definition, clinical diagnosis is less reliable than classification criteria. However, classification criteria are developed to increase specificity in order to include patients in clinical trials and not for epidemiologic purpose.[32] We only adopted classification criteria[22] to validate unclear diagnoses. This might have introduced a verification bias in our study, slightly increasing the specificity of the algorithms. Differential misclassification may take place based on disease duration, since the probability to have diagnostic codes and DMARD prescription may increase with disease duration, leading to underrepresentation of incident RA.

In conclusion this study shows the accuracy of administrative data algorithms for identifying RA patients both in rheumatology clinics and general population in Italy. This study also supports the usefulness of misclassification data to adjust estimates and to drive the decision of the appropriate algorithm to adopt based on the study objectives. Beyond the content of the applied classification criterion, validation data are useful to select homogeneous cohorts of patients with RA across countries and health care systems, making feasible the implementation of surveillance systems aiming to improve care of patients with RA.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of the entire clinical staff including the rheumatologists of the clinical centres for their former data collection and for making available their registers of patients.

CONTRIBUTORSHIP STATEMENT

CAS, AZ, MAC, GiCo, GM, CM planned the study. CC, MC, GiCa and SM collected data. GrCa, FN, AA analysed data. GrCa wrote the first draft, and all the authors critically revised and approved the final manuscript. GM was involved in obtaining of funding.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

The RECORD study is funded by the Italian Society for Rheumatology (SIR) as part of the Epidemiology Unit development program.

DATA SHARING

No additional data available.

REFERENCE LIST

1 Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. *BullWorld Health Organ* 2003;**81**:646–56.

2 Woolf AD, Vos T, March L. How to measure the impact of musculoskeletal conditions. *BestPractResClinRheumatol* 2010;**24**:723–32.

3 Kim SY, Solomon DH. Use of administrative claims data for comparative effectiveness research of rheumatoid arthritis treatments. *Arthritis ResTher* 2011;**13**:129.

4 Corrao G, Cesana G, Merlino L. Pharmacoepidemiological research and the linking of electronic healthcare databases available in the Italian region of Lombardy. *BioMed Stat Clin Epidemiol* 2008;**2**:117–25.

5 Scotti L, Arfè A, Zambon A, *et al.* Cost-effectiveness of enhancing adherence with oral bisphosphonates treatment in osteoporotic women: an empirical approach based on healthcare utilisation databases. *BMJ Open* 2014;**4**:e003758. doi:10.1136/bmjopen-2013-003758

6 Gabriel SE. The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. *Arthritis Rheum* 1994;**37**:821–3.

7 Katz JN, Barrett J, Liang MH, *et al.* Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. *Arthritis Rheum* 1997;**40**:1594–600.

8 MacLean CH, Park GS, Traina SB, *et al.* Positive predictive value (PPV) of an administrative data-based algorithm for the identification of patients with rheumatoid arthritis (RA). *Arthritis Rheum* 2001;**44**:1–520.

9 Losina E, Barrett J, Baron JA, *et al.* Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. *JClinEpidemiol* 2003;**56**:515–9.

10 Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum* 2004;**51**:952–7.

11 Pedersen M, Klarlund M, Jacobsen S, *et al.* Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. *EurJEpidemiol* 2004;**19**:1097–103.

12 Thomas SL, Edwards CJ, Smeeth L, *et al.* How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? *Arthritis Rheum* 2008;**59**:1314–21.

13 Kim SY, Servi A, Polinski JM, *et al.* Validation of rheumatoid arthritis diagnoses in health care utilization data. *Arthritis Res Ther* 2011;**13**:R32. doi:10.1186/ar3260

14 Widdifield J, Bernatsky S, Paterson JM, *et al.* Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: A validation study using the medical records of rheumatologists. *Arthritis Care ResHoboken* Published Online First: 16 April 2013.PM:23592598

15 Widdifield J, Labrecque J, Lix L, *et al.* Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases. *Arthritis Care Res* 2013;**65**:1490–503. doi:10.1002/acr.21993

16 Chung CP, Rohan P, Krishnaswami S, *et al.* A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine* 2013;**31, Supplement 10**:K41–K61. doi:10.1016/j.vaccine.2013.03.075

17 Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;**64**:821–9. doi:10.1016/j.jclinepi.2010.10.006

18 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. doi:10.7326/0003-4819-138-1-200301070-00012-w1

19 Biesheuvel CJ, Vergouwe Y, Oudega R, *et al.* Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;**8**:48. doi:10.1186/1471-2288-8-48

20 Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. *Biostatistics* 2009;**10**:94–105.

21 Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. New York: : Springer 2009.

22 MacGregor AJ, Bamber S, Silman AJ. A comparison of the performance of different methods of disease classification for rheumatoid arthritis. Results of an analysis from a nationwide twin study. *JRheumatol* 1994;**21**:1420–6.

23 Ng B, Aslam F, Petersen NJ, *et al.* Identification of rheumatoid arthritis patients using an administrative database: a Veterans Affairs study. *Arthritis Care ResHoboken* 2012;**64**:1490–6.

24 Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978;**107**:71–6.

25 Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *AmJEpidemiol* 1995;**141**:263–72.

26 Cimmino MA, Parisi M, Moggiana G, *et al.* Prevalence of rheumatoid arthritis in Italy: the Chiavari Study. *AnnRheumDis* 1998;**57**:315–8.

27 Salaffi F, De AR, Grassi W. Prevalence of musculoskeletal conditions in an Italian population sample: results of a regional community-based study. I. The MAPPING study. *ClinExpRheumatol* 2005;**23**:819–28.

28 Marotto D, Nieddu ME, Cossu A, *et al.* [Prevalence of rheumatoid arthritis in North Sardinia: the Tempio Pausania's study]. *Reumatismo* 2005;**57**:273–6.

29 Della Rossa A, Neri R, Talarico R, *et al.* Diagnosis and referral of rheumatoid arthritis by primary care physician: results of a pilot study on the city of Pisa, Italy. *Clin Rheumatol* 2010;**29**:71–81. doi:10.1007/s10067-009-1285-x

30 Lunn DJ, Thomas A, Best N, *et al.* WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput* 2000;**10**:325–37.

31 Gladen B, Rogan WJ. Misclassification and the design of environmental studies. *AmJEpidemiol* 1979;**109**:607–16.

32 Sakellariou G, Scire CA, Zambon A, *et al.* Performance of the 2010 classification criteria for rheumatoid arthritis: a systematic literature review and a meta-analysis. *PLoSOne* 2013;**8**:e56528.

Table 1. Characteristics of the training sample.

Characteristic	RA (n=301)	Non-RA (n=526)
Age (years) - mean (SD)	66.8 (13.1)	57.7 (15.7)
Female gender- n (%)	218 (72.4)	405 (77)
Disease duration <2years – n (%)	81* (27.6)	
Rheumatoid Factor positive – n (%)	151** (54.4)	
NSAID or COX-2 Inhibitor - n (%)	198 (65.8)	298 (56.7)
Glucocorticoids - n (%)	228 (75.8)	178 (33.8)
DMARDs		
Methotrexate - n (%)	182 (60.5)	31 (5.9)
Antimalarials - n (%)	153 (50.8)	67 (12.7)
Sulfasalazine - n (%)	14 (4.7)	24 (4.6)
Leflunomide - n (%)	12 (4)	0 (0)
Other DMARDs - n (%)	5 (1.7)	7 (1.3)
Any DMARD - n (%)	271 (90)	114 (21.7)
Biologic	30 (10)	7 (1.3)

*Data available on *293 and **277subjects*

Table 2. Empirical values of sensitivity and specificity of candidate items to be included in the algorithm in the first rheumatologic sample.

Variable	Cases (N=301)		Controls (N=526)		Sensitivity (95% CI)	Specificity (95% CI)
	+	-	+	-		
RA certification by Rheumatologist	232	69	19	507	77.08 (71.91-81.70)	96.39 (94.42-97.81)
Absence of certification for other autoimmune diseases*	294	7	449	77	97.67 (95.27-99.06)	14.64 (11.73-17.95)
ICD9-CM code 714 in HDF	57	244	2	524	18.94 (14.67-23.83)	99.62 (98.63-99.95)
Methotrexate	182	119	31	495	60.47 (54.69-66.03)	94.11 (91.74-95.96)
Antimalarials	153	148	67	459	50.83 (45.03-56.61)	87.26 (84.11-89.99)
Sulfasalazine	14	287	24	502	4.65 (2.57-7.68)	95.44 (93.29-97.06)
Leflunomide	12	289	0	526	3.99 (2.08-6.86)	100 (99.30-100)
Azathioprine	1	300	4	522	0.33 (0.01-1.84)	99.24 (98.06-99.79)
Cyclosporine	4	297	3	523	1.33 (0.36-3.37)	99.43 (98.34-99.88)
Anti-Tumor Necrosis Factor alpha	29	272	5	521	9.63 (6.55-13.54)	99.05 (97.80-99.69)
Abatacept	4	297	0	526	1.33 (0.36-3.37)	100 (99.30-100)
Rituximab	0	301	2	524	0	99.62 (98.63-99.95)
RA certification by Rheumatologist + ICD9 code 714 in HDF	41	260	1	525	13.62 (9.96-18.02)	99.81 (98.95-100)
RA certification by Rheumatologist + any DMARD	211	90	14	512	70.10 (64.58-75.22)	97.34 (95.57-98.54)
RA certification by Rheumatologist + ICD9 code 714 in HDF + any DMARD	38	263	1	525	12.62 (9.09-16.91)	99.81 (98.95-100)

*ICD: international classification of diseases; HDF: hospital discharge form; * ankylosing spondylitis, psoriatic arthritis and psoriasis, connective tissue diseases, systemic vasculitis, inflammatory bowel diseases*

Table 3. Accuracy of the algorithm in the training sample by step

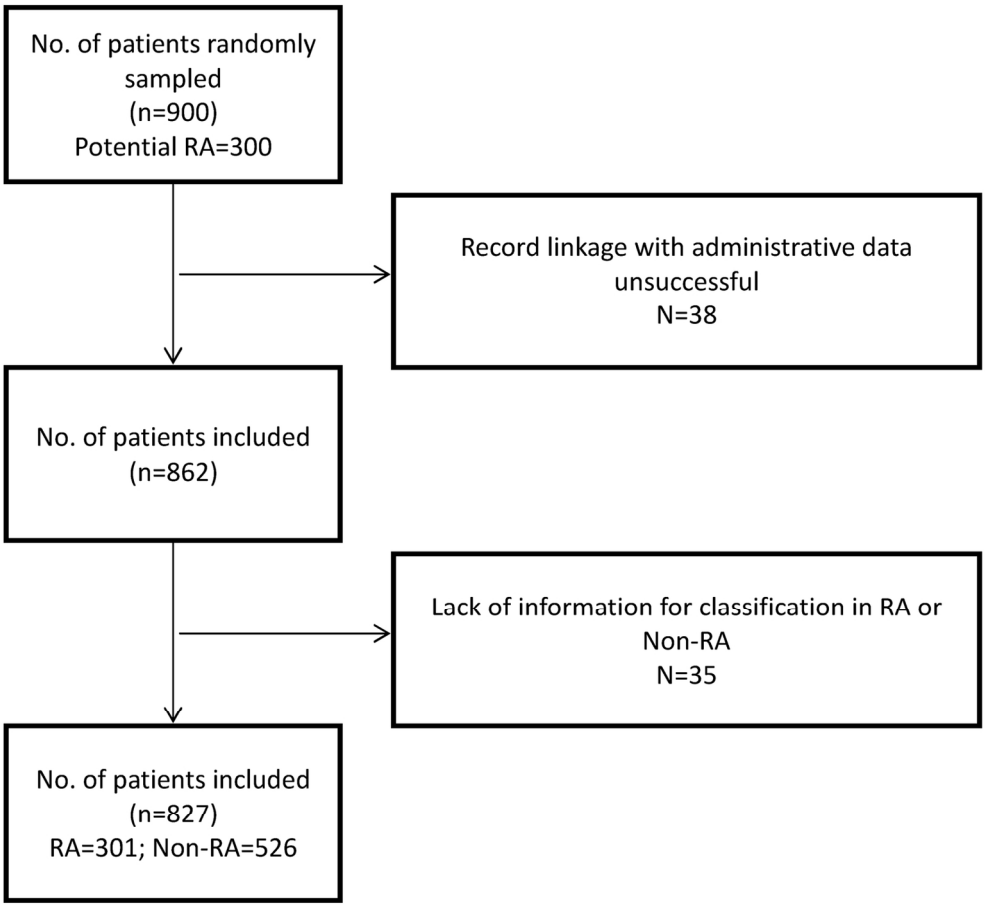
Step	Sensitivity (95% CI)	Specificity (95% CI)
Step 1: RA certification by rheumatologist OR ICD9-CM code 714 in HDF OR leflunomide OR tocilizumab OR abatacept OR Gold Salts	82.39 (77.61-86.52)	96.20 (94.19-97.66)
Step 2: Step 1 OR (methotrexate AND antimalarials AND no certification for other autoimmune diseases))	85.38 (80.88-89.17)	95.63 (93.51-97.21)
Step 3: Step 2 OR (glucocorticoids ≥ 3 prescriptions AND antimalarials AND no certification for other autoimmune diseases))	91.36 (87.60-94.28)	92.21 (89.57-94.35)
Step 4: Step 3 OR (methotrexate ≥ 3 prescriptions AND no certification for other autoimmune diseases)*	96.35 (93.56-98.16)	90.30 (87.45-92.70)

*the final algorithm used in the analysis

Table 4. Accuracy of the algorithm in the validation samples.

	Training set- Rheumatologic sample*	Validating set - Rheumatologic sample	Validating set – General population
Sensitivity (95% CI)	96.32 (96.25-96.38)	93.75 (79.19-99.23)	92.50 (79.61-98.43)
Specificity (95% CI)	90.33 (90.24-90.41)	90.54 (81.48-96.11)	99.77 (99.61-99.87)
PPV (95% CI)	85.04 (80.81–88.66)	81.08 (64.84–92.04)	72.55 (58.26-84.11)
NPV (95% CI)	97.74 (95.99–98.86)	97.10 (89.92-99.65)	99.95 (99.85-99.99)

* bootstrap estimates



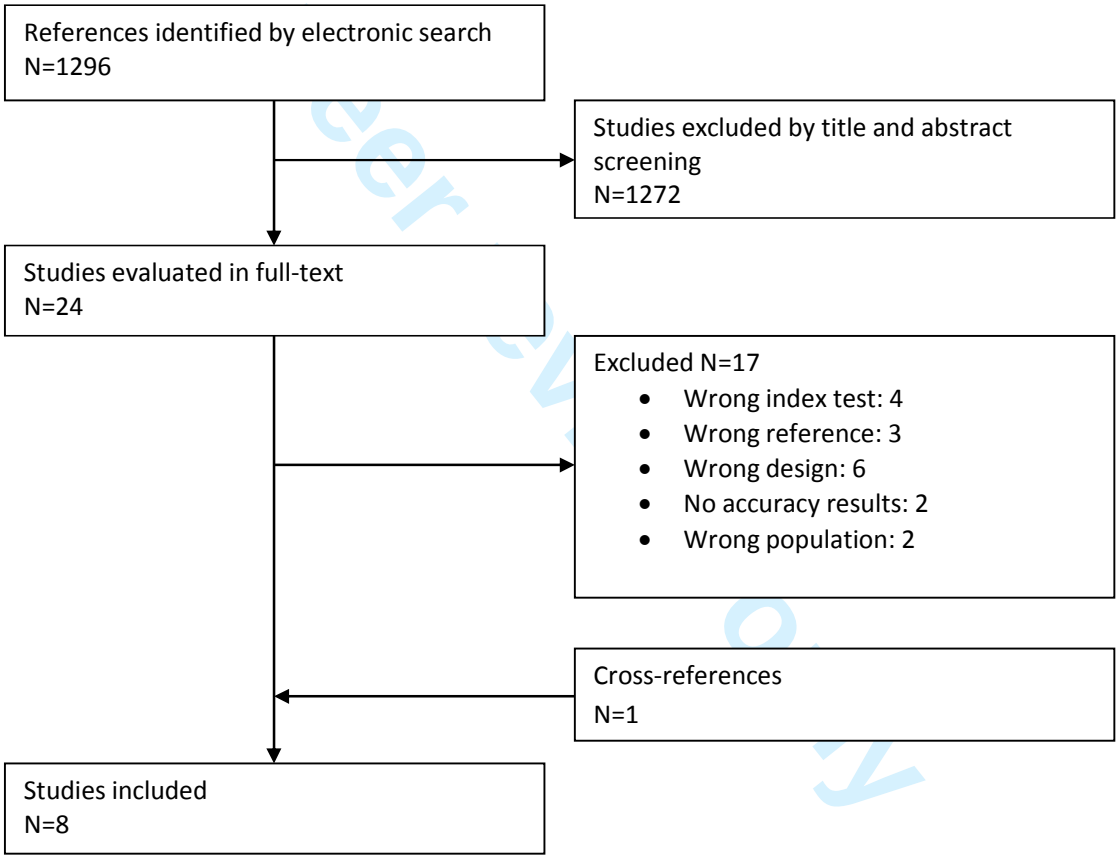
Flow-chart of the training set sample
138x128mm (300 x 300 DPI)

APPENDIX 1– LITERATURE SEARCH

Search Strategy MEDLINE (Pubmed)

("rheumatoid arthritis"[All Fields] OR "arthritis, rheumatoid"[Mesh]) AND ("Algorithms"[MeSH Terms] OR "Clinical Coding"[MAJR] OR "Databases, Factual"[MeSH Terms] OR "International Classification of Diseases/statistics and numerical data"[MAJR] OR "Insurance Claim Review/statistics and numerical data"[MAJR] OR "Medical Records"[MeSH Terms] OR "Medical Records Systems, Computerized"[MeSH Terms] OR "insurance claim reporting"[MeSH Terms] OR "Registries"[Mesh] OR "administrative database*"[All Fields] OR "healthcare database*"[All Fields] OR "claims database*"[All Fields] OR "medical records"[TIAB]) AND (("0001/01/01"[PDAT] : "2011/12/31"[PDAT]) AND "humans"[MeSH Terms])

Flow-chart



Included Studies

1. Gabriel SE. The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. Arthritis Rheum. 1994 Jun;37(0004-3591 (Print)):821–3.

2. Katz JN, Barrett J, Liang MH, Bacon AM, Kaplan H, Kieval RI, et al. Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. Arthritis Rheum. 1997 Sep;40(0004-3591 (Print)):1594–600.

3. MacLean CH, Park GS, Traina SB, Liu HH, Hahn BH, Paulus HE, et al. Positive predictive value (PPV) of an administrative data-based algorithm for the identification of patients with rheumatoid arthritis (RA). Arthritis Rheum. 2001;44(S9):1–520.

4. Losina E, Barrett J, Baron JA, Katz JN. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. JClinEpidemiol. 2003 Jun;56(0895-4356 (Print)):515–9.

5. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Rheum. 2004 Dec 15;51(0004-3591 (Print)):952–7.

6. Pedersen M, Klarlund M, Jacobsen S, Svendsen AJ, Frisch M. Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. EurJEpidemiol. 2004;19(0393-2990 (Print)):1097–103.

7. 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 Sep 15;59(0004-3591 (Print)):1314–21.

8. Kim SY, Servi A, Polinski JM, Mogun H, Weinblatt ME, Katz JN, et al. Validation of rheumatoid arthritis diagnoses in health care utilization data. Arthritis Res Ther. 2011 Feb 23;13(1):R32.

Summary table

Year	Author	Country	Index	Reference	N (setting)	Se	Sp	PPV	NPV
1994	Gabriel S	USA	Electronic DB	Chart review (ACR1987)		0.89	0.74	0.57	
1997	Katz JN	USA	Physician's claim ICD	Medical records	153 / 150	0.90		0.95	
2001	MacLean C (abstract)	USA	≥65years, ≥2 visits with 714.x	Self reported diagnosis	924 RA			0.92	
2003	Losina C	USA	Physician's claims ICD	Medical records	37 RA	0.65		0.86-0.89	
2004	Singh J	USA	Veterans amministrative data a) ICD codes b) ICD + ≥3mo DMARDs c) ICD + RF d) ≥3mo DMARD + RF e) IC-D + ≥3mo DMARD + RF	Two rheum's visits	184	a)1.00 b)0.84 c)0.88 d)0.76 e)0.76	a)0.55 b)0.82 c)0.91 d)0.95 e)0.97	a)0.66 b)0.81 c)0.92 d)0.95 e)0.97	a)1.00 b)0.86 c)0.86 d)0.77 e)0.77
2004	Pedersen M	DK	ICD 8-10 codes	Chart review a) Clinical dg b) ACR 1987	217 RA (inpts)			a)0.59 b)0.46	
2008	Thomas SL	UK	Algorithm including • >1 RA code • RA group (GP code) • ≥1 DMARD • Other diagnosis	1987 ACR (MacGregor)	223 - 112 probable RA	0.87 - 0.84	0.88 - 0.86		
2011	Kim SY	USA	a) ≥2 -714 code b) ≥3 -714 code c) ≥2 -714 code by rheumatologist ' + 1 DMARD	Chart review (RA per rheumatologist)	131 RA (>65 yrs)			a)0.55 b)0.65 c)0.66 a')0.86 b')0.87 c')0.88	

APPENDIX 2 -ADMINISTRATIVE DATA INCLUDED

Diagnostic codes for exemption

Life-long certification codes for chronic diseases by specialist for RA (006.714.0) and other systemic autoimmune diseases (arteritis 002.447.6; Crohn disease 009.555; ulcerative colitis 009.556; systemic lupus erythematosus 028.710.0; Sjögren syndrome 030.710.2; psoriasis or psoriatic arthritis 045.696; systemic sclerosis 047.710.1; ankylosing spondylitis 054.720.0; Behçet disease RC0210; IgA vasculitis RD0030; microscopic polyangiitis RG0020; polyarteritis nodosa RG0030; eosinophilic granulomatosis with polyangiitis RG0050; granulomatosis with polyangiitis RG0070; giant cell (temporal) arteritis RG0080; Takayasu arteritis RG0090; dermatomyositis RM0010; polymyositis RM0020; mixed connective tissue disease RM0030; undifferentiated connective tissue disease RMG010; relapsing polychondritis RM0060);

Hospital Discharge Form

ICD9-CM 714.0 code (RA)

Drug prescriptions

Anatomical Therapeutic Chemical (ATC) codes for disease modifying anti-rheumatic drugs (DMARD) (methotrexate L01BA01; antimalarials P01BA01-02; sulfasalazine A07EC01; leflunomide L04AA13; azathioprine L04AX01; cyclosporine A L04AD01; gold salts M01CB01-03; anti-tumour necrosis factor L04AB; anakinra L04AC03; tocilizumab L04AC07; abatacept L04AA24; rituximab L01XC02) and glucocorticoids (H02AB).

STARD checklist for reporting of studies of diagnostic accuracy
(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	2
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	3
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	4
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	4
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	4
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	4
Test methods	7	The reference standard and its rationale.	5
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	5
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	NA
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	5
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	5
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	6
	13	Methods for calculating test reproducibility, if done.	NA
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	4
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	Tab1
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	Fig1
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	NA
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	Tab1
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	Tab2
	20	Any adverse events from performing the index tests or the reference standard.	NA
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	TAB2-4
	22	How indeterminate results, missing data and outliers of the index tests were handled.	8
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	9
	24	Estimates of test reproducibility, if done.	NA
DISCUSSION	25	Discuss the clinical applicability of the study findings.	12