PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	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
AUTHORS	Scirè, Carlo; Carrara, Greta; Zambon, Antonella; Cimmino, Marco; Cerra, Carlo; Caprioli, Marta; Cagnotto, Giovanni; Nicotra, Federica; Arfè, Andrea; Migliazza, Simona; Corrao, Giovanni; Minisola, Giovanni; Montecucco, carlomaurizio

VERSION 1 - REVIEW

REVIEWER	Hjalmar Wadström Karolinska Institutet, Sweden
REVIEW RETURNED	02-Oct-2014

GENERAL COMMENTS	I found the manuscript well written and easy to follow. There are some minor issues that might need clarification. 1. In figure 1, 35 patients are excluded on the basis of “Not meeting criteria for classification in RA and Non-RA”, on the other hand on page 10 under “study samples” it seems that these patients were excluded because of lacking information. Please clarify what you mean by this. 2. This study is based on register data and the quality of such studies hinges upon good data quality and coverage. Therefore, it might improve the paper if some information or maybe a reference on the quality of these registers was added. Also, are missing values prevalent in these registers and how were they handled? 3. To measure the prevalence of RA is not mentioned in the title or the objectives, should it have been? Or was this just a side note? 4. When entering the numbers from table 4 in formula 1 on page 9 it seems the result differs from that listed on page 11 (0.31%). Maybe this could be explained further? adjusted prevalence= 0.52+0.9977-1 / 0.9250+0.9977-1 5. The final algorithm is not specified in the results section, probably because of the complexity of it, instead it is listed in table 3. It might be helpful for the reader if it was emphasized in Table 3 that the algorithm listed is the final algorithm that was used or if it was specified in the results section. 6. In table 2, the total number of cases and controls are not listed. It would improve readability if this was added.

REVIEWER	Iosief Abraha Health Planning Service Regional Health Authority of Umbria Perugia, (Italy)
REVIEW RETURNED	14-Oct-2014

GENERAL COMMENTS	The proposed article is about the development and validation of algorithms necessary to identify rheumatoid arthritis (RA) patients in administrative health databases. The authors randomly identified a sample of patients with RA from a clinical database and linked them to an administrative regional database. Accuracy of four-steps algorithm was assessed and subsequently the prevalence of RA at a regional level was evaluated. The article is interesting in its field and it might increase the interest of the validation process to be applied in the Italian administrative databases (http://dx.doi.org/10.6000/1929-6029.2014.03.03.10). However, the following are critical points that authors need to address: 1. While it clear the reference standard is represented by the clinical diagnosis from the medical records, it not completely clear what information from the administrative databases should function as an index test. The only 'clinical' information administrative databases have is the ICD9 code, thus, it is unclear how the two validation sets (a secondary rheumatology centre and primary care) could be related with administrative database. In other words is the administrative database used as a third validation set? Shouldn't it be considered as the main target for the validation given that it appears in the title and it was used to calculate the RA prevalence in the Lombardy region? This is the most critical point that authors need to address. 2. RA patients are often outpatients. Administrative databases usually does not have outpatient information. If this is not the case, authors need to clarify which information of outpatients have the AHD since the ICD9 code is related only to inpatients. 3. The flow of the description of the methodology should be clearly presented and avoid placing descriptions in the results section. Other, suggestions can be found below. Background Page 4, lines 44-49. The authors might report a reference for the statement or provide a further description of the RECORD study. Patients and Methods Explanation about the training set should be given. Page 6, lines 14-22. While the cases are intuitively patients with RA it is unclear who the controls are. In other words clarification as to what types of patients are recorded in the medical record databases should be given. Are the controls patients with rheumatic diseases but without RA diagnosis? Furthermore are the patients (cases and controls) recorded in the databases only for ambulatory visit or are they also inpatients? Lines 33-46. A clear information about inpatients and outpatients

	should be provided when dealing about the population in the validation set. This point is relevant because it is related with the next. Page 6, lines 33-46. Description of content of the two registries (or references if described in other articles) should be provided. Page 7, lines 18-25. The statement about the AHD is correct. However, generally administrative databases in Italy do not provide clinical diagnosis for outpatients. Here, authors should clarify what information from AHD used for the outpatients. This issue is relevant also for the next paragraph that deals with RA cases. Lines 27-31. This point is very interesting. The authors state that they reviewed the literature for RA case definition. They should report either any publication they may have produced or provide a description of the process and submit it with the paper as a supplemental file. Page 7 line 52-53. “For each variable identified”: please mention here the list of the variables. Page 7 Lines 55- Page 8 Line 7. This point is extremely important for the understanding of the algorithm construction. Authors should describe clearly the different steps by mentioning the variables with which they started the assessment of the SE and SP and the variables added in the subsequent assessment. The use of the examples can make the issue easier to understand. Results Page 9, lines 5-16. This paragraph of the variable selection should go in the methods section and presented appropriately. Lines 18-20. The listed candidates not included in the analysis should be removed or moved in the Methods section if there were interesting points to mention. After revising the flow of the description of the methods and clarifying the issue of the validation set, authors may consider writing the results following the amendment in the methods. The proposed article is about the development and validation of algorithms necessary to identify rheumatoid arthritis (RA) patients in administrative health databases. The authors randomly identified a sample of patients with RA from a clinical database and linked them to an administrative regional database. Accuracy of four-steps algorithm was assessed and subsequently the prevalence of RA at a regional level was evaluated. The article is interesting in its field and it might increase the interest of the validation process to be applied in the Italian administrative databases (http://dx.doi.org/10.6000/1929-6029.2014.03.03.10). However, the following are critical points that authors need to address: 1. While it clear the reference standard is represented by the clinical diagnosis from the medical records, it not completely clear what information from the administrative databases should function as an index test. The only ‘clinical’ information administrative databases have is the ICD9 code, thus, it is unclear how the two validation sets (a secondary rheumatology centre and primary care) could be

	related with administrative database. In other words is the administrative database used as a third validation set? Shouldn't it be considered as the main target for the validation given that it appears in the title and it was used to calculate the RA prevalence in the Lombardy region? This is the most critical point that authors need to address. 2. RA patients are often outpatients. Administrative databases usually does not have outpatient information. If this is not the case, authors need to clarify which information of outpatients have the AHD since the ICD9 code is related only to inpatients. 3. The flow of the description of the methodology should be clearly presented and avoid placing descriptions in the results section. Other, suggestions can be found below. Background Page 4, lines 44-49. The authors might report a reference for the statement or provide a further description of the RECORD study. Patients and Methods Explanation about the training set should be given. Page 6, lines 14-22. While the cases are intuitively patients with RA it is unclear who the controls are. In other words clarification as to what types of patients are recorded in the medical record databases should be given. Are the controls patients with rheumatic diseases but without RA diagnosis? Furthermore are the patients (cases and controls) recorded in the databases only for ambulatory visit or are they also inpatients? Lines 33-46. A clear information about inpatients and outpatients should be provided when dealing about the population in the validation set. This point is relevant because it is related with the next. Page 6, lines 33-46. Description of content of the two registries (or references if described in other articles) should be provided. Page 7, lines 18-25. The statement about the AHD is correct. However, generally administrative databases in Italy do not provide clinical diagnosis for outpatients. Here, authors should clarify what information from AHD used for the outpatients. This issue is relevant also for the next paragraph that deals with RA cases. Lines 27-31. This point is very interesting. The authors state that they reviewed the literature for RA case definition. They should report either any publication they may have produced or provide a description of the process and submit it with the paper as a supplemental file. Page 7 line 52-53. "For each variable identified": please mention here the list of the variables. Page 7 Lines 55- Page 8 Line 7. This point is extremely important for the understanding of the algorithm construction. Authors should describe clearly the different steps by mentioning the variables with which they started the assessment of the SE and SP and the variables added in the subsequent assessment. The use of the examples can make the issue easier to understand. Results Page 9, lines 5-16. This paragraph of the variable selection should go in the methods section and presented appropriately.

	Lines 18-20. The listed candidates not included in the analysis should be removed or moved in the Methods section if there were interesting points to mention. After revising the flow of the description of the methods and clarifying the issue of the validation set, authors may consider writing the results following the amendment in the methods.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name Hjalmar Wadström
Institution and Country Karolinska Institutet, Sweden

• In figure 1, 35 patients are excluded on the basis of “Not meeting criteria for classification in RA and Non-RA”, on the other hand on page 10 under “study samples” it seems that these patients were excluded because of lacking information. Please clarify what you mean by this.

We modified the sentence in the figure 1, to make clearer this point as suggested.

• This study is based on register data and the quality of such studies hinges upon good data quality and coverage. Therefore, it might improve the paper if some information or maybe a reference on the quality of these registers was added. Also, are missing values prevalent in these registers and how were they handled?

The validity and completeness of the AHD of the Lombardy Region is based on a wide literature. We added two references to support this.
We included the sentence “Only subjects successfully linked were retained for the analyses” in the methods section to underline the complete case analysis we performed.
Since we assume that missing were at random, it is unlikely that this choice generated bias.

• To measure the prevalence of RA is not mentioned in the title or the objectives, should it have been? Or was this just a side note?

The objective is stated in the introduction but not in the abstract. We included this objective also in the abstract.

• When entering the numbers from table 4 in formula 1 on page 9 it seems the result differs from that listed on page 11 (0.31%). Maybe this could be explained further?
adjusted prevalence= 0.52+0.9977-1 / 0.9250+0.9977-1

We agree with the reviewer that the formula might be misleading, because we report prevalence (and Se, Sp…) as % in the text while we require the same not in % in the formula. The formula is now modified using only %. The right adjusted prevalence is now= 0.52 +99.77-100 / 92.5+99.77-100, as now reported also in the text.

• The final algorithm is not specified in the results section, probably because of the complexity of it, instead it is listed in table 3. It might be helpful for the reader if it was emphasized in Table 3 that the algorithm listed is the final algorithm that was used or if it was specified in the results section.

Thank you to the reviewer for this suggestion. We agree that the final algorithm was not enough emphasized. In the results we declared the 4th step of the algorithm as the final one and we also

modified the table to visually emphasize this aspect.

• In table 2, the total number of cases and controls are not listed. It would improve readability if this was added.

We apologise for the lack of these useful data, which are now reported.

Reviewer: 2
Reviewer Name Iosief Abraha
Institution and Country Health Planning Service
Regional Health Authority of Umbria
Perugia, (Italy)

The article is interesting in its field and it might increase the interest of the validation process to be applied in the Italian administrative databases (http://dx.doi.org/10.6000/1929-6029.2014.03.03.10).

1. While it clear the reference standard is represented by the clinical diagnosis from the medical records, it not completely clear what information from the administrative databases should function as an index test. The only 'clinical' information administrative databases have is the ICD9 code, thus, it is unclear how the two validation sets (a secondary rheumatology centre and primary care) could be related with administrative database. In other words is the administrative database used as a third validation set? Shouldn't it be considered as the main target for the validation given that it appears in the title and it was used to calculate the RA prevalence in the Lombardy region? This is the most critical point that authors need to address.

We agree with the general comment of the reviewer that more clarity is needed in reporting the applied methodology in order to avoid misunderstanding.
In our design we tested administrative variables [exemption codes, hospital discharge form diagnoses, drug prescription data] as candidate index texts against the clinical diagnosis (reference standard), connecting administrative data and clinical diagnoses by deterministic record linkage using tax code as unique identifier. The record linkage was performed for both the training and validating sets. The developed algorithm combines candidate variables, and it represents our final index text for validation. The index text is now explicitly reported in the method section.

To improve clarity, we made specific changes in response to this comment highlighted in the text.

2. RA patients are often outpatients. Administrative databases usually does not have outpatient information. If this is not the case, authors need to clarify which information of outpatients have the AHD since the ICD9 code is related only to inpatients.

No 'clinical' outpatient information is retrieved from administrative databases.

3 The flow of the description of the methodology should be clearly presented and avoid placing descriptions in the results section. Other, suggestions can be found below.

Descriptions were moved from the results to the methods section as suggested.

4 Page 4, lines 44-49. The authors might report a reference for the statement or provide a further description of the RECORD study.

More details on the RECORD Project are now reported in the Introduction.

• Patients and Methods: Explanation about the training set should be given. Page 6, lines 14-22. While the cases are intuitively patients with RA it is unclear who the controls are. In other words clarification as to what types of patients are recorded in the medical record databases should be given. Are the controls patients with rheumatic diseases but without RA diagnosis? Furthermore are the patients (cases and controls) recorded in the databases only for ambulatory visit or are they also inpatients?

The controls are only rheumatologic outpatients without diagnosis of RA (other rheumatological diagnosis or no rheumatological diagnosis) for rheumatologic samples (training and first validating set) and all the subjects without RA in the general population / primary care sample (second validating set). These definitions are now more explicit in the method section.

• Patients and Methods: Lines 33-46. A clear information about inpatients and outpatients should be provided when dealing about the population in the validation set. This point is relevant because it is related with the next.

We agree with the reviewer that this point was obscure. We hope that in the new version this point has made clearer: we started with a list of outpatients that did not include inpatient subjects.

• Patients and Methods: Page 6, lines 33-46. Description of content of the two registries (or references if described in other articles) should be provided.

We apologies for the misleading wording. We used the term 'registry' to refer to the list of subjects along with their diagnoses, which were retrievable from the electronic medical record. Now we resolved this misunderstanding in the text.

• Patients and Methods: Page 7, lines 18-25. The statement about the AHD is correct. However, generally administrative databases in Italy do not provide clinical diagnosis for outpatients. Here, authors should clarify what information from AHD used for the outpatients. This issue is relevant also for the next paragraph that deals with RA cases.

The reviewer is right: no ICD9-CM codes for outpatients were available. We only used available information from AHD [exemption codes, hospital discharge form, drug prescription data] to build up our algorithm (index text), to be tested against the clinical diagnoses (reference) obtained by the record linkage between administrative data and clinical data. We also made explicit the source of exemption codes.

• Patients and Methods: Lines 27-31. This point is very interesting. The authors state that they reviewed the literature for RA case definition. They should report either any publication they may have produced or provide a description of the process and submit it with the paper as a supplemental file.

Key words and mesh terms are now reported in the text and the process of the literature search reported in a supplementary material. We did not emphasize this point since it was not a systematic literature review (we searched only Medline) we carried out to identify potential items to be included in the algorithm. Moreover, to date two well designed and reported SLR are available on this topic, and we think that it is not useful to report similar results to those obtained by an extensive search:
• Widdifield, Jessica, Jeremy Labrecque, Lisa Lix, J. Michael Paterson, Sasha Bernatsky, Karen Tu, Noah Ivers, and Claire Bombardier. “Systematic Review and Critical Appraisal of Validation Studies to Identify Rheumatic Diseases in Health Administrative Databases.” Arthritis Care & Research 65, no. 9 (2013): 1490–1503. doi:10.1002/acr.21993.
• Chung, Cecilia P., Patricia Rohan, Shanthi Krishnaswami, and Melissa L. McPheeters. “A Systematic Review of Validated Methods for Identifying Patients with Rheumatoid Arthritis Using

Administrative or Claims Data." Vaccine, Active Surveillance of Vaccine Safety in the US Food and Drug Administration's Mini-Sentinel Program: Identification of Exposures and Outcomes, 31, Supplement 10 (December 30, 2013): K41–61. doi:10.1016/j.vaccine.2013.03.075.

• Patients and Methods: Page 7 line 52-53. "For each variable identified": please mention here the list of the variables.

Now we added the list in the Methods section

• Patients and Methods: Page 7 Lines 55- Page 8 Line 7. This point is extremely important for the understanding of the algorithm construction. Authors should describe clearly the different steps by mentioning the variables with which they started the assessment of the SE and SP and the variables added in the subsequent assessment. The use of the examples can make the issue easier to understand.

We now more clearly explain this aspect in the text, as suggested.

• Results: Page 9, lines 5-16. This paragraph of the variable selection should go in the methods section and presented appropriately.

We moved this paragraph in the Methods section.

• Results: Lines 18-20. The listed candidates not included in the analysis should be removed or moved in the Methods section if there were interesting points to mention.

We removed this list because it is deducible from the list of candidate items subtracting the selected ones.

* After revising the flow of the description of the methods and clarifying the issue of the validation set, authors may consider writing the results following the amendment in the methods.

In the reporting of the methods and results we followed the STARD statement - as requested by the Journal. After the higlighted amendments, the structure of methods and results are now consistent.

VERSION 2 – REVIEW

REVIEWER	Iosief Abraha Health Planning Service Regional Health Authority of Umbria Perugia, Italy
REVIEW RETURNED	15-Dec-2014

GENERAL COMMENTS	The revised version of the paper is satisfactory. However, in the astract, results section, I suggest the authors to specify the content of the 4-step algorithm for which they provide sensitivity and specificity.