PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	An algorithm to identify rheumatoid arthritis in primary care: a Clinical Practice Research Datalink study
AUTHORS	Muller, Sara; Hider, Samantha; Raza, Karim; Stack, Rebecca; Hayward, Richard; Mallen, Christian

VERSION 1 - REVIEW

REVIEWER	Elizabeth Ford Research Fellow in Primary Care Epidemiology, Brighton and Sussex Medical School UK
REVIEW RETURNED	04-Aug-2015

GENERAL COMMENTS	Many thanks for the opportunity to read this paper. I commend the authors on tackling an important topic and presenting their work clearly and concisely. I have a few comments which need addressing before I can recommend acceptance. It seems to me that the utility of this work is in allowing other researchers to use this algorithm for their GP database studies, however, the description of the algorithm is lacking. I would like to see the code lists for the algorithm supplied in an appendix or at least put in a respository such as clinicalcodes.org so they are easily accessible. Knowing which codes make up the algorithm would also allow the reader (and reviewer) to assess the quality of the algorithm and its relevance to their particular study. Without seeing the code list I cannot really assess the new algorithm and its importance in the field. Secondly this study develops the algorithm by first choosing patients with one Read code for RA, and testing the algorithm in this cohort. Therefore, the sensitivity of the algorithm cannot be addressed, but this weakness is not mentioned throughout the manuscript. No weaknesses of the algorithm are properly discussed - what is the rate of true vs false negatives? what about delay in diagnosis? Who could form a suitable control group to test sensitivity? Even if you are not going to address the issue of sensitivity, because your algorithm is only to create a very specific and well defined patient group, this must be explicitly stated. For example, a researcher who wanted to do a prevalence study, a clinical audit, or plan health services for RA using CPRD, would not want to use this algorithm as they would not know how many RA patients they were missing. A useful discussion point would also be what weaknesses of such algorithms are inherent to CPRD algorithms and which are particular to this algorithm. Specific comments:

	1) Methods page 7. What was the length of each patient's record obtained from CPRD? 2) Page 8. Give a code list for defining each condition which was an alternative indication for DMARDs in an appendix. 3) Results page 9. Over what time period did the algorithm have to be fulfilled? Add to methods 4) Discussion page 10. The first paragraph of the discussion needs rethinking. You say accurate diagnosis of RA is of important clinically, to initiate treatment. However, your algorithm incorporates treatment as part of the case definition. So the argument is circular. If cases are defined on having received treatment, they can't be useful in identifying people who need to start treatment. 5) page 11, you say a valid diagnosis and a clean sample of patients is needed for analysis - this is only true for some types of studies. For other studies, as mentioned above, increased sensitivity is more important. 6) Only 3 years within CPRD were used (2010-2012), so it is hardly worth mentioning time trends in the discussion. 7) Page 12 - discussion section on the differences between this and previous algorithm: Follow up time is mentioned for the first time here - please make it clear in the methods. Implementation of NICE guidelines in 2009, and other recommendations to treat RA early with DMARDs may have affected diagnosis and practice - would be good to mention in discussion. Could the new algorithm be more specific within the new cohort because more RA patients are missing any code for the disease? Are absolute numbers of RA patients in the CPRD going up or down with time? 8) page 13 - here you mention the follow up time available for fulfilling the definition of RA but we still don't know what the timing is. 9) page 13 - you say a strength of the study is that it excludes the period during which RA was added to QOF but this is hardly a strength, as all future research will probably incorporate the later period and the algorithm may need revising again in light of this. This would be better mentioned as a limitation.

REVIEWER	Carlo Alberto Scirè Research Coordinator Epidemiology Unit Italian Society for Rheumatology Milan, Italy
REVIEW RETURNED	03-Oct-2015

GENERAL COMMENTS	This is a methodological work evaluating the validity of an updated definition of RA in a primary care database in the UK for future studies. I would suggest to address some general aspects. The study does not follow the usual ‘diagnostic accuracy’ design required of validation of algorithms, and therefore reporting does not follow STARD recommendations. The authors do no refer to relevant standards of reporting for validation studies of algorithms for identification of patients from administrative healthcare databases. I would suggest to include relevant references. Furthermore it does not follow the STROBE of reporting for observational studies, for example sample size seems to be missing is missing at page 7. I would suggest to report these results according to strongly justify the study design beyond the simple statement of impossibility to access to patients to perform a true validation. The users of the algorithm might be interested in its accuracy (Se/Sp), or at least

	PPV. Study sample selected subjects with at least one of the RA-related Read Codes. Looking at this study as a validation study, given that the reference standard includes these codes, this approach could lead to a verification bias. The updated algorithm combining RA-related Read Codes and drug prescription was used as reference standard. Again, in the framework of validation studies, this approach may lead to incorporation bias. Furthermore this approach does not validate the accuracy of the new algorithm, because it is the reference standard. If the true objective of the paper is limited to a description of the updating process more details on the items, the process should be provided, though this could affect the relevance of the research question.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name
Elizabeth Ford
Institution and Country
Research Fellow in Primary Care Epidemiology, Brighton and Sussex Medical School
UK
Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below

Many thanks for the opportunity to read this paper. I commend the authors on tackling an important topic and presenting their work clearly and concisely. I have a few comments which need addressing before I can recommend acceptance.

Major Comment A
It seems to me that the utility of this work is in allowing other researchers to use this algorithm for their GP database studies, however, the description of the algorithm is lacking. I would like to see the code lists for the algorithm supplied in an appendix or at least put in a repository such as clinicalcodes.org so they are easily accessible. Knowing which codes make up the algorithm would also allow the reader (and reviewer) to assess the quality of the algorithm and its relevance to their particular study. Without seeing the code list I cannot really assess the new algorithm and its importance in the field.

Response: We thank the reviewer for highlighting this point. We debated the inclusion of the codes as an appendix and would be happy to take the editor's guidance on this point. We have included them with our resubmission for the reviewer to see and have amended the manuscript to state that they will be available at clinicalcodes.org as well as in our institutional repository. In addition, we would be quite willing to share Stata syntax so that others can reproduce the algorithm more easily.

Changes to manuscript (page 9):
"Full lists of the codes used to define RA, DMARDs and their alternative indications, and alternative diagnoses are available from the clinicalcodes.org website and in the authors' institutional repository (keele.ac.uk/mrr)."

Major Comment B
Secondly this study develops the algorithm by first choosing patients with one Read code for RA, and testing the algorithm in this cohort. Therefore, the sensitivity of the algorithm cannot be addressed,

but this weakness is not mentioned throughout the manuscript. No weaknesses of the algorithm are properly discussed - what is the rate of true vs false negatives? what about delay in diagnosis? Who could form a suitable control group to test sensitivity? Even if you are not going to address the issue of sensitivity, because your algorithm is only to create a very specific and well defined patient group, this must be explicitly stated. For example, a researcher who wanted to do a prevalence study, a clinical audit, or plan health services for RA using CPRD, would not want to use this algorithm as they would not know how many RA patients they were missing. A useful discussion point would also be what weaknesses of such algorithms are inherent to CPRD algorithms and which are particular to this algorithm.

Response: We apologise if the aim of the paper is not clear. As the reviewer suggests, we do not intend to study the diagnostic accuracy of this definition of RA, or for this algorithm to be used all studies of RA in the CPRD (or any other database). Rather, we intended to update the original piece of work, which also set out to create a specific definition of the condition opposed to a highly sensitive definition. We also wanted to describe the characteristics of the RA population captured using this definition in our sample, compared to the original definition. We have amended the final paragraph of the Introduction accordingly.
Furthermore, we have added to the Discussion to highlight that from our study design, it is not possible to assess the formal performance of the algorithm in terms of sensitivity, specificity or true and false positive rates. We have also added discussion points around the ideas of how these concepts might be tested and in what circumstances the algorithm would and would not be useful. Delays in diagnosis are already discussed on page 14. Additionally, we have added a general discussion of the strengths and weakness of routinely collected data in research and how these relate to the current study.

Changes to manuscript - Introduction (page 5):
“Therefore the aim of this study was to describe our updating of the definition of Thomas et al[5] in order to create an up-to-date algorithm to identify highly probable RA cases in the CPRD and to compare the characteristics of the sample and the algorithm’s performance to the original.”

Changes to manuscript – Discussion (page 13-14):
“For the reasons discussed above, those wishing to apply the updated algorithm should do so with caution, particularly in the situation where a sensitive definition of RA is required (e.g. prevalence study, clinical audit). The current algorithm is likely to be unsuitable for such studies, as it is designed to find those with highly probable RA. However, if changes in coding practice have occurred in the manner discussed above with GPs more certain of a diagnosis before entering a code, the updated algorithm may be more specific than the original. Before the algorithm is used in settings where a less specific definition of RA is required, it would be sensible to formally test its performance, by comparing to full medical records, as was the case in its original development. However, this was beyond the scope of the current study.
In addition to the potential weaknesses of this study discussed above, there are some limitations to the use of clinical databases in general that should be considered in all such studies. These include a reliance on what is coded by the general practioner, which may be different to what the patient perceived they were consulting with, and indeed may not reflect the entire content of a consultation (if only the main clinical problem was coded). This is particularly the case when considering symptoms, as opposed to clear-cut diagnoses, but is less of a problem with prescriptions, which are generally issued electronically and therefore accurately recorded by default. In addition, it is usually not possible to understand the reasons for a particular diagnostic code or prescription being recorded and one must rely on what is in the record having been a true event and assume that anything that is not present did not happen.”

Specific comments:

1) Methods page 7. What was the length of each patient's record obtained from CPRD?

Response: The full record available was downloaded from the CPRD dataset. This was variable across individuals. Each patient was required to have at least 3 years up-to-standard data before the index date. No requirements were imposed after the index date, as the data were extracted as the basis of a further study. However, empirically, consultation data were available for a median of 3.25 years (IQR 2.5, 4.1) after the index date and 37.7 years (25.4, 49.0) prior to the index date. The time frame before the index date is due to the nature of the CPRD data whereby practices can enter information held on paper records into the electronic system. We have added the following to the manuscript to describe the time frame of the dataset.

Changes to manuscript - Methods (page 7):
"The full period of the record held by the CPRD was downloaded for all individuals in the sample, before and after their first RA code."

Changes to manuscript – Results (page 9):
"The median length of time from the index date (date of first RA code) to the final consultation in the record of these patients was 3.25 years (interquartile range 2.5, 4.1), and the median length of the consultation record prior to the index date was 37.7 years (25.4, 49.0)."

2) Page 8. Give a code list for defining each condition which was an alternative indication for DMARDs in an appendix.

Response: Please see response to Comment 1 above. These code lists are provided for the reviewer and will be archived in clinicalcodes.org and our institutional repository.

3) Results page 9. Over what time period did the algorithm have to be fulfilled? Add to methods

Response: We did not impose a time-scale in which the algorithm was required to be fulfilled, as there was no such time scale in the original algorithm. However, the maximum follow-up time was a maximum of approximately 5 years, as data were extracted from the CPRD in April 2015. We have added a sentence to the Methods section to clarify this.

Changes to manuscript (page 9):
"Searches for appropriate codes to implement the algorithm were conducted in all available data for each individual."

4) Discussion page 10. The first paragraph of the discussion needs rethinking. You say accurate diagnosis of RA is of important clinically, to initiate treatment. However, your algorithm incorporates treatment as part of the case definition. So the argument is circular. If cases are defined on having received treatment, they can't be useful in identifying people who need to start treatment.

Response: We thank the reviewer for pointing out the potential confusion here. However, we did not suggest that the algorithm should be used to identify patients for early treatment. Rather we said that because early treatment is now regarded as a gold standard, more research is needed to enable clinicians to accurately identify those with early RA and that the algorithm could aid this research. As

this was obviously unclear, we have reword the first paragraph of the Discussion as follows:

Changes to manuscript (page 11):
Accurate diagnosis of RA is of paramount importance clinically, as current guidelines recommend early and aggressive treatment with DMARDs. In order to take this approach clinically, further research will be necessary to accurately identify patients with probable RA in primary care. This updated algorithm could contribute to this research. Without suitable means of defining an RA cohort that has a high probability of being true RA, such studies would be of poorer quality.

5) page 11, you say a valid diagnosis and a clean sample of patients is needed for analysis - this is only true for some types of studies. For other studies, as mentioned above, increased sensitivity is more important.

Response: Again we thank the reviewer for pointing out this inaccuracy. Please see response to Comment B above.

6) Only 3 years within CPRD were used (2010-2012), so it is hardly worth mentioning time trends in the discussion.

Response:
Our intention here was really just to show that the proportion of people meeting the algorithm was relatively stable, rather than to conduct a comparison. We have added to the Results sections to make this clearer and have also amended our mention of this in the Discussion.

Changes to manuscript - Results (page 10):
“On the whole, the proportion of people with a single RA code meeting the updated definition of RA was relatively stable across the three years included in this study, although the definition of RA was less likely to be met in those receiving their first RA code in 2011 (88.0%), with slightly lower rates of confirmed diagnosis in earlier and later years. (p=0.029). This difference is driven by a combination of differences in the number of people with a suitable DMARD and the number of people with multiple RA codes (Table 3). “

Changes to manuscript - Discussion (page 14):
“Our investigation of the proportion of people fulfilling the definition RA according to the year was intended to investigate the algorithm’s stability over time. However, it also gave some insight into the time required to fulfil the criteria (e.g. second RA code). The stability of the proportion fulfilling the definition over time suggests that 12 months seems a reasonable time frame in which to consider follow-up after the first RA code, in order to apply this definition.”

7) Page 12 - discussion section on the differences between this and previous algorithm: Follow up time is mentioned for the first time here - please make it clear in the methods. Implementation of NICE guidelines in 2009, and other recommendations to treat RA early with DMARDs may have affected diagnosis and practice - would be good to mention in discussion. Could the new algorithm be more specific within the new cohort because more RA patients are missing any code for the disease? Are absolute numbers of RA patients in the CPRD going up or down with time?

Response: We have added information on the time frame of the study to the Methods and Results sections (see responses to previous comments). We have also added discussion points regarding the likely meaning of an RA code in the record now compared to the time of the original study and also

relating to the introduction of the NICE guidelines. As mentioned in response to Comment 9 below, we believe the number of RA cases is stable over the relatively short time period that we have considered in this study, with the exception of the year that it was fully included in the QOF. This number was also stable for some years prior to this, but there has been grown in the number of new RA cases since the original algorithm was developed (see below). However, we do not have denominator data with which to work out rates of diagnoses in each year and the CPRD has grown considerably in size over this time period.

Financial year: number of incident RA cases
1990-91: 13
1999-2000: 826
2007-08: 1559
2010-11: 1448
2011-12: 1435
2012-13: 1540
2013-14: 2821
2014-15: 1496

Changes to manuscript (page 12-13):
“This may mean that a single code for RA is now a more accurate reflection of a true diagnosis of RA than was previously the case.”
“Similarly, the introduction of the National Institute for Health and Care Excellence Rheumatoid Arthritis Guideline in 2009 should have prompted faster referral by GPs of suspected RA patients to secondary care. This may have increased the speed of secondary care diagnosis, potentially increasing the accuracy of the diagnoses recorded in primary care records such as the CPRD.”

8) page 13 - here you mention the follow up time available for fulfilling the definition of RA but we still don't know what the timing is.

Response: Please see response to Comments 1 and 3.

9) page 13 - you say a strength of the study is that it excludes the period during which RA was added to QOF but this is hardly a strength, as all future research will probably incorporate the later period and the algorithm may need revising again in light of this. This would be better mentioned as a limitation.

Response: We think that our original description of the inclusion of RA in the QOF was not sufficient and have added to it in the manuscript. For information for the reviewer, this was our reasoning for excluding this single year from our work.
RA was included in the QOF only as a pilot in 2013-14, and subsequently removed. As is well known, the inclusion of a condition in QOF can affect the way it is recorded in primary care data. This is what we found for RA: there was a stable number of newly coded cases of RA in each year up to its inclusion in QOF. The year it was included, there was a large jump in the number recorded, with almost double the number of incident cases. The year after (2014-15), this number returned to its previous, stable level (see numbers in response to Comment 7). We cannot explain this other than by the inclusion of RA in QOF. This is the reason that we feel it was a strength not to include the data from 2013-14 in our analysis, even though this may seem counter-intuitive. We have not added the above data to the paper, as we felt it would over-complicate the discussion. However, we have added a fuller explanation of our choice of time frame to the analysis and would reiterate the caution about using this year in any analysis of RA data in the CPRD.

Changes to manuscript (page 15):
“A strength of the current study was that it was careful to exclude the period when RA was included in the Quality and Outcomes Framework (QOF) a set of quality standards by which UK GPs receive some of their funding. In 2013-14, RA was included in the QOF, requiring GP to maintain a register of patients, provide them with a face-to-face review and dependent on their age, screen them for cardio-vascular disease and fracture risk. This package of care was worth 18 QOF points. In the following and subsequent years, this was reduced to only the register and review and worth only 6 points. The inclusion of a condition in QOF has this has been known to alter the way in which GPs code the conditions and indeed we found that the number of individuals with a new RA code was considerably higher in this 2013-14 than in the years before or after. Future studies should exercise caution if including this one year period in their work, as the algorithm has not been tested in this setting.”

Reviewer: 2
Reviewer Name
Carlo Alberto Scirè
Institution and Country
Research Coordinator
Epidemiology Unit
Italian Society for Rheumatology
Milan, Italy
Please state any competing interests or state ‘None declared’:
None declared

Please leave your comments for the authors below

This is a methodological work evaluating the validity of an updated definition of RA in a primary care database in the UK for future studies.
I would suggest to address some general aspects.

Comment 1
The study does not follow the usual ‘diagnostic accuracy’ design required of validation of algorithms, and therefore reporting does not follow STARD recommendations. The authors do no refer to relevant standards of reporting for validation studies of algorithms for identification of patients from administrative healthcare databases. I would suggest to include relevant references. Furthermore it does not follow the STROBE of reporting for observational studies, for example sample size seems to be missing is missing at page 7.

Response: As the reviewer notes, we did not follow the STARD recommendations for the reporting of diagnostic accuracy, as the aim of this paper was not to test the diagnostic accuracy of this algorithm. Rather it was simply to update it, accepting that the original algorithm had been tested in this way. We did attempt to follow the STROBE guidelines, however, as this is a rather unusual study design, not all were applicable. We have however revisited the manuscript to ensure that we have followed these recommendations as closely as possible in this situation. Exceptions to fully following STROBE are as follows:
For Items 7 and 8 (variables and data sources/ measurements), we do not have any variables as such, but we have referenced to where we discuss the updating of the components needed for the algorithm (pages 6-7).
As the reviewer notes, we did not conduct a sample size calculation. These data are taken from a larger study that is still in progress and required this updated algorithm. The sample size was therefore calculated for that study. For the purposes of this algorithm update, we feel it is sufficient to state the number of cases we achieved in the three years studied. In the main study, we chose those

three years as they gave us an appropriate sample size based on our power calculation. We have amended the checklist to indicate that the sample size is not applicable in this study.
For item 15 (outcome data), we have considered meeting the algorithm to be our outcome.

Comment 2
I would suggest to report these results according to strongly justify the study design beyond the simple statement of impossibility to access to patients to perform a true validation. The users of the algorithm might be interested in its accuracy (Se/Sp), or at least PPV.

Response:
Whilst we appreciate that the sensitivity, specificity, PPV and NPV may be of interest to readers, these are not values we can calculate in the study, as we do not have appropriate data. In response to comments from Reviewer 1, we have added discussion of this to the manuscript (pages 13-14). We have clearly stated the sensitivity and specificity of the original algorithm when that was developed and fully tested.

Comment 3
Study sample selected subjects with at least one of the RA-related Read Codes. Looking at this study as a validation study, given that the reference standard includes these codes, this approach could lead to a verification bias.

Response: Whilst this may be the case if this were a diagnostic study, we do not think it to be true in the current study. In the original algorithm, which we are attempting to update, the starting point was people with a diagnostic Read code for RA. Therefore, this is the starting point that we have also adopted. This is common in studies using routinely collected data, as may studies would simple define anyone with a single Read code for RA as having RA. The idea of this algorithm is to improve on this starting point for cases when this is necessary (e.g. when a very specific definition of RA is needed). There would be no sense in applying the algorithm in those without a code. Indeed if one were to do so, everyone without one or more RA codes would be excluded at the start.

Comment 4
The updated algorithm combining RA-related Read Codes and drug prescription was used as reference standard. Again, in the framework of validation studies, this approach may lead to incorporation bias. Furthermore this approach does not validate the accuracy of the new algorithm, because it is the reference standard.

Response: As we have stated in response to previous comments, the aim was not to validate the new algorithm, more simply to update it and to describe it's characteristics in a sample of those with one or more RA codes. Hence we do not think that incorporation bias is relevant here.

Comment 5
If the true objective of the paper is limited to a description of the updating process more details on the items, the process should be provided, though this could affect the relevance of the research question.

Response: The reviewer is correct that this was our primary aim. Although we did also seek to describe the characteristics of the algorithm in those with one or more RA codes, so that in this way we could compare it to the sample used in the original study that devised the algorithm. We accept that this is a very different question to one of diagnostic accuracy. However, we still feel this is relevant and important, as a suitably specific definition of RA is required in a range of studies and an

up-to-date way to define RA in these circumstances is therefore useful to future research. This is a major advance on the current situation, as the previously published algorithm is out of date and a full specification of the algorithm was never available, something we have rectified.
In terms of fully describing the process of updating the algorithm, we describe the updating of the lists of codes and the processes used to do this in the section entitled “Updating the Thomas algorithm” (pages 6-7).

VERSION 2 – REVIEW

REVIEWER	Elizabeth Ford Brighton and Sussex Medical School, UK
REVIEW RETURNED	10-Nov-2015

GENERAL COMMENTS	I commend the authors on the improvements to their paper, I am happy for it to be published.

REVIEWER	Carlo Alberto Scirè Epidemiology Unit Italian Society for Rheumatology Milan, Italy
REVIEW RETURNED	15-Nov-2015

GENERAL COMMENTS	I understand that it is not a study addressing the accuracy of the proposed algorithm but only a description of a new algorithm for RA with unknown accuracy, based on available data. I would suggest at least to remove the word specificity from the conclusions of the abstract. I also suggest to quote more literature on the topic of healtcare database use in epidemiological research.

VERSION 2 – AUTHOR RESPONSE

We have made the additional revisions suggested by Reviewer 2, as outlined below.

I would suggest at least to remove the word specificity from the conclusions of the abstract.
We have reworded the final sentence of the abstract to read:
“Those wishing to define RA in the CPRD, should consider using this updated algorithm, rather than a single RA code, if they wish to identify only those who are most likely to have RA.”

I also suggest to quote more literature on the topic of healthcare database use in epidemiological research.
Approximately half of our references already cite papers that have used health care databases in epidemiological research. We have however amended the second paragraph of the Introduction as follows, citing the CPRD website, which details the number of papers published using its data in 2014.
“One potential way of investigating RA in primary care is the use of health care databases, for example the Clinical Practice Research Datalink (CPRD), QResearch or The Health Improvement Network (THIN). Use of such databases in epidemiological research is increasing, with CPRD data used in over 190 studies in 2014.[4] These data sources include data recorded in routine clinical practice, such as information regarding symptoms, diagnoses, prescriptions and referrals. These large databases are highly generalizable, because they cover large numbers of people from the general population (e.g. CPRD covers approximately 6% of the UK population [5]), meaning that they can be used efficiently in epidemiological studies.”