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)	What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text
AUTHORS	Ford, Elizabeth; Carroll, John; Smith, H; Davies, Kevin; Koeling, Rob; Petersen, Irene; Rait, Greta; Cassell, Jacqueline

VERSION 1 - REVIEW

REVIEWER	Jessica Widdifield McGill University
REVIEW RETURNED	23-Nov-2015

GENERAL COMMENTS	The Study: The results of this study suggest that free text information relevant to RA diagnosis can often be found under correspondence codes, general consultation codes and general symptom codes. While this current study is potentially useful data, I have several general comments. Results & Conclusions: The findings suggest that up to 22% of patients may experience a delay in the coding of their RA diagnosis. However, the critical issue is that the authors fail to quantify the number of RA patients who experience no diagnostic coding for RA. The authors make an excellent case regarding issues of “missing or delayed diagnostic codes” in the introduction, but the work is only focused on delayed coding (which has less impact on researchers and clinical care than if RA codes are missing altogether). Given that the study sample is drawn from all cases with at least one diagnostic code of RA, the results reported may be biased if the text information captured in these patients’ EHRs is different from those patients who are completely missing diagnostic codes for RA. Moreover, recent studies from the UK CPRD report that the prevalence of RA (by diagnostic codes) is much lower (~0.5%) in comparison to recent international studies (~1.0%). This raises concerns regarding the poor sensitivity of RA diagnostic codes in the CPRD & how free text information could possibly improve case detection in the CRPD. While the delay in diagnostic coding may be important, it would be much more useful if the authors could also contribute information on the impact of free text diagnoses for those with missing RA codes. General comments: Discussion line 45 “Very little previous work has sought to use text to establish delayed recording of diagnosis in electronic health records research”. This is a rapidly growing field and numerous studies are incorporating free text into EHR research. Conclusion – “Natural language processing techniques have the

	potential to extract information from text at scale." The authors mention this methodology in the conclusion but fail to comment on this methodology in the previous sections. Rather, the discussion mentions 'automated text analysis systems', which is not always 'Natural language processing techniques'. Figure 1 – label axes Tables – when reporting '% of all strings", perhaps incorporate "% of unique patients". Limitations – RA is a heterogeneous disease, and there may be delays in establishing a diagnosis; DMARDs initiated for inflammatory arthritis, which evolved into RA, may not necessarily reflect a delay in diagnosis code, but rather a delay in the clinician establishing a definitive diagnosis. Title – "Rheumatoid arthritis in UK general practice – what can free text tell us about diagnostic recording?" Perhaps the authors could incorporate a more accurate description of the study with respects to the "delay of diagnostic coding".

REVIEWER	Mark Nielen NIVEL, Utrecht, the Netherlands
REVIEW RETURNED	14-Dec-2015

GENERAL COMMENTS	Major comments: - This is an interesting study which describes the additional value of using free text fields when determining the date of diagnosis of RA. The research question is very clear, however, the main issue of the manuscript is that there are a large number of very detailed tables which are not needed to answer the research questions. This not only resulted in a very complicated manuscript, but it is also difficult to find the main message: does using free text form EHRs increase the number of incident RA cases? And to what extent does this influence the date of diagnosis? The manuscript would improve when only results are shown that can be used to answer the research question. - I do not understand why in table 3 the author report on text level instead of patient level. The most important outcome of the study is whether free text fields influence the date of diagnosis in newly diagnosed RA patients. - Introduction: there is so much information given in the introduction that the purpose of the study is not completely clear: why did the authors use rheumatoid arthritis as an example? Why is it so important to study this subject? Is it only a problem for selecting study populations for research or does it also influence care? - I have some privacy concerns. Is it allowed to extract free text fields from EHRs in the UK? What happened when the GP entered for instance a name or phone number of a patient in their HER? - The authors used data from 2005-2008. In the Netherlands data quality improved a lot the last five years. This issue of data quality should be added to the discussion section. When recording diagnosis by the GP improves, it is very likely that in 2015 using free text field does not add as much as in the period 2005-2008. Minor comments: - I do not know the concept of Read codes. Could you please add more information about this method in the method section? Also, please remove the example about these code from the introduction to the method section.

	- Why did the authors randomly select 294 cases? Was this based on a power analysis? - The study is mainly focused on the UK. A comparison with studies in primary care of other countries is missing.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

3) Title – "Rheumatoid arthritis in UK general practice – what can free text tell us about diagnostic recording?" Perhaps the authors could incorporate a more accurate description of the study with respects to the "delay of diagnostic coding".

- We have changed the title on page 1 to: "What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text"

4) Results: The findings suggest that up to 22% of patients may experience a delay in the coding of their RA diagnosis. However, the critical issue is that the authors fail to quantify the number of RA patients who experience no diagnostic coding for RA…the work is only focused on delayed coding (which has less impact on researchers and clinical care than if RA codes are missing altogether). Results reported may be biased if the text information captured in these patients' EHRs is different from those patients who are completely missing diagnostic codes for RA.

- Many thanks for your comments. We agree that the idea that codes may be missing for RA is mere speculation based on this data, so we have re-written the introduction to reflect that the study only evaluates delayed coding (pages 4-7). We have speculated on missing codes in the discussion and suggested further work should be done to quantify this (page 16).

5) Moreover, recent studies from the UK CPRD report that the prevalence of RA (by diagnostic codes) is much lower (~0.5%) in comparison to recent international studies (~1.0%). This raises concerns regarding the poor sensitivity of RA diagnostic codes in the CPRD & how free text information could possibly improve case detection in the CRPD. While the delay in diagnostic coding may be important, it would be much more useful if the authors could also contribute information on the impact of free text diagnoses for those with missing RA codes.

- We would love to reference this prevalence study in the discussion but after an extensive search, we cannot find the paper you refer to.

Discussion:

6) This is a rapidly growing field and numerous studies are incorporating free text into EHR research.

- Yes we agree and have included a new paragraph on this topic (page 15-16). However, the bulk of these free text studies are being performed on US-based hospital records data, (see my 2016 JAMIA review of the literature for more detail; https://jamia.oxfordjournals.org/content/early/2016/02/04/jamia.ocv180). However, very few research groups are using CPRD free text due to cost and governance issues.

7) "Natural language processing techniques have the potential to extract information from text at scale." The authors mention this methodology in the conclusion but fail to comment on this methodology in the previous sections. Rather, the discussion mentions 'automated text analysis systems', which is not always 'Natural language processing techniques'.

- We have homogenised terms for consistency (page 16-18).

8) Limitations – RA is a heterogeneous disease, and there may be delays in establishing a diagnosis; DMARDs initiated for inflammatory arthritis, which evolved into RA, may not necessarily reflect a delay in diagnosis code, but rather a delay in the clinician establishing a definitive diagnosis.

- We investigated the idea that patients receiving DMARDS before RA code may have had an inflammatory arthritis diagnosis. We found they had more text entries regarding inflammatory arthritis but found no more inflammatory arthritis codes than the comparison group (added to results, page 18). We included this possibility in the discussion (page 17).

9) Figure 1 – label axes

- We have added labels to both axes of the table.
10) Tables – when reporting '% of all strings", perhaps incorporate "% of unique patients".
- As per reviewer 2's comments we have changed table 3 and integrated the information from tables 3 and 4 to form a single table, thus streamlining presentation of the results.

Reviewer 2
11) The main issue of the manuscript is that there are a large number of very detailed tables which are not needed to answer the research questions
- Thank you for your comment. We have integrated tables 3 and 4 into a single table and removed the supplementary tables from the paper (page 11-12).
12) It is also difficult to find the main message: does using free text form EHRs increase the number of incident RA cases? And to what extent does this influence the date of diagnosis?
- Thank you for your comments, we have changed the rationale for the paper, and the discussion, the rationale is now both clearer and more succinct.
13) Introduction: the purpose of the study is not completely clear: why did the authors use rheumatoid arthritis as an example? Why is it so important to study this subject? Is it only a problem for selecting study populations for research or does it also influence care?
- As above, we have taken on board your comments and have substantially re-written the introduction and discussion, we feel this is a better reflection of the analysis undertaken and makes the main message of the paper clear.
14) I do not understand why in table 3 the author report on text level instead of patient level. The most important outcome of the study is whether free text fields influence the date of diagnosis in newly diagnosed RA patients.
- We have changed table 3 to report on the patient level statistics and have combined this with table 4 (page 11-12).
15) Is it allowed to extract free text fields from EHRs in the UK? What happened when the GP entered for instance a name or phone number of a patient in their EHR?
- The governance issues around extracting free text from GP records are complicated. Currently (since 2013) CPRD are not allowed to extract any information from GP practices which contains patient identifiers, this means the free text can no longer be extracted. However, historically, the free text was extracted to CPRD and only released to researchers after a 3 stage anonymization process, hence the substantial cost of accessing the data. This is made clear in the paragraph "Free text data" in the methods on page 9.
16) The authors used data from 2005-2008. In the Netherlands data quality improved a lot the last five years. This issue of data quality should be added to the discussion section. When recording diagnosis by the GP improves, it is very likely that in 2015 using free text field does not add as much as in the period 2005-2008.
- We agree this is likely to be true so have added a sentence suggesting this in the discussion (page 20).
17) I do not know the concept of Read codes. Could you please add more information about this method in the method section? Also, please remove the example about these code from the introduction to the method section.
- A section has been included in the methods to describe Read codes, however we have decided to leave the examples about the codes and free text in the introduction as we feel it illustrates our introductory rationale that the code may only be a partial summary of the clinical information.
18) Why did the authors randomly select 294 cases? Was this based on a power analysis?
- The rationale for the sample size is given on page 8 in the paragraph "Identification of cases".
19) The study is mainly focused on the UK. A comparison with studies in primary care of other countries is missing.
- We have added this point to the discussion as a limitation (page 17-18) as these results cannot inform us about delays in coding in other primary care systems.

VERSION 2 – REVIEW

REVIEWER	Mark Nielen NIVEL, Utrecht, the Netherlands
REVIEW RETURNED	07-Mar-2016

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.