# 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)	Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database.
AUTHORS	Tate, Anne; Dungey, Sheena; Glew, Simon; Beloff, Natalia; Williams, Rachael; Williams, Tim

#### **VERSION 1 - REVIEW**

REVIEWER	Dipak Kalra
	The EuroRec Institute, Belgium
REVIEW RETURNED	30-Jun-2016

GENERAL COMMENTS	Overall comments
	This is an interesting and well conducted study that is valuable in relation to the incidence of diabetes, but more generally highlighting the challenges with taking GP data at face value without an understanding of the historic evolution of reimbursement models and other changes to the clinical, managerial and ICT ecosystems. Reminding me of the "old" days of PRIMIS, it seems still necessary to fish for the truth through contextual codes about clinical activity and treatment to correct for the patchy recording of a diagnosis. This study is especially valuable given the wider interest amongst many stakeholders to reuse routinely collected clinical data for public health, epidemiology and tother kinds of research (kinds of big health data). They need to learn from this work.
	I have made a few specific remarks below, but overall this paper reports its research methodically and details its result clearly. I do not have much to recommend by way of changes to the core of the paper.
	Given the expertise of the authors, I would have liked to see an additional section to the discussion to propose some implications from this research, since it is likely that the observed coding behaviour, trends, errors and variations would be true for multiple conditions, not just diabetes. (1) what are the ways in which diagnostic coding practice might be improved for the future, through changes to the systems, through education, or other initiatives and (2) what cautions or specific measures might users of CPRD and other equivalent data sources adopt in order to improve the accuracy of analytic interpretations made from such data sets?
	Specific remarks

summary of the key features of the data quality approach of citations 2 and 3 might be helpful to the reader.
Page 6 regarding misclassification. Just a question, not a critical view: is it not possible that a person is initially diagnosed Type 2, and control attempted with oral medication, but proves problematic to stabilise and so is moved to insulin? This raises the question of whether you looked for a subsequent Type 1 diabetes code even if a Type 2 code had been entered some months earlier. (We know that GP systems and GPs are not generally good at deleting codes from problem summaries.)
The discussion section suggests that some spurious date peaks for incidence of diabetes might be due to retrospective data entry (of data from a former GP practice). Might it be a recommendation arising from this research that GP systems need to improve the ability to attribute an entry to a prior date, and that educational investments are needed to encourage practice staff to enter such retrospective dates when they are known? Does the import of data via GP2GP message automatically assign entries to their original date of recording?

REVIEWER	Mireia Raluy-Callado
	Evidera, United Kingdom
REVIEW RETURNED	18-Jul-2016

GENERAL COMMENTS	This article focuses on the quality of recording in the UK Primary
	Care setting using data capture by CPRD through the Vision system
	on diabetes. This is a condition highly prevalent and mainly
	managed by Primary Care clinicians so it is of relevance to
	investigate patterns of coding to inform future studies. Therefore,
	this article is of interest and suitable for publication in the BMJ.
	There are two major areas where the paper would benefit from
	providing more clarification, one in the Introduction section and one
	in the Discussion. Comments are provided below by section of the
	manuscript:
	INTRODUCTION
	- The last paragraph in this section mentions the objectives of this
	study. However, the rational for using only Read codes in
	determining the diabetes diagnosis is not well explained. Given other
	publications in the past have highlighted the need of using additional
	medication and test codes to validate the diagnosis, it is unclear why
	the authors then start again from Read codes only.
	- In the second part of the study, the sensitivity analyses is only
	performed in type 2 diabetes, and it would be good to understand
	why this type only was chosen.
	METHODS
	- Under 'Measuring incidence', it could be clarified that any Read
	code from the first one till the end of follow up was used, and not
	only the index code
	RESULTS.
	- Under 'Practice Variation' 4a are all those in 4 that are not "seen
	in "? Afterwards in Table 3 are called suggested please clarify
	- Under 'Overall incidence in each category' type 1 natients are
	investigated to see if they also had a record for type 2 diabetes. Was
	medication not used to assess notential misdiagnosis?
	Figure 2: the title is unclear. In a) all codes these are all diabetes
	patients in all 4 study categories, and b) without 4b2
	patients in all 4 study categories, and b) without 4b?

- Figure 4: the size of the sample for the 2 groups is very different. Suggest adding confidence intervals around the estimated incidence to understand potential overlap. In the paragraph below describing results in this figure it is mentioned that "… 'good' practices shows a less marked downward trend, as does the incidence rate for all." If the incidence rate for all is the product of good and bad and good represents 85% of the patients, then it is no surprise the line follows a similar slope. Suggest removing this one. DISCUSSION
<ul> <li>DISCUSSION</li> <li>In the results section, the authors observe increases and decreases in the incidence rate. In the discussion last paragraph before 'Strengths and limitations', it is stated that the decrease in type 1 was found to be due to better coding. I believe the evidence for this decrease being the product of better coding instead of a reflection of true lower incidence it is not well explained and suggest adding some points to the discussion.</li> <li>Similarly, the regression model confirms this finding but there is no further information on which these 'poor quality' practices are. It would help to better understand the findings if there were some more context on what do we know about these 'poor' practices and if by removing them from the analyses we're still obtaining estimates that are representative of the UK population.</li> <li>MINOR COMMENTS:</li> <li>Different terminology is used throughout the manuscript to refer to the same thing: 'poor' quality of recording, inaccurate Read code use, miscoded, misclassified or misdiagnosed, it makes it</li> </ul>
confusing to the reader and could be streamlined. - If the authors plan for the codes in the Appendix to be used for future studies, it would help that 4a and 4b are differentiated.

REVIEWER	Zoe Morrison
	University of Aberdeen
	UK
REVIEW RETURNED	22-Jul-2016

GENERAL COMMENTS	There are a number of technical errors in this paper: The literature reviewed does not include the most recent relevant publications – see for example Robertson, ARR, Fernando B, Morrison Z, Kalra, D, and Sheikh, A. (2015) Structuring and coding in health care records: a qualitative analysis using diabetes as a case study. Inf Primary Care, 22(2):275-283. Available at http://hijournal.bcs.org/index.php/jhi/article/view/90.
	The analysis is incomplete and in places incorrect. The discussion does not follow from the findings and the throwaway recommendation at the end of the conclusion is unfounded and unexplained. The study is not replicable, particularly as not all information has been provided (there is no Appendix 1). On this basis I suggest rejection of this paper.

# **VERSION 1 – AUTHOR RESPONSE**

Reviewer Name: Dipak Kalra Institution and Country: The EuroRec institute, Belgium Competing Interests: None declared

# **Overall comments**

This is an interesting and well conducted study that is valuable in relation to the incidence of diabetes, but more generally highlighting the challenges with taking GP data at face value without an understanding of the historic evolution of reimbursement models and other changes to the clinical, managerial and ICT ecosystems. Reminding me of the "old" days of PRIMIS, it seems still necessary to fish for the truth through contextual codes about clinical activity and treatment to correct for the patchy recording of a diagnosis. This study is especially valuable given the wider interest amongst many stakeholders to reuse routinely collected clinical data for public health, epidemiology and tother kinds of research (kinds of big health data). They need to learn from this work.

I have made a few specific remarks below, but overall this paper reports its research methodically and details its result clearly. I do not have much to recommend by way of changes to the core of the paper.

\*\* Many thanks for these positive comments.

Given the expertise of the authors, I would have liked to see an additional section to the discussion to propose some implications from this research, since it is likely that the observed coding behaviour, trends, errors and variations would be true for multiple conditions, not just diabetes. (1) what are the ways in which diagnostic coding practice might be improved for the future, through changes to the systems, through education, or other initiatives and (2) what cautions or specific measures might users of CPRD and other equivalent data sources adopt in order to improve the accuracy of analytic interpretations made from such data sets?

\*\*We have added an extra section to the discussion entitled "Implications"

### Specific remarks

Introduction paragraph 3: Although published elsewhere, a two line summary of the key features of the data quality approach of citations 2 and 3 might be helpful to the reader. \*\*We have added a summary as requested.

Page 6 regarding misclassification. Just a question, not a critical view: is it not possible that a person is initially diagnosed Type 2, and control attempted with oral medication, but proves problematic to stabilise and so is moved to insulin? This raises the question of whether you looked for a subsequent Type 1 diabetes code even if a Type 2 code had been entered some months earlier. (We know that GP systems and GPs are not generally good at deleting codes from problem summaries.) \*\*We categorise a patient with a type 1 code at any point as having a diagnosis of type 1 so these patients would not have been included in this step of the misclassification analysis. \*\*I agree that the description is ambiguous have clarified what we mean by diagnosis of type 2 in a footnote.

The discussion section suggests that some spurious date peaks for incidence of diabetes might be due to retrospective data entry (of data from a former GP practice). Might it be a recommendation arising from this research that GP systems need to improve the ability to attribute an entry to a prior date, and that educational investments are needed to encourage practice staff to enter such retrospective dates when they are known? Does the import of data via GP2GP message automatically assign entries to their original date of recording?

\*\*We have expanded the discussion by adding a section "Implications" to address these comments.

Reviewer: 2

Reviewer Name: Mireia Raluy-Callado Institution and Country: Evidera, United Kingdom Competing Interests: None declared

This article focuses on the quality of recording in the UK Primary Care setting using data capture by CPRD through the Vision system on diabetes. This is a condition highly prevalent and mainly managed by Primary Care clinicians so it is of relevance to investigate patterns of coding to inform future studies. Therefore, this article is of interest and suitable for publication in the BMJ. There are two major areas where the paper would benefit from providing more clarification, one in the Introduction section and one in the Discussion. Comments are provided below by section of the manuscript:

\*\*Many thanks for the helpful comments.

INTRODUCTION

- The last paragraph in this section mentions the objectives of this study. However, the rational for using only Read codes in determining the diabetes diagnosis is not well explained. Given other publications in the past have highlighted the need of using additional medication and test codes to validate the diagnosis, it is unclear why the authors then start again from Read codes only.

\*\*I agree that researchers have used many workarounds and highlighted the need to use additional medication and test codes etc. But our aim is different: it is to examine how the use of Read codes to record a diagnosis of diabetes has changed since 1995 and 2014 and also the effect of "poor" coding will affect incidence estimates (with diabetes as a test case). Most official stats are based on Read codes only – such as those of Diabetes UK – as are many research studies, so we think this is an important piece of work. We have added this sentence to make things clearer "Most research studies and official statistics using GP records are based on the Read codes so in this paper we focus our investigation on the quality of coding."

- In the second part of the study, the sensitivity analyses is only performed in type 2 diabetes, and it would be good to understand why this type only was chosen.

\*\*Diabetes 2 has a much higher incidence, which is reported to have been increasing, whereas type 1 has remained fairly static. We have clarified this in the text

### METHODS

- Under 'Measuring incidence', it could be clarified that any Read code from the first one till the end of follow up was used, and not only the index code

\*\*We have expanded on the example to make this clearer.

# **RESULTS**:

- Under 'Practice Variation' 4a are all those in 4 that are not "seen in..."? Afterwards in Table 3 are called suggested, please clarify.

\*\*We have clarified this in the table

- Under 'Overall incidence in each category', type 1 patients are investigated to see if they also had a record for type 2 diabetes. Was medication not used to assess potential misdiagnosis?

\*\*No, in this part of the analysis we only used the codes as the aim is to investigate coding practice (and not misclassification).

- Figure 2: the title is unclear. In a) all codes these are all diabetes patients in all 4 study categories, and b) without 4b?

\*\*We have clarified this in the title

- Figure 4: the size of the sample for the 2 groups is very different. Suggest adding confidence intervals around the estimated incidence to understand potential overlap. In the paragraph below describing results in this figure it is mentioned that "… 'good' practices shows a less marked downward trend, as does the incidence rate for all." If the incidence rate for all is the product of good and bad and good represents 85% of the patients, then it is no surprise the line follows a similar slope. Suggest removing this one.

\*\*This is a good point. Adding confidence intervals would make the figure very messy, so we have removed the figure (and accompanying text) as you suggested.

#### DISCUSSION

- In the results section, the authors observe increases and decreases in the incidence rate. In the discussion last paragraph before 'Strengths and limitations', it is stated that the decrease in type 1 was found to be due to better coding. I believe the evidence for this decrease being the product of better coding instead of a reflection of true lower incidence it is not well explained and suggest adding some points to the discussion.

\*\*We have changed this sentence to make it clearer. "The decrease in incidence in category 1 (type 1) was found to be due to an inflated incidence in earlier years due to miscoding type 2 cases as type 1."

- Similarly, the regression model confirms this finding but there is no further information on which these 'poor quality' practices are. It would help to better understand the findings if there were some more context on what do we know about these 'poor' practices and if by removing them from the analyses we're still obtaining estimates that are representative of the UK population.

\*\*Unfortunately investigating the characteristics of the `poor' quality practices was beyond the scope of this study. What we do know, from our previous work on data quality, is that practices that are `poor' at coding one particular disease are almost always `good' at recording other diseases. MINOR COMMENTS:

- Different terminology is used throughout the manuscript to refer to the same thing: 'poor' quality of recording, inaccurate Read code use, miscoded, misclassified or misdiagnosed, ... it makes it confusing to the reader and could be streamlined.

\*\*We agree this is a bit confusing although we do define misclassification and miscoding in the introduction, and then expand the definitions in the methods section. We have now clarified that we are using the terms re: misclassification and miscoding only for the sections on the investigation of the effect of poor quality coding by explicitly mentioning these terms in the section headings.

- If the authors plan for the codes in the Appendix to be used for future studies, it would help that 4a and 4b are differentiated.

\*\* Agreed we have modified the supplementary file accordingly

Reviewer: 3 Reviewer Name: Zoe Morrison Institution and Country: University of Aberdeen, UK Competing Interests: None declared

there are a number of technical errors in this paper:

The literature reviewed does not include the most recent relevant publications – see for example Robertson, ARR, Fernando B, Morrison Z, Kalra, D, and Sheikh, A. (2015) Structuring and coding in health care records: a qualitative analysis using diabetes as a case study. Inf Primary Care, 22(2):275-283. Available at http://hijournal.bcs.org/index.php/jhi/article/view/90.

\*\* Apologies for not citing your recent paper. We have now done so and also added some more recent references.

The analysis is incomplete and in places incorrect.

We disagree, the analysis was very thorough and painstaking and we do not believe it was incorrect. The discussion does not follow from the findings and the throwaway recommendation at the end of the conclusion is unfounded and unexplained.

\*\* We suggest this because these codes do not confirm a diagnosis. The increase in recent years of these codes suggests that there is more monitoring of diabetes, not that these patients actually have it, so it is erroneous to include these codes as a diagnosis unless backed up by other information. We have had experience of researchers using these codes to suggest a diagnosis and this may lead to erroneous results. So we think it is important to make this point. We have modified the sentence to make it clearer what we mean

\*\*We suggest that if these codes are to be included as indication of diagnosis, the diagnosis should be confirmed with test results and prescribing information as was done by Sadek et al. for diagnosis codes

The study is not replicable, particularly as not all information has been provided (there is no Appendix 1).

\*\* We apologise; the codelist was provided as a supplementary file, but was wrongly labeled in the text as an Appendix, we have now corrected this.

# **VERSION 2 – REVIEW**

REVIEWER	Dipak Kalra
	The EuroRec Institute
	Belgium
REVIEW RETURNED	22-Sep-2016

GENERAL COMMENTS	I would like to thank the authors for updating the manuscript to reflect the recommendations that I made in my original review. I now look forward to seeing this in print!

REVIEWER	Mireia Raluy-Callado Evidera, Inc Real World Evidence
	London, UK
REVIEW RETURNED	17-Oct-2016

GENERAL COMMENTS	The authors have addressed the concerns raised in the review
	adequately and the paper is clearer now and addressed the objectives correctly.