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)	Presentation of respiratory symptoms prior to diagnosis in general practice: a case control study examining free text and morbidity codes
AUTHORS	Hayward, Richard; Chen, Ying; Croft, Peter; Jordan, Kelvin

VERSION 1 - REVIEW

REVIEWER	William Hamilton University of Exeter, UK
REVIEW RETURNED	12-Jan-2015

GENERAL COMMENTS	I am conversant with the subject. Overall, the paper is methodologically sound, and addresses a relatively important problem. However, it has considerable limitations (not the fault of the researchers, but real, nonetheless) and the authors do not give sufficient consideration to these limitations in their discussion. The paper addresses the bugbear of 'free text' – viz, uncoded material in electronic health records (EHRs) which may render studies which omit it unreliable. The methods were simple; three cardio-respiratory conditions were identified, and prior complaints of respiratory symptoms identified both in main coded form, and in free text. Discrepancies between the two dates of onset of symptoms were identified, with – not surprisingly – additional symptoms being found in free text preceding coded text in a significant proportion of cases. As I state above, the methods were fine and well performed. However, the study still leaves major questions to be answered. Here are my concerns… 1) The date of diagnosis is considered to be 'correct' – however, we know that analysis of free text can change the date of diagnosis.1 They cite this paper, but ignore the logical conclusion methodologically. One way the authors could considerably improve this paper would be to look for relevant therapies as proxies for a diagnosis. I'm sure they will find examples of β-agonist usage before a disease code, for instance. 2) They use Read codes (the EHR they study mandates use of these). However Read is very disease dominated, and symptom reporting in Read is very crude. It is not surprising that symptoms may be 'relegated' to free text given the constraints of Read. This is actually a massive problem, exemplified by comparing QCancer studies (which use Read) and CAPER studies (which use CPRD, with clinical data stored as medcodes, which are much more broad – there are over 100,000 of them). Comparing symptom prevalence in QCancer with CAPER shows that Read based studies only capture less than half of symptoms. See the two pancreas studies for

	example.2 3 3) The possibility that early symptoms may be unrelated to the later diagnosis is not given sufficient prominence. 4) This is a preliminary study, so this criticism is perhaps harsh. However, their methods are very labour-intensive, and so would be very costly to reproduce. It's very helpful to say 'omit free text and you will miss something important' but it's not realistic to expect all studies henceforth to use free text. This was a small study (with liberal access to free text) – larger studies simply couldn't do what's been done here. 5) The Price paper, which had similar conclusions on free text was larger and is already pulblished.4 This paper also raises the issue of differential recording styles, with text records more common in those who transpire to have less serious disease. This is important, as it weakens lines 21 onwards on p21; more isn't necessarily better! Minor points 6) I prefer numbers plus percentages to be quoted, especially in the abstract. 7) Tables 4 and 5 are very similar and I had to concentrate hard – too hard – to distinguish them. I think they could be merged, which would actually increase the likelihood of a reader capturing the difference! 8) Frequently the text simply repeats tabular results. I appreciate space is less of a concern with e-publishing, but redundancy is still redundancy. 9) Table 3 could also have the reference rows removed, then it would fit on one page. References 1. Tate A, Martin A, Murray-Thomas T, Anderson S, Cassell J. Determining the date of diagnosis - is it a simple matter? The impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK primary care. BMC medical research methodology 2009;9(1):42. 2. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. British Journal of General Practice 2012;62(594):38-45. 3. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, Hamilton W. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. Br J Cancer 2012;106(12):1940-44. 4. Price S, Shephard E, Stapley S, Barraclough K, Hamilton W. Non-visible vs visible haematuria and bladder cancer risk: a primary care electronic record study BJGP 2014; DOI: 10.3399/bjgp14X681409.

REVIEWER	Elizabeth Ford Brighton and Sussex Medical School, UK
REVIEW RETURNED	14-Jan-2015

GENERAL COMMENTS	It is very nearly ready for publication and it makes an important contribution to the field of electronic health research, namely that when researchers take into account the information in the free text, a different picture of clinical care or disease presentation emerges. The study has been done to a high quality and the write up is of publishable standard. I just have a few minor comments:

	1) The title given in the submission and that on the manuscript are not the same. Overall I prefer "Presentation of respiratory symptoms prior to diagnosis in general practice: a case control study examining free text and morbidity codes". Please ensure consistency. 2) Some thoughts on wording (which I acknowledge is not yet standardised): In the literature on research using medical records "notes" tends to mean narrative text, particularly in US studies using free text. On page 5 line 40 you say "GPs in the UK can record symptoms or diagnoses or both in the medical notes" - I would take this to mean the free text, although you go on to explain you mean both codes or free text. I prefer the term "records" to mean the complete record. 3) When searching the free text you appear to have done a keyword search (page 7 line 54). It would be good to know exactly what your search terms are, could you put them in a table? Also to know if you accounted for spelling mistakes and other errors? If your keyword list is very long it could be supplied as an appendix. 4) I would like to see the code lists you used either supplied as an appendix or placed in a repository such as clinicalcodes.org so they are easily available for scutiny and onward use. 5) It does not look like your logistic regression was multinomial (page 8 line 20) It looks like a series of hierarchical binomial logistic regressions to me. Please clarify. 6) In Table 1 (page 9) it appears that patients could be included if they had more than one diagnosis. This is not clear in the methods. Please could you clarify if patients were included or excluded if they had more than one of the diagnoses of interest. How many patients did this affect? Did you do a sub-analysis for these patients? 7) In your subgroup analysis on page 15, it is not clear what the purpose of this analysis is or what it shows. Why would we be more interested in this group of patients? Please clarify the purpose. 8) The discussion is good and covers some interesting points.

REVIEWER	Anoop Shah University College London, UK
REVIEW RETURNED	25-Jan-2015

GENERAL COMMENTS	This is an interesting evaluation of GP diagnosis and previously recorded symptoms in Read codes or free text. I think this article will be suitable for publication after clarifying some of the methodological details. The value of this research is in showing out that GP-recorded symptoms in free text and Read codes are a useful resource for finding out ways to help differentiate between people who go on to be diagnosed with a serious illness and those who do not. This could lead to better decision algorithms to help GPs to target investigations more effectively. Study population: Please supply the list of Read codes for identifying patients as supplementary material to aid replication of this study in other general practice databases. What practice management software do the practices use? I note that all cases were in one of the three diagnostic categories.

	How did you handle patients with more than one of these diagnoses? Statistical analysis: Please explain what was the exposure and outcome in the multinomial logistic regression. Presumably this model was used with so that four outcomes (control, asthma, COPD, IHD) could be modelled simultaneously. Presentation of results: Table 3 - odds ratio results - it would be clearer to show a single odds ratio labelled '... with vs. without ...' rather than two rows, one of which contains just a row of ones. State in the table that it is an age and gender adjusted odds ratio, otherwise readers might expect it to be crude odds ratio. Table 5 - type and number of patients column is duplicated; it could be shown just once. It would be useful also to include the results from phase 2 in the abstract, if possible. Discussion: Another limitation is that the general practices were likely not representative of practices in the UK in terms of the quality of their medical record keeping, as they receive special training and are encouraged to keep high quality records.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

1)The date of diagnosis is considered to be 'correct' – however, we know that analysis of free text can change the date of diagnosis.1 They cite this paper, but ignore the logical conclusion methodologically. One way the authors could considerably improve this paper would be to look for relevant therapies as proxies for a diagnosis. I'm sure they will find examples of β-agonist usage before a disease code, for instance.

Authors' response
The reviewer makes a good point that analysis of free text could potentially change the date of diagnosis (although this will have all the issues regarding the complexity of interrogation of free text that we bring out in our discussion). This does strengthen one of our messages that free text searching may aid investigation of early presentation of long-term conditions. Assessing therapies at time of symptom recording would be an important next step in this research, to see if the GP was managing the patient in line with the future diagnosis despite not recording a diagnostic code (and we acknowledge this may be the case on p.19). This does not change a central message and implication from this study for GP database research that symptoms are often recorded before the coding of a formal diagnosis and this may be in the form or coded information or free text.

We have added to the limitations section (p.19):
It is possible that the GP diagnosed the condition prior to the first coded mention in the records and that interrogation of free text may indicate an earlier date of diagnosis.[18,19] This does though further emphasise the importance of considering the use of free text in investigation of early

presentation of long-term conditions using GP databases. Alternative definitions of morbidity than recorded diagnosis codes using relevant prescriptions and managements are possible and further research could explore whether GPs were managing patients optimally despite a lack of formal recorded diagnosis.

2)They use Read codes (the EHR they study mandates use of these). However Read is very disease dominated, and symptom reporting in Read is very crude. It is not surprising that symptoms may be 'relegated' to free text given the constraints of Read. This is actually a massive problem, exemplified by comparing QCancer studies (which use Read) and CAPER studies (which use CPRD, with clinical data stored as medcodes, which are much more broad – there are over 100,000 of them). Comparing symptom prevalence in QCancer with CAPER shows that Read based studies only capture less than half of symptoms. See the two pancreas studies for example.2 3

Authors' response
We slightly disagree with the reviewer here. It is our understanding that CPRD converts diagnosis and symptom codes into their own coding system (Medcodes) but that the most common underlying coding system of information taken from primary care to be converted to Medcodes in CPRD is the Read Code. There is a broad section of symptom coding available to the recorder within the Read Code hierarchy, but we appreciate there is wide variation in the use of symptom versus diagnostic coding between clinicians, practices and databases. Part of this may be due to the guidance (or lack of guidance) given to practitioners when recording morbidity as we have shown previously.[27]

We have added to the limitations section: (p.19)
All practices providing data to the CiPCA database use the Read code system, a widely used system by GPs in the UK. There may be variation in extent of symptom coding between practices and also between databases of health care information depending on coding system used (for example Read Codes or ICD codes) and guidance given on recording.[27]

3) The possibility that early symptoms may be unrelated to the later diagnosis is not given sufficient prominence.

Authors' response
In response to this comment and to comment 5 the authors recognise that only a prospective study can answer these questions. The last sentence at the end of the first paragraph on p21 under the section Implications for practice and research now reads:

"Other information from prospective studies will be needed to discriminate between those with symptoms who will and will not develop long-term conditions."

4) This is a preliminary study, so this criticism is perhaps harsh. However, their methods are very labour-intensive, and so would be very costly to reproduce. It's very helpful to say 'omit free text and you will miss something important' but it's not realistic to expect all studies henceforth to use free text. This was a small study (with liberal access to free text) – larger studies simply couldn't do what's been done here

Authors' response
The reviewer is of course correct and we have made an addition to our paragraph on this (p.21):
"Diagnostic and symptom codes in GP electronic health records form a ready source of data for research but much useful information exists in the free text which is harder to extract. Studies of

earliest presenting symptoms in national databases are challenging because manual free text searching on a large scale is difficult. However this would improve if GPs were to code symptoms more readily when unable or not wishing to make a diagnosis and with expanding investment in research resources, technological barriers to textual analysis are likely to be solved."

5) The Price paper, which had similar conclusions on free text was larger and is already pulblished.4 This paper also raises the issue of differential recording styles, with text records more common in those who transpire to have less serious disease. This is important, as it weakens lines 21 onwards on p21; more isn't necessarily better!

Authors' response:
This is a good point and we have added to the discussion around use of identifying symptoms in text and codes. (p.21):
"Our study suggests that a combination of symptom codes and text records would provide the optimal sample for such studies by ensuring the highest proportion possible of all those who present with such symptoms will be included. However, the weaker association of symptoms with a future diagnosis when including text recorded symptoms suggests it is possible that those with purely textual information have less serious symptoms.[32] Other information from prospective studies will be needed to discriminate between those with recorded symptoms who will and will not develop long-term conditions."

Minor points
6) I prefer numbers plus percentages to be quoted, especially in the abstract.

Authors' response:
We have now included both.

7) Tables 4 and 5 are very similar and I had to concentrate hard – too hard – to distinguish them. I think they could be merged, which would actually increase the likelihood of a reader capturing the difference!

Authors' response:
We have now combined the tables.

8) Frequently the text simply repeats tabular results. I appreciate space is less of a concern with e-publishing, but redundancy is still redundancy.

Author's response:
We would prefer to keep the information in the text as it highlights the main findings.

9) Table 3 could also have the reference rows removed, then it would fit on one page.

Authors' response:

Thank you for the suggestion. We have now done this.

Reviewer 2

1)The title given in the submission and that on the manuscript are not the same. Overall I prefer "Presentation of respiratory symptoms prior to diagnosis in general practice: a case control study examining free text and morbidity codes". Please ensure consistency.

Authors' response:
We apologise for the confusion here. We have used the reviewer's suggested title.

2) Some thoughts on wording (which I acknowledge is not yet standardised): In the literature on research using medical records "notes" tends to mean narrative text, particularly in US studies using free text. On page 5 line 40 you say "GPs in the UK can record symptoms or diagnoses or both in the medical notes" - I would take this to mean the free text, although you go on to explain you mean both codes or free text. I prefer the term "records" to mean the complete record.

Authors' response:
We have amended the above on p5 to read "medical records"

3) When searching the free text you appear to have done a keyword search (page 7 line 54). It would be good to know exactly what your search terms are, could you put them in a table? Also to know if you accounted for spelling mistakes and other errors? If your keyword list is very long it could be supplied as an appendix.

Authors' response:
We have included the search strategy in an appendix and added to the text (p8):
The search strategy is given in the appendix.

4) I would like to see the code lists you used either supplied as an appendix or placed in a repository such as clinicalcodes.org so they are easily available for scutiny and onward use.

Authors' response:
The code lists will be available from our website www.keele.ac.uk/mrr. This is now added to the paper (p6):
The code lists used for this study are available from www.keele.ac.uk/mr.

5) It does not look like your logistic regression was multinomial (page 8 line 20) It looks like a series of hierarchical binomial logistic regressions to me. Please clarify.

Authors' response:
The analysis was multinomial logistic regression but the results incorrectly described as ORs rather than RRRs. We apologise for this and have amended the text and tables.

6) In Table 1 (page 9) it appears that patients could be included if they had more than one diagnosis. This is not clear in the methods. Please could you clarify if patients were included or excluded if they had more than one of the diagnoses of interest. How many patients did this affect? Did you do a sub-analysis for these patients?

Authors' response:
Patients were indeed included if they had more than 1 diagnosis. However only three patients were in multiple case groups. These were put in most "severe" category (in order of IHD, COPD, asthma). This is indicated in the footnote to table 1.

7) In your subgroup analysis on page 15, it is not clear what the purpose of this analysis is or what it shows. Why would we be more interested in this group of patients? Please clarify the purpose.

Authors' response:

The purpose was to allow a direct comparison in time interval from 1st recorded symptom to diagnosis between i) using coded symptoms only and ii) using code or text recorded symptoms. The best comparison of this is in those who had both a coded and text recorded symptom prior to diagnosis.

Reviewer 3

1)Please supply the list of Read codes for identifying patients as supplementary material to aid replication of this study in other general practice databases. What practice management software do the practices use?

Authors' response
As in our response to Reviewer 2, comment no. 4, the code lists will be made available from our website www.keele.ac.uk/mrr. This is now added to the paper. The practices used EMIS software.

2) I note that all cases were in one of the three diagnostic categories. How did you handle patients with more than one of these diagnoses?

Authors' response:
As in our response to Reviewer 2, comment no. 6: Patients were indeed included if they had more than 1 diagnosis. However only three patients were in multiple case groups. These were put in most "severe" category (in order of IHD, COPD, asthma). This is indicated in the footnote to table 1.

3) Please explain what was the exposure and outcome in the multinomial logistic regression. Presumably this model was used with so that four outcomes (control, asthma, COPD, IHD) could be modelled simultaneously.

Authors' response:
The referee is correct. The outcome was a diagnosis of asthma, COPD or IHD with control as the reference group. This has been clarified on p.8:

Association between a prior coded symptom of breathlessness or wheeze (exposure status from phase 1) and the outcome of later diagnosis (IHD, COPD or asthma) was assessed using multinomial multivariable logistic regression, adjusting for age and gender and reported as relative risk ratio (RRR) with 95% confidence intervals (95% CI) with the controls as the reference group.

4) Table 3 - odds ratio results - it would be clearer to show a single odds ratio labelled '... with vs. without ...' rather than two rows, one of which contains just a row of ones. State in the table that it is an age and gender adjusted odds ratio, otherwise readers might expect it to be crude odds ratio.

Authors' response:
Thank you for the suggestion, we have now amended table 3 as suggested and in line with Reviewer 1's comment no.9.

5) Table 5 - type and number of patients column is duplicated; it could be shown just once.

Authors' response:
We have now combined tables 4 and 5 in line with Reviewer 1's comment no.7. Due to this we would prefer to keep the apparent duplication for overall clarity of the table.

6) It would be useful also to include the results from phase 2 in the abstract, if possible.

Authors' response:
Results from Phase 2 are recorded in the Results section of the abstract.

7) Another limitation is that the general practices were likely not representative of practices in the UK in terms of the quality of their medical record keeping, as they receive special training and are encouraged to keep high quality records

Authors' response:
We have added to the limitations paragraph around the use of the CiPCA database (pp. 19 –20) see also response to Reviewer 1's Comment no. 2):

There may be variation in extent of symptom coding between practices and also between databases of health care information depending on coding system used (for example Read Codes or ICD codes) and guidance given on recording. [27] One possible limitation was the localised nature of the CiPCA database It has been used for example in studies of gout, dementia and frequent consulters, and provided comparable musculoskeletal prevalence figures to national UK and international databases.[28-31] North Staffordshire is a deprived area, but the participating practices were socially and economically diverse. GPs in the practices undergo some training in morbidity recording, but whilst encouraged to use diagnostic codes, symptom codes may also be used

.

VERSION 2 – REVIEW

REVIEWER	William Hamilton University of Exeter, UK
REVIEW RETURNED	02-Mar-2015

GENERAL COMMENTS	I'm happy with all the responses to my review. We differ on my point 2 - I've no problem with that as science would be bring if we all agreed. It's their paper not mine!

REVIEWER	Anoop Shah University College London, United Kingdom
REVIEW RETURNED	01-Mar-2015

GENERAL COMMENTS	The authors have attended to the reviewers points and the article is much improved. There are a few minor points which should be corrected before publication. Please define the acronym RRR in the abstract. I was unable to log in to the Keele / mrr website -- the site asks for a username and password in order to see the list of Read codes. I would like to see the list of codes used. There is a typographical error in the last sentence of the Discussion -- it should be two sentences: "However this would improve if GPs were to code symptoms more readily when unable or not wishing to make a diagnosis. With expanding investment in research resources, technological barriers to textual analysis are likely to be solved."

VERSION 2 – AUTHOR RESPONSE

The authors have defined RRR in the abstract and have corrected the typographical error in the discussion. Unfortunately the Keele.mrr website -our code repository does not go live until the end of this month. It is at the moment a matter of policy of our institution that all our codes should be sent to this site. I can only apologise for the problem and ask forbearance by the reviewers. I hope this is satisfactory.
I shall now upload v4