Using free text information to explore misclassification in dating diagnoses of ovarian cancer. Observational study using the General Practice Research database.

BMJ Open-2010-000025

Reviewer 1: Hamilton, William

Peninsula College of Medicine and Dentistry, Primary Care

The Study	**Yes**	**No**
Is the research question clearly defined?	✓	
Is the overall study design appropriate and adequate to answer the research question?		✓
Are the participants adequately described, their conditions defined, and the inclusion and exclusion criteria described?	✓	
Are the patients representative of actual patients the evidence might affect?	✓	
Are the methods adequately described?	✓	
Is the main outcome measure clear?	✓	
Are the abstract/summary/key messages/limitations accurate?	✓	
Are the statistical methods described?	✓	
Are they appropriate?	✓	
Is the standard of written English acceptable for publication?	✓	
Are the references up to date and relevant? (If not, please provide details of significant omissions below.)	✓	
Do any supplemental documents e.g. a CONSORT checklist, contain information that should be better reported in the manuscript, or raise questions about the work?	✓	

If you answered No to any of the above, please supply details below.
See my 'traditional report' it's a single example chosen to quantify a potential problem with database studies. It's fine as it goes, just one cannot know how generalisable it is.

RESULTS AND CONCLUSION (For articles reporting research findings only)	**Yes**	**No**
Do the results answer the research question?		✓
Are they credible?	✓	
Are they well presented?	✓	
Are the interpretation and conclusions warranted by and sufficiently derived from/focused on the data?	✓	
Are they discussed in the light of previous evidence?	✓	
Is the message clear?	✓	

If you answered No to any of the above, please supply details below.
Again I have the issue that a single metric may not tell the whole story.

REPORTING AND ETHICS	Yes	No
Is the article reported in line with the appropriate reporting statement or checklist (e.g. CONSORT)?	✓	
Are research ethics (e.g. consent, ethical approval) addressed appropriately?	✓	
Is the article free from any concerns about publication ethics (e.g. plagiarism, fabrication, redundant publication, undeclared conflicts of interest)?	✓	

If you answered No to any of the above, please supply details below or contact the editorial office.

In compliance with the BMJ Open system of open peer review, please sign your review in the box below. Include your name, position, institution and country. Please also include a statement of competing interests. If you have filled out an ICMJE Conflicts of Interests form, please attach this using the box beneath instead.

Prof Willie Hamilton, PCMD, UK

req **Recommendation**

Accept

Minor Revision

✓ Major Revision

Reject

Would you be willing to review a revision of this manuscript?

✓ Yes

No

Comments

If you have any further comments for the authors please enter them below.

Dear Editor,

Using free text information in the GPRD....
by Tate, Martin et al

Thanks for asking me to review this paper. I am very conversant with the clinical problem, and with the GPRD, having published a few papers based on it, and having recently analysed a dataset from the GPRD on the features of brain tumours. I have also published in the BMJ research into early diagnosis of ovarian cancer. I have no conflicts of interest in doing the review.

Overall

This is a methodological paper addressing the important question of reliability of data in the GPRD. Many studies now use the GPRD, and any 'warts' on the GPRD may affect their value. It needs to be remembered that originally the GPRD was established to be a pharmaco-epidemiology database, which would provide high quality data of particular interest to pharmaceutical companies. It is only in the last decade or so that non-pharmacological studies have utilised the GPRD, and there is a considerable under-appreciation of its quirks and its problems.

The researchers use a specific example of dating ovarian cancer diagnoses to identify missing data. Arguably, that's a tricky example to have chosen as international consensus is yet to be achieved on precisely which milestone on the cancer diagnostic journey counts as 'diagnosis'. Even so, it's adequate for the purpose here: to see if free-text data changes the date of diagnosis (defined as the first substantive code for ovarian cancer in the main code file).

And the free-text did change the date of diagnosis – a bit. The importance of the paper has to be decided upon the following questions

1) Was dating of ovarian cancer diagnosis a good choice?
2) Can the findings from ovarian cancer mis-dates be generalised to other data fields?
3) Did the study design omit other possible areas of mis-dating (in particular, are there ovarian cancers without a substantive code at all – the literature suggests there are)
4) Is the percentage of mis-dating enough to matter?

Q1. It may have been a good choice, but it's quite likely other benchmarks may have been different. I have identified colon polyps and colonoscopies in the free-text in the THIN database which were uncoded in the main text, so not identifiable in the conventional way.
Q2. It's hard to know if other items will have been 'lost' at a higher or lower rate in the free text. Arguably, ovarian cancer diagnosis may be better recorded than more primary-care based items, as secondary care data is generally entered onto the system by an administrator (who follows the rules much better than us GPs). A primary care 'disease' say IBS may have much more 'hidden' data in the free-text.
Q3. I have almost answered myself. The free-text is only one area where data may be lost. GP letters (as the authors accept) are a rich source.
Q4. I can't be prescriptive here. For some studies, having 10% of the 'diagnosis dates' out by over a month would be very relevant; for others a 10% miss rate is acceptable. It has to be remembered that there are >30,000 GPs each with their own recording idiosyncrasies, so the GPRD (and GP casenotes studies as a whole) have considerable blurriness inherently. Even so, this paper does at least quantify to some extent the amount of blurriness!

Reviewer 2: Neal, Richard

North Wales Clinical School, Primary Care

The Study	Yes	No
Is the research question clearly defined?	✓	

If you answered No to any of the above, please supply details below.

Abstract
The use of ‘etc’ generally assumes that the reader implies the same as the author, or that the author cannot state what they mean accurately. I suspect neither is true here, but it is to be avoided.
The first sentence of ‘results’ does not scan

Summary
Bullet point 1 - This is important, but only if the findings suggest that this is important
Bullet point 2 - This is important as long date of diagnosis is an understood and defined concept
Bullet point 3 - This is important, but only if the findings suggest that this is important

Key messages
Bullet point 1 - Yes, but what how and why?
Bullet point 2 - As long as the date of diagnosis has meaning

Strengths and limitations
This needs some reference to what is actually meant by ‘date of diagnosis’, and what is a gold satndard to measure against.

should be better reported in the manuscript, or raise questions about the work?

RESULTS AND CONCLUSION (For articles reporting research findings only)	Yes	No
Do the results answer the research question?	✓	
Are they credible?	✓	
Are they well presented?	✓	
Are the interpretation and conclusions warranted by and sufficiently derived from/focused on the data?		✓
Are they discussed in the light of previous evidence?		✓
Is the message clear?		✓

If you answered No to any of the above, please supply details below.

Results
Table 1 is excellent (and gets to the crux of defining date of diagnosis), but does not relate to the rest of the paper, in that date of diagnosis is taken as a straightforward concept. There is an issue about how a 'definitive' ovarian cancer diagnosis is made and how may this may differ from other cancers.
Classification scheme for free text. General, these definitions are helpful, but as mentioned above need to feature more strongly in the intro/methods/discussion. Point 4 - 'metastatic cancer' is hardly ambiguous.
Figure 1 is helpful, with the area below the dotted line being the area of interest. The main finding is that 33/340 patients had a 'definite' text entry for diagnosis greater than one month prior to the coded date of diagnosis (still not entirely sure what this was) with 9/340 greater than 6 months before.

Discussion
The discussion is wordy and could be severely edited to improve impact and readability. It would also be improved by:
Greater discussion about the cause of the differences in dates the implications of these differences. This needs to be placed within the wider concept of what is meant by and what is recorded as a 'diagnosis'. Reference to the cancer registry definitions of date of diagnosis help here. The authors touch on the process of adding coded data but this could be improved by expanding this section. There is much variability in the process both between and within practices. It is also a process that is open to error both in terms of the coding (which they touch upon) and incorrect entry of dates (which they do not). How does one decide when there has been an error? Or which one is more truthful / valid?
Greater discussion about the methodological issues, including the additional time and resource needed to analyse free text comments, compared with just coded data.
Some comment about changes over time. With increasing levels and standards of computerization, will these findings have implications for the future, or in 5-10 years will the use of free text / coded data have changed?
Some comment perhaps about the implications of this for other diseases / conditions?
Lastly, should the authors be making a plea for less / more free text entries in order to improve the quality of routinely collected general practice data for epidemiological research?

REPORTING AND ETHICS	Yes	No
Is the article reported in line with the appropriate reporting statement or checklist (e.g. CONSORT)?		✓
Are research ethics (e.g. consent, ethical approval) addressed appropriately?	✓	
Is the article free from any concerns about publication ethics (e.g. plagiarism, fabrication, redundant publication, undeclared conflicts of interest)?	✓	

If you answered No to any of the above, please supply details below or contact the editorial office.

No appropriate statement to use

In compliance with the BMJ Open system of open peer review, please sign your review in the box below. Include your name, position, institution and country. Please also include a statement of competing interests. If you have filled out an ICMJE Conflicts of Interests form, please attach this using the box beneath instead.

Richard Neal

req Recommendation

Accept

Minor Revision

[✓ Major Revision]

Reject

Would you be willing to review a revision of this manuscript?

[✓ Yes]

No

Comments

If you have any further comments for the authors please enter them below.

This is a well-conducted, but not terribly well-reported study that has implications for the methodological community.
The main issues (that I think can be dealt with easily) is that what 'should', and 'in an ideal world' be the date of diagnosis, and what actually is a 'gold standard' date of diagnosis.

Introduction
'e.g. in the hospital letters..' - these are not free text in GP records, these are totally different
Need something about the concept of date of diagnosis

Methods
3rd para 'the GP is required' perhaps should read 'the GP may'?
Last para – mention of the term 'definitive diagnosis' – what do they mean by this? Clinical? Tissue? Date communicated to patient? Date communicated to GP? Date in cancer registry? Date of MDT? Date of first record in general practice record.......?
Date 2 – how to distinguish 'definite' from 'ambiguous'?
Date 3 – maybe this equates to 'diagnosis'
Date 4 – I have no idea how this can be determined as it is such an abstract concept
The finding that dates 1 & 4 were different in 73% cases is no surprise to UK GPs.... (and why are results presented in the methods?)
Extraction of information. The word 'texts' is used throughout the paper. This is confusing and should be changed to textual, non-coded data or similar.

Authors Response to Decision Letter for (BMJ Open-2010-000025)

Using free text information to explore misclassification in dating diagnoses of ovarian cancer. Observational study using the General Practice Research database.

Dear Mr Sands,
Thank-you very much for provisionally accepting our paper for publication in BMJ Open. I very much appreciate the helpful comments of the two reviewers. I have amended the paper as suggested (changes are in bold in the attached revision).
Detailed answers to the reviewers' comments are provided (starred) below
Yours sincerely
Rosemary Tate
02-Dec-2010

Dear Dr. Tate:

Manuscript ID BMJ Open-2010-000025 entitled "Using free text information to explore misclassification in dating diagnoses of ovarian cancer. Observational study using the General Practice Research database." which you submitted to BMJ Open, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

In the light of the reviewers' comments you may also wish to consult the reporting statements available at www.equator-network.org for further ideas.

The reviewer(s) have recommended publication, but also suggest some substantive revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript. IMPORTANT: please remember that the reviewers' comments and the previous drafts of your manuscript will be published as supplementary information alongside the final version.

To revise your manuscript, log into http://mc.manuscriptcentral.com/bmjopen and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

You may also click the below link to start the revision process (or continue the process if you have already started your revision) for your manuscript. If you use the below link you will not be required to login to ScholarOne Manuscripts.

http://mc.manuscriptcentral.com/bmjopen?URL_MASK=6trPPjcY5HJ9hP5HwYMm

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text. Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s).

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

Because we are trying to facilitate timely publication of manuscripts submitted to BMJ Open, your revised manuscript should be submitted by 01-Jan-2011. If it is not possible for you to submit your revision by this date, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to BMJ Open and I look forward to receiving your revision.

Sincerely,
Mr. Richard Sands
Managing Editor, BMJ Open
rsands@bmjgroup.com

Reviewer(s)' Comments to Author:
Reviewer: Prof Willie Hamilton, PCMD, UK
See my 'traditional report' [below] it's a single example chosen to quantify a potential problem with database studies. It's fine as it goes, just one cannot know how generalisable it is. I have the issue that a single metric may not tell the whole story.
Dear Editor,

Using free text information in the GPRD....
by Tate, Martin et al

Thanks for asking me to review this paper. I am very conversant with the clinical problem, and with the GPRD, having published a few papers based on it, and having recently analysed a dataset from the GPRD on the features of brain tumours. I have also published in the BMJ research into early diagnosis of ovarian cancer. I have no conflicts of interest in doing the review.

Overall

This is a methodological paper addressing the important question of reliability of data in the GPRD. Many studies now use the GPRD, and any ‘warts’ on the GPRD may affect their value. It needs to be remembered that originally the GPRD was established to be a pharmaco-epidemiology database, which would provide high quality data of particular interest to pharmaceutical companies. It is only in the last decade or so that non-pharmacological studies have utilised the GPRD, and there is a considerable under-appreciation of its quirks and its problems.

The researchers use a specific example of dating ovarian cancer diagnoses to identify missing data. Arguably, that's a tricky example to have chosen as international consensus is yet to be achieved on precisely which milestone on the cancer diagnostic journey counts as 'diagnosis'. Even so, it's adequate for the purpose here: to see if free-text data changes the date of diagnosis (defined as the first substantive code for ovarian cancer in the main code file).

And the free-text did change the date of diagnosis – a bit. The importance of the paper has to be decided upon the following questions

1) Was dating of ovarian cancer diagnosis a good choice?
2) Can the findings from ovarian cancer mis-dates be generalised to other data fields?
3) Did the study design omit other possible areas of mis-dating (in particular, are there ovarian cancers without a substantive code at all – the literature suggests there are)
4) Is the percentage of mis-dating enough to matter?

Q1. It may have been a good choice, but it's quite likely other benchmarks may have been different. I have identified colon polyps and colonoscopies in the free-text in the THIN database which were uncoded in the main text, so not identifiable in the conventional way.
Q2. It's hard to know if other items will have been 'lost' at a higher or lower rate in the free text. Arguably, ovarian cancer diagnosis may be better recorded than more primary-care based items, as secondary care data is generally entered onto the system by an administrator (who follows the rules much better than us GPs). A primary care 'disease' say IBS may have much more 'hidden' data in the free-text.
Q3. I have almost answered myself. The free-text is only one area where data may be lost. GP letters (as the authors accept) are a rich source.
Q4. I can't be prescriptive here. For some studies, having 10% of the 'diagnosis dates' out by over a month would be very relevant; for others a 10% miss rate is acceptable. It has to be remembered that there are >30,000 GPs each with their own recording idiosyncrasies, so the GPRD (and GP casenotes studies as a whole) have considerable blurriness inherently. Even so, this paper does at least quantify to some extent the amount of blurriness!

** Many thanks to Dr Hamilton for his positive review.
I have added an extra point to the strengths and limitations section to address the comment about generalisability.
"We have only looked at ovarian cancer, and cannot say whether our findings will generalise to other diseases."

Some of the answers to Dr Neal (below) also address Dr Hamilton's comments. **

Reviewer: Richard Neal
North Wales Clinical School, Primary Care

** Many thanks to Dr Neal for the helpful and constructive comments. Many of the points relate to our definition of the date of diagnosis. I apologise if this was ambiguous. The main aim of this study was not to determine the "true" date of diagnosis, but rather to determine whether and how often the coded date of a diagnosis differed from the date that a definite diagnosis was first recorded in the text field (i.e. when the GP first knew of the diagnosis). So there were two definitions of date of diagnosis: 1. The date when there was a record of a definitive diagnosis in the letter or free text and 2. The coded date of diagnosis.
In order to clarify this we have changed any sentences where this term is ambiguous. I hope this is now a lot clearer.
I have also changed the title to make the purpose of the work clearer.
"Using free text information to explore how and when GPs code a diagnosis of ovarian cancer. Observational study using the General Practice Research database."**

Abstract
The use of 'etc' generally assumes that the reader implies the same as the author, or that the author cannot state what they mean accurately. I suspect neither is true here, but it is to be avoided.
I have changed all 3 sentences in the text that use etc.
The first sentence of 'results' does not scan

I have amended this sentence.

Summary
Bullet point 1 - This is important, but only if the findings suggest that this is important
Bullet point 2 - This is important as long date of diagnosis is an understood and defined concept
Bullet point 3 - This is important, but only if the findings suggest that this is important

Key messages
Bullet point 1 - Yes, but what how and why?
I have added this phrase to clarify this point "including grade and stage of the cancer and the date that the patient was seen in and diagnosed in secondary care"

Bullet point 2 - As long as the date of diagnosis has meaning
I have clarified this

Strengths and limitations
This needs some reference to what is actually meant by 'date of diagnosis', and what is a gold standard to measure against.
**I hope that our clarifications above will address this. I have also added (which will also address Dr Hamilton's comment)
"We only looked at cases which had been assigned an unambiguous Read code for ovarian cancer and thus will have missed cases with no code or an ambiguous code. " **

Results
Table 1 is excellent (and gets to the crux of defining date of diagnosis), but does not relate to the rest of the paper, in that date of diagnosis is taken as a straightforward concept. There is an issue about how a 'definitive' ovarian cancer diagnosis is made and how may this may differ from other cancers.
Classification scheme for free text. General, these definitions are helpful, but as mentioned above need to feature more strongly in the intro/methods/discussion.
**I have clarified the two concepts for the date of diagnosis.
And have added this sentence to the abstract.
"We investigate how much information on ovarian cancer diagnosis is ``hidden" in the free text and the time lag between a diagnosis being described in the text or in a hospital letter and the patient being given a READ code for that diagnosis."**

Point 4 - 'metastatic cancer' is hardly ambiguous.
**We meant by this that it is ambiguous in the sense that we don't know if it cancer of the ovary or not. It could be a metastatic ovarian cancer or it could be cancer in the ovary with primary in the breast. I have rewritten this sentence to make it clearer
"an ambiguous diagnosis, e.g. ``tumour", which could be benign or could be a primary or secondary cancer, or ``metastatic cancer" which could be a primary or secondary ovarian cancer or another type of cancer."**

Figure 1 is helpful, with the area below the dotted line being the area of interest. The main finding is that 33/340 patients had a 'definite' text entry for diagnosis greater than one month prior to the coded date of diagnosis (still not entirely sure what this was) with 9/340 greater than 6 months before.

Hopefully the extra sentences added about the diagnosis date (above) will clarify

Discussion
The discussion is wordy and could be severely edited to improve impact and readability. It would also be improved by:
Greater discussion about the cause of the differences in dates the implications of these differences. This needs to be placed within the wider concept of what is meant by and what is recorded as a 'diagnosis'. Reference to the cancer registry definitions of date of diagnosis help here. The authors touch on the process of adding coded data but this could be improved by expanding this section. There is much variability in the process both between and within practices. It is also a process that is open to error both in terms of the coding (which they touch upon) and incorrect entry of dates (which they do not). How does one decide when there has been an error? Or which one is more truthful / valid?
Greater discussion about the methodological issues, including the additional time and resource needed to analyse free text comments, compared with just coded data.
Some comment about changes over time. With increasing levels and standards of computerization, will

these findings have implications for the future, or in 5-10 years will the use of free text / coded data have changed?
Some comment perhaps about the implications of this for other diseases / conditions?
Lastly, should the authors be making a plea for less / more free text entries in order to improve the quality of routinely collected general practice data for epidemiological research?

This is a well-conducted, but not terribly well-reported study that has implications for the methodological community.
The main issues (that I think can be dealt with easily) is that what 'should', and 'in an ideal world' be the date of diagnosis, and what actually is a 'gold standard' date of diagnosis.

I have rewritten this section to address most of these points and have made it more concise. I have clarified the text throughout the paper to make it clear that the main issue is using free text to determine how and when GPs and practice staff code a notified diagnosis, rather than addressing the more thorny issue of when the diagnosis is actually made.

Introduction
'e.g. in the hospital letters..' - these are not free text in GP records, these are totally different
**I take Dr Neal's point, but many letters are provided in the form of free text records to the researcher, where no distinction is made between doctors' notes and letters. I have amended the sentence to clarify.
"Free text records, as distinct from coded records, may contain further information on diagnosis (e.g. which have been copied or imported from hospital letters)"**

Need something about the concept of date of diagnosis
I have amended the term "coded date" to "the event date for which the diagnosis is coded"

Methods
3rd para 'the GP is required' perhaps should read 'the GP may'?
done
Last para – mention of the term 'definitive diagnosis' – what do they mean by this? Clinical? Tissue? Date communicated to patient? Date communicated to GP? Date in cancer registry? Date of MDT? Date of first record in general practice record.......?

**I couldn't find this term anywhere in the manuscript and there is no mention of definite diagnosis in the last para. However, to clarify I have amended "date of diagnosis" to "coded date of diagnosis".
I have also amended the definition of Date 1. To "Date of first definite diagnosis - referred to in this paper as date of coded diagnosis. "**

Date 2 – how to distinguish 'definite' from 'ambiguous'?
Date 3 – maybe this equates to 'diagnosis'
Date 4 – I have no idea how this can be determined as it is such an abstract concept
The finding that dates 1 & 4 were different in 73% cases is no surprise to UK GPs.... (and why are results presented in the methods?)
All of the above refer to our previous paper – where the reader can find full details. The reason that I put this last sentence in (which is a result of the previous study) is because we only use dates1 and dates 4 in this work (to select the time period) and I thought it was important that the reader knows that date1 and date 4 were the same in 27% of cases). I included the definition of dates 2 and 3 because they are needed to explain date 4.

Extraction of information. The word 'texts' is used throughout the paper. This is confusing and should be changed to textual, non-coded data or similar.

**Done **

Using free text information to explore how and when GPs code a diagnosis of ovarian cancer.
Observational study using the General Practice Research database.

BMJ Open-2010-000025.R1

Reviewer 1: Hamilton, William

Peninsula College of Medicine and Dentistry, Primary Care

The Study	**Yes**	**No**
Is the research question clearly defined?	✓	
Is the overall study design appropriate and adequate to answer the research question?	✓	
Are the participants adequately described, their conditions defined, and the inclusion and exclusion criteria described?	✓	
Are the patients representative of actual patients the evidence might affect?	✓	
Are the methods adequately described?	✓	
Is the main outcome measure clear?	✓	
Are the abstract/summary/key messages/limitations accurate?	✓	
Are the statistical methods described?	✓	
Are they appropriate?	✓	
Is the standard of written English acceptable for publication?	✓	
Are the references up to date and relevant? (If not, please provide details of significant omissions below.)	✓	
Do any supplemental documents e.g. a CONSORT checklist, contain information that should be better reported in the manuscript, or raise questions about the work?	✓	

If you answered No to any of the above, please supply details below.

It's improved considerable (I assume the boldface is the correction). The key message box gives info not available in the abstract on staging/grade (and not that much in the manuscript) so should be rewritten.

RESULTS AND CONCLUSION (For articles reporting research findings only)	**Yes**	**No**
Do the results answer the research question?	✓	
Are they credible?	✓	
Are they well presented?	✓	
Are the interpretation and conclusions warranted by and sufficiently derived from/focused on the data?	✓	
Are they discussed in the light of previous evidence?	✓	
Is the message clear?	✓	

If you answered No to any of the above, please supply details below.

I still have the concerns about generalisability that I aired in my first review. Yes, they've done a fair job in quantifying the level of inaccuracies in ovarian cancer diagnositics. the $64m question is whether this is a specific thing for ovary, or a specific thing for cancer. or a generalised issue with electronic records (in which case we GPs should be very worried).

The new conclusion has a few grammatical errors (if split infinitives still count..) but these will be picked up later if accepted.

REPORTING AND ETHICS	**Yes**	**No**
Is the article reported in line with the appropriate reporting statement or checklist (e.g. CONSORT)?	✓	
Are research ethics (e.g. consent, ethical approval) addressed appropriately?	✓	
Is the article free from any concerns about publication ethics (e.g. plagiarism, fabrication, redundant publication, undeclared conflicts of interest)?	✓	

If you answered No to any of the above, please supply details below or contact the editorial office.

req **BMJ Open uses compulsory open peer review. Your name and institution will be returned to the authors and will be published with this review if the article is accepted. Therefore please sign your review in the box below. Include your name, position, institution and country. Please also include a statement of competing interests. If you have filled out an ICMJE Conflicts of Interests form - please attach this using the box beneath instead.**

Prof Willie Hamilton, Peninsula College of Medicine and Dentistry.

No conflicts of interest.

req **Recommendation**

Accept

✓ Minor Revision

Major Revision

Reject

Would you be willing to review a revision of this manuscript?

✓ Yes

No

Comments

If you have any further comments for the authors please enter them below.

Reviewer 2: Neal, Richard

North Wales Clinical School, Primary Care

The Study	Yes	No
Is the research question clearly defined?	✓	
Is the overall study design appropriate and adequate to answer the research question?	✓	
Are the participants adequately described, their conditions defined, and the inclusion and exclusion criteria described?	✓	
Are the patients representative of actual patients the evidence might affect?	✓	
Are the methods adequately described?	✓	
Is the main outcome measure clear?	✓	
Are the abstract/summary/key messages/limitations accurate?	✓	
Are the statistical methods described?	✓	
Are they appropriate?	✓	
Is the standard of written English acceptable for publication?	✓	
Are the references up to date and relevant? (If not, please provide details of significant omissions below.)	✓	
Do any supplemental documents e.g. a CONSORT checklist, contain information that should be better reported in the manuscript, or raise questions about the work?		✓

If you answered No to any of the above, please supply details below.
No supplemental documents

RESULTS AND CONCLUSION (For articles reporting research findings only)	Yes	No
Do the results answer the research question?	✓	
Are they credible?	✓	
Are they well presented?	✓	
Are the interpretation and conclusions warranted by and sufficiently derived from/focused on the data?	✓	
Are they discussed in the light of previous evidence?	✓	
Is the message clear?	✓	

If you answered No to any of the above, please supply details below.

REPORTING AND ETHICS	Yes	No
Is the article reported in line with the appropriate reporting statement or checklist (e.g. CONSORT)?	✓	
Are research ethics (e.g. consent, ethical approval) addressed appropriately?	✓	
Is the article free from any concerns about publication ethics (e.g. plagiarism, fabrication, redundant publication, undeclared conflicts of interest)?	✓	

If you answered No to any of the above, please supply details below or contact the editorial office.

req **BMJ Open uses compulsory open peer review. Your name and institution will be returned to the authors and will be published with this review if the article is accepted. Therefore please sign your review in the box below. Include your name, position, institution and country. Please also include a statement of competing interests. If you have filled out an ICMJE Conflicts of Interests form - please attach this using the box beneath instead.**

Richard Neal
Cardiff University

req **Recommendation**

✓ Accept

Minor Revision

Major Revision

Reject

Would you be willing to review a revision of this manuscript?

✓ Yes

No

Comments

If you have any further comments for the authors please enter them below.

Authors Response to Decision Letter for (BMJ Open-2010-000025.R1)

Using free text information to explore how and when GPs code a diagnosis of ovarian cancer. Observational study using the General Practice Research database.

Many thanks. I have removed the reference to the stage and grade of the tumour in the key message box, as this is not of key interest in this paper. We have corrected the split infinitives and other grammatical errors in the Conclusions.

I agree with Professor Hamilton re: generalisability, and indeed we have mentioned this in the

Limitations. However, we do not have the data here to address this $64m question. This would be a very interesting topic for another study.