Reviewer’s report

Title: ‘Caveat emptor’: the cautionary tale of endocarditis and the potential pitfalls of clinical coding data: an electronic health records study

Version: 2 Date: 24 Jun 2019

Reviewer: Martin Thornhill

Reviewer's report:

Thank you for asking me to review this revised manuscript. The authors have clearly made a number of changes and have attempted to address some of the issues raised by both reviewers - generally improving the paper.

On this occasion, the line numbering referred to in the authors response to the reviewers' comments seemed to align much more closely with the text of the manuscript. Thank you. Thank you also for the clean version of the manuscript.

There are still residual issue of concern, however:

1. The Title.
I remain unconvinced by the authors response to my previous comments regarding this. Now that the authors seem to have finally concluded that they are writing a scientific paper rather than an opinion piece, a journalistic title that expresses an opinion rather than describing the study, is inappropriate. "'Caveat emptor': the cautionary tale" etc is better suited for a newspaper than a scientific journal. But I am happy to let the editor decide.

Using the same kind of phraseology for the concluding sentence of the 'Conclusions' is also inappropriate and opinionated. Rather than starting the concluding sentence with "The cautionary tale of infective endocarditis should serve as a learning point for anyone wishing to use diagnostic codes....". Why not use "This study should serve as a learning point for anyone wishing to use diagnostic codes....".

2. There are still some problems with the referencing e.g. reference 24 (line 102 of the tracked changes version) references an opinion paper in the British Dental Journal but the proceeding sentence talks about a US health insurance data study - reference 40.

3. There is a terminological confusion that runs throughout the manuscript. You talk about ICD-10 diagnostic codes and then refer to primary and secondary codes. i.e. meaning respectively the single ICD-10 code recorded as the primary diagnosis for that hospital admission and the secondary diagnosis codes of which there can be several. This is correct. However, you also call what the WHO refers to as supplemental or additional codes, used to identify the possible causal agent for diagnosis codes, as secondary codes. This is inaccurate and confusing. While diagnostic codes, such as those you list in
Table 1, are often referred to as primary or secondary diagnosis codes to denote their diagnostic ranking, codes such as, the B95-B96 codes, are not diagnosis codes. They are supplementary or additional codes used to provide further information on the infectious agent likely to be responsible for the diagnosis codes recorded. As the WHO descriptor for these codes says "These categories should never be used in primary coding. They are provided for use as supplementary or additional codes when it is desired to identify the infectious agent(s) in diseases classified elsewhere."

Unfortunately, throughout the manuscript, the term 'secondary' is used to describe both secondary diagnosis codes and supplementary/additional codes. And this is both confusing and unhelpful. It would make the whole study easier to follow if different terms were used for these. Perhaps, primary and secondary when referring to diagnoses codes e.g. I33.0, I38, I339 etc, and supplementary codes when referring to supplementary/additional codes e.g. B95 - B96 etc.

The distinction is important since recording of a primary diagnosis code - is a requirement for all hospital admissions. Recording of secondary codes provides a means for recording additional diagnoses where needed. While the recording of supplementary codes is not required and is purely to provide additional further descriptive information relating to the causal organisms for the main diagnosis. Thus, recording of it is much more variable and differs considerably from institution to institution and over time according to different imperatives and the time available to recorders.

The distinction between diagnosis and supplementary codes, should be made clear in the methods section and a list of the supplementary codes provided somewhere (it may already be present in one of the many annexes). At present only primary and secondary diagnosis codes are described in the methods section.

Throughout the text, as indicated above, different terms should be used when referring to diagnosis codes and supplementary codes. The term 'secondary' is used to describe both secondary diagnosis codes and supplementary codes at present.

This causes considerable confusion, particularly in the final section of the Results and the Discussion.

4. At the end of the first paragraph of the discussion, you mention that the supplementary organism codes might allow changes in the proportions of differently coded organisms over time to be tracked. This idea repeated in lines 478-9. I disagree, for all the reasons mentioned in 6 below. Furthermore, we clearly demonstrated that this was not the case in the results of reference 14 where we reported: "Pathogen-specific secondary or supplementary coding of causal organism in cases of infective endocarditis was unevenly distributed and increased from 32% to 49% during the study period (data not shown). The rate of increase was also uneven and diminished over the last 4 years of the study (April, 2009 to March, 2013). Furthermore, no specific codes could be used to identify oral viridans group streptococci. As a result, we could not obtain any meaningful information from this data with respect to the effect of the change in prescription of antibiotic prophylaxis on the nature of the organisms that caused the reported cases of infective endocarditis." Clearly demonstrating the limitations of supplementary codes to identify trends in causal organisms even in a much larger data set than used in this study.

Furthermore, you state yourself in lines 427-8 "using secondary codes to estimate incidence of streptococcal endocarditis can give misleading incidence trends". All the limitations that affect the use of supplementary codes mean that the data is far too compromised to be used with any confidence to track changes in the pattern of organisms responsible for IE over time - as you state yourself with regard to Streptococcal IE.
Lines 429 through 432 should be deleted as should lines 478-9 (except the word 'endocarditis' at the start of line 478).

5. In the Discussion, there should be a clear "Limitations" section, rather than a "Strength and Weaknesses" section. The current section does not clearly lay out all the potential limitations of the study as should be the case. And, indeed, as you advocate for others report coding data studies. In particular the casual use of the term 'secondary codes' causes confusion and lack of clarity regarding potential limitations. Again, you should specify if you are talking about secondary diagnosis codes or supplementary codes for organisms. At present, little is said about the limitations of the study and what is said is scattered through the discussion e.g. a bit in the "Strength and weaknesses section" a bit in the para starting line 533, but these should be gathered more clearly into a proper "Limitations" section. The strengths are well covered but even then, they are scattered through the discussion some in the "Strength and weaknesses" section and some in other sections of the Discussion e.g. the paragraph starting at line 468. The strengths could be left scattered as they are through the Discussion or grouped under a heading, but the limitations must be clearly grouped under a "Limitations" heading.

6. On several occasions now I have pointed out a list of potential limitations to the use of supplementary causal organism codes in any coding data study, some of which were clearly demonstrated in the results of reference 14. But the authors have continued to ignore these. However, they need to be fully listed in a section of the Discussion headed 'Limitations' (along with other limitations such as those listed in the para starting line 533). Although there is mention of some of these, it needs to be clearly stated that the following Limitations are an issue with the use of supplementary codes to identify the causal organism for IE admissions. These include:

* Lack of supplementary codes to identify some important causal organisms for IE.
* The fact that recording of supplementary codes is voluntary.
* And they are, therefore, only recorded in a proportion of IE hospital admissions
* Furthermore, that proportion can change over time due to political imperatives, administrative, resourcing and other factors (and varies enormously between hospitals) completely unrelated to epidemiological or disease related factors.
* Such factors can result in the changes affecting certain species and not others
* Furthermore, the use of supplementary codes means that the organisms identified may have no relationship to the disease being studied but could instead relate to intercurrent infections and other pathologies and co-morbidities unrelated to the primary diagnosis.

You should also separately list the limitations that relate to the use of diagnostic codes.

Some of these points are mentioned in passing but there is not a clear list of the limitations associated with the use of both diagnostic and supplementary codes that clearly separates and distinguishes the issues that afflict each.

7. Under 'Clinical and policy implications' you say "Regarding the clinical concern that infective endocarditis, and specifically streptococcal endocarditis, increased in England after changes in antibiotic dental prophylaxis around 200714, our work suggests that the major studies examining endocarditis incidence have not used any poorly-predictive codes, but that the algorithms used could nevertheless have overestimated incidence trends by including short admissions/readmissions." The only reference attached to this statement is reference 14 and so the implication is that these statements relate to that study i.e. the paper by Dayer et al published in the Lancet. However, while the Dayer
paper collected causal organism data using supplementary codes, we reported "we could not obtain any meaningful information from this data with respect to the effect of the change in prescription of antibiotic prophylaxis on the nature of the organisms that caused the reported cases of infective endocarditis." and said nothing about streptococcal endocarditis. Like-wise, we made clear in the last reviewer's comments that we had examined the effect of removing short admissions/readmissions and found that it had no effect on the outcome of our study. It is therefore totally incorrect to imply that as we had (i) look specifically at streptococcal endocarditis, (ii) "overestimated incidence trends by including short admissions/readmissions. This may be true of other studies, but it was not true of ours - but it is our study that is referenced, and this should be corrected.

Other specific corrections (line numbers refer to those in the tracked changes version):

1. Line 57. Replace "secondary" with "supplementary". There are many other such occurrences that need attending to throughout the text.
2. Line 352. Change "using all admissions with any endocarditis code (as defined in Table 1)…" to "using all admissions and all endocarditis codes (as defined in Table 1)…"
3. Line 365. Change "…use only the I33.0 and I33.9 codes…" to "…use only the I33.0 code or the I33.0 and I33.9 codes…"
4. Line 565. Change "The cautionary tale of infective endocarditis should serve as a learning point for anyone wishing to use diagnostic codes…". To "This study should serve as a learning point for anyone wishing to use diagnostic codes…".
5. Line 403. Replace "secondary" with "supplementary".
6. Line 405. Change "….moderate agreement between streptococcal codes and Streptococcus spp….." To "….moderate agreement between supplementary streptococcal codes and Streptococcus spp….."
7. Line 407. Change "…had an associated streptococcal code; 94 (30%) had no organism code and 19 (6%) had a different organism code…” to "…had an associated streptococcal code; 94 (30%) had no supplementary organism code and 19 (6%) had a different supplementary organism code…”
8. Line 411. Change "…had no code…” to "…had no supplementary code…”
9. Line 413. Change "Use of secondary organism codes, and secondary codes overall, increased…” to "Use of supplementary organism codes, and secondary diagnosis codes, increased…”
10. Line 427. Replace "secondary" with "supplementary"
11. Line 428. Replace "…likely due to increasing use of secondary codes…” with "…likely due to increasing use of supplementary codes…”
12. Line 429. Replace "organism codes" with "supplementary organism codes"

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes
Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

'I declare that I have no competing interests'

Statement on potential review bias
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?
2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria”. If your reply is yes to any, please give details below.

'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria”.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal