Reviewer's report

Title: Case Identification of Mental Health and Related Problems in Children and Young People using the New Zealand Integrated Data Infrastructure

Version: 0 Date: 09 Sep 2019

Reviewer: Elizabeth Ford

Reviewer's report:

I congratulate the authors on tackling an important topic and using a strong data source to do so. There are many strengths to this paper, but also some lack of clarity and depth of thought which I feel should be rectified before I could recommend publication, and which preclude the use of the method they present in further data sources. Detailed comments listed below:

1) The introduction makes a good case and rationale for the study and is concise and clearly written.

2) The methods sections on the data source are well written and clear.

3) The description of the development of the case definitions needs more work. For example:
   a) How were the 13 MH problems of interest established; by what process and using what justifications were these reached? What did each of the headings encompass, for example, were OCD, phobias, PTSD included under the anxiety heading?
   b) Table 1 is blank - please supply a new version
   c) It appears that the second stage of work involved drawing up code lists of codes or drug names which were agreed by consensus to represent the MH problems of interest. Given that the authors recognise that different sources of data are more or less reliable, and more or less indicative of a definitive diagnosis, was any attempt made to weight the codes by how much it was likely to indicate a diagnosis? E.g. were medications given a lower weight than ICD diagnostic codes in specialist mental health care? If not, why was this not considered?
   d) On the basis of what evidence were the age categories created and restrictions applied for age threshold of diagnosis?
e) In the section on data management page 13. Can you clarify that a single code from the code list in any of the categories was sufficient to get a 1 in the dichotomous mental health problem indicator variable? If there were more than one code per indicator, this was disregarded? Could you give a good justification for this approach?

4) Results. As Pharms identifies the most individuals, it flags to me that there may be a problem with using Pharm data as a proxy for diagnosis. I understand it may be capturing primary care diagnoses, which the other datasets cannot do. But it may also be throwing up a lot of false positives.

5) Discussion. Page 16. This seems to be the first time you have described your approach as an algorithm. I think this is overstating the approach and is misleading. It seems like case-finding was just matching single codes from a code list? It would be more likely to be called an algorithm if various codes had different weights in terms of determining whether a diagnosis was present, or if different combinations of codes indicated a diagnosis (e.g. secondary diagnosis code from one dataset + prescription from another). Suggest you change Algorithm to another term.

6) What is the likely quality of your case finding method in terms of sensitivity and specificity? This has not been addressed. How does your method work as a whole? What is the rate of false positives from this method? How many cases might you miss? Does having a code on the list from one data source make you more likely to be a true case than if the code came from another data source? Are the codes within the code lists all equally likely to be representative of a true diagnosis? Even if you can't actually answer these questions, a discussion of the limitation and much more detailed suggestions for future validation should be included in the discussion. You could include here literature on validation methods, and literature on influences on recording of conditions by medical practitioners (e.g. stigma, uncertainty), which may inform our understanding of what might be missing.

7) Page 17 you say "The assignment of diagnostic categories using medications should be considered an informed guess, rather than a definitive classification". It's great that you recognise this but what should be done about it? How will it affect the next stages of development and validation of your method?
8) Page 18 you say "administrative data lacks clinical detail and have known quality issues which affect accuracy of case identification" - so what should be done about this for your case definition? what implications does this have for using the method you present?

**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.

No

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

Yes

**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

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