Reviewer’s report

Title: Chart Validation of an Algorithm for Identifying Hereditary Progressive Muscular Dystrophy in Healthcare Claims

Version: 0 Date: 12 Mar 2019

Reviewer: Kelvin P. Jordan

Reviewer's report:

Please include all comments for the authors in this box rather than uploading your report as an attachment. Please only upload as attachments annotated versions of manuscripts, graphs, supporting materials or other aspects of your report which cannot be included in a text format.

Please overwrite this text when adding your comments to the authors.

Routinely recorded health information is of great value to research but there needs to be confidence in the research quality of this information which is not recorded for the purposes of research. This study examines the ability of an algorithm to detect patients with muscular dystrophy (MD).

Whilst the aim of the study is important, and results are promising, there are several weaknesses which lessen its impact, and means the conclusion that this algorithm "…accurately identifies patients with MD…" is over the top.

1) The methods - examining only records of people identified as positive by the algorithm - means the study cannot examine sensitivity or NPV, therefore it is not known whether the algorithm is missing a large percentage of true cases. This is very important for studies using claims data to recruit patients, or using such data for observational research. I appreciate this would be difficult to design but I wonder if the authors could also have assessed those with only one visit coded with MD, or a relevant "control" population, to get an indication whether only absolutely certain cases were identified and others may be missed?

2) The problem above is compounded in that records were only examined for less than 25% of patients identified by the algorithm. How do we now interpret the quality of the algorithm? Does this mean that we can only have confidence in being true cases of MD for less than a quarter of patients identified by the algorithm or do the authors feel the algorithm is accurate for all patients it identifies? How do the remaining patients (~400) differ from those examined? If patients do not have adequate information in their records to be assessed, what does this indicate about the quality of their records and likelihood of having MD, or the quality of their data for research?

3) The design as it stands also removes any element of blinding by the reviewers. They know the patients were all identified by the algorithm and therefore likely had MD.

4) There is no detail on how the algorithm was developed, and whether others were attempted and fared less well. The algorithm itself is a little unclear (p.7). Is it simply relevant ICD code, male, and aged less than 18 at diagnosis or does it also include specialist?
5) It would be interesting to know the level of agreement by reviewers on the first 10 cases which were jointly reviewed.

**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.

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

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.