Reviewer’s report

Title: Validation of diagnosis codes to identify side of colon in an electronic health record registry

Version: 0 Date: 14 Jan 2019

Reviewer: Reviewer 2

Reviewer's report:

PEER REVIEWER ASSESSMENTS:

OBJECTIVE - Full research articles: is there a clear objective that addresses a testable research question(s) (brief or other article types: is there a clear objective)?

Yes - there is a clear objective

DESIGN - Is the current approach (including controls and analysis protocols) appropriate for the objective?

Not sure - key details are missing from the manuscript

EXECUTION - Are the experiments and analyses performed with technical rigor to allow confidence in the results?

No - there are minor issues

INTERPRETATION - Is the current interpretation/discussion of the results reasonable and not overstated?

No - there are minor issues

OVERALL MANUSCRIPT POTENTIAL - Could an appropriately REVISED version of this work represent a technically sound contribution?

Maybe - with major revisions
GENERAL COMMENTS: This paper examines the feasibility of using structured diagnostic codes in EHRs to determine tumour location for patients with mCRC. This study is well written and potentially relevant to other researches that may wish to leverage coding in EHRs for secondary uses. As the authors stated, it is indeed important to understand the reliability and completeness of EHR-based variables, although this would not be a novel assertion.

REQUESTED REVISIONS:

My major concern with this study design is the validation sample size of 100 cases. Even though demographic and clinical characteristics seem to be comparable, could these have happened by chance? If another random sample of 100 patients was drawn, would the same characteristics be observed? Did the authors draw any other samples before selecting this specific one? Ideally if checks were not carried out manually, a larger sample size (thousands) would be preferred.

Still related to the above comment, in Table 1, could the authors also compare the age distribution across both groups using averages with standard deviations (another metric of skewness could also be helpful)?

Secondly, could the authors please expand the discussion section to focus on the reproducibility of this work both in statistical terms (i.e. validation sample size, already mentioned above) but also by describing coding practices at their institution and how this may differ across different institutions/countries. For instance, if the primary purpose of coding diagnoses is billing, would there be any particular bias towards specific codes if they have similar costs or if some are considerably more expensive? A discussion on the origins and purpose of this data would be most helpful - this could also be done in the introduction section.

ADDITIONAL REQUESTS/SUGGESTIONS:

None

Note: This reviewer report can be downloaded - see attached pdf file.

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.

Not relevant to this manuscript

**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.
This reviewer has been recruited by a partner organization, Research Square. Reviewers with declared or apparent competing interests are not utilized for these reviews. This reviewer has agreed to publication of their comments online under a Creative Commons Attribution License attributed to Research Square and was paid a small honorarium for completing the review within a specified timeframe. Honoraria for reviews such as this are paid regardless of the reviewer recommendation.

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.