Reviewer's report

Title: Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records

Version: 0 Date: 11 May 2017

Reviewer: Simon Poon

Reviewer's report:

This paper addresses an important issue relating to data quality and completeness in the electronic patient records. This research project aimed to back fill information regarding comorbidities by integrating additional databases and human judgement in the process. In particular, Enrichment of the CUD (Theriaque pharmaceutical database) with the French National Comorbidities List seemed to be the critical part of this study.

The study proposed several ways to identify the additional comorbidities:

1. Via prescribed/administered drug in EHR - extracting missing ICD-10 codes suggested by the Theriaque database for each prescribed/administered drug

2. Via ICD-10 billing codes in EHR - extracting ICD-10 codes suggested by the Theriaque database for each responding drug in the Theriaque database

3. Using Laboratory Results (only case of hyperkalemia was included)

Identified comorbidities (ICD-10 codes) were subsequently examined by human expert judgement with examination of clinical notes.

In general, the paper was well written and followed a well thought-out procedure. I would like to congratulate the author(s) in regards to the implementation of this research project and would like to address the following concerns from the analytical viewpoints

* Page 10, line 212, the agreement is 98.2% with CI between 0.58 and 0.90. Please check the solution again.

* Are supplementary Tables S1 and S2 presented as Appendix 1 and 2 in the paper? Otherwise these tables are missing.

* As recall could not be derived, performance of the procedure relied only on precision. The quality of the suggested ICD-10 codes would rely on the quality of the Theriaque database
and the coverage/completeness of the French National Comorbidities List. There is a potential of bias towards certain ICD-10 codes, particularly the ones in recorded in the List

* I would like to see more details about way human evaluation step was designed and executed. For example, experience & background of the human experts, level of independence of the experts in evaluating, etc… More information will help us determining the repeatability or reliability of this evaluation task.

I believe this paper has achieved the goal of back filling some of the missing ICD-10 codes for certain EHR through the procedure proposed in this paper. My reservations are more related to the context of robustness & implications:

1. Any health data added to patient record would require high objectivity and scrutiny to ensure no errors is inadvertently added to EHRs. Some assurance need to built-in.

2. The enrichment of Theriaque database by the French Comorbidity List could have also introduced errors, i.e. has the French Comorbidity List has been validated for correctness?

3. Objectivity of the human evaluation step can be improved by involving more human experts, for example some randomisation can be added when allocating evaluators to the suggested ICD-10 codes.

4. It will be interesting to explore from the Recall perspective to evaluation this procedure, hence examining the number ICD-10 codes can be identified from the total number of missing ICD-10 codes. One way to achieve this would be by removing some ICD-10 codes from the EHR and see how many ICD-codes this procedure could correctly retrieve.

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

Yes

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.

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