Reviewer’s report

Title: Methods for enhancing the reproducibility of biomedical research findings using electronic health records.

Version: 2 Date: 18 Aug 2017

Reviewer: Jessica Tenenbaum

Reviewer's report:

The authors give a valuable and practical overview of best practices in software development that could be very beneficial to researchers who seek to perform reproducible data mining on EHR data.

Overall, they authors have done a good job addressing the concerns raised by previous reviewers.

I heartily agree with the authors that it is worth mentioning version control tools. While such tools are pervasive in the software development community, they are surprisingly underutilized in the academic environment. The authors make an interesting point regarding the false distinction made between analytic code and software, and the consequential discounting of the importance of best practices in software engineering. I've always thought of it more as an academia vs. industry distinction. I wonder what the breakdown in use is on the bioinformatics side, and how it differs between industry and academia. I digress.

Here are a few additional suggestions to strengthen the paper:

Figures, tables, boxes:

1. The various "sibling" concepts in Figure 1 feel rather apples-to-oranges, particularly "phenotyping" and "instrumental" under data, and "connectivity" under Translational

2. Figure 3 is helpful, though minor issue- the text in the code doesn't seem to match the text in the output. That might be confusing to someone who has not seen this before.

3. I agree with the previous reviewer's comment that Box 1 feels somewhat unnecessary, but it ties nicely in with Box 2, which is more useful, so I think it makes sense to keep.

4. I'm no Git expert, but is line 21 in Box 3 necessary ("$git add defineCohort.R") given that it's the same as Line 13?
5. Most of Box 3 is reasonably self-explanatory, but it might be useful to explain lines 52-57-what do +++, ---, @@, -1,2 +1 mean? Of the PatientLowerAge lines, which is being added and which taken away, and how can one tell?

Miscellaneous:

* Page 6- "design by contract"- would take this out, or give EHR data example.

* An example UML diagram would be particularly helpful for those unfamiliar with the concept.

* Page 11- "While some libraries for manipulating and analysing EHR data exist[64], their adoption is not widespread"- why? Based on the findings of this paper, it would seem that libraries for manipulating EHR data, for R or otherwise, is precisely what's needed. So if they already exist, why have they not been adopted? And why should someone try to follow the practices described herein? Would such contributions also not be adopted? Even if the authors don't know why, they might speculate and suggest how would-be package writers could avoid the same fate.

A number of awkward and difficult to parse sentences that would benefit from copy-editing:

* "The process is subsequently converted to machine readable instructions, executed and data are usually exported from a relational database"

* "(e.g. diagnostic code position has to be located in or if all children terms of the ontology are included"

* "in terms of reproducibility but equally of building and sustaining a research community around the use of EHR data"

* Page 11 not clear what's meant by "that warrant standardised coding practices"

* "Future changes in the underlying statistical packages or operating systems (OS), such as for example back-compatibility of very old proprietary binary file formats used in commercial statistical packages or the ability to execute older versions of a library developed for an OS that is no longer actively supported, can potentially be mitigated"

Minor:

1. Table 1- reference to it has an extra period- "Table. 1"
2. Page 3 line 37 extra e- "guidelines focuse"

3. Table 1- the seemingly out of nowhere numbers are somewhat jarring. Looking at the reference, I get it now, but providing some context might be helpful.

4. Page 4 line 52, "subtle prevalent" should be "subtle, prevalent" or "subtle but prevalent"

5. Page 5
   a. line 19 " several of the most critical ones," should be "several of the most critical ones:" (colon vs. comma)
   b. line 34 "a collection OF functions"
   c. Line 48- instead of hyphen should just be period?

6. Page 9 line 8 "isol ation"

7. Page 11 line 6 "Scalable"


9. Page 11 line 37 "e.g.matching" needs space

10. Other various grammatical errors at bottom of page 11 and top of 12- disagreement of tense, plurality

11. Page 12 line 42 "mechanism FOR backwards-compatibility"

12. Page 13 line 61 "its" not "it's"; "closing" instead of "to close"

Level of interest
Please indicate how interesting you found the manuscript:

An article of importance in its field

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

Needs some language corrections before being published

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.

Was hosted by Jason Moore at U Penn and received reimbursement for travel expenses.

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

Do you want to get recognition for reviewing this manuscript? Add a record of this review to Publons to track and showcase your reviewing expertise across the world’s journals. Signing up is quick, easy and free!

No