Reviewer’s report

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

Version: 0 Date: 13 Jun 2017

Reviewer: Andreas Bietenbeck

Reviewer's report:

I have read the manuscript „Methods for enhancing the reproducibility of biomedical research findings using electronic health records" with great interest. It explains some useful programming techniques to make research with data extracted from electronic health records reproducible. Very valuable examples are provided. This is a very important topic as reproducibility should be the basis of all science.

However, this manuscript lacks the necessary clarity and structure to be published. The authors should state clearly the intended meaning of "reproducibility" in the context of this work. Is reproducibility just providing complete information? Or is "reproducibility" the ability to execute the same code on the same data and reach the same results? Or do the authors intent something along the lines of transferability (e.g. when the code is executed on similar data similar results are obtained)? After a clear definition has been provided, the authors should state which aspect of reproducibility is supported by which recommendation.

Major points:

- STROBE is only mentioned in the abstract.

- (p. 4 l. 53 ff.) The explanation of modular programming is misleading. Modularity refers to the logical structure of the code (e.g. separation of concerns). The physical layout into one or many files is only secondary.

- (p. 5 l. 46 ff.) Test driven development is not a category of unit testing. You seem to refer to test automation.

- (p. 10, l. 15 ff., table 3) In-line code documentation is intermixed with report generation. In R package roxygen2 is used to create in-line code documentation like you display in Box 1. (Check for example how help files are generated for CRAN.) Other languages use similar doxygen dialects.
Minor points:

- (p. 4, l. 39) Your description of code development implies stages that follow after each other (e.g. waterfall model). Modern software development is usually more iterative.

- (p. 4, l. 49) The whole sections seems to be limited to functional programming languages. Features of object oriented languages like classes are not mentioned. This matters as modular programming is more evolved in object oriented programming.

- (p. 4, l. 50) The list of methods and approaches is not taken up in the subsections of the methods section. In table 2 this the order is changed.

- (p. 8, l. 11) How does version control relate to metadata? What implementation details are released by CALIBER and again how is this related to version control?

- (p. 10, l. 17ff.) Here you mention several properties of bioconductor that enable reproducible research. (e.g. high quality documentation, programming conventions). These properties could be discussed in the introduction and possibly also in more depth in the main part after the meaning of "reproducibility" is clarified.

- (p. 10, l. 45f.) Why do platforms such as galaxy increase scalability? Why does virtualization increase scalability?

- (p. 11, l. 60) Why is an extraction process imported into a database? This sounds like an implementation detail specific to your setting.

- (p. 12, l. 28) Very minor, but because this work recommends best practices, I would not use a plain text editor for any serious programming. Even very basic IDEs make one's life much easier simply because of syntax highlighting.

- (p. 12, l. 48) You are probably referring to Table 3 (and not Table 2).

### Level of interest

Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

### 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

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!

Yes