Author’s response to reviews

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

Authors:

Spiros Denaxas (s.denaxas@ucl.ac.uk)
Kenan Direk (k.direk@ucl.ac.uk)
Arturo Gonzalez-Izquierdo (a.gonzalez-izquierdo@ucl.ac.uk)
Maria Pikoula (m.pikoula@ucl.ac.uk)
Aylin Cakiroglu (aylin.cakiroglu@crick.ac.uk)
Jason Moore (jhmoore@aupenn.edu)
Harry Hemingway (h.hemingway@ucl.ac.uk)
Liam Smeeth (Liam.Smeeth@lshtm.ac.uk)

Version: 3 Date: 22 Aug 2017

Author’s response to reviews:

20/08/2017

Dear Editor; Dear Reviewer,

Re: Methods for enhancing the reproducibility of biomedical research findings using electronic health records (BIDM-D-17-00013R2)

Thank you for considering our manuscript and for the helpful feedback and suggestions. We have addressed the entirety of the reviewer’s comments and suggestions as indicated below. Please note, previous revisions are marked in blue and red; current revisions are marked in green.

Please do not hesitate to get in touch if you require any additional information or changes.
** Reviewer #3

** - 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.

Thank you!

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

Thank you!

** 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.

Good point!

Our assumption and experience was always that the bioinformatics community has embraced and used similar technologies from early on, probably due to the size and complexity of the datasets that existed from the start.

** 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

Good point, we agree, we have revised the figure.

** 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.

Thanks, we have expanded the text in the caption to further explain why some things are not seen in the output and what things are rendered and shown to the user.

** 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.

Thanks! We also like Box 1 very much.

** 4. I'm no Git expert, but is line 21 in Box 3 necessary ("$git add define Cohort.R") given that it's the same as Line 13?

It is the same command yes, but one needs to commit the file every time it changes, hence the duplication.

We have further revised the comments in the code to explain things and hopefully make it clearer to follow the example.

** 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?
We have further revised the comments in the code to explain things.

**Miscellaneous:**

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

Thank you, we have removed the term from the manuscript.

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

We have included a simple example of a UML case diagram as Figure 5 in the manuscript.

**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.**

Good point – we have revised the discussion part of the manuscript to reflect some further thoughts on this.

**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"

Thank you – we have reworded all of the above sentences and made them clearer.

** 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 "Sscalable"

** 8. Page 11 line 7 "Scientific software,"- no comma needed.

** 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"

Thank you – we have corrected the typographical errors in the manuscript.