Reviewer’s report

Title: A systematic review identifying common data items in neonatal trials and assessing their completeness in routinely recorded United Kingdom national neonatal data

Version: 0 Date: 15 Jul 2019

Reviewer: Andrew McWilliams

Reviewer's report:

The authors describe a study with 2 aims. First in a systematic review of 4 journals, they identified randomized trials with neonates as the target population. From these trials, they identified a set of common variables used for baseline, stratification, and adjusting outcomes. They then looked at a research registry to report on the completeness of these variables. This effort sets the stage for such data elements to be used both in trial design and in off-setting some of the primary data collection efforts needed in trial conduct.

Major:

Introduction:
Authors need to make a more compelling case for why this exercise to use real-world data elements is necessary. Comment on costs and efficiency. Can you cite literature and actual dollars. If I do not have to collect 9 data elements as part of a trial…is that going to really be more efficient (how much), more cost effective (how much)…

Also, are there references you can add from the pragmatic literature that help frame what level of completeness is desired to make a variable useful e.g. is 90% good enough?

Discussion needs added commentary on how this study adds significance despite no study using all 14 of most common variables and most variables were not common. This might suggest unique data from primary data collection are actually needed for studies, an idea counter to the suggestion that a common data core will be used and thus drive efficiency in study conduct.

Lastly - Intro framed that using common data core would allow more efficient, cost-effective study conduct, but this point is absent from discussion. Authors need to frame how this would happen, and look at other non-neonatal areas where this has successfully been done, framing their results and suggestions in such context.

Minor:

Introduction:
Page 4 Line 12: add ref and more detail around cost.

Page 4 Line 19: Sentence is redundant.
Page 4 Line 42-44: This area is redundant to prior.

Add sentences or paragraph with references defining pragmatic research, which is where this study’s results could be most applicable.

Page 4 Line 55. Comment on approach to identify non-outcome related variables. Are you defining the approach of 'how' to define here as well as defining a set for neonates? IF the intent is to highlight an approach (steps to take)…this needs to be defined clearly throughout.

Word choice --Could you refer to 'data item' as "variable" throughout?

Method:

Page 5 Line 52: In checking your online prospective registry, it says inclusion population &lt;34 wks receiving special, high-dependency, or intensive care. Was there a change from what was proposed? Need to add that descriptor here if changed for a reason.

Page 7. Line 7. Aggregating many different types of support and counting presence of any one variable as complete for a very important clinical trial variable like respiratory support is a bit of a leap that should be discussed in limitations.

Page 7 line 10. 2 errors. We = was. Or = nor…ie …neither…nor

Page 7: Need definition for implausible, how was determination made?

Results:
Page 7 Line 36. Avoid 'nearly' in results…put actual number

Page 7 Line 45 - need to clean up stratification variable paragraph…too easy to get lost in numbers in current state. Can you try using the same cadence? 35 trials reported n=x stratification items with y and z being most common. 33 trials reported n=y adjustment items. Then keep same cadence/terminology as baseline data paragraph….of these x# defined as common (&gt;20% of trials) and y# were only present in single studies.
It would be helpful for reader to walk away from these 2 paragraphs with clear understanding of how many variables meet definition of 'common', broken out by baseline, stratification, and adjusting categories.

Page 8 Line 2. Would be helpful to know how demographics and other characteristics of NNRD population lines up to trials…but this may be too difficult to obtain.

Page 8 Line 14. Avoid 'nears' - put result in.

Page 8 Line 22. Can you put in avg for completeness by term/preterm here?
Discussion

Page 8 Line 37: Needs to be tempered. I'm not so sure that the assertion is supported by your finding…rather the NNRD may serve as a reliable source for some important variables in neonatal clinical trials.

Page 8 Line 43: How would using the common data elements help one determine suitability and feasibility of NNRD for a specific study…more so I think you are suggesting that given the completeness of data element in the NNRD, it is suitable for investigators to use in planning trials?

Page 8 Line 50: First mention that these are 'simple' trials? Do you mean this?

Page 8 Line 64: Are researchers really going to have much sway in changing the culture of routinely collected data across multiple countries.

Page 9 Line 5: recorded at baseline? Some of the adjustment variables are not baseline e.g respiratory support.

Page 9 Line 17: 'avoiding error in trials’? what is meant here. Error in design, bias, misinterpretation of results?

Page 9 Line 19: second reference to existing methodology for outcomes variables. Might be helpful to compare/contrast that methodology to what you have done in discussion. You do a little bit of this in line 50…but bring them together and add to.

Limitations: would move to last paragraph before conclusion.

Page 9 Line 22: main limitation…this sentence is choppy and needs to be reworked. Also I would not say this is the main limitation, actually seems rather minor that studies may have not reported all variables collected. Main limitation is probably that there is no way in current study to evaluate accuracy of data in addition to completeness.

Page 9 Line 33 - would expound on this limitation. This is actually a limitation to the whole concept of using something like NNRD in lieu of primary data collection…there will frequently be some data elements that are difficult to capture in real-world data.

Also in limitations, you need to comment on the aggregation of data elements, which could have biased your results towards completeness.

Additionally, recognizing that some of the NNRD data elements, may not align exactly with those used in trials.

Lastly -commentary on completeness is all positive, but somewhere in limitations or discussion, authors should comment on the implications of some of these variables having 10-20% missingness...that is actually a pretty big deal for analyses.
Page 9 line 50: Not sure I follow why doing this for laboratory tests is not relevant to clinical trial conduct, the authors just chose not to look at labs…but those might be very important data elements for stratification and adjusting.

Page 9 line 52: Not sure I see relevance of Sheehan ref here?

Page 9 line 59: add more on chari comparing what you did.

Conclusion:

Page 11 Line 35. First mention of accuracy here…which is actually not appropriate as your study did not look at accuracy at all. Remove and discuss in limitations.

Page 11 Line 38: Reference to suggestion for lower impact studies seems random and not supported by study.

Conclusion needs to be crisper. Is conclusion focused on NNRD can be used for trial design only or are you also suggesting that trials should use NNRD for conduct of trial instead of primary data collection.

Page 11 Line 45: point is not really when planning EMR systems…but rather that care providers/health systems continue to work towards a culture of completeness in primary data collection because the completeness and accuracy of such data are important for research.

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

Quality of figures

All images and figures within the manuscript should be genuine i.e. without evidence of manipulation. No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. If you have concerns about the veracity of the figures you should choose the first option below.
**Statistical review**
Is it essential that this manuscript is seen by an expert statistician? If so, please give your reasons in your report.

**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

Were you mentored through this peer review?

No