Reviewer’s report

Title: A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus)

Version: 0 Date: 07 Jan 2019

Reviewer: Marc Avey

Reviewer's report:

This is an important and timely RCT on a question that has broader implications than simply the effectiveness of reporting guidelines to improve the completeness of reporting in preclinical research. Efforts to improve the completeness of reporting are not new to preclinical or other domains of research, but actual tests of the effectiveness of the interventions (guidelines) are scant to non-existent as the authors note. The authors have done what the rest of the meta-research community so often fail to do: Conduct an actual RCT of the intervention (reporting guideline) instead of banging out another descriptive study on what we already know (reporting is poor). I strongly support this paper being published and only recommend changes to improve the clarity of the methods/results and rationale for the outcomes reported.

Throughout: Is there any rationale for switching between 'manuscript' and 'paper' throughout? Consider using reporting results a statistically significant or no statistically difference to make it clear to the reader you are referring to the p-value and not whether there was a potentially editorially meaningful difference.

Comments/recommendations by line number:

Line 129: Please consistently describe this as in vivo animal to make it clear.

Line 134: This is written vaguely. By quality, you mean completeness of reporting? Is the intervention the request (i.e. an email or pop-up box in editorial manager), or the checklist itself?

Line 136: probably true, but not necessary to state.

Line 147: In some jurisdictions if this were seen as quality improvement and not research it would not require ethics approval despite the use of deception. I believe the QI nature of the project is why the authors had trouble getting an IRB to review the project?

Line 161: Does PLOS screening determine in vivo animal vs ex vivo animal vs ex vivo human? This seems important to clarify given the extremely large number of exclusions the authors had.
Line 206: Was the primary outcome assessed at the level of the 38 subitems or at the level of the 108 questions? Did they have to score perfect on all questions in a subitem to be fully compliant with that subitem?

Line 208: Do you mean outcomes or outcome measures? I think the construct here is completeness of reporting and is measured using the ARRIVE guidelines at either the sub-item or question level (dependent variables)?

Line 209: Did they have to score perfect on all questions in a subitem to be fully compliant with that subitem?

Line 209: Are the Landis assessed at the subitem level (if so which ones) or at the question level (if so which ones). Also, the authors refer to these as the Landis risk of bias items?

Line 211: This is only a partial list of the Landis items (not all; see box 1 of the original paper) which are referred as reporting items in the original Landis paper. What rationale was there for selecting these 4 items from Landis' list? The authors also refer to these as risk of bias items but it is unclear how sample size calculations and inclusion/exclusion criteria are related to risk of bias.

Line 209-245: There is a primary outcome, secondary outcomes, tertiary outcomes, feasibility outcomes, and a posteriori exploratory outcomes. Listing the outcomes not in the original protocol is excellent, but it's unclear if these are the only outcomes that are exploratory. Please indicate which, if any, outcomes are confirmatory and which are exploratory since the language is used by the authors.

Since there are so many outcome categories, consider explaining what the categories mean for the reader and how they impact the analysis/interpretation. For instance, what is the difference between the secondary and tertiary outcomes here? The study was powered for both the primary outcome and 1st secondary outcome (aren’t they both primary outcomes then?).

Consider clearly articulating the unit of analysis for each of the 19 listed outcomes so it's clearer for the reader (especially since the IICARUS team had a miserable time identifying this consistently in the primary studies.).

Line 217-225: Is there a rationale for the stratification of these outcomes (e.g. by experimental animal, by country)? Do you expect to find differences and if so why? The human data stratification seems quite odd without an explanation?

Line 227: Which of these feasibility outcome measures relates to the duration it took authors to complete the ARRIVE checklist? Do these outcome measures relate to the construct of complete reporting or these a completely separate set of feasibility measure related to duration/cost to the journal of implementation?
Line 235: No bullet, this is a subheading.

Line 247: Here an in the manuscript you refer to 38 subitems but in appendix 1 they are listed as items. Please use a consistent ontology throughout the manuscript and tables/appendices. Also, it would be clearer to say that you operationalized the 20 ARRIVE items (or headings) into 38 sub-items since the numbers in appendix only count up to 20.

Line 247: What was the rationale/method for operationalizing the ARRIVE items into sub-items and questions (e.g. see methods in your reference Avey et al. 2016)? Operationalizing the ARRIVE guidelines into a proper survey for research purposes is crucial to the analysis plan.

Line 248: When was this determined by the steering committee? Before or after data was collected?

Line 249: Which seven questions?

Line 283: "reviewer quality" or "review quality"?

Line 288: What questions match the Landis criteria?

Line 292: How did the authors arrive at a difference from 1% to 10% and 30% to 50% for their power calculations?

For authors this will be a single trial of the intervention with the expectation of a large behaviour change given previous research on poor completeness of reporting. There is also essentially no penalty or benefit to the individual for completing the checklist (e.g. increase acceptance chance for completion or worst case scenario is resubmitting their paper to Scientific Reports). These seem like large differences to expect between groups here.

Line 311: What was the method for solving disagreement amongst two reviewers for a particular question (e.g. a third reviewer who did not review the question)?

Line 341: The number of outcomes and statistical options is so large here it would honestly help to have a table to know what test was done for which outcome/sub-item

Line 356: "Largely" is vague here. What are the n & % were ex vivo instead of in vivo? This seems like a large number and percentage to be mistakenly included during what I think is the PLOS ONE screening phase. 420 is ~25% of the 1689 randomized manuscripts which are mis-identified at a very basic level of in vivo animal vs ex vivo (animal/human?). This result deserves more attention, despite not being one of the 19 listed outcomes, as it seems absurd that PLOS ONE’s screening process can’t identify such a basic aspects of research reliably.

Please also list what the other exclusion were and the n & % for each.
Consider including the n as well as % when reporting this information in text. For the 28.4\% that fell below the 80\% threshold (also the 80\% threshold for reviewer training) were these manuscripts dropped from the analysis? What is the 80\% threshold important for the results?

Was there an a priori cut off for the kappa level for a question to be included in the analysis? I'm not sure what the kappa values are used for here?

Consider always reporting the control and intervention groups in the same order throughout the manuscript (i.e. line 381 the authors switch the order). Is the 'no difference' statistical (i.e. p-value)? If the author conducted a chi-square analysis then please reported it in full in the text (e.g. follow American Psychological Association guidelines).

Given that there are 19 outcomes in this paper, it would make the results easier to follow if the results sub-headings aligned with the outcome names as listed in the methods (e.g. logistic regression is not an outcome, it's an analysis on the data for the outcome).

Here 9b improved in the intervention group, but in the abstract it's 9a. Is this a difference between the logistic regression and chi-squared analysis? Also in the abstract it's listed as an item and here a sub-item? Are subitems the letters (e.g. b) and the numbers the items? Are the log odds and p value the only important information to report for a logistic regression (e.g. APA format)?

Ok, here's the chi-square for 9b but it's 9a in the abstract and referred to as an item and not a sub-item. I think the 20 ARRIVE items are made into 38 items and some sub-items have letter like a/b/c which are sub-sub-items?

What is ARRIVE item 9b? Please list the sub-item in the text everywhere you discuss the item so the reader knows what the content of the sub-item is.

In the figure caption consider rephrasing to read: "* denotes [statistical] significance"

Do you mean a statistically significant or just a large proportion in this sentence? Also, is there any explanation for why such a large proportion of the remaining manuscripts?

How were data skewed?

Since colour is used to differentiate the countries here, consider using colour blind friendly option for the labels.

why mice and rats? Were these the only species?

Table 3a (mice) and Table 3b (rats) consider reporting the exact p-value regardless of whether it is statistically significant.
Line 455/Table 3: The caption for Table 3 list abbreviations that don't appear in it.

Line 456: Here the authors explore meaningful differences for individual items. I'm not sure what the authors mean here though by "we did not analyse these comparisons statistically" since they report percentages and Cohen's H values. From a quality improvement perspective (which I think that this RCT is designed for), consider reporting what a meaningful difference (improvement) would be for editors between the two arms.

Line 475: Isn't this one subitem which is actually many items?

Discussion in general:

Did the authors consider whether their expectation of increased compliance was realistic given that for most authors in their sample this would be a single trial intervention to change a behaviour?

The authors exploratory analysis and tables indicate that some subitems had increases of ~5% which was sometimes a doubling in the intervention arm (i.e. sample size went from 3.5% to 7.6%). Are these meaningful for quality improvement even if they are not statistically significant?

Why do the authors think Animal Husbandry reporting improved statistically? Is this an artifact of testing so many outcomes?

Line 479: Did the authors consider whether or not the lack of any penalty for not completing the ARRIVE checklist reduced compliance (e.g. reject without resubmission option at PLOS).

Line 544: I don't see where else this is discussed so it caught me off guard here. Is this discussed in the results?

Line 558-561: The RCT did not evaluate whether a smaller number of items would improve compliance compared to larger number of items so I am not sure why this is your conclusion.

Since animal husbandry was improved do the author consider this as important area for the ARRIVE guidelines and journal to focus on in their implementation efforts?

Table 2: This is not referenced in the text. Consider reporting the exact p-value regardless of whether it is statistically significant.

Appendix 1 Heading: The heading indicates "operationalised checklist number" but the authors did not operationalize the checklist numbers (they operationalized the constructs); also no heading in the table corresponds to "operationalized checklist numbers". Are the "questions not
referred to explicitly in the ARRIVE guidelines are shown in grey" the same the same seven that were removed from data analysis?

Appendix 1: The ARRIVE Item #s are referred to as subitems in the text. Consider listing the 38 subitems numerically from 1-38 so it is easier to follow for the reader.

The question numbers are not in order and the lead number does not always match the ARRIVE subitem (item?) number it nests under making it confusing. I think the question numbers are supposed to align with the original ARRIVE numbers but they are not included so it's simply confusing why the 108 questions don't all nest numerically with the 38 subitem and appear in numerical order. Sometimes the authors uses a two digit scheme and sometimes a three digit scheme but it is unclear why or how this helps the reader.

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

An exceptional article

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 organisation 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