Reviewer’s report

Title: A Patient-Centered Composite Endpoint Weighting Technique for Orthopaedic Trauma Research

Version: 0 Date: 17 Jul 2019

Reviewer: Janet Turk Wittes

Reviewer's report:

General comments

This interesting paper has major strengths and major weaknesses. Let me start with the major strengths. Use of the best-worst scaling to elicit preferences is an attractive approach when outcomes can differ greatly in severity and clinical meaningfulness. The large sample size, the care with which the authors chose the outcomes and the population, the fractional factorial design to limit participant burden, the analysis of baseline predictors of utility, and the fact that the utilities follow an order that has face validity are major strengths of the paper. I must confess never to have heard of this method but the paper has led me to think of it for diseases with which I am involved. Also, the paper is clearly written, a feature that has positive utility for the reviewer.

The major weaknesses stem from the magnitude of the utilities themselves. The authors set perfect health at 0 and then use some Bayesian methods to establish utilities. They consistently refer to "relative" importance of the outcomes - these utilities do not fall on a relative scale - they are absolute numbers. While I find the order of the outcomes satisfy my personal sense of face validity, the numerical values do not. Here is the scale the authors develop.

% of Eeath

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Utility</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>-8.91</td>
<td>0.86</td>
</tr>
<tr>
<td>Above knee</td>
<td>-7.66</td>
<td>0.78</td>
</tr>
<tr>
<td>Below knee</td>
<td>-6.9</td>
<td>0.66</td>
</tr>
<tr>
<td>Severe pain</td>
<td>-5.9</td>
<td>0.64</td>
</tr>
<tr>
<td>Deep SSI</td>
<td>-5.69</td>
<td>0.58</td>
</tr>
<tr>
<td>Bone healing complications</td>
<td>-5.2</td>
<td>0.52</td>
</tr>
<tr>
<td>Moderate pain</td>
<td>-4.59</td>
<td>0.37</td>
</tr>
<tr>
<td>Mild pain</td>
<td>-3.3</td>
<td>0.37</td>
</tr>
<tr>
<td>Superficial SSI</td>
<td>-3.29</td>
<td>0</td>
</tr>
</tbody>
</table>
Perfect health

These make no sense to me. That table says that the relative disutility for above the knee amputation is closer to that for death than superficial SSI is to perfect health. Given the centrality of these utilities to all the analysis that follows, I encourage the authors to explain what these numbers mean. If I were to ask people to scale these outcomes on a scale of 0 (perfect health) to 100 (death), where would the numbers fall. I recognize that that proposal is not consistent with best-worst scaling but without some intuitive calibration, I would not know how to interpret results from a trial that used this method.

Lines 234 to 243. (this is rather major). The paper uses a weighting scheme to facilitate comparisons between groups. Why is this a sensible scheme? How dependent is it on the particular utilities?

Line 249. I am uncomfortable with the use of FET here. If you allow more than one outcome per person, you don't have independent data which FET requires. Also, the fact that you are using the weights as fixed numbers without variability means that you are underestimating the strength of evidence. You show 95% confidence intervals for the utilities in Table 2, but you don't use those ranges anywhere in the weighted test.

Minor comments

First, stay away from "relative" unless you really mean relative. Nearly every time you use the term you are referring to numerical utilities.

Line 70. The definition of composite endpoint is only one of a number of definitions in the clinical trial literature. I would say "Braunwald et al. have defined…" or "A commonly used definition of …"

Line 72. A composite outcome reduces sample size only if the treatment accrues benefit on each of the endpoints. Then, depending on the degree of benefit, inclusion of different endpoints may increase or decrease the power.

Line 74. Adding a rare outcome does not confer power for that endpoint.

Line 77. I would say, "Composite endpoints have limitations" or something like that.

Line 80, 81, …"only one event; therefore,…

Line 84. I believe later papers by the Armstrong - Westerhout team do incorporate patient values specific to the target population.

Line 105. "called a choice set" seems to be misplaced.

Line 116. What was the study location. Later the paper says it was at a single location in Baltimore. Stating that here would help the reader.

Line 119. Delete comma after "levels"
124. Not "minimized" - minimum burden would be no questionnaire. Say something like "reduce" or "limit"

125. What is D-efficiency? If that is a well known quantity, at least reference it.

130-133. I found this confusing. What did the respondents do first? And then what? And then what?

138-141. If I understand what you did, there were 10 attributes and you made choice sets of 3. That means there were 3C10 or 80 possible sets. But the experiment used only 40. How did you choose the 40?

145. Delete comma after "older"

148. Change "patients" to "they"

150. Delete "the" before "adequate"

165. Define what hierarchical Bayes model you used.

171-173. Suggest, "….preference for a given outcome. We set the mean utility at zero for perfect health; all other possible outcomes have negative…"

175, Delete "of" and change "outcomes" to "outcome"

177. I would say "To adjust for the 10 statistical tests, we set the level of significance for the interaction terms at \( \alpha = 0.05/10 = 0.005 \)"

182. Please give a reference for the Tukey-Kramer test

184. Suggest, "….respondents by those who had and had not experienced.."

197 - 198. Change to "Nearly half (47.5%) of respondents…". (47.5% is not "most")

217. "was" should be "were"

221-225. Yes, some of these are nominally significant, but if you test 72 comparisons, you should not be surprised by a 0.02. And -7.63 seems trivially different form -7.67. What is interesting to me is the -9.50 and -8.91 for those with an above the knee amputation. That difference suggests to me that the outcomes are not properly calibrated with respect to death. I find it not surprising that people with an amputation think of death as farther from where they are than people without such an amputation. We know that people say, "I'd rather be dead than have x," but when they in fact develop x, they want to live.

245. Suggest, "trial illustrates"

285. See comment above about rare events.

287. Add "or" after fracture" and change "minimal" to "only small"
Lines 293 - 295. See comments above about power.

Line 324. See comments above about power and where does "net clinical benefit" fit in?

Line 333. Change "that" to "who"

Table 1. Nice to see these data.

Table 2. If perfect health is set at 0, why does it have a confidence interval? Also, how is the reader to interpret the -2970 log likelihood?

Table 3. I found this not interesting. These numbers don't look importantly different to me even with the stars.

Exhibit 1. This is not useful. I recommend just including in the text the attributes that are not defined by their name.

Exhibit 2. I am curious - for the choice triplets that included death, was death ever chosen as not the worst outcome? It would be good to know those data to help the reader understand why death is not farther away from upper knee amputation than it is.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**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 I have no competing interest.

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.