Reviewer’s report

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

Version: 0 Date: 22 Apr 2019

Reviewer: Paul Brown

Reviewer's report:

Review of "A Patient-Centered Composite Endpoint Weighting Technique for Orthopaedic Trauma Research", BMC MRM

* I enjoyed reading this paper by Udogwu, Howe, Frey et al. I was pleased to see the authors describe some limitations of composites in the introduction and, given the widespread use of composites, I'm with them when they express a desire to address these limitations.

* Regarding the introductory paragraph which defines composites and lists their benefits: "Composite endpoints in clinical research are defined as the occurrence of any one of several study events ..." I would suggest expanding this definition and using a more up-to-date reference (the reference is from 1992). The definition insists on a particular algorithm for combining outcomes i.e. the any-versus-none type composite. But there are other algorithms for combining outcomes and an expanded definition should acknowledge them (the global rank is later referred to in the text and it would not fall under this definition).

* "Incorporating multiple endpoints into a single metric ... [reduces] the required sample size" (this is reiterated in the Discussion's final sentence). This is a common claim but it is not necessarily true (see e.g. 10.1161/CIRCOUTCOMES.113.000149, and doi 10.1016/j.ajh.2008.05.018). I'd worry that we are pushing researchers towards composites with a promise of maximising power. The claim would have to be more precise and backed up by data (e.g., 10.1016/j.cjca.2016.02.067, or doi 10.1161/CIRCHEARTFAILURE.112.969154). Looking at the summary table in Sun et al., note that a single outcome (dyspnea) beat the 'clinical composite' on power in a majority of assumed scenarios. Power may also be sensitive to the weighting of outcomes, and weights in turn may be data-dependent (doi 10.1161/CIRCOUCEOMES.117.004419). Thus, a blanket claim of increased power feels like marketing. And power is irrelevant as Senn says: https://www.ncbi.nlm.nih.gov/pubmed/2790115. The authors also note that a composite yields "an estimate of the net clinical benefit". I'm not sure I believe this either (see arguments below). Although it might be worth noting that the multiplicity issue is avoided; if composites offer a gain in power (successful trial) it is achieved by not carving up the alpha into unattainable criteria.

* "Composite endpoints enable the inclusion of rare but clinically important outcomes that the study would otherwise be statistically underpowered to detect". But this is just a sleight of hand (i.e. there is a gain in power while appearing to use clinical outcomes) as the authors immediately acknowledge in the subsequent paragraph: "The treatment effect of an outcome of high importance but low frequency, such as death, may be muted by the inclusion of more common outcomes of lesser importance". The reverse could be claimed i.e. optimistic surrogate/subjective endpoints are restrained, rightly so, by the inclusion of clinical outcomes
such as death (that was insinuated by Felker & Maisel who hoped to make phase II results mimic phase III: doi 10.1161/CIRCHEARTFAILURE.109.926030). Incidentally, we illustrated this graphically by defining the Influence of outcomes here: doi 10.1161/CIRCOUTCOMES.117.004419.

* "in traditional analyses with composite endpoints, each study participant can have only one event, therefore censoring subsequent events biases treatment effects to earlier outcomes". No reference is given to support this and it isn't really true e.g. an unmatched win-ratio or global rank prioritises 'important' outcomes, rather than 'early' outcomes. But it depends on what is meant by 'traditional analyses' I guess. A global rank type composite was described by Finkelstein & Schoenfeld in 1999 but maybe the time-to-first composite of survival outcomes has been more common because survival outcomes are more common? That might require a reference because it is used to suggest that the authors' composite is superior by addressing some limitation of 'traditional analyses'.

* Further on in the Introduction the authors describe the aims of the paper: "The first aim was to quantify the relative importance .... of clinical outcomes ... The second aim was to ... develop a patient-centered composite endpoint weighting technique ..." I can see aim 1 covered at length, but I cannot see aim 2 clearly detailed (we are only told how to derive the weights from the mean utilities arrived at in previous sections). As I understand it, there are several parts to the research described by the authors: 1) the elicitation of component weights, 2) the construction of the composite, and 3) data simulations for illustration.

* The bulk of the paper concerns 1), but I am more curious about 2) and 3). The real effort for the authors is to persuade the reader that their composite should be favoured over alternatives (which falls under 2 and 3), because ultimately the relative importance weights prove to be intuitive and non-controversial, or as the authors say, the "clinical outcomes followed a logical gradient". Or is the aim of the paper to discern these weights (calculable from the Excel spreadsheet provided) and then leave it up to the reader to decide how to incorporate them in their analysis (hence the emphasis on 1)? E.g. it is extremely vague in the Discussion when they allude to the possibility of a recurrent events analysis (and the corresponding reference #26 is missing). Sorry, but it's not clear to me. The section titled "Composite outcome weighting: An example", as far as I can tell, doesn't say anything about how the weights were included in the analysis and it's a little perplexing to me that Fisher's exact test was used to compare treatment groups.

* Regarding 2), the more 'precise' the weights are, the more tenuous they will seem. The advantage of a global rank is that we are only obliged to order the outcomes from the most definitive to the least definitive; we are not required to specify numerical coefficients that quantify their importance. In other words we only need to believe the ordering of outcomes and not the distance between them. The problem however is that weights become somewhat data-dependent (as noted above re Influence: doi 10.1161/CIRCOUTCOMES.117.004419). Thus there might be reason to favour the authors' composite over the obvious global rank alternatives (the unmatched win-ratio or the clinical composite promoted by Dr Packer) and it might be worth making that more explicit.

* The authors make the important point that their composite "maintains the total number of study events" (which a global rank does not, leading to inadvertent and unpredictable Influence of component outcomes). But what about e.g. Berger's "information preserving" composite endpoint (doi 10.1016/S0197-2456(02)00233-7). Or Mascha et al. who adopt a
multivariate modelling approach (doi 10.1213/ANE.0b013e31821796d3). Then events are retained and, in addition, weights can be applied post-hoc and are explicit and not hidden within the mechanism of the calculation. Thus, perceived importance of outcomes is not embedded in the results, but merely a layer of interpretation superimposed on the estimates (maybe this is what the authors intend; it is not clear). And one can easily evaluate heterogeneity of the effect across outcomes by including an interaction term in the model (see e.g. doi10.1186/1471-2288-10-49).

* Mascha et al. inferred a gain in power for the GEE model over the composite. It seems whenever a multivariate approach is compared with a composite the former wins on power, yet researchers still claim that the benefit of the composite is a gain in power (example comparisons between approaches: doi 10.1161/CIRCOUTCOMES.116.003382, and doi 10.2147/CLEP.S153196). With the authors' composite, and in general, there seems to be an inverse relationship between the incidence rate for the outcome and its perceived importance and thus weighting (the obvious example is death). Thus the composite may lean on the 'weakest' outcome. If this is death, power can then become, incidentally, sensitive to study duration which is dictated by extraneous factors (i.e. if follow-up is brief you see few deaths and power is depleted). The limitations with regard to power need to be understood and emphasised because a power calculation for a composite, which has so many moving parts, is likely to be crude and unreliable. Hence a costly, failed trial may be the result. In that case you might advise the reader to plan an interim, blinded reassessment of power.

* It is considered good practice to analyse the component outcomes in a supplementary analysis. Also, we can expect sensitivity to treatment to vary across the outcomes. I suppose, given the weighting mechanism, it is not unlikely that a statistically significant result may be seen on an individual outcome and suppressed in the composite (I think the authors allude to such a possibility). This would certainly complicate the interpretation of results. It would be reassuring to see some assessment of the correlations among outcomes from previous studies; effects should be congruent if we are to assert that a composite of the outcomes will yield an accumulation of power and interpretation of results will not be complicated by a treatment by outcome interaction or null effects on some outcomes. But that will be dependent on the clinical question and the study set-up.

* Regarding 3), if the authors want the reader to adopt the composite we must know how to plan a study around it. For example, how might we do an approximate power calculation (e.g. doi 10.22237/jmasm/1509495120)? How would we handle missing data (a particular concern for composites)? How would we analyse the resulting metric? As an amalgamation of outcomes, a composite can be a rather opaque endpoint. Can we even specify a clinically meaningful difference on this scale a priori? Thus, how will the results be interpreted (see doi 10.1161/CIRCHEARTFAILURE.116.003222)? Hence the approach may inadvertently compel us towards significance testing and away from estimation at a time when statistical significance and dichotomisation of results is heavily criticised (https://www.nature.com/articles/d41586-019-00857-9). This indicates the value of e.g. the probability index.

* Can the data simulations used by the authors help us evaluate the performance of the composite? I.e. are they useful in power estimation? The authors claim that their composite avoids "the loss of study power observed with other weighting techniques". This requires a reference e.g. doi 10.1016/j.cjca.2016.02.067, or illustration within the paper. Maybe these issues are too onerous to be addressed in this paper and some comment should be made
about future research. In any case, if the authors can make their programming code available, i.e. the code for data simulations, it might provide the reader with a useful starting point and would allow them to scrutinise the calculations. They would help me to understand, for example, why Fisher's exact test was used for the analysis.

* The data simulations at the moment are likely tautologous and don't illustrate much: the results just repeat the assumptions fed into them i.e. we assume congruent, positive, moderate effects and then report that a difference between treatments was detected. What we really want to know is what happens when the data are not well-behaved and how precarious is the hoped-for power gain that is claimed?

* The alternatives described above (global rank, Mascha et al.) are potentially less computationally intensive, and hence easier to validate and more cogent. Perhaps a macro (in R or SAS) would be ideal if the reader is to adopt the authors' composite? (It depends on how cumbersome the required calculations are.) The main concern is the likelihood of coding errors: the analyst needs to code by hand, from scratch, and few academic researchers have the resources for truly independent validation of code. Standardisation of application would be enhanced by making validated programming code available, if possible (I appreciate that making code user-friendly and flexible requires additional effort).

* Given all the talk about reproducibility, it seems ironic that composites are becoming popular and there is limited sharing of programming code among researchers. There appears to be little standardisation of application with researchers often nominating an ad hoc, arbitrarily constructed composite as the primary outcome for a study. In my view then, the value of this paper is its potential to affect some consensus or standardisation of application within orthopaedic research, to whatever extent, by offering an "off-the-shelf" solution. Although they will need to persuade the reader of the composite's benefits over alternatives. There is likely a reason that Bakal & Armstrong's analogous incorporation of weights (that the authors reference) is not in favour. Yet the authors make the claim in the discussion: "This study's patient-centered composite endpoint weighting technique represents an improvement on previous weighted composite endpoint techniques" and it "addresses many of the challenges inherent to composite endpoints". These are bold claims and at the moment I do not believe they have established this. If they could make just one improvement that would be significant and they should make the claim explicit and precise in the paper.

* Finally, some brief notes about the statistical analyses presented in the paper: I would not adjust for multiple testing. I would not report a result as "P>0.1", especially given the current overwhelming pushback against dichotomising results (e.g. the see Nature weblink above). There is too much reliance on p-values in the paper in my view, although there are no p-values in Table 2 which is a good thing. It is not clear to me what a sentence like this means: "The relative importance for each of the included clinical outcomes was significantly different (P < 0.001)". Significantly different from what? I would hope that we are not comparing e.g. the mean utility for death versus amputation. Significance testing has its place, but it feels excessive here (we know the null hypothesis of patient equipoise is absurd for most comparisons). The authors note that they did not detect heterogeneity but I guess there is no power to test the interaction (that is usually the case).

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 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