Author’s response to reviews

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

Authors:

Ugochukwu Udogwu (uudogwu@gmail.com)
Andrea Howe (ahowe@som.umaryland.edu)
Katherine Frey (kparris1@jhu.edu)
Marckenley Isaac (marckenley.isaac@gmail.com)
Daniel Connelly (connellyd@icloud.com)
Dimitrius Marinos (dmarinos@liberty.edu)
Mitchell Baker (mbaker@som.umaryland.edu)
Renan Castillo (rcastil1@jhu.edu)
Gerard Slobogean (gslobogean@som.umaryland.edu)
Robert O’Toole (rotoole@som.umaryland.edu)
Nathan O’Hara (nohara@som.umaryland.edu)

Version: 1 Date: 02 Oct 2019

Author’s response to reviews:

Reviewer 1

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.

Response: We thank the reviewer for their kind remarks and comprehensive review. We believe comments regarding the methods and suggestions on how the weighting technique may be applied in practice have greatly improved the quality of this manuscript.

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).
Response: We agree with the reviewer’s comments and have revised the sentence to acknowledge that this is but one definition of a composite endpoint, as suggested by the other reviewer.

Revision: A commonly used definition of a composite endpoint in clinical research is the occurrence of any one of several study events of interest [1]. (Introduction)

* "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/CIRCOUOUTCOMES.113.000149, and doi 10.1016/j.ahj.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/CIRCOUOUTCOMES.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.

Response: We agree with the feedback from the reviewer and have revised the sentence accordingly.

Revision: Incorporating multiple endpoints into a single metric increases the number of observed events, can avoid issues pertaining to multiplicity, and thus, may increase statistical power [1,2,3]. (Introduction)

* "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/CIRCOUOUTCOMES.117.004419.

Response: We thank the reviewer for providing some very pertinent references. We have revised the sentence accordingly.
Revision: Composite endpoints also enable the inclusion of rare, but clinically important, outcomes; therefore, providing a broader interpretation of the net clinical benefit of a treatment [1]. (Introduction)

* "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'.

Response: Thank you for the helpful comments. We have clarified that this sentence was intending to refer to a traditional time to first event analysis or other analyses of frequency that only consider the first event.

Revision: Additionally, in studies that analyze composite endpoints using a traditional time to first event analysis or other analyses of frequency that only consider the first event, each study participant can have only one event; therefore, censoring subsequent events biases treatment effects to earlier outcomes.

* 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.

Response: We acknowledge that the three aims are ambitious for a single paper. The primary focus of this manuscript is to describe the elicitation of the component weights. However, we felt that the component weights would be meaningless without describing a vision for the construction of the composite and a brief example of how the composite could be applied to trial data.
Based on the excellent feedback from the reviewers, we have made minor updates to the weighting formula and revised and expanded examples of how the weights may be applied in practice. In future publications, we aim to investigate how the technique performs under a variety of scenarios related to data (e.g., missing data, competing risk, opposing effects), statistical power, rare outcomes, and the number of included components.

Revisions:

A hypothetical pilon fracture trial was used to illustrate the application of the proposed weighting technique (Table 4). In this hypothetical trial, 1000 patients are randomized to hypothetical Treatment A (n=498) or Treatment B (n=502). Three components (deep surgical site infection, bone healing complication, and superficial surgical site infection) were included in the hypothetical trial’s primary composite endpoint. The effect of Treatment A versus Treatment B on the composite endpoint was then calculated using several unweight methods, including a Fisher’s Exact Test, time to first event analysis, and a random effects model. For comparison, the treatment effect was also calculated using several methods that accounted for the proposed component weights, including a Wilcoxon Rank Sums test, time to event allowing for weighted repeated events, and a random effects model that accounted for component weights [32]. The effect size for the random effects models are reported as odds ratios, and hazard ratios are used for the time to event models [33]. The Probability Index was used to report the treatment effect for the Wilcoxon Rank Sums test [32-34]. These analyses were performed using R Version 3.6.1 (Vienna, Austria). All of the data and code for the models are included in Supplementary Appendix B and C. However, for simplicity, only the unweighted and weighted time to event analysis are reported in the results section.

(Methods)

For the hypothetical pilon fracture trial, the results with the unweighted composite endpoint using a time to first event analysis would have determined that there was no difference between the two treatments (hazard ratio (HR): 1.02, 95% CI 0.83 – 1.27, P = 0.83) (Figure 3). When weights are applied to the included component outcomes, and the analysis allows for patients to have more than one event, Treatment A is superior (HR: 0.72, 95% CI 0.57 – 0.90, P < 0.01). A similar difference in effect size was observed when the data were analyzed using unweighted and weighted random effects models (Supplementary Appendix C). However, the treatment effect was not statistically significant when the weights were applied using a global rank approach, and treatment groups were compared using a Wilcoxon Rank Sums test and Probability Index Model. (Results)

* 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.
Response: The reviewer raises an important point regarding the implications of a precise weight. In the revised version of the analysis, we demonstrate how our weighting technique can be applied to several different methods of analysis, including a global rank test. We believe that precision in the weights is valuable for all of the analysis methods presented, including the global rank test. In a global rank analysis with components of similar weight, the precision of the weights will be important in determining the rank when multiple endpoints are observed in a single patient.

Revision: In the weighting formula, the weights adjust relative to the components that are included in the composite. The precision of the weights is useful in distinguishing order in a global rank test with several components of similar weight [32,33].

* 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).

Response: The reviewer’s comment caused us to rethink the importance of preserving the total number of study events. In our original weighting formula, the relative difference between the component weights was reduced in an attempt to preserve the total number of study event. Based on your comments, we have adjusted the formula so that the weights are now on a 0 to 1 scale and therefore more applicable to different types of post-hoc analysis.

Revision:

Derivation of composite outcome weights

An orthopaedic trauma composite endpoint weighting technique based on the mean utilities of the component outcomes and a modified version of the conditional logit formula described by McFadden [23] is provided below:

\[ W_a = \frac{e^{u_b} + e^{u_i}}{e^{u_a} + e^{u_b} + e^{u_i}} \]

The weight \( W \) is calculated separately for each included outcome \( a \) where \( u \) is the mean utility of each included outcome. \( b \) and \( i \) note the component outcomes included in the composite. A weight calculator, with sub-group adjustment, is included in the Supplementary Appendix A.

A hypothetical pilon fracture trial was used to illustrate the application of the proposed weighting technique (Table 4). In this hypothetical trial, 1000 patients are randomized to hypothetical Treatment A (n=498) or Treatment B (n=502). Three components (deep surgical
site infection, bone healing complication, and superficial surgical site infection) were included in the hypothetical trial’s primary composite endpoint. The effect of Treatment A versus Treatment B on the composite endpoint was then calculated using several unweighted methods, including a Fisher’s Exact Test, time to first event analysis, and a random effects model. For comparison, the treatment effect was also calculated using several methods that accounted for the proposed component weights, including a Wilcoxon Rank Sums test, time to event allowing for weighted repeated events, and a random effects model that accounted for component weights [32]. The effect size for the random effects models are reported as odds ratios, and hazard ratios are used for the time to event models [33]. The Probability Index was used to report the treatment effect for the Wilcoxon Rank Sums test [32-34]. These analyses were performed using R Version 3.6.1 (Vienna, Austria). All of the data and code for the models are included in Supplementary Appendix B and C. However, for simplicity, only the unweighted and weighted time to event analysis are reported in the results section.

(Methods)

* 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/CIRCOOUTCOMES.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.

Response: These are all excellent points. We have retracted our claims regarding power, revised the weighting formula, and included a multivariate application of the technique to the Appendix.

Revision: Supplementary Appendix C

* 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.

Response: The reviewer raises a number of excellent points and a topic that we plan to pursue in a future publication. However, we believe this is beyond the scope of the current manuscript.
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.

Response: These are all excellent suggestions. We have expanded our example on how the weighting technique may be applied in a global rank test, time to event analysis, or a random effects model. We believe the other topics are better suited to a subsequent publication, where they issues can be fully explored.

Revision: Supplementary Appendix C.

* 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.

Response: Based on comments from the reviewers, we have removed our claims regarding the power associated with the composite weighting technique and replaced the Fisher’s Exact Test example for the weighting. We have added the simulated data used for the examples (Supplementary Appendix B) and provided more detail on how the technique can be applied to different types of analysis along with the R Code used (Supplementary Appendix C).

Revision: Supplementary Appendix B and C.

* 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?

Response: We agree with the reviewer that the single illustrative example is limited and that the true test of the technique will be how it performs under various constraints. We plan to continue to apply the technique under various conditions and report its performance in future publications.
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).

Response: We hope that the simple Excel macros supplied in Supplementary Appendix A will make the component weights very easy for researchers to calculate, assuming their study includes those component outcomes. Based on the comments from the reviewer, we have revised the demonstrated applications of the technique to three different types of analysis and included the R code used for the included example to hopefully increase adoption of the technique.

Revision: Supplementary Appendix A and C

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 any 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.

Response: We appreciate the reviewer’s feedback and have modified our conclusion based on the revised weighting formula and specific benefits gained by this approach.

Revision: This composite endpoint technique applies weights to the component outcomes based on orthopaedic trauma patient preferences and can be applied to several types of statistical comparisons to estimate the clinical benefit of a treatment. (Conclusion)

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 >&lt; 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 &lt; 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).

Response: We appreciate the reviewer’s comments regarding the use of p-values, particularly in light of the current debate on null hypothesis significance testing. We have revised the many based on these comments.

Revision: Death was the outcome of greatest importance (mean utility = -8.91), followed by above knee amputation (-7.66), below knee amputation (-6.97), severe pain (-5.90), deep surgical site infection (SSI) (-5.69), bone healing complications (-5.20), and moderate pain (-4.59). Mild pain (-3.30) and superficial SSI (-3.29), on the other hand, were the outcomes of least importance to respondents. (Abstract)

The mean utility for each of the included clinical outcomes was scaled relative to “perfect health” (referenced at zero) (Table 2). Of the ten included clinical outcomes, the greatest importance was associated with death (mean utility = -8.91, 95% CI -9.23 - -8.65), followed by an above knee amputation (AKA) (-7.66, 95% CI -7.83 - -7.48]). Mild pain (-3.30, 95% CI -3.46 - -3.13) and a superficial surgical site infection (-3.29, 95% CI -3.39 to -3.16) were determined to be the outcomes of least importance to the respondents. The was no overlap in the confidence intervals of the clinical outcomes, except for those of superficial surgical site infection and mild pain, where considerable overlap in their utilities was observed.

(Results)

Reviewer 2

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.

Response: We thank the reviewer for their positive comments.

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.
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.

Response: We thank the reviewer for their comments and appreciate confusion regarding the interpretation of the reported utilities. The Bayesian methods are applied to a multinomial logit model. The utilities are only interpretable relative to other utilities derived in the same model.

Based on some of the other comments from the reviewers, we have slightly revised our weighting formula. We had initially intended to have the weights of the included components average to one. However, given other changes to the manuscript, we no longer feel that is necessary and have slightly adjusted our weighting formula. In both versions for the formula, the weight associated with each component outcome is dependent upon which other components included in the composite.

If you had a composite of death and bone healing complications, the weights would be 0.98 and 0.02, respectively. However, if the composite included death and a below knee amputation, the weights would be 0.87 and 0.13, respectively.

We have clarified these changes in the Methods.

Revision: The utility estimates for a specific outcome derived in the model have no direct interpretation, and can only be interpreted relative to another utility estimate in the model.

An orthopaedic trauma composite endpoint weighting technique based on the mean utilities of the component outcomes and a modified version of the conditional logit formula described by McFadden [23] is provided below:

\[
W_a = \frac{(e^{u_b} + e^{u_i})}{(e^{u_a} + e^{u_b} + e^{u_i})}
\]

The weight (W) is calculated separately for each included outcome a where u is the mean utility of each included outcome. b and i note the component outcomes included in the composite. A weight calculator, with sub-group adjustment, is included in the Supplementary Appendix A. (Methods)

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?

Response: Currently, a study could only use the weighting technique if their study included the described component outcomes in their composite endpoint. However, the process for deriving the utilities could easily be replicated in other health conditions. Furthermore, we are
working on expanding the suite of component outcomes to included other psychosocial outcomes common to orthopaedic trauma research.

Revision: This weighting technique could be easily expanded to other outcomes and replicated in other health conditions. However, at present, the application of this weighting technique is limited to studies with component outcomes included in our model.

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.

Response: We agree with the reviewer’s criticisms regarding the application of the technique to FET. In the revised version, we have removed the application of the technique to FET and demonstrated how the weighting technique could be applied to a global rank test, time to event analysis, or a multivariate model.

The primary aim of this paper was to describe the BWS process for eliciting the weights. We believe there are several options to incorporate the variance of the weights into the analysis. However, we believe that more work is required to understand the implications of different approaches under different data conditions and thus, is outside the scope of this paper.

Revision: A hypothetical pilon fracture trial was used to illustrate the application of the proposed weighting technique (Table 4). In this hypothetical trial, 1000 patients are randomized to hypothetical Treatment A (n=498) or Treatment B (n=502). Three components (deep surgical site infection, bone healing complication, and superficial surgical site infection) were included in the hypothetical trial’s primary composite endpoint. The effect of Treatment A versus Treatment B on the composite endpoint was then calculated using several unweight methods, including a Fisher’s Exact Test, time to first event analysis, and a random effects model. For comparison, the treatment effect was also calculated using several methods that accounted for the proposed component weights, including a Wilcoxon Rank Sums test, time to event allowing for weighted repeated events, and a random effects model that accounted for component weights [32]. The effect size for the random effects models are reported as odds ratios, and hazard ratios are used for the time to event models [33]. The Probability Index was used to report the treatment effect for the Wilcoxon Rank Sums test [32-34]. These analyses were performed using R Version 3.6.1 (Vienna, Austria). All of the data and code for the models are included in Supplementary Appendix B and C. However, for simplicity, only the unweighted and weighted time to event analysis are reported in the results section.

(Methods)

For the hypothetical pilon fracture trial, the results with the unweighted composite endpoint using a time to first event analysis would have determined that there was no difference between the two treatments (hazard ratio (HR): 1.02, 95% CI 0.83 – 1.27, P = 0.83) (Figure 3). When weights are applied to the included component outcomes, and the analysis allows
for patients to have more than one event, Treatment A is superior (HR: 0.72, 95% CI 0.57 – 0.90, P < 0.01). A similar difference in effect size was observed when the data were analyzed using unweighted and weighted random effects models (Supplementary Appendix C). However, the treatment effect was not statistically significant when the weights were applied using a global rank approach, and treatment groups were compared using a Wilcoxon Rank Sums test and Probability Index Model. (Results)

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.

Response: We have rewritten the manuscript to ensure that the word “relative” is only used when appropriate.

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 …"

Response: We have revised as suggested.

Revision: A commonly used definition of a composite endpoint in clinical research is the occurrence of any one of several study events of interest [1]. (Introduction)

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.

Response: We thank the reviewer for this important comment and have revised the sentence accordingly.

Revision: Incorporating multiple endpoints into a single metric increases the number of observed events, can avoid issues pertaining to multiplicity, and thus, may increase statistical power [1,2,3]. (Introduction)

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

Response: Thank you for this comment. We have revised the sentence.

Revision: Composite endpoints also enable the inclusion of rare, but clinically important, outcomes; therefore, providing a broader interpretation of the net clinical benefit of a treatment [1]. (Introduction)

Line 77. I would say, "Composite endpoints have limitations" or something like that.
Response: Revised as suggested.

Revision: Composite endpoints have several limitations [4-7]. (Introduction)

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

Response: Revised as suggested.

Revision: Additionally, in studies that analyze composite endpoints using a traditional time to first event analysis or other analyses of frequency that only consider the first event, each study participant can have only one event; therefore, censoring subsequent events biases treatment effects to earlier outcomes. (Introduction)

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

Response: We thank you for bringing that to our attention. While we were unable to find an example from Armstrong and Westerhout, we were able to find two other examples of studies that applied patient preferences to the composite weights. This includes a recent publication from the Netherlands that used a Best-Worst Scaling experiment for adverse outcomes of revascularization procedures. The range of the derived weights were similar in scale to our study.

Revision: However, weighting methods which incorporated patient values specific to the target patient population are lacking [15,16]. (Introduction)

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

Response: The sentence has been revised.

Revision: In a Best-Worst Scaling experiment, respondents are presented with a set of three or more attribute levels and then asked to select the best and worst attribute level in each choice set. (Methods)

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.

Response: We have added that information to the paragraph, as suggested.

Revision: The study was performed at a single Level-1 trauma center in Baltimore and followed the International Society for Pharmacoeconomics and Outcomes Research conjoint analysis practice guidelines [21]. (Methods)

Line 119. Delete comma after "levels"

Response: Revised as suggested.
Revision: Finally, semi-structured interviews were conducted with three orthopaedic trauma patients for additional perspective on plausible clinical outcomes. Information gathered from this work informed the final selection of the included attributes and levels deemed most important by our patient and clinician stakeholders. (Methods)

Line 124. Not "minimized" - minimum burden would be no questionnaire. Say something like "reduce" or "limit"

Response: Revised as suggested.

Revision: The respondent burden was reduced using a blocked, balanced, fractional factorial design, based on optimal D-efficiency. (Methods)

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

Response: D-efficiency is a measure of design efficiency and is the relative number of runs (expressed as a percent) required by a hypothetical orthogonal design to achieve the same determinant value on a scale of 0 (one or more parameters cannot be estimated) to 100 (balanced and orthogonal design). This is a common measure to evaluate the choice set design in a conjoint analysis (best-worst scaling or discrete choice experiment). We have added a reference to the text.


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

Response: We have revised the sentence to clarify the sequence.

Revision: Prior to completing the Best-Worst Scaling questionnaire, respondents answered several demographic questions and indicated which orthopaedic complications they had experienced during their post-operative clinical course. (Methods)

Lines 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?

Response: If one assumes the order of the attributes in a choice may create bias, there are 1000 possible combinations of the 10 attributes into choice sets of three. The subset of 40 choice sets were selected using the D-efficiency process described above, whereas the design aims to get as close as possible to an orthogonal design to minimize any design bias for the included attributes.

Line 145. Delete comma after "older"
Response: Revised as suggested.

Revision: The Best-Worst Scaling questionnaire was administered to English-speaking patients, 18 years of age or older with a surgically treated appendicular fracture from November 2017 to March 2018 (Methods)

Line 148. Change "patients" to "they"

Response: Revised as suggested.

Revision: Patients were enrolled in the study at an outpatient follow-up appointment, at which time they provided written informed consent and completed the written questionnaires. (Methods)

Line 150. Delete "the" before "adequate"

Response: Revised as suggested.

Revision: To ensure adequate statistical power for an a priori defined sub-group analysis by injury location, study participants were purposely sampled to ensure at least 50 participants with each of the following fractures: hand/wrist; upper extremity (proximal to distal ¼ radius/ulna); hip (pelvis, acetabulum, femoral neck, and greater/lesser trochanter), tibia/femur (distal to lesser trochanter and proximal to ankle fractures), and foot/ankle. (Methods)

Line 165. Define what hierarchical Bayes model you used.

Response and Revision: The respondent-level covariates are estimated based on the algorithm described by Train, which incorporates Adaptive Bayes and Metropolis-Hastings approaches [28]. The likelihood function for the utility parameters for a given respondent is based on a model for each subject’s preference within a choice set, given the attributes in the choice set [29]. (Methods)

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

Response: Revised as suggested.

Revision: We set the mean utility at zero for perfect health; all other possible outcomes are then presented as negative utilities. (Methods)

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

Response: Revised as suggested.
Revision: To test heterogeneity in respondents’ utility for each included clinical outcome, ten demographic and injury-specific covariates were independently tested as interaction terms in the primary model. (Methods)

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

Response: Revised as suggested.

Revision: To adjust for ten statistical tests, we set the level of significance for the interaction terms at $\alpha=0.05/10=0.005$. [Methods]

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

Response: Reference added.

Revision:


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

Response: Revised as suggested.

Revision: To determine if experiencing a clinical outcome is associated with a different utility for that outcome, we stratified respondents by those who had and had not experienced the outcome. (Methods)

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

Response: Revised as suggested.

Revision: Nearly half (47.5%) of respondents had a tibia or femur fracture below the lesser trochanter. (Methods)

Line 217. "was" should be "were"

Response: Revised as suggested.
Statistically significant interactions based on age, race, education level, income level, and health insurance status were observed. (Results)

Lines 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 from -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.

Response: We agree that some of these comparisons are more clinically important than others. However, to avoid any perception of selective reporting, we have included all comparisons that were less than the p=0.05 threshold.

Line 245. Suggest, "trial illustrates"

Response: Revised as suggested.

Revision: A hypothetical pilon fracture trial illustrates the application of the proposed weighting technique (Table 1). (Methods)

Line 285. See comment above about rare events.

Response: We appreciate the reviewer’s comments on rare events and statistical power, and have modified the sentence accordingly.

Revision: For the orthopaedic community, the technique provides a set of ten common clinical outcomes researchers may incorporate into future composites endpoints. (Discussion)

Line 287. Add "or" after fracture" and change "minimal" to "only small"

Response: Revised as suggested.

Revision: The limited heterogeneity in observed preferences suggests a common value gradient for clinical outcomes that is not altered by the type of fracture, or the time since injury, and only a small variation based on outcomes experienced. (Discussion)

Lines 293 - 295. See comments above about power.

Response: We have removed that sentence based on the reviewer’s comments.

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

Response: Based on this and previous comments, we have revised the conclusion.
Revision: This composite endpoint technique applies weights to the component outcomes based on orthopaedic trauma patient preferences and can be applied to several types of statistical comparisons to estimate the clinical benefit of a treatment. (Conclusion)

Line 333. Change "that" to "who"

Response: Revised as suggested.

Revision: The authors would like to thank the study participants who participated in the survey. (Acknowledgments)

Table 1. Nice to see these data.

Response: Thank you.

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?

Response: The mean utility for perfect health is set to zero, but there is still variance around the mean estimate. We agree that the log likelihood reported in the Table is not interpretable, and we have removed it from the Table.

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

Response: We appreciate the reviewer’s concern that many of these differences will have negligible impact when applied to the weighting of a composite used to compare treatments. However, as previously mentioned, depending on the components that are included in the composite, accounting for the variation in utilities based on the characteristics of the sample provides some additional precision in the weights.

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

Response: We appreciate that this information may not be of interest to the majority of BMC MRM readers. However, for researchers that would like to utilize the weighting technique in their analysis, we believe it is important to have a clear definition for the attribute levels.

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.

Response: Death was included in 1190 analyzed triplets and was selected as the worst of the three options in 1038 of those triplets (87%). In 141 of 1190 triplets, death was selected as neither the best or worst option. For the vast majority of those triplets, severe pain or an
amputation was selected as worse than death. In 5 of the 1190 triplets, death was selected as the best of the three options. We suspect this to be an error on the part of the participant. However, removing these responses does not qualitatively change the findings.