PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study (GRADE-THRESHOLD)
AUTHORS	Morgano, Gian Paolo; Mbuagbaw, Lawrence; Santesso, Nancy; Xie, Feng; Brozek, Jan; Siebert, Uwe; Bognanni, Antonio; Wiercioch, Wojtek; Piggott, Thomas; Darzi, Andrea; Akl, Elie; Verstijnen, Ilse; Parmelli, Elena; Saz-Parkinson, Zuleika; Alonso-Coello, Pablo; Schünemann, Holger J

VERSION 1 – REVIEW

REVIEWER	Raittio, Eero University of Eastern Finland
REVIEW RETURNED	17-Aug-2021

GENERAL COMMENTS	This protocol deals with an important topic and tests a method to outline thresholds behind communication and consideration of effect-sizes in the GRADE Evidence to Decision framework. The task is very challenging, but study is well planned and would provide valuable insights on this topic. I have some comments which I hope the authors could consider or discuss them under limitations. 1) Authors could shortly introduce theoretical background of division of effect-sizes to 'Trivial or None', 'Small', 'Moderate' and 'Large'. Why this is better than some other forms, or communication with numbers only? 2) If I understood correctly, participants evaluate benefits and harms separately from each other. In my view, benefits or harms related to a treatment are one of the most important determinants of deciding a threshold for benefits or harms related to treatment, and thus I am not convinced how a person can evaluate benefits and harms of a intervention separately. It is even a prerequisite in some situations that the individual has not died (a harm) in order to get some benefits (avoid stroke, example), or vice versa. It is a fundamental assumption that we could evaluate benefits and harms separately and then consider their thresholds and wordings together in communication and decision-making related to effect-sizes in the GRADE Evidence to Decision framework. I think reasoning behind this needs clarification. 3) Death, major ischemic stroke, pulmonary embolism of moderate severity, diarrhea of moderate severity, and mild nausea/vomiting

	were the outcomes in the case-scenarios. How duration of the beneficial or harmful effects were considered in the cases? To my knowledge, death is a permanent phenomenon, but what about others? It is important whether you have nausea forever or just few days. 4) Figure 1 case-scenario example: I do not understand the sentences "This outcome..." and "In other words..." Has the hypothetical individual utility of 1 at the time of intervention (or if outcome would not appear)? Is this a reasonable assumption? What utility levels were in other case-scenarios? Thresholds respondents would give may be sensitive to these values. How outcome values were chosen? 5) It is great that authors randomize for the order effects. I would also consider (in future) also randomize assessors to different certainty of evidence categories. In my opinion, it possible that the certainty of evidence affects on thresholds a reader gives. We are prone to assume that a high certainty of evidence corresponds a large effect. Likewise, clinical and statistical significance in research setting. As many studies have shown, average clinicians are not very well aware of effect-sizes or risks related to health risks and interventions, thus even indirect hints/cues/nudge towards words such as "high" or "low" may affect responses about thresholds, which may affect generalizability of findings to other settings. 6) For clarity, I suggest authors provide complete set of case-scenarios for the reasons I have noted above. Or provide summary of factors which are different between the scenarios and which are not.

REVIEWER	Abraham, Ivo Matrix45 / Univ of Arizona
REVIEW RETURNED	11-Oct-2021

GENERAL COMMENTS	It would be helpful if you could address the following: What is the status of the study? Has it started; if not when will it launch? If it has started, what was the start date? For either scenario, what is the anticipated close-of-trial date? Why not randomize the case scenarios, if need be in stratified fashion, as a means of pre-empting order effects, as opposed to testing for order effects post facto with ANOVA. P15 L9: I don't see how a paired t-test will enable you to test if the DTs are different. Are you planning on pairwise comparisons only? How will you address multiplicity of your statistical significance testing.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

This protocol deals with an important topic and tests a method to outline thresholds behind communication and consideration of effect-sizes in the GRADE Evidence to Decision framework. The task is very challenging, but study is well planned and would provide valuable insights on this topic. I have some comments which I hope the authors could consider or discuss them under limitations.

Comment 1.1:
Authors could shortly introduce theoretical background of division of effect-sizes to 'Trivial or None', 'Small', 'Moderate' and 'Large'. Why this is better than some other forms, or communication with numbers only?

Reply 1.1:
We thank you for this comment. While we agree that alternative approaches for categorization and/or communication are possible, our study builds on the extensive research conducted to develop the GRADE EtD frameworks and its wide use by organizations producing guidelines. The judgements have also been integrated in a newer version of it, the WHO Integrate framework. This work focuses on use of the EtDs and the theoretical background on the rationale is addressed elsewhere. So, that the GRADE judgments will be use is a given but this study focuses on how to use them best, not if they should be used. We have provided ample background for this rationale in the introduction. We clarified this in the revised manuscript to say:

“The guidance from the GRADE Working Group includes expressing and facilitating these judgments by assigning the health benefits or health harms of some intervention under evaluation to one of the following four categories: 'Trivial or None', 'Small', 'Moderate' and 'Large'.3,4 To be useful, however, this simplification requires that EtD users have a similar understanding of what magnitude of health benefits or health harms belong into which category and are consistent in their judgments. A similar common understanding is also important between those assigning a category and those interpreting the meaning of a category that is communicated to them (i.e. “imagining” how substantial is an effect based on the category).” ….

Comment 1.2:
If I understood correctly, participants evaluate benefits and harms separately from each other. In my view, benefits or harms related to a treatment are one of the most important determinants of deciding a threshold for benefits or harms related to treatment, and thus I am not convinced how a person can evaluate benefits and harms of a intervention separately. It is even a prerequisite in some situations that the individual has not died (a harm) in order to get some benefits (avoid stroke, example), or vice versa. It is a fundamental assumption that we could evaluate benefits and harms separately and then consider their thresholds and wordings together in communication and decision-making related to effect-sizes in the GRADE Evidence to Decision framework. I think reasoning behind this needs clarification.

Reply 1.2:
We thank you for your comment. The study does not evaluate benefits and harms separately. The study focuses, like in many studies of utilities on individual outcomes and includes those that are benefits and those that are harms. Furthermore, whether an outcome is a benefit or harm depends on the direction of the effect (increase in mortality = harm; decrease in mortality = benefit) for nearly all outcomes. Judging the magnitude of these benefits and harms separately is critical in the GRADE

Evidence to Decision framework. Then, decision makers judge them together by balancing the benefits and harms. By explicitly marking separately the judgment deemed appropriate for the desirable and the undesirable effects, decision-makers pave their way towards a consistent and well-informed decision about the balance between health benefits and harms. Decision-thresholds would assist this approach in terms of consistency, mitigating the possibility that effects having similar impact could be judged differently, and allowing to differentiate between effects that could be receiving the same judgment but that are unavoidably not equal. The feasibility has been demonstrated in the use of the EtDs in 1000's of recommendations (a large number of guidelines with many recommendations where this approach has been used). Furthermore, although not cited here because until now anecdotal, our user testing with various groups shows it also works with the decision-thresholds but the latter anecdotal evidence is the reason for the trial.

Comment 1.3:
Death, major ischemic stroke, pulmonary embolism of moderate severity, diarrhea of moderate severity, and mild nausea/vomiting were the outcomes in the case-scenarios. How duration of the beneficial or harmful effects were considered in the cases? To my knowledge, death is a permanent phenomenon, but what about others? It is important whether you have nausea forever or just few days.

Reply 1.3:
We thank you for this comment. We agree that duration is a key aspect. We use health outcome descriptors to provide survey participants with a structured description of the outcomes' key attributes (Symptoms, Time Horizon, Treatment, Consequences). Our survey describes the outcome as a point in time but provides information about consequences that may occur in the future. So, the utilities are point in time utilities.

Following your comment:

a) we have modified page 9 of the manuscript as follows:
"Each case-scenario will include …. …. a Health Outcome Descriptor 15 describing key attributes of the outcome under consideration including symptoms, time horizon, testing and treatment, and consequences; (3)…."

Comment 1.4:
Figure 1 case-scenario example: I do not understand the sentences "This outcome..." and "In other words..." Has the hypothetical individual utility of 1 at the time of intervention (or if outcome would not appear)? Is this a reasonable assumption? What utility levels were in other case-scenarios? Thresholds respondents would give may be sensitive to these values. How outcome values were chosen?

Reply 1.4:
We thank you for the comment. To minimize the bias that could stem from having respondents assign different weights to the same outcome, we provide survey participants with outcome values. As described in the survey, the outcome value is expressed on a scale from 0 (being dead) and 1 (perfect health) and is a measure of how much people value an outcome in comparison to other outcomes. Therefore, outcome values concern the health state (as described through the corresponding health outcome descriptor) associated with the outcome occurring. The values presented in the survey are: 0 for Death, 0.14 for major Ischemic Stroke, 0.42 for Pulmonary

Embolism of moderate severity, 0.90 for Diarrhea of moderate severity, and 0.95 for mild Nausea and/or Vomiting. These outcomes were selected to achieve a broad representation of values in the spectrum between 0 and 1. Since matching values were not available from the literature, approximate values were defined by study authors based on considerations on outcomes. This has no impact on the validity of the findings because the key factors are “a” fixed utility to derive the thresholds. They just need to be realistic enough and have face validity.

Comment 1.5:
It is great that authors randomize for the order effects. I would also consider (in future) also randomize assessors to different certainty of evidence categories. In my opinion, it possible that the certainty of evidence affects on thresholds a reader gives. We are prone to assume that a high certainty of evidence corresponds a large effect. Likewise, clinical and statistical significance in research setting. As many studies have shown, average clinicians are not very well aware of effect-sizes or risks related to health risks and interventions, thus even indirect hints/cues/nudge towards words such as "high" or "low" may affect responses about thresholds, which may affect generalizability of findings to other settings.

Reply1.5:
Thank you for your comment. We agree that the certainty of the evidence of effects could influence raters and introduce a bias in their judgments about how substantial anticipated effects are. The GRADE EtD frameworks aim at disentangling these different aspects (size of effects and CoE of effects) through judgments on separate criteria. To avoid uncertainty around the point estimate, in our survey we assigned the rating of HIGH CoE of effects to all case-scenarios and did not provide any 95% confidence interval. We agree, in future steps of this research we may consider assessing the impact of different CoE ratings on the thresholds.

Comment 1.6:
For clarity, I suggest authors provide complete set of case-scenarios for the reasons I have noted above. Or provide summary of factors which are different between the scenarios and which are not.

Reply 1.6:
We thank you for your comment. We believe that we should avoid this for the time being but will consider publishing them with the final study report (the reason is that we would not want respondents to see them before they participate in the trial or make them public before and we may want to use them in the future). We have however provided supplement 2 that describes the scenarios hypothetically and the related survey. We also added: These scenarios differ in the description of the severity of the outcome and the consequences to represent clearly different values.

Reviewer: 2

Comment 2.1:
What is the status of the study? Has it started; if not when will it launch? If it has started, what was the start date? For either scenario, what is the anticipated close-of-trial date?

Reply 2.1:
We thank you for your comment. Please note that we started the recruitment on June 9th, 2020 after completing the protocol (which we submitted shortly thereafter). We will continue recruitment until the sample size will be met or until December 31, 2022 (this is arbitrary but we feel that if we have not completed recruitment for this particular trial by then, it will be unlikely that we will complete sample size requirements through further efforts). There are other minor changes to the protocol (apart from the analytical suggestions made by reviewer 2) which include explorative subgroup analysis by language used for the survey (e.g. Spanish versus English), that we will use results to inform guideline panels in the interim but this will have no impact on stopping the trial or drawing final conclusions and that we will recruit participants through our work with guideline producers.

The changes are described as follows in the protocol:
We will continue recruitment for this trial until reaching our anticipated sample size (see below) or until December 31, 2022 as it is unlikely that we will meet the sample size through additional recruitment efforts beyond then.

Furthermore, we will use periodic interim analyses to inform judgments by guideline groups that develop recommendations but will not use these to draw conclusions about the trial until it is stopped formally.

Comment 2.2:
Why not randomize the case scenarios, if need be in stratified fashion, as a means of pre-empting order effects, as opposed to testing for order effects post facto with ANOVA.

Reply 2.2:
We thank you for your comment. We decided not to stratify randomization of case-scenarios by order of outcomes because it would have resulted in a large number of combinations (5! N=120). We also could not stratify by judgment given the design of the surveymonkey, but will certainly consider this in follow up validation studies. Thank you for this suggestion!

Comment 2.3:
P15 L9: I don't see how a paired t-test will enable you to test if the DTs are different. Are you planning on pairwise comparisons only?

2.3 Reply:
We thank your comment. We agree that if we are comparing T1 vs T2 vs T3 (within the same comparison) we must do ANOVA first, and then a post-hoc paired t-test.
Following your comment:

a) we have modified page 13 of the manuscript as follows:

We will conduct an ANOVA to determine if there are any differences between the thresholds (T1□T2□T3). If we identify a difference, since each participant will contribute data to each threshold, we will employ a post-hoc paired sample t-test to assess if the which of the DTs are different i.e., (T1□T2; T2□T3; T1□ T3 T1□ T2□T3).

Comment 2.4:
How will you address multiplicity of your statistical significance testing.

Reply 2.4:

We thank you for your comment. We will use the Bonferroni correction and use a stricter p-value. For 5 secondary outcomes we will divide 0.05/5 and declare significance at 0.01.

Following your comment, we have added the following sentence (page 15):

We will use the Bonferroni correction for multiple testing in all secondary analyses
(Abdi H. Holm's sequential Bonferroni procedure. Encyclopedia of research design 2010, 1(8):1-8.)

VERSION 2 – REVIEW

REVIEWER	Raittio, Eero University of Eastern Finland
REVIEW RETURNED	29-Dec-2021

GENERAL COMMENTS	I am quite happy with authors's responses and revision. I would still suggest authors' state the most important limitations of the trial in the discussion. In my opinion, there is so many moving parts in the case-scenarios, that as an author, I would be worried about the generalizability of the findings to other settings with different effect-sizes, certainty of evidence, outcomes, exposures, wordings, utilities, populations, time frames etc.. For instance related to the authors' response to my comment 1.4 in the previous round, I guess there will be variation in responses due to how participants feel the health utility at time zero in the case, which may be hard to move to other settings.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

I am quite happy with authors's responses and revision.

Comment 1.1:

I would still suggest authors' state the most important limitations of the trial in the discussion. In my opinion, there is so many moving parts in the case-scenarios, that as an author, I would be worried about the generalizability of the findings to other settings with different effect-sizes, certainty of evidence, outcomes, exposures, wordings, utilities, populations, time frames etc...

For instance, related to the authors' response to my comment 1.4 in the previous round, I guess there will be variation in responses due to how participants feel the health utility at time zero in the case, which may be hard to move to other settings.

Reply 1.1:

We have now added the following paragraph under discussion:

However, our trial will not be free of limitations. Generalizability of the findings may be limited by the use of the case scenarios we chose and the limited number of effect sizes we include in the trial. Generalizability may also be limited by the type of participants we will be able to recruit. Therefore, we plan, following the completion of this trial, to conduct further research with additional case scenarios and different target populations.