BMJ Open

Development and user testing of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-056400
Article Type:	Original research
Date Submitted by the Author:	22-Aug-2021
Complete List of Authors:	Phillips, Mark ; McMaster University, Health Research Methods, Evidence, and Impact Sadeghirad, Behnam; McMaster University, Health Research Methods, Evidence, and Impact; McMaster University, Anesthesia Busse, Jason; McMaster University, Anesthesia; McMaster University, Health Research Methods, Evidence, and Impact Brignardello-Petersen, Romina ; McMaster University, Health Research Methods, Evidence, and Impact (HEI) Cuello-Garcia, Carlos ; McMaster University, Health Research Methods, Evidence, and Impact Kenji Nampo, Fernando ; Federal University of Latin-American Integration Guo, Yu Jia; McMaster University, Health Sciences Bannuru, Raveendhara R.; Tufts Medical Center, Center for Treatment Comparison and Integrative Analysis Bzovsky, Sofia; McMaster University, Department of Surgery - Division of Orthopaedics Thabane, Lehana; McMaster University, Health Research Methods, Evidence, and Impact; St. Joseph's Healthcare, Biostatistics Unit Bhandari, Mohit; Mcmaster University, Division of Orthopaedic Surgery; McMaster University, Health Research Methods, Evidence, and Impact Guyatt, Gordon; McMaster University, Health Research Methods, Evidence, and Impact
Keywords:	QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS, EDUCATION & TRAINING (see Medical Education & Training)

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Development and user testing of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Mark R Phillips[1], Behnam Sadeghirad[1,2], Jason W Busse[1,2], Romina Brignardello-Petersen[1], Carlos Cuello[1], Fernando Kenji Nampo[3], Yu Jia Guo, Sofia Bzovsky[5], Raveendhara R Bannuru[6], Lehana Thabane[1,7], Mohit Bhandari[1,5], Gordon H Guyatt[1]

1. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
2. Department of Anesthesia, McMaster University, Hamilton, ON, Canada
3. Latin-American Institute of Life and Nature Sciences/Evidence-Based Public Health Research Group, Federal University of Latin-American Integration, Foz do Iguassu, Brazil
4. Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
5. Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada
6. Center for Treatment Comparison and Integrative Analysis, Tufts Medical Center, Boston, MA, USA
7. Biostatistics Unit, St Joseph's Healthcare, Hamilton, ON, Canada

Correspondence: Mark Phillips phillimr@mcmaster.ca

1280 Main Street West, Hamilton, ON, CAN L8S 4L8

Keywords: Network Meta-Analysis, User-testing, Qualitative Research, Statistics and Research Methods

Word Count: 3589

Abstract

Objective The GRADE working group recently developed an innovative approach to interpreting results from network meta-analyses (NMA) through minimally and partially contextualized methods; however, the optimal method for presenting results for multiple outcomes using this approach remains uncertain. We therefore developed and iteratively modified a presentation method that effectively summarizes NMA results of multiple outcomes for clinicians using this new interpretation approach.

Design Qualitative descriptive study

Setting A steering group of 7 individuals with experience in NMA and user testing studies developed two colour-coded presentation formats for evaluation. Through an iterative process, we user-tested both formats to maximize their clarity and ease of interpretation.

Participants 26 participants including 20 clinicians who routinely provide patient care, 3 research staff/research methodologists, and 3 residents.

Main Outcome Measures Two team members used qualitative content analysis to independently analyze transcripts of all interviews. The steering group reviewed the analyses and responded with serial modifications of the presentation format.

Results To ensure that readers could easily discern the benefits and safety of each included treatment across all assessed outcomes, participants primarily focused on simple information presentations, with intuitive organizational decisions and colour coding. Feedback ultimately resulted in two presentation versions, each preferred by a substantial group of participants, and development of a legend to facilitate interpretation.

Conclusion Iterative user testing facilitated the development of two novel formats for presenting minimally or partially contextualized NMA results for multiple outcomes. These presentation approaches appeal to audiences that include clinicians with limited familiarity with NMAs.

Strengths and Limitations of this Study

- Extensive user-testing in a targeted audience has validated the NMA presentation approaches within this study; something that has not been done for other presentation formats
- Structured qualitative research methodology has ensured accurate use of user feedback to develop and refine the NMA presentation formats
- Limited by the omission of some information within the presentation formats in order to achieve simplicity and interpretability, such as greater detail for individual outcomes, absolute effects, or specifics about the certainty of evidence assessments.
- The aforementioned information should still be included in NMA manuscripts, but cannot be feasibly fit within the presentation formats.

Introduction

Network meta-analysis (NMA) provides an increasingly popular approach to evidence synthesis that allows comparison between multiple competing treatment options within a single analysis.[1,2] Although NMA is an important tool for clinicians, patients, and other stakeholders, results involve multiple treatments and outcomes, and as a result are complex and difficult to interpret.[3]

Common methods for presenting NMA results include the use of forest plots, league tables, and surface under the cumulative ranking curve (SUCRA).[1,4] The key limitation with these options is that they can only provide results of a single outcome.[5] NMAs often compare multiple benefit and harm outcomes, resulting in challenges for NMA authors seeking to avoid presentation methods that are onerous for clinicians to review and challenging for them to understand.[6]

There are a number of novel approaches that have been suggested for presenting NMA results for multiple outcomes[7,8]; however, these approaches lack key information, present challenges to interpretation, and have not undergone user testing with their target audiences. While some previously suggested approaches have merit for a limited number of outcomes,[4,6,9–12] although not all taking certainty of evidence into account, they have serious limitations for simultaneous presentation of multiple outcomes.

Recently, the GRADE working group has suggested two variations on a new methodology that places interventions in categories from best to worst considering the estimates of effect and certainty of the evidence for each comparison.[13,14] We therefore developed interpretable presentation approaches for NMAs with multiple outcomes that builds on GRADE guidance and

effectively summarizes results for clinicians and other relevant audiences.

Methods

Study Design

A 7-member steering committee (MRP, BS, JWB, RB-P, CC, FKN, GHG) oversaw study design and implementation. The committee generated two initial presentation formats and chose a combination of large group sessions and individual user testing interviews to inform iterative modifications of the two initial formats. The presentation format consisted of treatment options in rows and outcomes in columns, with colour-coded shading of cells to identify the magnitude and certainty of the treatment effect in relation to the reference treatment.

Initial large group testing with two groups of methodologists, graduate students in health research-focussed programs, and statisticians, as well as presentation at a national conference (2019 Canadian Pain Society annual scientific meeting), provided the foundational feedback for modifications of the initial presentation versions. The steering committee reviewed input from four rounds of user-testing, iteratively modifying the formats after each round and presenting updated options of the presentation versions to subsequent participants.

For the user testing interviews, the committee chose a qualitative descriptive study approach that focuses on creating a close description of the information that participants provide.[17] This is ideal for user testing that, without interpretive direction, aims to optimize the understandability of a tool within the target population.[17] This study involves human participants but an Ethics Committee exempted this study. After reviewing the protocol, the Hamilton Integrated Research Ethics Board (HiREB) committee and chair, judging the study to be a quality

improvement investigation within the methodology and knowledge translation field, provided an exemption from formal ethics approval. We followed, when applicable, the consolidated criteria for reporting qualitative research (COREQ) checklist in reporting our findings.[16]

Sampling and Recruitment

Target users for this study included academic and non-academic clinicians, research staff/research methodologists, and residents. The steering committee, through their professional contacts, provided a pool of initial possible participants that the principal investigator supplemented using snowball sampling technique. [17] Specifically, we asked individuals who agreed to participate for contact information of any colleagues whom we could approach to interview. Prior to their interviews, each participant received information outlining the purpose of the study. Study recruitment ceased when data collection reached redundancy – the point at which there were no further refinements requested to improve the interpretability of the presentation formats.[17]

Data Collection

The principal investigator (MRP) conducted all user testing interviews either in-person or through video teleconferencing. Interviews followed a flexible interview guide (**Appendix A**) to leave the conversation open for participants to explore any topics they felt were relevant and important.[15] Throughout the study, the principal investigator iteratively updated the interview guide to explore areas of importance that emerged. Interviews began with a brief introduction to NMA methods, followed by questions regarding the participant's familiarity and experience with NMA.

Participants then viewed the current versions of the NMA presentation formats and provided feedback. YJG or MRP transcribed all interviews verbatim. Transcripts were not returned to participants and interviewers did not conduct follow up interviews. The steering committee incorporated all feedback to arrive at two final presentation versions.

Patient and Public Involvement

Not applicable; No patients involved.

NMA for User Testing

The steering committee developed five core criteria to which the example NMA must adhere: (1) variability in quality of evidence (2) variability in magnitudes of effect; (3) assessment of both benefits and harms; (4) inclusion of both continuous and binary outcomes; and (5) including at least 5 outcomes and 5 interventions. Based on these criteria the steering committee chose, for user testing, a recent NMA that used a minimally contextualized approach to address acute pain management in patients experiencing non-low back acute musculoskeletal injuries.[18]

Based on the GRADE approach[13] this NMA categorized, for each benefit outcome, interventions as among those with the largest benefit, those with intermediate benefit, and those with the least benefit. For each harm outcome, they categorized interventions as among the least harmful, intermediate harm, and the most harmful. They then categorized interventions as those for which there was high or moderate certainty evidence, and those for which there was low or very low-quality evidence.[18] These results provided the example for user testing.

Data Analysis

Two reviewers (MP and SB) independently conducted data analysis, in duplicate, using a qualitative content analysis approach.[6] The study team recruited participants, collected data, and conducted data analysis in parallel. As new data became available, the reviewers coded and grouped similar phrases, patterns, and themes.[15] The steering committee met four times over a period of 14 months to review the collected data and made iterative changes to the presentation formats as dictated by feedback, initially from large group presentations and subsequently from user testing. When analysis of the data provided actionable feedback, the reviewers presented their findings to the steering committee who ranked feedback as a "large change required", "moderate change required", or "minor change required" and then revised the presentation format(s) accordingly.

Subsequent participants provided input on the modified versions of the NMA results presentations. Participants commented regarding their interpretation of the data within the presentation format; the team considered study objectives met once participants consistently reported a clear interpretation of the results with no or minimal suggested modifications. Reviewers documented all changes to the presentation format in a study audit trail.[19,20] Reviewers conducted all qualitative analysis using RQDA software (R version 3.5.0).

Results

Study Sample

Two focus groups, both of which included methodologists, graduate students, and statisticians, participated in the initial large group testing: the first, a critical care guideline development group

(GUIDE: https://guidecanada.org/) many of whose members have NMA expertise (65 attendees); the second, a research group (CLARITY: http://www.clarityresearch.ca/) who meet regularly at McMaster University to discuss current methodological and statistical topics (20 attendees).

The user testing portion of this study included 26 participants of mean (standard deviation [SD]) age of 47.6 (13.9) years, 20 of whom were clinicians whose primary activity involved direct patient care (77%); 3 research staff/research methodologists (12%); and 3 residents (12%). Typical participants were male (73%) physicians in clinical practice for almost two decades (mean [SD]: 19.5 [14.3] years) with no prior involvement with conducting an NMA (58%) (**Table 1**).

Content Analysis Themes

Main themes that arose from the content analysis conducted on interview transcripts of participant interviews included "organizational", "language/terminology", "included information" and "colour options". Respondents also provided feedback regarding necessary details to include in the presentations' footnote. The following sections provide details regarding the most important feedback and how this feedback informed choices regarding presentation format. The fourth round of user testing resulted in minimal new information, resulting in two presentation versions that participants deemed satisfactory.

Final Presentation Versions

Ultimately, respondents proved equally enthusiastic about two options; the steering group, therefore, chose to offer both as alternative presentations. **Figure 1** summarizes the development process from conceptualization to the final presentation versions. We will refer to the presentation in **Figure 2** as the "colour gradient" version and the presentation in **Figure 3** as

the “stoplight” version. Each presentation has a legend and footnote with pertinent information that the user-testing process demonstrated necessary to include.

Figure Organization

User-testing identified a number of key components that aid in interpreting presentation formats. Within the organizational theme, the use of a bolded vertical line to separate benefit and adverse event outcomes, as well as the header and results data (horizontal), proved desirable. Regarding the ordering of interventions from top to bottom in the rows, participants preferred ordering treatment options at the top with high/moderate certainty evidence of maximal benefit and minimal harm to those with high/moderate certainty evidence of minimal or no benefits and significant harms placed in the bottom rows. Respondents provided mixed feedback regarding the organization of the presentation within the middle section, with no consistent guidance that could be applied across all NMAs. This leaves the optimal ordering within the middle rows that include treatments that have low/very low certainty evidence, treatments with high/moderate certainty evidence of intermediate effects, and treatments with trade-offs between both large benefits and large harms, uncertain (or perhaps there is no single optimal ordering). **Figure 4** provides an overview of guidance regarding intervention order within the rows.

Presentation Terminology

Respondents indicated that the presentation should clearly and succinctly label outcomes with specification of the measure of treatment effect (e.g. odds ratios, mean differences) and that the

header of each column should include these labels. Participants had no strong preference regarding the terminology of "benefit" and "adverse events" outcome categories; options discussed included "effectiveness/efficacy outcomes" and "harms outcomes". Whatever option investigators choose, the terminology should remain consistent across the presentation, legend, and manuscript text.

Presentation Included Information

Participants considered the magnitude of treatment effect, confidence/credible intervals, certainty of evidence, and statistical significance to be the four important elements that should be included in each comparison cell. Possibilities explicitly discussed but rejected included sample size, patient characteristics, and heterogeneity/incoherence estimates. Respondents considered these items as important elements of the NMA, but felt they would be better suited within another section of the manuscript rather than within this summary presentation.

Footnote Included Information

Participants felt that footnotes should include: an indication of a dash representing no available evidence (- *: no evidence*); designation of the reference group (e.g. *Reference Group: Placebo*); and labelling of how statistical significance within the presentation is identified (i.e. *Bold = statistically significant, p<0.05*); as well as all abbreviations used within the presentation.

Legend Organization

Participants felt that benefit outcomes should be located in the left columns, with a bold vertical line separating the benefit and adverse event outcomes within the legend – similar to the structure of the main presentation. They also suggested a bold horizontal line separating the header from the legend in a similar format as within the main presentation. Within the benefit and adverse event sections, respondents preferred that high/moderate certainty evidence categories should be presented in the left column, and low/very low certainty in the right column. High and moderate certainty evidence, as well as low and very low certainty evidence were grouped together to simplify the presentation format into two groups (high/moderate, and low/very low), as participants perceived these groupings to hold similar weight in clinical decision making.

Legend Terminology

Participants encouraged the use of simple language within the legend. Participants preferred legend rows organized from "among the best" to "among the worst" vertically down the first column of the legend, with the middle category labelled as "intermediate". Terms such as "better" and "worse" were clearer to participants than terminology such as "statistically significant"; specifically, respondents favored "better than placebo" over "statistically significant over placebo".

The language used for our NMA example, in accordance with the minimally contextualized approach, contained treatments that were "better than placebo and some other interventions", "better than placebo, but no better than any other interventions", and "no better than placebo" for high/moderate certainty evidence of benefit outcomes. For high/moderate

certainty evidence of harm outcomes, the corresponding language was “no more harmful than placebo”, “more harmful than placebo, but no worse than other interventions”, and “more harmful than placebo and some other interventions”. Participants felt that, with respect to category of magnitude of effect low/very low certainty evidence descriptions should be the same as those of the high/moderate certainty evidence categories, with the included qualifier of “may be” at the beginning of the description of low to very low certainty evidence.

Gradient Colour-Coding

The gradient colour-coding scheme utilizes three shades of green for the high/moderate certainty benefit outcomes (**Figure 5: cells 1-3)**, and three shades of red for the high/moderate certainty adverse events (**Figure 5: cells 7-9**). The use of three-shade grey gradient for low/very low certainty evidence is consistent for both beneficial outcomes and adverse events (**Figure 5: cells 4-6, 10-12**). Participants preferred dark grey be used for the “among the worst” category (least beneficial or most harmful) and light grey be used for the “among the best” category (most beneficial or least harmful), when presenting low/very low certainty of evidence results.

Participants had clear views regarding the colour shades used in **Figure 5: cell 3** (among the least beneficial; high/moderate certainty), and **Figure 5: cell 7** (among the least harmful; high/moderate certainty)**:** because green is intuitively associated with positive results, they suggested caution regarding the use of a green shade for treatments categorized as “among the worst” in benefit outcomes supported by high/moderate certainty evidence (**Figure 5: cell 3**). Participants strongly suggested that the shade of green used in this cell should, as a result, be a pale and faint green. Similarly, **Figure 5: cell 7** utilizes a shade of red, despite being within the

"among the best" category in adverse events supported by high/moderate certainty evidence. Intuitively, participants noted that red is associated with poorer results. In order to avoid this inappropriate association, they suggested **Figure 5: cell 7** should utilize a pale and faint shade of red. Other options tested used white for **Figure 5: cell 3**, and **Figure 5: cell 7**; however, participants ultimately believed that faint colouring within the respective colour gradients was most appropriate and did not hinder interpretation.

Stoplight Colour-Coding

Because it dealt with the aforementioned concerns of the gradient colour-coding, participants also expressed enthusiasm for the stoplight colour-coding. The use of the same colour scheme across **Figure 6: cells 1-3** and **Figure 6: cells 7-9** simplifies the interpretation based on colour. Although the stoplight colour-coding addressed concerns with the gradient option, some participants preferred the gradient colour-coding due to the clear distinction between benefit and harms outcomes. Others also felt that the stoplight colour-coding looked distracting due to the inclusion of 3 bold colours, while the gradient colour-coding reserves bold colours that "stand out" for the comparisons with large benefits or large harms.

Discussion

The GRADE working group has developed methodologically coherent and innovative approaches to rating treatments within NMAs, including both benefits and harms, as "among the best", "intermediate" and "among the worst".[13,14] This may represent an important advance in the interpretation of the results of NMAs for clinicians using findings to guide clinical care. Clinicians,

however, need to apply this rating for all outcomes of importance to patients. Rigorously developed, user-friendly, intuitive, and user tested approaches to simultaneous presentation of rated treatments across multiple outcomes has thus far been unavailable for either the new GRADE rating approach or prior approaches to enhance interpretability.[4–6,9,12]

This study has addressed existing limitations by developing presentation methods that summarize NMA results for multiple outcomes in clear and interpretable formats. Although previous methods may still be useful in presenting the results of individual outcomes in greater detail with certainty of evidence incorporated[4–6,9], the current presentation method allows for a clear and succinct summary of all outcomes considered within an NMA in a single presentation that our user testing has found both appealing and understandable to clinicians, many with limited prior exposure to NMAs.[6]

Strengths and Limitations

Extensive user-testing in a targeted audience has validated our NMA presentation approaches, allowing future NMA's to enhance the ease with which clinicians can interpret their results. Additional strengths of this study include consultation with individuals involved in the process of developing and disseminating systematic reviews and clinical practice guidelines, and extensive user testing that included the careful selection of a study population that reflects the broader clinical audience who will be making use of NMA results. The use of structured qualitative research methods including duplicate data analysis allowed the accurate and appropriate incorporation of user feedback to be incorporated into iterative presentation development.

Our study does have limitations. First, although the simplicity of the developed presentations represents a strength, achieving that simplicity required the omission of data that some audiences may consider important.[6] For instance, the previous development of an NMA summary of findings table for individual outcomes provides greater detail for each treatment comparison that cannot feasibly fit within a multiple outcome presentation.[6] A particularly important omission may be the absolute effects of interventions that sometimes become crucial in trading off benefits and harms.[8] For this reason, authors may find it most appropriate to include both the multiple outcome presentation from this investigation, as well as additional outcome summaries suggested by other investigators.[4,6–11] Finally, we did not implement member checking. We did, however, employ data source triangulation to ensure that the findings of our study were robust.

Relation to Prior Work

Recent publications have addressed the issue of presenting NMA results for multiple outcomes, but have limitations that our proposal has addressed.[7,8] First, and crucially important, other options do not address the certainty of the evidence.[7,8] The Kilim plot provides a measure of the "strength of statistical evidence", which equates to the magnitude of the p-value.[8] Considerations of risk of bias, inconsistency, indirectness, publication bias, intransitivity, and incoherence may, however, reduce certainty in treatment effects with low p-values (which may or may not represent large effects). Additionally, the lack of user testing precludes confidence in how target users will understand these formats. For these reasons, the presentation versions

proposed in the current study represent important improvements on previous tools for reporting NMA results for multiple outcomes.

Choosing a Presentation Variation

Authors can, based on the appropriateness of the colour-coding and the corresponding categorization, choose between the two presentation versions in this manuscript. For example, the stoplight colour-coding variation may be most suitable when some treatments are better than the reference for some outcomes, while other treatments are worse for some outcomes. The three categories and explanations for benefit outcomes would then be “among the best – better than reference (colour: green)”, “intermediate – same as reference (colour: yellow)”, “among the worst – worse than reference (colour: red)”. Intuitively, these descriptions and colours align. **Appendix B** provides an example of this scenario, with suggested details on the appropriate language to use within the legend.

The colour-gradient variation of the presentation may be most appropriate when the reference treatment is the worst (or best) treatment option across all outcomes. This would typically occur when placebo is the reference treatment, as placebo would likely be the worst treatment for benefit outcomes and the best treatment option for adverse event outcomes. The acute pain NMA used for our presentation formats fits this scenario. Although typically occurring with a placebo reference treatment, there may also be NMAs with other reference treatments that would intuitively follow this gradient colour-coding. **Appendix C** provides an example with suggested details on the appropriate language to use within the legend.

Additional Considerations

There is no single set of legend terminologies that universally apply to all NMAs, so authors must use their discretion to determine the most applicable and intuitive terminology. Authors may use the general guidance provided in this study in conjunction with categorization recommendations of the minimally or partially contextualized approach.[13,14] The minimally and partially contextualized approaches to NMA treatment categorization have the potential for more than three categories, which would require an adaptation to the colour schemes we identified. The appropriate title for this presentation format represents another consideration that this study did not test. We would encourage authors to be explicit in defining the patient population assessed within the presentation.

Methodologists and statisticians have long bemoaned an excessive focus on statistical significance, in particular through the use of p-values.[21–24] Notwithstanding, our participants felt it was important to highlight results indicating statistical significance, and our view is that there is considerable merit in the suggestion. Bolding or italics would be two possible ways of such highlighting, and the choice may depend on a journal's particular font suggestions.

A final consideration is the use of colours in the presentation methods. Participants believed that green, yellow, and red were the most intuitive colours for the table colour-coding; however, these colours may be problematic for colour-blind individuals. Authors who want to ensure colour-blind accessibility may consider using blue instead of green, and orange instead of red.

Conclusion

This study utilized user-testing to develop easily interpretable presentation formats for reporting NMA results with multiple outcomes, with a focus both on relative magnitude of effects and certainty of evidence. If further empirical study verifies our finding that clinicians, and potentially patients - who are increasingly involved in clinical shared-decision making – who are naïve to NMAs find the presentation understandable and appealing, its wide implementation may enhance the impact and usefulness of NMAs.

Contributorship statement:

MRP, BS, JWB, RPB, CC, FKN, RRB, LT, MB, and GHG conceptualized the study

MRP, BS, JWB, and GHG recruited participants for the study.

MRP, YJG, and SB collected and analyzed data.

MRP, BS, JWB, RPB, CC, FKN, and GHG acted as the steering committee to interpret and implement data from participants.

MRP and GHG developed a first draft of the manuscript.

All authors reviewed, edited and approved the manuscript.

Competing Interests: Mohit Bhandari research grants from Pendopharm, Bioventus, and Acumed. All other authors have no conflicts of interest to disclose.

Funding: This study did not receive any funding.

Data sharing statement: Not Applicable

References

1. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med*. 2017;12(1):103-111. doi:10.1007/s11739-016-1583-7

2. Mills E, Thorlund K, Ioannidis J. Demystifying trial networks and network meta-analysis. *BMJ*. 2013;346:f2914. doi:10.1136/bmj.f2914

3. Ellis SG. Do we know the best treatment for in-stent restenosis via network meta-analysis (NMA)?: simple methods any interventionalist can use to assess NMA quality and a call for better NMA presentation. *JACC Cardiovasc Interv*. 2015;8(3):395-397. doi:10.1016/j.jcin.2014.11.012

4. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

5. Law M, Alam N, Veroniki AA, Yu Y, Jackson D. Two new approaches for the visualisation of models for network meta-analysis. *BMC Med Res Methodol*. 2019;19(1):61. doi:10.1186/s12874-019-0689-9

6. Yepes-Nuñez JJ, Li S-A, Guyatt G, et al. Development of the summary of findings table for network meta-analysis. *J Clin Epidemiol*. 2019;115:1-13. doi:10.1016/j.jclinepi.2019.04.018

7. Daly CH, Mbuagbaw L, Thabane L, Straus SE, Hamid JS. Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol*. 2020;20(1):266. doi:10.1186/s12874-020-01128-2

8. Seo M, Furukawa TA, Veroniki AA, et al. The Kilim plot: A tool for visualizing network meta-analysis results for multiple outcomes. *Res Synth Methods*. Published online June 10, 2020. doi:10.1002/jrsm.1428

9. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PloS One*. 2013;8(10):e76654. doi:10.1371/journal.pone.0076654

10. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Med Res Methodol*. 2013;13:35. doi:10.1186/1471-2288-13-35

11. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. *J Clin Epidemiol*. 2014;67(6):672-680. doi:10.1016/j.jclinepi.2013.11.006

12. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev*. 2017;6(1):79. doi:10.1186/s13643-017-0473-z

13. Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ*. 2020;371. doi:10.1136/bmj.m3900

14. Brignardello-Petersen R, Izcovich A, Rochwerg B, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. *BMJ*. 2020;371:m3907. doi:10.1136/bmj.m3907

15. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description – the poor cousin of health research? *BMC Med Res Methodol*. 2009;9(1):52. doi:10.1186/1471-2288-9-52

16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007;19(6):349-357. doi:10.1093/intqhc/mzm042

17. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant*. 2018;52(4):1893-1907. doi:10.1007/s11135-017-0574-8

18. Busse JW, Sadeghirad B, Oparin Y, et al. Management of Acute Pain From Non-Low Back, Musculoskeletal Injuries : A Systematic Review and Network Meta-analysis of Randomized Trials. *Ann Intern Med*. 2020;173(9):730-738. doi:10.7326/M19-3601

19. Maher C, Hadfield M, Hutchings M, Eyto A de. Ensuring Rigor in Qualitative Data Analysis: A Design Research Approach to Coding Combining NVivo With Traditional Material Methods. *Int J Qual Methods*. Published online July 10, 2018. doi:10.1177/1609406918786362

20. Morse JM. Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qual Health Res*. 2015;25(9):1212-1222. doi:10.1177/1049732315588501

21. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3

22. Gagnier JJ, Morgenstern H. Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing. *J Bone Joint Surg Am*. 2017;99(18):1598-1603. doi:10.2106/JBJS.16.01314

23. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

24. Phillips M. Letter to the Editor: Editorial: Threshold P Values in Orthopaedic Research-We Know the Problem. What is the Solution? *Clin Orthop*. 2019;477(7):1756-1758. doi:10.1097/CORR.0000000000000827

Tables and Figures

Table 1: Participant Demographics: n=26

Demographic	Value
Age (Mean, SD) years	47.6 (13.9)
Gender (Count, %)	
Male	19 (73.1%)
Female	7 (26.9%)
Primary Occupation (Count, %)	
Clinician	20 (76.9%)
Research Staff/ Methodologist	3 (11.5%)
Resident	3 (11.5%)
Highest Degrees Held (Count, %)	
MD	12 (46.2%)
MD, MSc/MPH	8 (30.8%)
PhD	3 (11.5%)
MD, PhD	2 (7.7%)
BSc	1 (3.9%)
Years in Practice (Mean, SD)	19.5 (14.3)
Previous involvement in an NMA? (Count, %)	
Yes	11 (42.3%)
No	15 (57.7%)
Used an NMA to inform practice? (Count, %)	
Yes	17 (65.4%)
No	9 (34.6%)

SD: Standard Deviation, MD: Doctor of Medicine, MSc: Masters of Science, MPH: Masters of Public Health, PhD: Doctor of Philosophy, BSc: Bachelor of Science, NMA: Network Meta-Analysis.

Figure 1: Study Overview

Figure 2: Gradient Colour Variation

Legend

Footnote

- : no evidence

Reference Group = Placebo
Bold = statistically significant (p<0.05)
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 3: Stoplight Colour Version

Legend

Footnote

- : no evidence

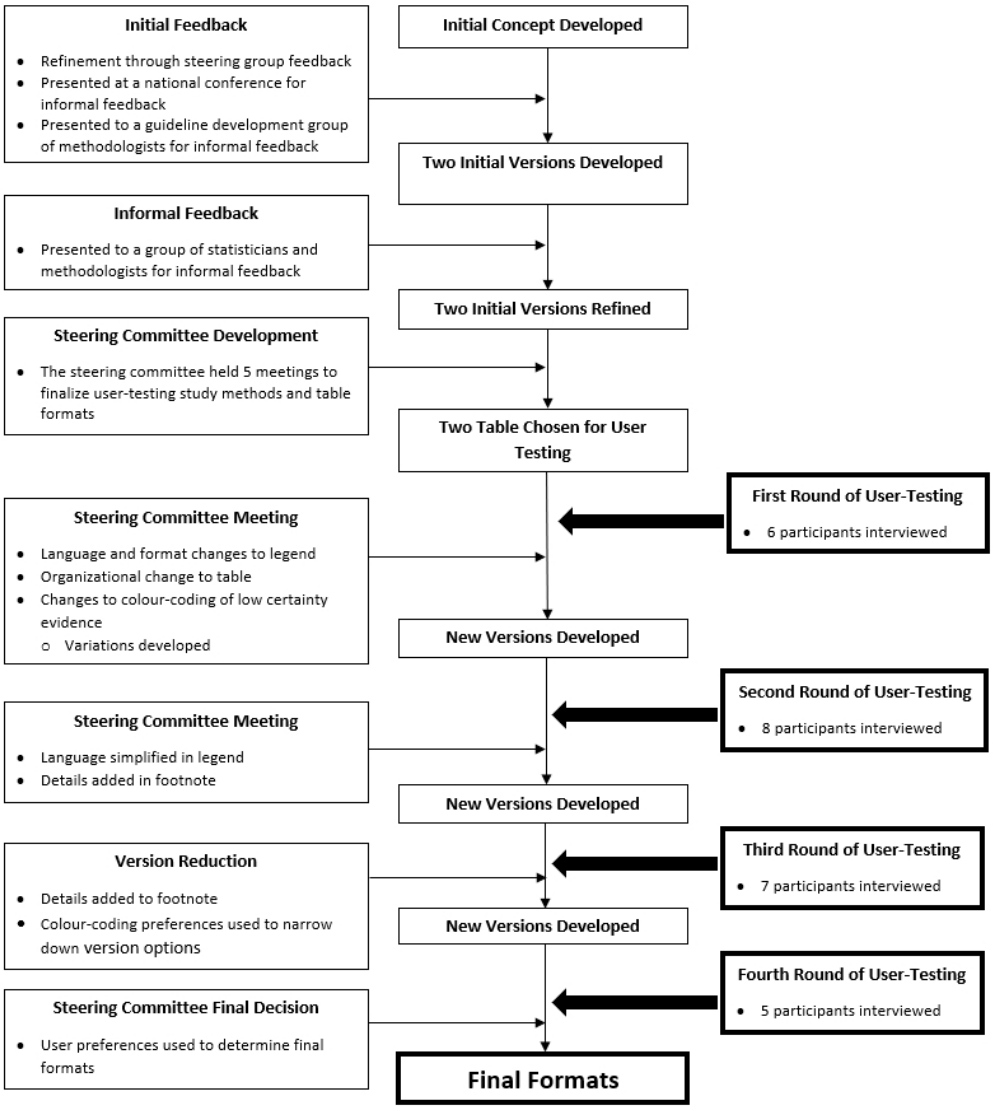
Reference Group = Placebo
Bold = statistically significant, p<0.05
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 4: Intervention Organizational Guide

Legend

Figure 5: Gradient Colour-Coding Legend

Figure 6: Stoplight Colour-Coding Legend



424x473mm (47 x 47 DPI)

Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	**-1.02 (-1.64,-0.39)**	**-1.08 (-1.40,-0.75)**	**1.66 (1.16,2.16)**	**5.20 (2.03,13.33)**	**6.39 (3.48,11.75)**	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	**-0.93 (-1.49,-0.37)**	**-0.99 (-1.46,-0.52)**	**0.73 (0.17,1.30)**	3.24 (0.43,24.70)	**3.10 (1.39,6.91)**	**1.77 (1.33,2.35)**	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	**-1.03 (-1.82,-0.24)**	**-1.07 (-1.89,-0.24)**	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	**-1.11 (-2.00,-0.21)**	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	**3.72 (1.02,13.52)**	-	-	-
Topical NSAID + Menthol Gel	**-1.68 (-0.27,-3.09)**	-0.89 (-2.33,0.54)	-	-	**13.34 (3.30,53.92)**	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	**-1.94 (-2.90,-0.98)**	**-1.18 (-2.09,-0.28)**	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	**-1.59 (-2.52,-0.66)**	**-2.09 (-3.86,-0.32)**	**1.51 (0.42,2.61)**	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	**-1.75 (-2.68,-0.81)**	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	**167.71 (6.67,4217.10)**	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	**-2.92 (-5.41,-0.43)**	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	**32.08 (4.93,208.60)**	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	**7.99 (1.29,49.41)**	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	**-1.71 (-2.97, -0.46)**	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	**5.63 (2.84,11.16)**	**3.53 (1.92,6.49)**	-
Acetaminophen, Ibuprofen + Codeine	**-1.36 (-2.49,-0.23)**	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	**4.91 (1.45,16.61)**	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	**-3.52 (-4.99,-2.04)**	-	-	-	-	**59.38 (6.21,567.71)**	**5.73 (1.20,27.47)**	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	**6.72 (1.24,36.39)**	-

354x445mm (47 x 47 DPI)

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	Better than placebo and some other interventions	May be better than placebo and some alternatives	No more harmful than placebo	May be no more harmful than placebo
INTERMEDIATE	Better than placebo, but no better than any other interventions	May be better than placebo, but no better than other interventions	More harmful than placebo, but no worse than other interventions	May be more harmful than placebo, but no worse than other interventions
AMONG THE WORST	No better than placebo	May be no better than placebo	More harmful than placebo and some other interventions	May be more harmful than placebo and some alternatives

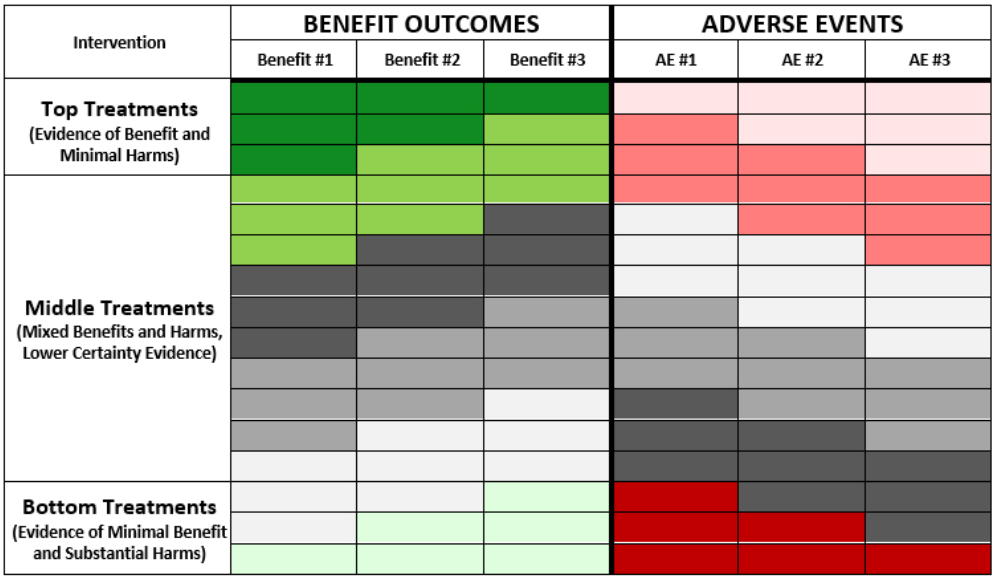
378x153mm (47 x 47 DPI)

Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	-1.02 (-1.64,-0.39)	-1.08 (-1.40,-0.75)	1.66 (1.16,2.16)	5.20 (2.03,13.33)	6.39 (3.48,11.75)	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	-0.93 (-1.49,-0.37)	-0.99 (-1.46,-0.52)	0.73 (0.17,1.30)	3.24 (0.43,24.70)	3.10 (1.39,6.91)	1.77 (1.33,2.35)	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	-1.03 (-1.82,-0.24)	-1.07 (-1.89,-0.24)	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	-1.11 (-2.00,-0.21)	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	3.72 (1.02,13.52)	-	-	-
Topical NSAID + Menthol Gel	-1.68 (-0.27,-3.09)	-0.89 (-2.33,0.54)	-	-	13.34 (3.30,53.92)	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	-1.94 (-2.90,-0.98)	-1.18 (-2.09,-0.28)	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	-1.59 (-2.52,-0.66)	-2.09 (-3.86,-0.32)	1.51 (0.42,2.61)	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	-1.75 (-2.68,-0.81)	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	167.71 (6.67,4217.10)	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	-2.92 (-5.41,-0.43)	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	32.08 (4.93,208.60)	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	7.99 (1.29,49.41)	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	-1.71 (-2.97, -0.46)	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	5.63 (2.84,11.16)	3.53 (1.92,6.49)	-
Acetaminophen, Ibuprofen + Codeine	-1.36 (-2.49,-0.23)	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	4.91 (1.45,16.61)	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	-3.52 (-4.99,-2.04)	-	-	-	-	59.38 (6.21,567.71)	5.73 (1.20,27.47)	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	6.72 (1.24,36.39)	-

355x442mm (47 x 47 DPI)

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	Better than placebo and some alternatives	May be better than placebo and some alternatives	No more harmful than placebo	May be no more harmful than placebo
INTERMEDIATE	Better than placebo, but no better than any alternatives	May be better than placebo, but no better than any alternatives	More harmful than placebo, but no worse than any alternatives	May be more harmful than placebo, but no worse than any alternatives
AMONG THE WORST	No better than placebo	May be no better than placebo	More harmful than placebo and some alternatives	May be more harmful than placebo and some alternatives

374x149mm (47 x 47 DPI)



408x241mm (47 x 47 DPI)

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST				
INTERMEDIATE				
AMONG THE WORST				

338x116mm (47 x 47 DPI)

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	1	4	7	10
INTERMEDIATE	2	5	8	11
AMONG THE WORST	3	6	9	12

488x185mm (47 x 47 DPI)

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	1	4	7	10
INTERMEDIATE	2	5	8	11
AMONG THE WORST	3	6	9	12

483x186mm (47 x 47 DPI)

Appendix A: Open-Ended Interview Guide

Part 1: Introductions

1. Introductions

Part 2: NMA Familiarity

To begin, we would like to understand your current knowledge of NMA:

2. How familiar are you with NMA?
3. Have you ever been part of an NMA project?
    a. If so, what was your role in the NMA project?
4. Have you ever read an NMA?

Part 3: Review of the table format

The table I am showing you summarizes the results of an NMA that assessed acute pain management treatment options.

Please think aloud as you interpret this table

Prompts regarding the legend:

5. Do you find the language within the legend to be understandable? If not, what is confusing?
6. Do you have any feedback regarding the format of the legend?
    i. Do you have feedback regarding the coloring used?
    ii. Do you have feedback regarding the language used?
    iii. Do you have feedback regarding the indication of the certainty of evidence component of the legend?

Prompts regarding the results table:

7. Now that you have reviewed the legend in more detail, does the legend accurately and completely summarize the results table?
    a. If not, what could be changed?
8. Please provide any feedback you have regarding the results within the table
    a. Are the results easily understandable? If not, what is confusing or could be changed?
9. Do you have any feedback regarding the format of the table?
    a. Do you have feedback regarding the coloring used?
    b. Do you have feedback regarding the language used?
    c. Do you have feedback regarding the outcome reporting within the table?
    d. Do you have feedback regarding the indication of the certainty of evidence component of the results?
10. Please provide any other feedback that you may have regarding the table

Part 4: Assessing Participant Interpretation

Based on the results within the table, please describe how you interpret the findings?

Prompts regarding interpretation:

11. Based on both the benefits and the harms, which treatment(s) do you consider to be the optimal choice(s)?
12. Which treatment(s) do you believe are the least optimal choices? What information is important for you in deciding this?
13. How confident are you in your interpretation?
    a. Why are you/aren't you confident in your interpretation?
    b. What would aid in improving your interpretation?

Part 5: Closing Remarks

We would like to ask if you have any colleagues that may be interested in participating in this study. Following this interview, it would be great if we could connect with anyone who you believe may be able to provide valuable insights to this project.

Appendix B: Example Legend When Active Treatment is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than reference**	**May be better than reference**	**Less harmful than reference**	**May be less harmful than reference**
INTERMEDIATE	**No better than reference**	**May be no better than reference**	**No more harmful than reference**	**May be no more harmful than reference**
AMONG THE WORST	**Worse than reference**	**May be worse than reference**	**More harmful than reference**	**May be more harmful than reference**

Appendix C: Example Legend When Placebo (Or Any Sham/Null Treatment Effect) is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than placebo and some other interventions**	**May be better than placebo and some alternatives**	**No more harmful than placebo**	**May be no more harmful than placebo**
INTERMEDIATE	**Better than placebo, but no better than any other interventions**	**May be better than placebo, but no better than other interventions**	**More harmful than placebo, but no worse than other interventions**	**May be more harmful than placebo, but no worse than other interventions**
AMONG THE WORST	**No better than placebo**	**May be no better than placebo**	**More harmful than placebo and some other interventions**	**May be more harmful than placebo and some alternatives**

BMJ Open

Development and design validation of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-056400.R1
Article Type:	Original research
Date Submitted by the Author:	18-Dec-2021
Complete List of Authors:	Phillips, Mark ; McMaster University, Health Research Methods, Evidence, and Impact Sadeghirad, Behnam; McMaster University, Health Research Methods, Evidence, and Impact; McMaster University, Anesthesia Busse, Jason; McMaster University, Anesthesia; McMaster University, Health Research Methods, Evidence, and Impact Brignardello-Petersen, Romina ; McMaster University, Health Research Methods, Evidence, and Impact (HEI) Cuello-Garcia, Carlos ; McMaster University, Health Research Methods, Evidence, and Impact Kenji Nampo, Fernando ; Federal University of Latin-American Integration Guo, Yu Jia; McMaster University, Health Sciences Bzovsky, Sofia; McMaster University, Department of Surgery - Division of Orthopaedics Bannuru, Raveendhara R.; Tufts Medical Center, Center for Treatment Comparison and Integrative Analysis Thabane, Lehana; McMaster University, Health Research Methods, Evidence, and Impact; St. Joseph's Healthcare, Biostatistics Unit Bhandari, Mohit; Mcmaster University, Division of Orthopaedic Surgery; McMaster University, Health Research Methods, Evidence, and Impact Guyatt, Gordon; McMaster University, Health Research Methods, Evidence, and Impact
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Evidence based practice, Qualitative research
Keywords:	QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS, EDUCATION & TRAINING (see Medical Education & Training)

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Development and design validation of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Mark R Phillips[1], Behnam Sadeghirad[1,2], Jason W Busse[1,2], Romina Brignardello-Petersen[1], Carlos Cuello[1], Fernando Kenji Nampo[3], Yu Jia Guo, Sofia Bzovsky[5], Raveendhara R Bannuru[6], Lehana Thabane[1,7], Mohit Bhandari[1,5], Gordon H Guyatt[1]

1. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
2. Department of Anesthesia, McMaster University, Hamilton, ON, Canada
3. Latin-American Institute of Life and Nature Sciences/Evidence-Based Public Health Research Group, Federal University of Latin-American Integration, Foz do Iguassu, Brazil
4. Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
5. Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada
6. Center for Treatment Comparison and Integrative Analysis, Tufts Medical Center, Boston, MA, USA
7. Biostatistics Unit, St Joseph's Healthcare, Hamilton, ON, Canada

Correspondence: Mark Phillips phillimr@mcmaster.ca

1280 Main Street West, Hamilton, ON, CAN L8S 4L8

Keywords: Network Meta-Analysis, Design Validation, Qualitative Research, Statistics and Research Methods

Word Count: 5292

Abstract

Objective The GRADE working group recently developed an innovative approach to interpreting results from network meta-analyses (NMA) through minimally and partially contextualized methods; however, the optimal method for presenting results for multiple outcomes using this approach remains uncertain. We therefore developed and iteratively modified a presentation method that effectively summarizes NMA results of multiple outcomes for clinicians using this new interpretation approach.

Design Qualitative descriptive study

Setting A steering group of 7 individuals with experience in NMA and design validation studies developed two colour-coded presentation formats for evaluation. Through an iterative process, we assessed the validity of both formats to maximize their clarity and ease of interpretation.

Participants 26 participants including 20 clinicians who routinely provide patient care, 3 research staff/research methodologists, and 3 residents.

Main Outcome Measures Two team members used qualitative content analysis to independently analyze transcripts of all interviews. The steering group reviewed the analyses and responded with serial modifications of the presentation format.

Results To ensure that readers could easily discern the benefits and safety of each included treatment across all assessed outcomes, participants primarily focused on simple information presentations, with intuitive organizational decisions and colour coding. Feedback ultimately resulted in two presentation versions, each preferred by a substantial group of participants, and development of a legend to facilitate interpretation.

Conclusion Iterative design validation facilitated the development of two novel formats for presenting minimally or partially contextualized NMA results for multiple outcomes. These presentation approaches appeal to audiences that include clinicians with limited familiarity with NMAs.

Strengths and Limitations of this Study

- Extensive design validation in a targeted audience has validated the NMA presentation approaches within this study; something that has not been done for other presentation formats
- Structured qualitative research methodology has ensured accurate use of user feedback to develop and refine the NMA presentation formats
- Limited by the omission of some information within the presentation formats in order to achieve simplicity and interpretability, such as greater detail for individual outcomes, absolute effects, or specifics about the certainty of evidence assessments.
- The aforementioned information should still be included in NMA manuscripts, but cannot be feasibly fit within the presentation formats.

Introduction

Network meta-analysis (NMA) provides an increasingly popular approach to evidence synthesis that allows comparison between multiple competing treatment options within a single analysis.[1,2] Although NMA is an important tool for clinicians, patients, and other stakeholders, results involve multiple treatments and outcomes, and as a result are complex and difficult to interpret.[3]

Common methods for presenting NMA results include the use of forest plots, league tables, and surface under the cumulative ranking curve (SUCRA).[1,4] The key limitation with these options is that they can only provide results of a single outcome.[5] NMAs often compare multiple benefit and harm outcomes, resulting in challenges for NMA authors seeking to avoid presentation methods that are onerous for clinicians to review and challenging for them to understand.[6]

There are a number of novel approaches that have been suggested for presenting NMA results for multiple outcomes[7,8]; however, these approaches lack key information, present challenges to interpretation, and have not undergone design validation with their target audiences. While some previously suggested approaches have merit for a limited number of outcomes,[4,6,9–12] although not all taking certainty of evidence into account, they have serious limitations for simultaneous presentation of multiple outcomes.

Recently, the GRADE working group has suggested two variations on a new methodology that places interventions in categories from best to worst considering the estimates of effect and certainty of the evidence for each comparison.[13,14] We therefore developed interpretable presentation approaches for NMAs with multiple outcomes that builds on GRADE guidance and

effectively summarizes results for clinicians and other relevant audiences.

Methods

Study Design

A 7-member steering committee (MRP, BS, JWB, RB-P, CC, FKN, GHG) oversaw study design and implementation. The committee generated two initial presentation formats and chose a combination of large group sessions and individual design validation interviews to inform iterative modifications of the two initial formats. The presentation format consisted of treatment options in rows and outcomes in columns, with colour-coded shading of cells to identify the magnitude and certainty of the treatment effect in relation to the reference treatment. The group developed the initial versions through a series of group discussions, which involved: determining the pertinent information for the presentation format to contain, options for how that information could be shown within a single presentation format, and draft presentation formats that may present this pertinent information. The group believed that the format should provide both relative treatment effects, as well as the certainty in those estimates for all outcomes, within a single presentation tool.

Initial large group testing with two groups of methodologists, graduate students in health research-focussed programs, and statisticians, as well as presentation at a national conference (2019 Canadian Pain Society annual scientific meeting), provided the foundational feedback for modifications of the initial presentation versions. The steering committee reviewed input from four rounds of design validation, iteratively modifying the formats after each round and presenting updated options of the presentation versions to subsequent participants.

For the user interviews, the committee chose a qualitative descriptive study approach that focuses on creating a close description of the information that participants provide.[15] This is ideal for design validation that, without interpretive direction, aims to optimize the understandability of a tool within the target population. This study involves human participants but an Ethics Committee exempted this study. After reviewing the protocol, the Hamilton Integrated Research Ethics Board (HiREB) committee and chair, judging the study to be a quality improvement investigation within the methodology and knowledge translation field, provided an exemption from formal ethics approval. Participants provided informed consent at the beginning of their interview. We followed, when applicable, the consolidated criteria for reporting qualitative research (COREQ) checklist in reporting our findings.[16]

Sampling and Recruitment

This study utilized purposeful sampling to identify participants who could provide information-rich interviews to inform the design validation process.[15,17] Target users for this study included academic and non-academic clinicians, research staff/research methodologists, and residents. The steering committee, through their professional contacts, provided a pool of initial possible participants that the principal investigator supplemented using snowball sampling technique. [18] Specifically, we asked individuals who agreed to participate for contact information of any colleagues whom we could approach to interview. Prior to their interviews, each participant received information outlining the purpose of the study. Study recruitment ceased when data collection reached redundancy – the point at which there were no further refinements requested to improve the interpretability of the presentation formats.[18]

Data Collection

The principal investigator (MRP) conducted all design validation interviews either in-person or through video teleconferencing. Interviews followed a flexible interview guide (**Appendix A**) to leave the conversation open for participants to explore any topics they felt were relevant and important.[15] Throughout the study, the principal investigator iteratively updated the interview guide to explore areas of importance that emerged. Interviews began with a brief introduction to NMA methods, followed by questions regarding the participant's familiarity and experience with NMA. Participants then viewed the current versions of the NMA presentation formats and provided feedback. YJG or MRP transcribed all interviews verbatim. Transcripts were not returned to participants and interviewers did not conduct follow up interviews. The steering committee incorporated all feedback to arrive at two final presentation versions.

Patient and Public Involvement

This study did not include patient or public involvement.

NMA for Design Validation

The steering committee developed five core criteria to which the example NMA must adhere: (1) variability in quality of evidence (2) variability in magnitudes of effect; (3) assessment of both benefits and harms; (4) inclusion of both continuous and binary outcomes; and (5) including at least 5 outcomes and 5 interventions. Based on these criteria the steering committee chose, for

design validation, a recent NMA that used a minimally contextualized approach to address acute pain management in patients experiencing non-low back acute musculoskeletal injuries.[19]

Based on the GRADE approach[13] this NMA categorized, for each benefit outcome, interventions as among those with the largest benefit, those with intermediate benefit, and those with the least benefit. For each harm outcome, they categorized interventions as among the least harmful, intermediate harm, and the most harmful. They then categorized interventions as those for which there was high or moderate certainty evidence, and those for which there was low or very low-quality evidence.[19] These results provided the example for design validation.

Data Analysis

Two reviewers (MP and SB) independently conducted data analysis, in duplicate, using a qualitative content analysis approach.[17] The study team recruited participants, collected data, and conducted data analysis in parallel. As new data became available, the reviewers coded and grouped similar phrases, patterns, and themes.[17] When discrepancies in feedback were identified, these would be noted and further elaborated on within future interviews. The feedback for this discrepancy would then be shared with the steering committee to review and identify if sufficient data had been captured to adequately determine a resolution for the discrepancy through consensus.[17] Data triangulation was utilized through multiple forms of data collection, as both large group and individual interview sessions were used. Additionally, data triangulation was provided through two forms of data analysis: independent qualitative content analysis, and group deliberation through steering committee meetings.[17,20] The steering committee met four times over a period of 14 months to review the collected data and made

iterative changes to the presentation formats as dictated by feedback, initially from large group presentations and subsequently from design validation. When analysis of the data provided actionable feedback, the reviewers presented their findings to the steering committee who ranked feedback as a "large change required", "moderate change required", or "minor change required" and then revised the presentation format(s) accordingly.

Subsequent participants provided input on the modified versions of the NMA results presentations. Participants commented regarding their interpretation of the data within the presentation format; the team considered study objectives met once participants consistently reported a clear interpretation of the results with no or minimal suggested modifications. Reviewers documented all changes to the presentation format in a study audit trail.[15,20] Reviewers conducted all qualitative analysis using RQDA software (R version 3.5.0).

Results

Study Sample

Two focus groups, both of which included methodologists, graduate students, and statisticians, participated in the initial large group testing: the first, a critical care guideline development group (GUIDE: https://guidecanada.org/) many of whose members have NMA expertise (65 attendees); the second, a research group (CLARITY: http://www.clarityresearch.ca/) who meet regularly at McMaster University to discuss current methodological and statistical topics (20 attendees).

The design validation portion of this study included 26 participants of mean (standard deviation [SD]) age of 47.6 (13.9) years, 20 of whom were clinicians whose primary activity involved direct patient care (77%); 3 research staff/research methodologists (12%); and 3

residents (12%). Typical participants were male (73%) physicians in clinical practice for almost two decades (mean [SD]: 19.5 [14.3] years) with no prior involvement with conducting an NMA (58%) (**Table 1**).

Content Analysis Themes

Main themes that arose from the content analysis conducted on interview transcripts of participant interviews included "organizational", "language/terminology", "included information" and "colour options". Respondents also provided feedback regarding necessary details to include in the presentations' footnote. The following sections provide details regarding the most important feedback and how this feedback informed choices regarding presentation format. The fourth round of design validation resulted in minimal new information, resulting in two presentation versions that participants deemed satisfactory.

Final Presentation Versions

Ultimately, respondents proved equally enthusiastic about two options; the steering group, therefore, chose to offer both as alternative presentations. **Figure 1** summarizes the development process from conceptualization to the final presentation versions. We will refer to the presentation in **Figure 2** as the "colour gradient" version and the presentation in **Figure 3** as the "stoplight" version. Each presentation has a legend and footnote with pertinent information that the design validation process demonstrated necessary to include.

Figure Organization

Design validation identified a number of key components that aid in interpreting presentation formats. Within the organizational theme, the use of a bolded vertical line to separate benefit and adverse event outcomes, as well as the header and results data (horizontal), proved desirable. Regarding the ordering of interventions from top to bottom in the rows, participants preferred ordering treatment options at the top with high/moderate certainty evidence of maximal benefit and minimal harm to those with high/moderate certainty evidence of minimal or no benefits and significant harms placed in the bottom rows. Respondents provided mixed feedback regarding the organization of the presentation within the middle section, with no consistent guidance that could be applied across all NMAs. This leaves the optimal ordering within the middle rows that include treatments that have low/very low certainty evidence, treatments with high/moderate certainty evidence of intermediate effects, and treatments with trade-offs between both large benefits and large harms, uncertain (or perhaps there is no single optimal ordering). **Figure 4** provides an overview of guidance regarding intervention order within the rows.

Presentation Terminology

Respondents indicated that the presentation should clearly and succinctly label outcomes with specification of the measure of treatment effect (e.g. odds ratios, mean differences) and that the header of each column should include these labels. Participants had no strong preference regarding the terminology of “benefit” and “adverse events” outcome categories; options discussed included “effectiveness/efficacy outcomes” and “harms outcomes”. Whatever option

investigators choose, the terminology should remain consistent across the presentation, legend, and manuscript text.

Presentation Included Information

Participants considered the magnitude of treatment effect, confidence/credible intervals, certainty of evidence, and statistical significance to be the four important elements that should be included in each comparison cell. Possibilities explicitly discussed but rejected included sample size, patient characteristics, and heterogeneity/incoherence estimates. Respondents considered these items as important elements of the NMA, but felt they would be better suited within another section of the manuscript rather than within this summary presentation.

Footnote Included Information

Participants felt that footnotes should include: an indication of a dash representing no available evidence (- *: no evidence*); designation of the reference group (e.g. *Reference Group: Placebo*); and labelling of how statistical significance within the presentation is identified (i.e. *Bold = statistically significant, p<0.05*); as well as all abbreviations used within the presentation.

Legend Organization

Participants felt that benefit outcomes should be located in the left columns, with a bold vertical line separating the benefit and adverse event outcomes within the legend – similar to the structure of the main presentation. They also suggested a bold horizontal line separating the header from the legend in a similar format as within the main presentation. Within the benefit

and adverse event sections, respondents preferred that high/moderate certainty evidence categories should be presented in the left column, and low/very low certainty in the right column. High and moderate certainty evidence, as well as low and very low certainty evidence were grouped together to simplify the presentation format into two groups (high/moderate, and low/very low), as participants perceived these groupings to hold similar weight in clinical decision making.

Legend Terminology

Participants encouraged the use of simple language within the legend. Participants preferred legend rows organized from “among the best” to “among the worst” vertically down the first column of the legend, with the middle category labelled as “intermediate”. Terms such as “better” and “worse” were clearer to participants than terminology such as “statistically significant”; specifically, respondents favored “better than placebo” over “statistically significant over placebo”.

The language used for our NMA example, in accordance with the minimally contextualized approach, contained treatments that were “better than placebo and some other interventions”, “better than placebo, but no better than any other interventions”, and “no better than placebo” for high/moderate certainty evidence of benefit outcomes. For high/moderate certainty evidence of harm outcomes, the corresponding language was “no more harmful than placebo”, “more harmful than placebo, but no worse than other interventions”, and “more harmful than placebo and some other interventions”. Participants felt that, with respect to category of magnitude of effect low/very low certainty evidence descriptions should be the same

as those of the high/moderate certainty evidence categories, with the included qualifier of "may be" at the beginning of the description of low to very low certainty evidence.

Gradient Colour-Coding

The gradient colour-coding scheme utilizes three shades of green for the high/moderate certainty benefit outcomes (**Figure 5: cells 1-3)**, and three shades of red for the high/moderate certainty adverse events (**Figure 5: cells 7-9**). The use of three-shade grey gradient for low/very low certainty evidence is consistent for both beneficial outcomes and adverse events (**Figure 5: cells 4-6, 10-12**). Participants preferred dark grey be used for the "among the worst" category (least beneficial or most harmful) and light grey be used for the "among the best" category (most beneficial or least harmful), when presenting low/very low certainty of evidence results.

Participants had clear views regarding the colour shades used in **Figure 5: cell 3** (among the least beneficial; high/moderate certainty), and **Figure 5: cell 7** (among the least harmful; high/moderate certainty)**:** because green is intuitively associated with positive results, they suggested caution regarding the use of a green shade for treatments categorized as "among the worst" in benefit outcomes supported by high/moderate certainty evidence (**Figure 5: cell 3**). Participants strongly suggested that the shade of green used in this cell should, as a result, be a pale and faint green. Similarly, **Figure 5: cell 7** utilizes a shade of red, despite being within the "among the best" category in adverse events supported by high/moderate certainty evidence. Intuitively, participants noted that red is associated with poorer results. In order to avoid this inappropriate association, they suggested **Figure 5: cell 7** should utilize a pale and faint shade of red. Other options tested used white for **Figure 5: cell 3**, and **Figure 5: cell 7**; however,

participants ultimately believed that faint colouring within the respective colour gradients was most appropriate and did not hinder interpretation.

Stoplight Colour-Coding

Because it dealt with the aforementioned concerns of the gradient colour-coding, participants also expressed enthusiasm for the stoplight colour-coding. The use of the same colour scheme across **Figure 6: cells 1-3** and **Figure 6: cells 7-9** simplifies the interpretation based on colour. Although the stoplight colour-coding addressed concerns with the gradient option, some participants preferred the gradient colour-coding due to the clear distinction between benefit and harms outcomes. Others also felt that the stoplight colour-coding looked distracting due to the inclusion of 3 bold colours, while the gradient colour-coding reserves bold colours that "stand out" for the comparisons with large benefits or large harms.

Discussion

The GRADE working group has developed methodologically coherent and innovative approaches to rating treatments within NMAs, including both benefits and harms, as "among the best", "intermediate" and "among the worst".[13,14] This may represent an important advance in the interpretation of the results of NMAs for clinicians using findings to guide clinical care. Clinicians, however, need to apply this rating for all outcomes of importance to patients. Rigorously developed, user-friendly, intuitive, and tested approaches to simultaneous presentation of rated treatments across multiple outcomes has thus far been unavailable for either the new GRADE rating approach or prior approaches to enhance interpretability.[4–6,9,12]

This study has addressed existing limitations by developing presentation methods that summarize NMA results for multiple outcomes in clear and interpretable formats. Although previous methods may still be useful in presenting the results of individual outcomes in greater detail with certainty of evidence incorporated[4–6,9], the current presentation method allows for a clear and succinct summary of all outcomes considered within an NMA in a single presentation that our design validation has found both appealing and understandable to clinicians, many with limited prior exposure to NMAs.[6]

Strengths and Limitations

Extensive design validation in a targeted audience has validated our NMA presentation approaches, allowing future NMA's to enhance the ease with which clinicians can interpret their results. Additional strengths of this study include consultation with individuals involved in the process of developing and disseminating systematic reviews and clinical practice guidelines, and extensive design validation that included the careful selection of a study population that reflects the broader clinical audience who will be making use of NMA results. The use of structured qualitative research methods including duplicate data analysis allowed the accurate and appropriate incorporation of user feedback to be incorporated into iterative presentation development.

Our study does have limitations. First, although the simplicity of the developed presentations represents a strength, achieving that simplicity required the omission of data that some audiences may consider important.[6] For instance, the previous development of an NMA summary of findings table for individual outcomes provides greater detail for each treatment

comparison that cannot feasibly fit within a multiple outcome presentation.[6] A particularly important omission may be the absolute effects of interventions that sometimes become crucial in trading off benefits and harms.[8] For this reason, authors may find it most appropriate to include both the multiple outcome presentation from this investigation, as well as additional outcome summaries suggested by other investigators.[4,6–11] Finally, we did not implement member checking. We did, however, employ data source triangulation to ensure that the findings of our study were robust.

Relation to Prior Work

Recent publications have addressed the issue of presenting NMA results for multiple outcomes, but have limitations that our proposal has addressed.[7,8] First, and crucially important, other options do not address the certainty of the evidence.[7,8] The Kilim plot provides a measure of the "strength of statistical evidence", which equates to the magnitude of the p-value.[8] Considerations of risk of bias, inconsistency, indirectness, publication bias, intransitivity, and incoherence may, however, reduce certainty in treatment effects with low p-values (which may or may not represent large effects). Additionally, the lack of design validation precludes confidence in how target users will understand these formats. For these reasons, the presentation versions proposed in the current study represent important improvements on previous tools for reporting NMA results for multiple outcomes.

Choosing a Presentation Variation

Authors can, based on the appropriateness of the colour-coding and the corresponding categorization, choose between the two presentation versions in this manuscript. For example, the stoplight colour-coding variation may be most suitable when some treatments are better than the reference for some outcomes, while other treatments are worse for some outcomes. The three categories and explanations for benefit outcomes would then be "among the best – better than reference (colour: green)", "intermediate – same as reference (colour: yellow)", "among the worst – worse than reference (colour: red)". Intuitively, these descriptions and colours align. **Appendix B** provides an example of this scenario, with suggested details on the appropriate language to use within the legend.

The colour-gradient variation of the presentation may be most appropriate when the reference treatment is the worst (or best) treatment option across all outcomes. This would typically occur when placebo is the reference treatment, as placebo would likely be the worst treatment for benefit outcomes and the best treatment option for adverse event outcomes. The acute pain NMA used for our presentation formats fits this scenario. Although typically occurring with a placebo reference treatment, there may also be NMAs with other reference treatments that would intuitively follow this gradient colour-coding. **Appendix C** provides an example with suggested details on the appropriate language to use within the legend.

Additional Considerations

There is no single set of legend terminologies that universally apply to all NMAs, so authors must use their discretion to determine the most applicable and intuitive terminology. Authors may use the general guidance provided in this study in conjunction with categorization recommendations

of the minimally or partially contextualized approach.[13,14] The minimally and partially contextualized approaches to NMA treatment categorization have the potential for more than three categories, which would require an adaptation to the colour schemes we identified. The appropriate title for this presentation format represents another consideration that this study did not test. We would encourage authors to be explicit in defining the patient population assessed within the presentation.

Methodologists and statisticians have long bemoaned an excessive focus on statistical significance, in particular through the use of p-values.[21–24] Notwithstanding, our participants felt it was important to highlight results indicating statistical significance, and our view is that there is considerable merit in the suggestion. Bolding or italics would be two possible ways of such highlighting, and the choice may depend on a journal's particular font suggestions.

A final consideration is the use of colours in the presentation methods. Participants believed that green, yellow, and red were the most intuitive colours for the table colour-coding; however, these colours may be problematic for colour-blind individuals. Authors who want to ensure colour-blind accessibility may consider using blue instead of green, and orange instead of red.

Conclusion

This study utilized end-user design validation to develop easily interpretable presentation formats for reporting NMA results with multiple outcomes, with a focus both on relative magnitude of effects and certainty of evidence. If further empirical study verifies our finding that clinicians, and potentially patients - who are increasingly involved in clinical shared-decision

making – who are naïve to NMAs find the presentation understandable and appealing, its wide implementation may enhance the impact and usefulness of NMAs.

Contributorship statement:

MRP, BS, JWB, RPB, CC, FKN, RRB, LT, MB, and GHG conceptualized the study

MRP, BS, JWB, and GHG recruited participants for the study.

MRP, YJG, and SB collected and analyzed data.

MRP, BS, JWB, RPB, CC, FKN, and GHG acted as the steering committee to interpret and implement data from participants.

MRP and GHG developed a first draft of the manuscript.

All authors reviewed, edited and approved the manuscript.

Competing Interests: Mohit Bhandari research grants from Pendopharm, Bioventus, and Acumed. All other authors have no conflicts of interest to disclose.

Funding: This study did not receive any funding.

Data sharing statement: Not Applicable

References

1. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med*. 2017;12(1):103-111. doi:10.1007/s11739-016-1583-7

2. Mills E, Thorlund K, Ioannidis J. Demystifying trial networks and network meta-analysis. *BMJ*. 2013;346:f2914. doi:10.1136/bmj.f2914

3. Ellis SG. Do we know the best treatment for in-stent restenosis via network meta-analysis (NMA)?: simple methods any interventionalist can use to assess NMA quality and a call for better NMA presentation. *JACC Cardiovasc Interv*. 2015;8(3):395-397. doi:10.1016/j.jcin.2014.11.012

4. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

5. Law M, Alam N, Veroniki AA, Yu Y, Jackson D. Two new approaches for the visualisation of models for network meta-analysis. *BMC Med Res Methodol*. 2019;19(1):61. doi:10.1186/s12874-019-0689-9

6. Yepes-Nuñez JJ, Li SA, Guyatt G, et al. Development of the summary of findings table for network meta-analysis. *J Clin Epidemiol*. 2019;115:1-13. doi:10.1016/j.jclinepi.2019.04.018

7. Daly CH, Mbuagbaw L, Thabane L, Straus SE, Hamid JS. Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol*. 2020;20(1):266. doi:10.1186/s12874-020-01128-2

8. Seo M, Furukawa TA, Veroniki AA, et al. The Kilim plot: A tool for visualizing network meta-analysis results for multiple outcomes. *Res Synth Methods*. Published online June 10, 2020. doi:10.1002/jrsm.1428

9. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PloS One*. 2013;8(10):e76654. doi:10.1371/journal.pone.0076654

10. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Med Res Methodol*. 2013;13:35. doi:10.1186/1471-2288-13-35

11. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. *J Clin Epidemiol*. 2014;67(6):672-680. doi:10.1016/j.jclinepi.2013.11.006

12. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev*. 2017;6(1):79. doi:10.1186/s13643-017-0473-z

13. Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ*. 2020;371. doi:10.1136/bmj.m3900

14. Brignardello-Petersen R, Izcovich A, Rochwerg B, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. *BMJ*. 2020;371:m3907. doi:10.1136/bmj.m3907

15. Morse JM. Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qual Health Res*. 2015;25(9):1212-1222. doi:10.1177/1049732315588501

16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007;19(6):349-357. doi:10.1093/intqhc/mzm042

17. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description – the poor cousin of health research? *BMC Med Res Methodol*. 2009;9(1):52. doi:10.1186/1471-2288-9-52

18. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant*. 2018;52(4):1893-1907. doi:10.1007/s11135-017-0574-8

19. Busse JW, Sadeghirad B, Oparin Y, et al. Management of Acute Pain From Non-Low Back, Musculoskeletal Injuries : A Systematic Review and Network Meta-analysis of Randomized Trials. *Ann Intern Med*. 2020;173(9):730-738. doi:10.7326/M19-3601

20. Maher C, Hadfield M, Hutchings M, Eyto A de. Ensuring Rigor in Qualitative Data Analysis: A Design Research Approach to Coding Combining NVivo With Traditional Material Methods. *Int J Qual Methods*. Published online July 10, 2018. doi:10.1177/1609406918786362

21. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3

22. Gagnier JJ, Morgenstern H. Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing. *J Bone Joint Surg Am*. 2017;99(18):1598-1603. doi:10.2106/JBJS.16.01314

23. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

24. Phillips M. Letter to the Editor: Editorial: Threshold P Values in Orthopaedic Research-We Know the Problem. What is the Solution? *Clin Orthop*. 2019;477(7):1756-1758. doi:10.1097/CORR.0000000000000827

Tables and Figures

Table 1: Participant Demographics: n=26

Demographic	Value
Age (Mean, SD) years	47.6 (13.9)
Gender (Count, %)	
Male	19 (73.1%)
Female	7 (26.9%)
Primary Occupation (Count, %)	
Clinician	20 (76.9%)
Research Staff/ Methodologist	3 (11.5%)
Resident	3 (11.5%)
Highest Degrees Held (Count, %)	
MD	12 (46.2%)
MD, MSc/MPH	8 (30.8%)
PhD	3 (11.5%)
MD, PhD	2 (7.7%)
BSc	1 (3.9%)
Years in Practice (Mean, SD)	19.5 (14.3)
Previous involvement in an NMA? (Count, %)	
Yes	11 (42.3%)
No	15 (57.7%)
Used an NMA to inform practice? (Count, %)	
Yes	17 (65.4%)
No	9 (34.6%)

SD: Standard Deviation, MD: Doctor of Medicine, MSc: Masters of Science, MPH: Masters of Public Health, PhD: Doctor of Philosophy, BSc: Bachelor of Science, NMA: Network Meta-Analysis.

Figure 1: Study Overview

Figure 2: Gradient Colour Variation

Legend

Footnote

- : no evidence

Reference Group = Placebo
Bold = statistically significant (p<0.05)
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 3: Stoplight Colour Version

Legend

Footnote

- : no evidence

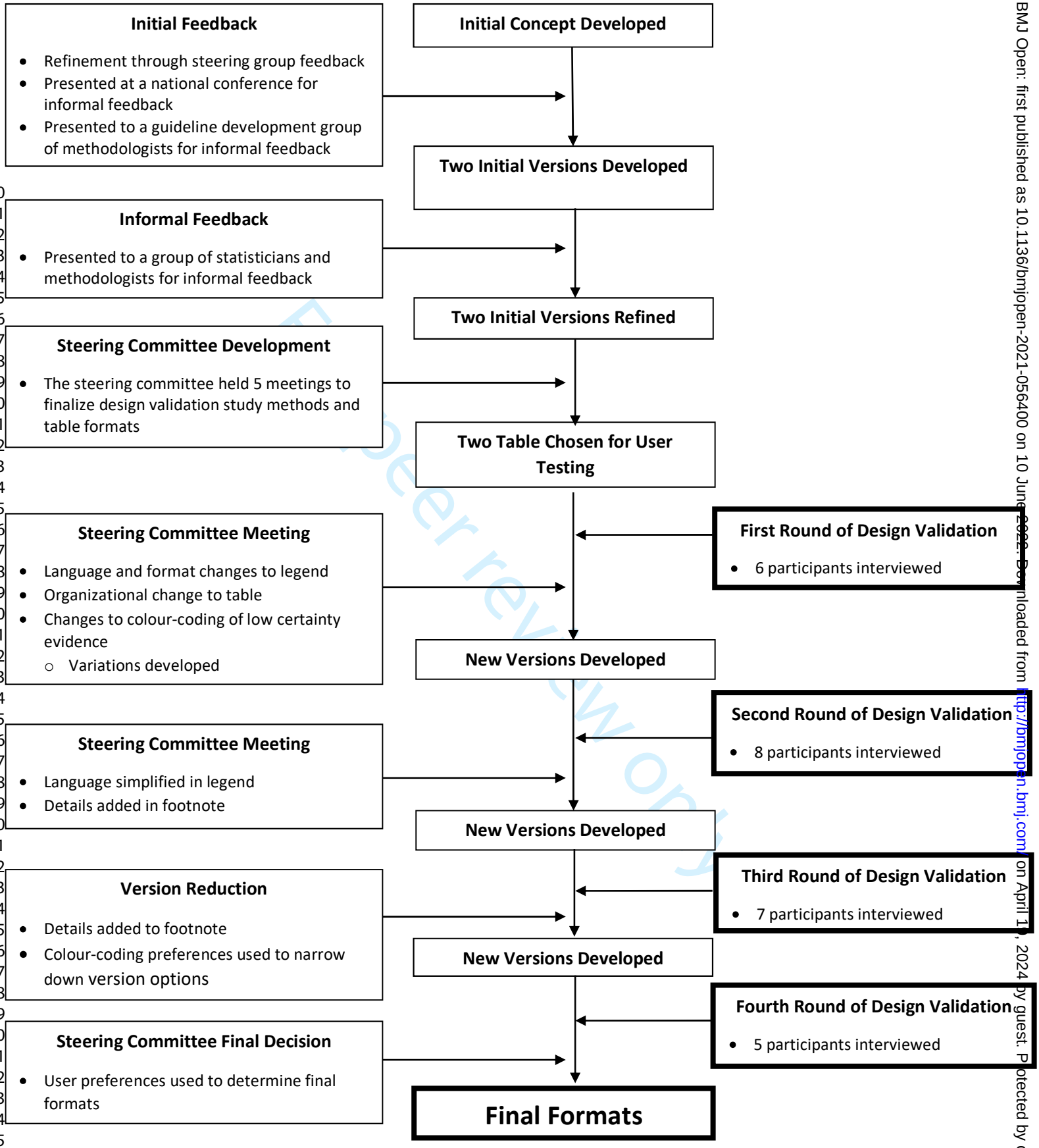
Reference Group = Placebo
Bold = statistically significant, p<0.05
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 4: Intervention Organizational Guide

Legend

Figure 5: Gradient Colour-Coding Legend

Figure 6: Stoplight Colour-Coding Legend

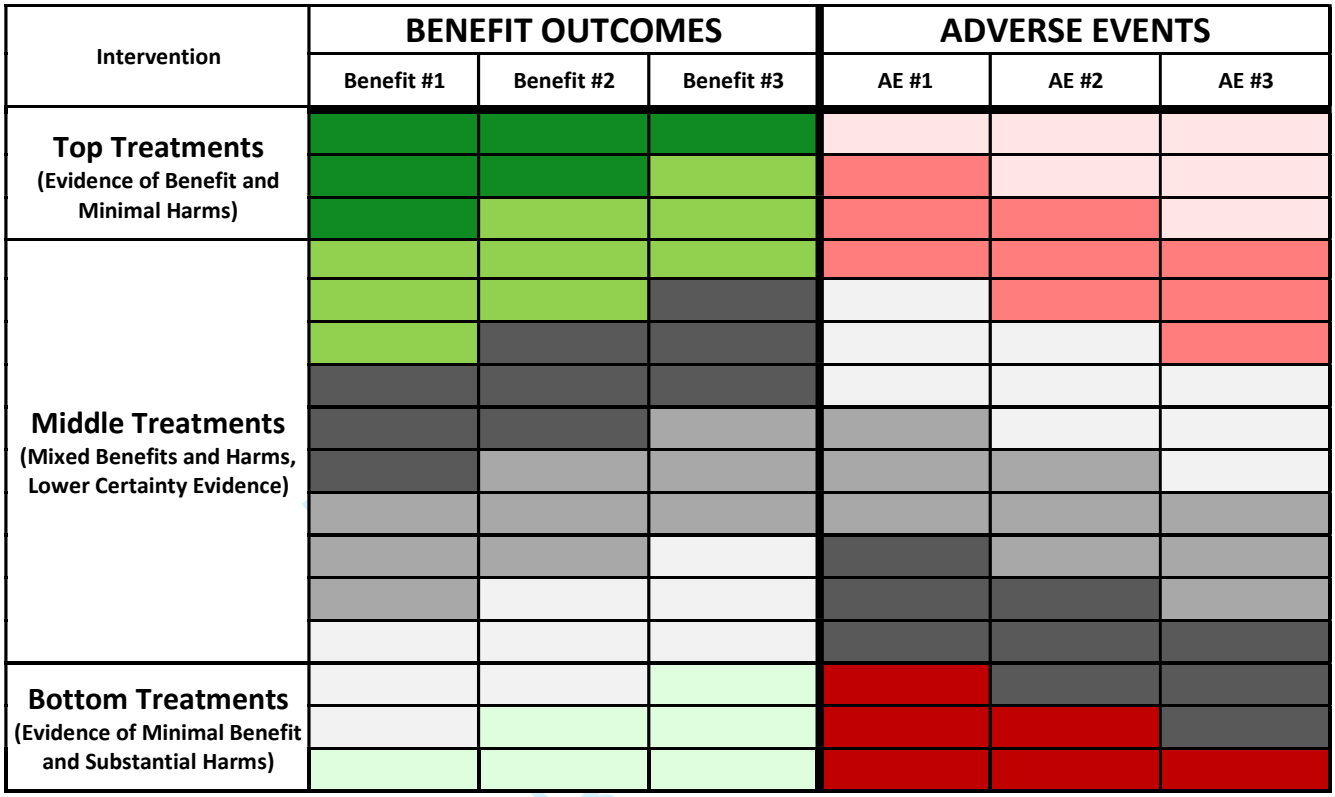


Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	**-1.02 (-1.64,-0.39)**	**-1.08 (-1.40,-0.75)**	**1.66 (1.16,2.16)**	**5.20 (2.03,13.33)**	**6.39 (3.48,11.75)**	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	**-0.93 (-1.49,-0.37)**	**-0.99 (-1.46,-0.52)**	**0.73 (0.17,1.30)**	3.24 (0.43,24.70)	**3.10 (1.39,6.91)**	**1.77 (1.33,2.35)**	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	**-1.03 (-1.82,-0.24)**	**-1.07 (-1.89,-0.24)**	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	**-1.11 (-2.00,-0.21)**	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	**3.72 (1.02,13.52)**	-	-	-
Topical NSAID + Menthol Gel	**-1.68 (-0.27,-3.09)**	-0.89 (-2.33,0.54)	-	-	**13.34 (3.30,53.92)**	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	**-1.94 (-2.90,-0.98)**	**-1.18 (-2.09,-0.28)**	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	**-1.59 (-2.52,-0.66)**	**-2.09 (-3.86,-0.32)**	**1.51 (0.42,2.61)**	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	**-1.75 (-2.68,-0.81)**	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	**167.71 (6.67,4217.10)**	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	**-2.92 (-5.41,-0.43)**	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	**32.08 (4.93,208.60)**	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	**7.99 (1.29,49.41)**	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	**-1.71 (-2.97, -0.46)**	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	**5.63 (2.84,11.16)**	**3.53 (1.92,6.49)**	-
Acetaminophen, Ibuprofen + Codeine	**-1.36 (-2.49,-0.23)**	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	**4.91 (1.45,16.61)**	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	**-3.52 (-4.99,-2.04)**	-	-	-	-	**59.38 (6.21,567.71)**	**5.73 (1.20,27.47)**	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	**6.72 (1.24,36.39)**	-

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than placebo and some other interventions**	**May be better than placebo and some alternatives**	**No more harmful than placebo**	**May be no more harmful than placebo**
INTERMEDIATE	**Better than placebo, but no better than any other interventions**	**May be better than placebo, but no better than other interventions**	**More harmful than placebo, but no worse than other interventions**	**May be more harmful than placebo, but no worse than other interventions**
AMONG THE WORST	**No better than placebo**	**May be no better than placebo**	**More harmful than placebo and some other interventions**	**May be more harmful than placebo and some alternatives**

Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	**-1.02 (-1.64,-0.39)**	**-1.08 (-1.40,-0.75)**	**1.66 (1.16,2.16)**	**5.20 (2.03,13.33)**	**6.39 (3.48,11.75)**	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	**-0.93 (-1.49,-0.37)**	**-0.99 (-1.46,-0.52)**	**0.73 (0.17,1.30)**	3.24 (0.43,24.70)	**3.10 (1.39,6.91)**	**1.77 (1.33,2.35)**	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	**-1.03 (-1.82,-0.24)**	**-1.07 (-1.89,-0.24)**	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	**-1.11 (-2.00,-0.21)**	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	**3.72 (1.02,13.52)**	-	-	-
Topical NSAID + Menthol Gel	**-1.68 (-0.27,-3.09)**	-0.89 (-2.33,0.54)	-	-	**13.34 (3.30,53.92)**	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	**-1.94 (-2.90,-0.98)**	**-1.18 (-2.09,-0.28)**	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	**-1.59 (-2.52,-0.66)**	**-2.09 (-3.86,-0.32)**	**1.51 (0.42,2.61)**	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	**-1.75 (-2.68,-0.81)**	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	**167.71 (6.67,4217.10)**	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	**-2.92 (-5.41,-0.43)**	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	**32.08 (4.93,208.60)**	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	**7.99 (1.29,49.41)**	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	**-1.71 (-2.97, -0.46)**	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	**5.63 (2.84,11.16)**	**3.53 (1.92,6.49)**	-
Acetaminophen, Ibuprofen + Codeine	**-1.36 (-2.49,-0.23)**	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	**4.91 (1.45,16.61)**	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	**-3.52 (-4.99,-2.04)**	-	-	-	-	**59.38 (6.21,567.71)**	**5.73 (1.20,27.47)**	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	**6.72 (1.24,36.39)**	-

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than placebo and some alternatives**	**May be better than placebo and some alternatives**	**No more harmful than placebo**	**May be no more harmful than placebo**
INTERMEDIATE	**Better than placebo, but no better than any alternatives**	**May be better than placebo, but no better than any alternatives**	**More harmful than placebo, but no worse than any alternatives**	**May be more harmful than placebo, but no worse than any alternatives**
AMONG THE WORST	**No better than placebo**	**May be no better than placebo**	**More harmful than placebo and some alternatives**	**May be more harmful than placebo and some alternatives**



	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST				
INTERMEDIATE				
AMONG THE WORST				

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**1**	**4**	**7**	**10**
INTERMEDIATE	**2**	**5**	**8**	**11**
AMONG THE WORST	**3**	**6**	**9**	**12**

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**1**	**4**	**7**	**10**
INTERMEDIATE	**2**	**5**	**8**	**11**
AMONG THE WORST	**3**	**6**	**9**	**12**

Appendix A: Open-Ended Interview Guide

Part 1: Introductions

1. Introductions

Part 2: NMA Familiarity

To begin, we would like to understand your current knowledge of NMA:

2. How familiar are you with NMA?
3. Have you ever been part of an NMA project?
    a. If so, what was your role in the NMA project?
4. Have you ever read an NMA?

Part 3: Review of the table format

The table I am showing you summarizes the results of an NMA that assessed acute pain management treatment options.

Please think aloud as you interpret this table

Prompts regarding the legend:

5. Do you find the language within the legend to be understandable? If not, what is confusing?
6. Do you have any feedback regarding the format of the legend?
    i. Do you have feedback regarding the coloring used?
    ii. Do you have feedback regarding the language used?
    iii. Do you have feedback regarding the indication of the certainty of evidence component of the legend?

Prompts regarding the results table:

7. Now that you have reviewed the legend in more detail, does the legend accurately and completely summarize the results table?
    a. If not, what could be changed?
8. Please provide any feedback you have regarding the results within the table
    a. Are the results easily understandable? If not, what is confusing or could be changed?
9. Do you have any feedback regarding the format of the table?
    a. Do you have feedback regarding the coloring used?
    b. Do you have feedback regarding the language used?
    c. Do you have feedback regarding the outcome reporting within the table?
    d. Do you have feedback regarding the indication of the certainty of evidence component of the results?
10. Please provide any other feedback that you may have regarding the table

Part 4: Assessing Participant Interpretation

Based on the results within the table, please describe how you interpret the findings?

Prompts regarding interpretation:

11. Based on both the benefits and the harms, which treatment(s) do you consider to be the optimal choice(s)?
12. Which treatment(s) do you believe are the least optimal choices? What information is important for you in deciding this?
13. How confident are you in your interpretation?
    a. Why are you/aren't you confident in your interpretation?
    b. What would aid in improving your interpretation?

Part 5: Closing Remarks

We would like to ask if you have any colleagues that may be interested in participating in this study. Following this interview, it would be great if we could connect with anyone who you believe may be able to provide valuable insights to this project.

Appendix B: Example Legend When Active Treatment is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than reference**	**May be better than reference**	**Less harmful than reference**	**May be less harmful than reference**
INTERMEDIATE	**No better than reference**	**May be no better than reference**	**No more harmful than reference**	**May be no more harmful than reference**
AMONG THE WORST	**Worse than reference**	**May be worse than reference**	**More harmful than reference**	**May be more harmful than reference**

Appendix C: Example Legend When Placebo (Or Any Sham/Null Treatment Effect) is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than placebo and some other interventions**	**May be better than placebo and some alternatives**	**No more harmful than placebo**	**May be no more harmful than placebo**
INTERMEDIATE	**Better than placebo, but no better than any other interventions**	**May be better than placebo, but no better than other interventions**	**More harmful than placebo, but no worse than other interventions**	**May be more harmful than placebo, but no worse than other interventions**
AMONG THE WORST	**No better than placebo**	**May be no better than placebo**	**More harmful than placebo and some other interventions**	**May be more harmful than placebo and some alternatives**

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

	Page/line no(s).
Title and abstract	
Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	1/3
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	2/28
Introduction	
Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	4/74
Purpose or research question - Purpose of the study and specific objectives or questions	4/86
Methods	
Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	6/112
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	5/102
Context - Setting/site and salient contextual factors; rationale**	5/105, 6/123, 7/149
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	6/122
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	5/114
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	7/134

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	9/186
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	5/105, 6/111
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	7/134, 8/163
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	8/163
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	8/171, 9/185

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	10/202
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	10/210

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	15/320, 16/328, 17/358
Limitations - Trustworthiness and limitations of findings	16/336

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	20/426
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	20/428

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014
DOI: 10.1097/ACM.0000000000000388

BMJ Open

Development and design validation of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-056400.R2
Article Type:	Original research
Date Submitted by the Author:	19-Mar-2022
Complete List of Authors:	Phillips, Mark ; McMaster University, Health Research Methods, Evidence, and Impact Sadeghirad, Behnam; McMaster University, Health Research Methods, Evidence, and Impact; McMaster University, Anesthesia Busse, Jason; McMaster University, Anesthesia; McMaster University, Health Research Methods, Evidence, and Impact Brignardello-Petersen, Romina ; McMaster University, Health Research Methods, Evidence, and Impact (HEI) Cuello-Garcia, Carlos ; McMaster University, Health Research Methods, Evidence, and Impact Kenji Nampo, Fernando ; Federal University of Latin-American Integration Guo, Yu Jia; McMaster University, Health Sciences Bzovsky, Sofia; McMaster University, Department of Surgery - Division of Orthopaedics Bannuru, Raveendhara R.; Tufts Medical Center, Center for Treatment Comparison and Integrative Analysis Thabane, Lehana; McMaster University, Health Research Methods, Evidence, and Impact; St. Joseph's Healthcare, Biostatistics Unit Bhandari, Mohit; Mcmaster University, Division of Orthopaedic Surgery; McMaster University, Health Research Methods, Evidence, and Impact Guyatt, Gordon; McMaster University, Health Research Methods, Evidence, and Impact
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Evidence based practice, Qualitative research
Keywords:	QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS, EDUCATION & TRAINING (see Medical Education & Training)

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Development and design validation of a novel network meta-analysis presentation tool for multiple outcomes: a qualitative descriptive study

Mark R Phillips[1], Behnam Sadeghirad[1,2], Jason W Busse[1,2], Romina Brignardello-Petersen[1], Carlos Cuello[1], Fernando Kenji Nampo[3], Yu Jia Guo, Sofia Bzovsky[5], Raveendhara R Bannuru[6], Lehana Thabane[1,7], Mohit Bhandari[1,5], Gordon H Guyatt[1]

1. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
2. Department of Anesthesia, McMaster University, Hamilton, ON, Canada
3. Latin-American Institute of Life and Nature Sciences/Evidence-Based Public Health Research Group, Federal University of Latin-American Integration, Foz do Iguassu, Brazil
4. Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
5. Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada
6. Center for Treatment Comparison and Integrative Analysis, Tufts Medical Center, Boston, MA, USA
7. Biostatistics Unit, St Joseph's Healthcare, Hamilton, ON, Canada

Correspondence: Mark Phillips phillimr@mcmaster.ca

1280 Main Street West, Hamilton, ON, CAN L8S 4L8

Keywords: Network Meta-Analysis, Design Validation, Qualitative Research, Statistics and Research Methods

Word Count: 5292

Abstract

Objective The GRADE working group recently developed an innovative approach to interpreting results from network meta-analyses (NMA) through minimally and partially contextualized methods; however, the optimal method for presenting results for multiple outcomes using this approach remains uncertain. We therefore developed and iteratively modified a presentation method that effectively summarizes NMA results of multiple outcomes for clinicians using this new interpretation approach.

Design Qualitative descriptive study

Setting A steering group of 7 individuals with experience in NMA and design validation studies developed two colour-coded presentation formats for evaluation. Through an iterative process, we assessed the validity of both formats to maximize their clarity and ease of interpretation.

Participants 26 participants including 20 clinicians who routinely provide patient care, 3 research staff/research methodologists, and 3 residents.

Main Outcome Measures Two team members used qualitative content analysis to independently analyze transcripts of all interviews. The steering group reviewed the analyses and responded with serial modifications of the presentation format.

Results To ensure that readers could easily discern the benefits and safety of each included treatment across all assessed outcomes, participants primarily focused on simple information presentations, with intuitive organizational decisions and colour coding. Feedback ultimately resulted in two presentation versions, each preferred by a substantial group of participants, and development of a legend to facilitate interpretation.

Conclusion Iterative design validation facilitated the development of two novel formats for presenting minimally or partially contextualized NMA results for multiple outcomes. These presentation approaches appeal to audiences that include clinicians with limited familiarity with NMAs.

Strengths and Limitations of this Study

- Extensive design validation in a targeted audience has validated the NMA presentation approaches within this study; something that has not been done for other presentation formats
- Structured qualitative research methodology has ensured accurate use of user feedback to develop and refine the NMA presentation formats
- Limited by the omission of some information within the presentation formats in order to achieve simplicity and interpretability, such as greater detail for individual outcomes, absolute effects, or specifics about the certainty of evidence assessments.
- The aforementioned information should still be included in NMA manuscripts, but cannot be feasibly fit within the presentation formats.

Introduction

Network meta-analysis (NMA) provides an increasingly popular approach to evidence synthesis that allows comparison between multiple competing treatment options within a single analysis.[1,2] Although NMA is an important tool for clinicians, patients, and other stakeholders, results involve multiple treatments and outcomes, and as a result are complex and difficult to interpret.[3]

Common methods for presenting NMA results include the use of forest plots, league tables, and surface under the cumulative ranking curve (SUCRA).[1,4] The key limitation with these options is that they can only provide results of a single outcome.[5] NMAs often compare multiple benefit and harm outcomes, resulting in challenges for NMA authors seeking to avoid presentation methods that are onerous for clinicians to review and challenging for them to understand.[6]

There are a number of novel approaches that have been suggested for presenting NMA results for multiple outcomes[7,8]; however, these approaches lack key information, present challenges to interpretation, and have not undergone design validation with their target audiences. While some previously suggested approaches have merit for a limited number of outcomes,[4,6,9–12] although not all taking certainty of evidence into account, they have serious limitations for simultaneous presentation of multiple outcomes.

Recently, the GRADE working group has suggested two variations on a new methodology that places interventions in categories from best to worst considering the estimates of effect and certainty of the evidence for each comparison.[13,14] We therefore developed interpretable presentation approaches for NMAs with multiple outcomes that builds on GRADE guidance and

effectively summarizes results for clinicians and other relevant audiences.

Methods

Study Design

A 7-member steering committee (MRP, BS, JWB, RB-P, CC, FKN, GHG) oversaw study design and implementation. The committee generated two initial presentation formats and chose a combination of large group sessions and individual design validation interviews to inform iterative modifications of the two initial formats. The presentation format consisted of treatment options in rows and outcomes in columns, with colour-coded shading of cells to identify the magnitude and certainty of the treatment effect in relation to the reference treatment. The steering committee developed the initial versions through a series of internal group discussions, which involved: determining the pertinent information for the presentation format to contain, options for how that information could be shown within a single presentation format, and draft presentation formats that may present this pertinent information. The group believed that the format should provide both relative treatment effects, as well as the certainty in those estimates for all outcomes, within a single presentation tool.

The steering committee developed initial versions of the presentation tool, which they then presented in separate large-group settings to gain outside insight. Initial large group testing with two groups of methodologists, graduate students in health research-focussed programs, and statisticians, as well as presentation at a national conference (2019 Canadian Pain Society annual scientific meeting), provided the foundational feedback for modifications of the initial presentation versions. After making iterative improvements from the group presentation

feedback, the steering committee began one-on-one interviews with clinicians to gain further insights for improvement. The steering committee reviewed input from four rounds of design validation individual interviews, iteratively modifying the formats after each round and presenting updated options of the presentation versions to subsequent participants.

For the user interviews, the committee chose a qualitative descriptive study approach that focuses on creating a close description of the information that participants provide.[15] This is ideal for design validation that, without interpretive direction, aims to optimize the understandability of a tool within the target population. This study involves human participants but an Ethics Committee exempted this study. After reviewing the protocol, the Hamilton Integrated Research Ethics Board (HiREB) committee and chair, judging the study to be a quality improvement investigation within the methodology and knowledge translation field, provided an exemption from formal ethics approval. Participants provided informed consent at the beginning of their interview. We followed, when applicable, the consolidated criteria for reporting qualitative research (COREQ) checklist in reporting our findings.[16]

Sampling and Recruitment

This study utilized purposeful sampling to identify participants who could provide information-rich interviews to inform the design validation process.[15,17] Target users for this study included academic and non-academic clinicians, research staff/research methodologists, and residents. The steering committee, through their professional contacts, provided a pool of initial possible participants that the principal investigator supplemented using snowball sampling technique. [18] Specifically, we asked individuals who agreed to participate for contact information of any

colleagues whom we could approach to interview. Prior to their interviews, each participant received information outlining the purpose of the study. Study recruitment ceased when data collection reached redundancy – the point at which there were no further refinements requested to improve the interpretability of the presentation formats.[18]

Data Collection

The principal investigator (MRP) conducted all design validation interviews either in-person or through video teleconferencing. Interviews followed a flexible interview guide (**Appendix A**) to leave the conversation open for participants to explore any topics they felt were relevant and important.[15] Throughout the study, the principal investigator iteratively updated the interview guide to explore areas of importance that emerged. Interviews began with a brief introduction to NMA methods, followed by questions regarding the participant's familiarity and experience with NMA. Participants then viewed the current versions of the NMA presentation formats and provided feedback. YJG or MRP transcribed all interviews verbatim. Transcripts were not returned to participants and interviewers did not conduct follow up interviews. The steering committee incorporated all feedback to arrive at two final presentation versions.

Patient and Public Involvement

This study did not include patient or public involvement.

NMA for Design Validation

The steering committee developed five core criteria to which the example NMA must adhere: (1) variability in quality of evidence (2) variability in magnitudes of effect; (3) assessment of both benefits and harms; (4) inclusion of both continuous and binary outcomes; and (5) including at least 5 outcomes and 5 interventions. Based on these criteria the steering committee chose, for design validation, a recent NMA that used a minimally contextualized approach to address acute pain management in patients experiencing non-low back acute musculoskeletal injuries.[19]

Based on the GRADE approach[13] this NMA categorized, for each benefit outcome, interventions as among those with the largest benefit, those with intermediate benefit, and those with the least benefit. For each harm outcome, they categorized interventions as among the least harmful, intermediate harm, and the most harmful. They then categorized interventions as those for which there was high or moderate certainty evidence, and those for which there was low or very low-quality evidence.[19] These results provided the example for design validation.

Data Analysis

Two reviewers (MP and SB) independently conducted data analysis, in duplicate, using a qualitative content analysis approach.[17] The study team recruited participants, collected data, and conducted data analysis in parallel. As new data became available, the reviewers coded and grouped similar phrases, patterns, and themes.[17] When discrepancies in feedback were identified, these would be noted and further elaborated on within future interviews. The feedback for this discrepancy would then be shared with the steering committee to review and identify if sufficient data had been captured to adequately determine a resolution for the discrepancy through consensus.[17] Data triangulation was utilized through multiple forms of data

collection, as both large group and individual interview sessions were used. Additionally, data triangulation was provided through two forms of data analysis: independent qualitative content analysis, and group deliberation through steering committee meetings.[17,20] The steering committee met four times over a period of 14 months to review the collected data and made iterative changes to the presentation formats as dictated by feedback, initially from large group presentations and subsequently from design validation. When analysis of the data provided actionable feedback, the reviewers presented their findings to the steering committee who ranked feedback as a "large change required", "moderate change required", or "minor change required" and then revised the presentation format(s) accordingly.

Subsequent participants provided input on the modified versions of the NMA results presentations. Participants commented regarding their interpretation of the data within the presentation format; the team considered study objectives met once participants consistently reported a clear interpretation of the results with no or minimal suggested modifications. Reviewers documented all changes to the presentation format in a study audit trail.[15,20] Reviewers conducted all qualitative analysis using RQDA software (R version 3.5.0).

Results

Study Sample

Two focus groups, both of which included methodologists, graduate students, and statisticians, participated in the initial large group testing: the first, a critical care guideline development group (GUIDE: https://guidecanada.org/) many of whose members have NMA expertise (65 attendees);

the second, a research group (CLARITY: http://www.clarityresearch.ca/) who meet regularly at McMaster University to discuss current methodological and statistical topics (20 attendees).

The design validation portion of this study included 26 participants of mean (standard deviation [SD]) age of 47.6 (13.9) years, 20 of whom were clinicians whose primary activity involved direct patient care (77%); 3 research staff/research methodologists (12%); and 3 residents (12%). Typical participants were male (73%) physicians in clinical practice for almost two decades (mean [SD]: 19.5 [14.3] years) with no prior involvement with conducting an NMA (58%) (**Table 1**).

Content Analysis Themes

Main themes that arose from the content analysis conducted on interview transcripts of participant interviews included "organizational", "language/terminology", "included information" and "colour options". Respondents also provided feedback regarding necessary details to include in the presentations' footnote. The following sections provide details regarding the most important feedback and how this feedback informed choices regarding presentation format. The fourth round of design validation resulted in minimal new information, resulting in two presentation versions that participants deemed satisfactory.

Final Presentation Versions

Ultimately, respondents proved equally enthusiastic about two options; the steering group, therefore, chose to offer both as alternative presentations. **Figure 1** summarizes the development process from conceptualization to the final presentation versions. We will refer to the presentation in **Figure 2** as the "colour gradient" version and the presentation in **Figure 3** as

the “stoplight” version. Each presentation has a legend and footnote with pertinent information that the design validation process demonstrated necessary to include.

Figure Organization

Design validation identified a number of key components that aid in interpreting presentation formats. Within the organizational theme, the use of a bolded vertical line to separate benefit and adverse event outcomes, as well as the header and results data (horizontal), proved desirable. Regarding the ordering of interventions from top to bottom in the rows, participants preferred ordering treatment options at the top with high/moderate certainty evidence of maximal benefit and minimal harm to those with high/moderate certainty evidence of minimal or no benefits and significant harms placed in the bottom rows. Respondents provided mixed feedback regarding the organization of the presentation within the middle section, with no consistent guidance that could be applied across all NMAs. This leaves the optimal ordering within the middle rows that include treatments that have low/very low certainty evidence, treatments with high/moderate certainty evidence of intermediate effects, and treatments with trade-offs between both large benefits and large harms, uncertain (or perhaps there is no single optimal ordering). **Figure 4** provides an overview of guidance regarding intervention order within the rows.

Presentation Terminology

Respondents indicated that the presentation should clearly and succinctly label outcomes with specification of the measure of treatment effect (e.g. odds ratios, mean differences) and that the

header of each column should include these labels. Participants had no strong preference regarding the terminology of "benefit" and "adverse events" outcome categories; options discussed included "effectiveness/efficacy outcomes" and "harms outcomes". Whatever option investigators choose, the terminology should remain consistent across the presentation, legend, and manuscript text.

Presentation Included Information

Participants considered the magnitude of treatment effect, confidence/credible intervals, certainty of evidence, and statistical significance to be the four important elements that should be included in each comparison cell. Possibilities explicitly discussed but rejected included sample size, patient characteristics, and heterogeneity/incoherence estimates. Respondents considered these items as important elements of the NMA, but felt they would be better suited within another section of the manuscript rather than within this summary presentation.

Footnote Included Information

Participants felt that footnotes should include: an indication of a dash representing no available evidence (- *: no evidence*); designation of the reference group (e.g. *Reference Group: Placebo*); and labelling of how statistical significance within the presentation is identified (i.e. *Bold = statistically significant, p<0.05*); as well as all abbreviations used within the presentation.

Legend Organization

Participants felt that benefit outcomes should be located in the left columns, with a bold vertical line separating the benefit and adverse event outcomes within the legend – similar to the structure of the main presentation. They also suggested a bold horizontal line separating the header from the legend in a similar format as within the main presentation. Within the benefit and adverse event sections, respondents preferred that high/moderate certainty evidence categories should be presented in the left column, and low/very low certainty in the right column. High and moderate certainty evidence, as well as low and very low certainty evidence were grouped together to simplify the presentation format into two groups (high/moderate, and low/very low), as participants perceived these groupings to hold similar weight in clinical decision making.

Legend Terminology

Participants encouraged the use of simple language within the legend. Participants preferred legend rows organized from "among the best" to "among the worst" vertically down the first column of the legend, with the middle category labelled as "intermediate". Terms such as "better" and "worse" were clearer to participants than terminology such as "statistically significant"; specifically, respondents favored "better than placebo" over "statistically significant over placebo".

The language used for our NMA example, in accordance with the minimally contextualized approach, contained treatments that were "better than placebo and some other interventions", "better than placebo, but no better than any other interventions", and "no better than placebo" for high/moderate certainty evidence of benefit outcomes. For high/moderate

certainty evidence of harm outcomes, the corresponding language was “no more harmful than placebo”, “more harmful than placebo, but no worse than other interventions”, and “more harmful than placebo and some other interventions”. Participants felt that, with respect to category of magnitude of effect low/very low certainty evidence descriptions should be the same as those of the high/moderate certainty evidence categories, with the included qualifier of “may be” at the beginning of the description of low to very low certainty evidence.

Gradient Colour-Coding

The gradient colour-coding scheme utilizes three shades of green for the high/moderate certainty benefit outcomes (**Figure 5: cells 1-3)**, and three shades of red for the high/moderate certainty adverse events (**Figure 5: cells 7-9**). The use of three-shade grey gradient for low/very low certainty evidence is consistent for both beneficial outcomes and adverse events (**Figure 5: cells 4-6, 10-12**). Participants preferred dark grey be used for the “among the worst” category (least beneficial or most harmful) and light grey be used for the “among the best” category (most beneficial or least harmful), when presenting low/very low certainty of evidence results.

Participants had clear views regarding the colour shades used in **Figure 5: cell 3** (among the least beneficial; high/moderate certainty), and **Figure 5: cell 7** (among the least harmful; high/moderate certainty)**:** because green is intuitively associated with positive results, they suggested caution regarding the use of a green shade for treatments categorized as “among the worst” in benefit outcomes supported by high/moderate certainty evidence (**Figure 5: cell 3**). Participants strongly suggested that the shade of green used in this cell should, as a result, be a pale and faint green. Similarly, **Figure 5: cell 7** utilizes a shade of red, despite being within the

"among the best" category in adverse events supported by high/moderate certainty evidence. Intuitively, participants noted that red is associated with poorer results. In order to avoid this inappropriate association, they suggested **Figure 5: cell 7** should utilize a pale and faint shade of red. Other options tested used white for **Figure 5: cell 3**, and **Figure 5: cell 7**; however, participants ultimately believed that faint colouring within the respective colour gradients was most appropriate and did not hinder interpretation.

Stoplight Colour-Coding

Because it dealt with the aforementioned concerns of the gradient colour-coding, participants also expressed enthusiasm for the stoplight colour-coding. The use of the same colour scheme across **Figure 6: cells 1-3** and **Figure 6: cells 7-9** simplifies the interpretation based on colour. Although the stoplight colour-coding addressed concerns with the gradient option, some participants preferred the gradient colour-coding due to the clear distinction between benefit and harms outcomes. Others also felt that the stoplight colour-coding looked distracting due to the inclusion of 3 bold colours, while the gradient colour-coding reserves bold colours that "stand out" for the comparisons with large benefits or large harms.

Discussion

The GRADE working group has developed methodologically coherent and innovative approaches to rating treatments within NMAs, including both benefits and harms, as "among the best", "intermediate" and "among the worst".[13,14] This may represent an important advance in the interpretation of the results of NMAs for clinicians using findings to guide clinical care. Clinicians,

however, need to apply this rating for all outcomes of importance to patients. Rigorously developed, user-friendly, intuitive, and tested approaches to simultaneous presentation of rated treatments across multiple outcomes has thus far been unavailable for either the new GRADE rating approach or prior approaches to enhance interpretability.[4–6,9,12]

This study has addressed existing limitations by developing presentation methods that summarize NMA results for multiple outcomes in clear and interpretable formats. Although previous methods may still be useful in presenting the results of individual outcomes in greater detail with certainty of evidence incorporated[4–6,9], the current presentation method allows for a clear and succinct summary of all outcomes considered within an NMA in a single presentation that our design validation has found both appealing and understandable to clinicians, many with limited prior exposure to NMAs.[6]

Strengths and Limitations

Extensive design validation in a targeted audience has validated our NMA presentation approaches, allowing future NMA's to enhance the ease with which clinicians can interpret their results. Additional strengths of this study include consultation with individuals involved in the process of developing and disseminating systematic reviews and clinical practice guidelines, and extensive design validation that included the careful selection of a study population that reflects the broader clinical audience who will be making use of NMA results. The use of structured qualitative research methods including duplicate data analysis allowed the accurate and appropriate incorporation of user feedback to be incorporated into iterative presentation development.

Our study does have limitations. First, although the simplicity of the developed presentations represents a strength, achieving that simplicity required the omission of data that some audiences may consider important.[6] For instance, the previous development of an NMA summary of findings table for individual outcomes provides greater detail for each treatment comparison that cannot feasibly fit within a multiple outcome presentation.[6] A particularly important omission may be the absolute effects of interventions that sometimes become crucial in trading off benefits and harms.[8] For this reason, authors may find it most appropriate to include both the multiple outcome presentation from this investigation, as well as additional outcome summaries suggested by other investigators.[4,6–11] This usability of this presentation tool was assessed specifically within the example NMA for pain management, which does not provide insights into the potential differences in usability for different future NMAs. Finally, we did not implement member checking. We did, however, employ data source triangulation to ensure that the findings of our study were robust.

Relation to Prior Work

Recent publications have addressed the issue of presenting NMA results for multiple outcomes, but have limitations that our proposal has addressed.[7,8] First, and crucially important, other options do not address the certainty of the evidence.[7,8] The Kilim plot provides a measure of the "strength of statistical evidence", which equates to the magnitude of the p-value.[8] Considerations of risk of bias, inconsistency, indirectness, publication bias, intransitivity, and incoherence may, however, reduce certainty in treatment effects with low p-values (which may or may not represent large effects). Additionally, the lack of design validation precludes

confidence in how target users will understand these formats. For these reasons, the presentation versions proposed in the current study represent important improvements on previous tools for reporting NMA results for multiple outcomes.

Choosing a Presentation Variation

Authors can, based on the appropriateness of the colour-coding and the corresponding categorization, choose between the two presentation versions in this manuscript. For example, the stoplight colour-coding variation may be most suitable when some treatments are better than the reference for some outcomes, while other treatments are worse for some outcomes. The three categories and explanations for benefit outcomes would then be "among the best – better than reference (colour: green)", "intermediate – same as reference (colour: yellow)", "among the worst – worse than reference (colour: red)". Intuitively, these descriptions and colours align. **Appendix B** provides an example of this scenario, with suggested details on the appropriate language to use within the legend.

The colour-gradient variation of the presentation may be most appropriate when the reference treatment is the worst (or best) treatment option across all outcomes. This would typically occur when placebo is the reference treatment, as placebo would likely be the worst treatment for benefit outcomes and the best treatment option for adverse event outcomes. The acute pain NMA used for our presentation formats fits this scenario. Although typically occurring with a placebo reference treatment, there may also be NMAs with other reference treatments that would intuitively follow this gradient colour-coding. **Appendix C** provides an example with suggested details on the appropriate language to use within the legend.

Additional Considerations

There is no single set of legend terminologies that universally apply to all NMAs, so authors must use their discretion to determine the most applicable and intuitive terminology. Authors may use the general guidance provided in this study in conjunction with categorization recommendations of the minimally or partially contextualized approach.[13,14] The minimally and partially contextualized approaches to NMA treatment categorization have the potential for more than three categories, which would require an adaptation to the colour schemes we identified. The appropriate title for this presentation format represents another consideration that this study did not test. We would encourage authors to be explicit in defining the patient population assessed within the presentation.

Methodologists and statisticians have long bemoaned an excessive focus on statistical significance, in particular through the use of p-values.[21–24] Notwithstanding, our participants felt it was important to highlight results indicating statistical significance, and our view is that there is considerable merit in the suggestion. Bolding or italics would be two possible ways of such highlighting, and the choice may depend on a journal's particular font suggestions.

A final consideration is the use of colours in the presentation methods. Participants believed that green, yellow, and red were the most intuitive colours for the table colour-coding; however, these colours may be problematic for colour-blind individuals. Authors who want to ensure colour-blind accessibility may consider using blue instead of green, and orange instead of red; although this was not specifically tested within this investigation.

Conclusion

This study utilized end-user design validation to develop easily interpretable presentation formats for reporting NMA results with multiple outcomes, with a focus both on relative magnitude of effects and certainty of evidence. If further empirical study verifies our finding that clinicians, and potentially patients - who are increasingly involved in clinical shared-decision making – who are naïve to NMAs find the presentation understandable and appealing, its wide implementation may enhance the impact and usefulness of NMAs.

Contributorship statement:

MRP, BS, JWB, RPB, CC, FKN, RRB, LT, MB, and GHG conceptualized the study

MRP, BS, JWB, and GHG recruited participants for the study.

MRP, YJG, and SB collected and analyzed data.

MRP, BS, JWB, RPB, CC, FKN, and GHG acted as the steering committee to interpret and implement data from participants.

MRP and GHG developed a first draft of the manuscript.

All authors reviewed, edited and approved the manuscript.

Competing Interests: Mohit Bhandari research grants from Pendopharm, Bioventus, and Acumed. All other authors have no conflicts of interest to disclose.

Funding: This study did not receive any funding.

Data sharing statement: Not Applicable

References

1. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med*. 2017;12(1):103-111. doi:10.1007/s11739-016-1583-7

2. Mills E, Thorlund K, Ioannidis J. Demystifying trial networks and network meta-analysis. *BMJ*. 2013;346:f2914. doi:10.1136/bmj.f2914

3. Ellis SG. Do we know the best treatment for in-stent restenosis via network meta-analysis (NMA)?: simple methods any interventionalist can use to assess NMA quality and a call for better NMA presentation. *JACC Cardiovasc Interv*. 2015;8(3):395-397. doi:10.1016/j.jcin.2014.11.012

4. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

5. Law M, Alam N, Veroniki AA, Yu Y, Jackson D. Two new approaches for the visualisation of models for network meta-analysis. *BMC Med Res Methodol*. 2019;19(1):61. doi:10.1186/s12874-019-0689-9

6. Yepes-Nuñez JJ, Li SA, Guyatt G, et al. Development of the summary of findings table for network meta-analysis. *J Clin Epidemiol*. 2019;115:1-13. doi:10.1016/j.jclinepi.2019.04.018

7. Daly CH, Mbuagbaw L, Thabane L, Straus SE, Hamid JS. Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol*. 2020;20(1):266. doi:10.1186/s12874-020-01128-2

8. Seo M, Furukawa TA, Veroniki AA, et al. The Kilim plot: A tool for visualizing network meta-analysis results for multiple outcomes. *Res Synth Methods*. Published online June 10, 2020. doi:10.1002/jrsm.1428

9. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PloS One*. 2013;8(10):e76654. doi:10.1371/journal.pone.0076654

10. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Med Res Methodol*. 2013;13:35. doi:10.1186/1471-2288-13-35

11. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. *J Clin Epidemiol*. 2014;67(6):672-680. doi:10.1016/j.jclinepi.2013.11.006

12. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev*. 2017;6(1):79. doi:10.1186/s13643-017-0473-z

13. Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ*. 2020;371. doi:10.1136/bmj.m3900

14. Brignardello-Petersen R, Izcovich A, Rochwerg B, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. *BMJ*. 2020;371:m3907. doi:10.1136/bmj.m3907

15. Morse JM. Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qual Health Res*. 2015;25(9):1212-1222. doi:10.1177/1049732315588501

16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007;19(6):349-357. doi:10.1093/intqhc/mzm042

17. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description – the poor cousin of health research? *BMC Med Res Methodol*. 2009;9(1):52. doi:10.1186/1471-2288-9-52

18. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant*. 2018;52(4):1893-1907. doi:10.1007/s11135-017-0574-8

19. Busse JW, Sadeghirad B, Oparin Y, et al. Management of Acute Pain From Non-Low Back, Musculoskeletal Injuries : A Systematic Review and Network Meta-analysis of Randomized Trials. *Ann Intern Med*. 2020;173(9):730-738. doi:10.7326/M19-3601

20. Maher C, Hadfield M, Hutchings M, Eyto A de. Ensuring Rigor in Qualitative Data Analysis: A Design Research Approach to Coding Combining NVivo With Traditional Material Methods. *Int J Qual Methods*. Published online July 10, 2018. doi:10.1177/1609406918786362

21. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3

22. Gagnier JJ, Morgenstern H. Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing. *J Bone Joint Surg Am*. 2017;99(18):1598-1603. doi:10.2106/JBJS.16.01314

23. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

24. Phillips M. Letter to the Editor: Editorial: Threshold P Values in Orthopaedic Research-We Know the Problem. What is the Solution? *Clin Orthop*. 2019;477(7):1756-1758. doi:10.1097/CORR.0000000000000827

Tables and Figures

Table 1: Participant Demographics: n=26

Demographic	Value
Age (Mean, SD) years	47.6 (13.9)
Gender (Count, %)	
Male	19 (73.1%)
Female	7 (26.9%)
Primary Occupation (Count, %)	
Clinician	20 (76.9%)
Research Staff/ Methodologist	3 (11.5%)
Resident	3 (11.5%)
Highest Degrees Held (Count, %)	
MD	12 (46.2%)
MD, MSc/MPH	8 (30.8%)
PhD	3 (11.5%)
MD, PhD	2 (7.7%)
BSc	1 (3.9%)
Years in Practice (Mean, SD)	19.5 (14.3)
Previous involvement in an NMA? (Count, %)	
Yes	11 (42.3%)
No	15 (57.7%)
Used an NMA to inform practice? (Count, %)	
Yes	17 (65.4%)
No	9 (34.6%)

SD: Standard Deviation, MD: Doctor of Medicine, MSc: Masters of Science, MPH: Masters of Public Health, PhD: Doctor of Philosophy, BSc: Bachelor of Science, NMA: Network Meta-Analysis.

Figure 1: Study Overview

Figure 2: Gradient Colour Variation

Legend

Footnote

- : no evidence

Reference Group = Placebo
Bold = statistically significant (p<0.05)
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 3: Stoplight Colour Version

Legend

Footnote

- : no evidence

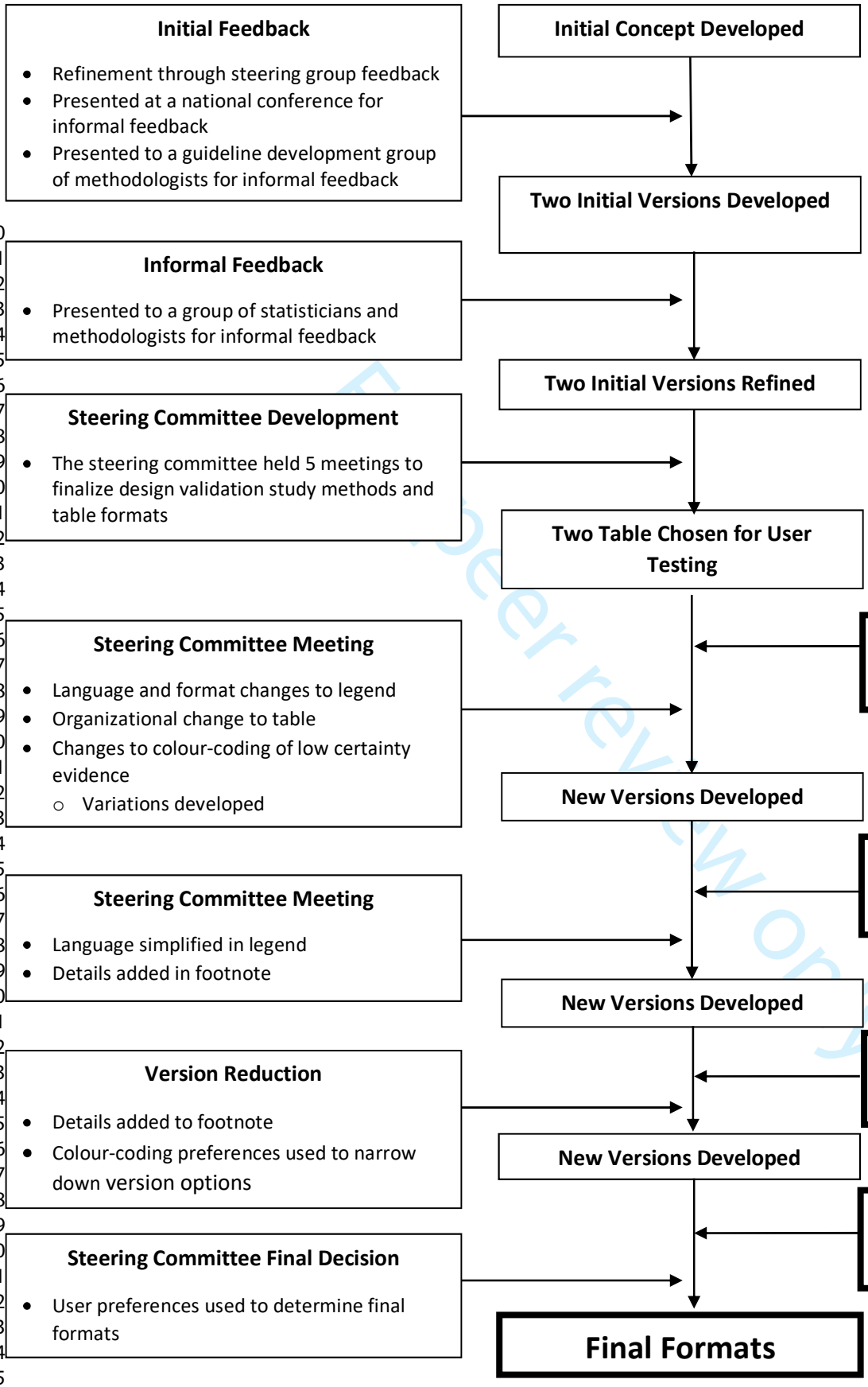
Reference Group = Placebo
Bold = statistically significant, p<0.05
MD: Mean Difference
OR: Odds Ratio
CI: Confidence Interval
h: hours
d: days
tx: treatment
AE: adverse event
NSAID: non-steroidal anti-inflammatory drug
TENS: transcutaneous electrical nerve stimulation

Figure 4: Intervention Organizational Guide

Legend

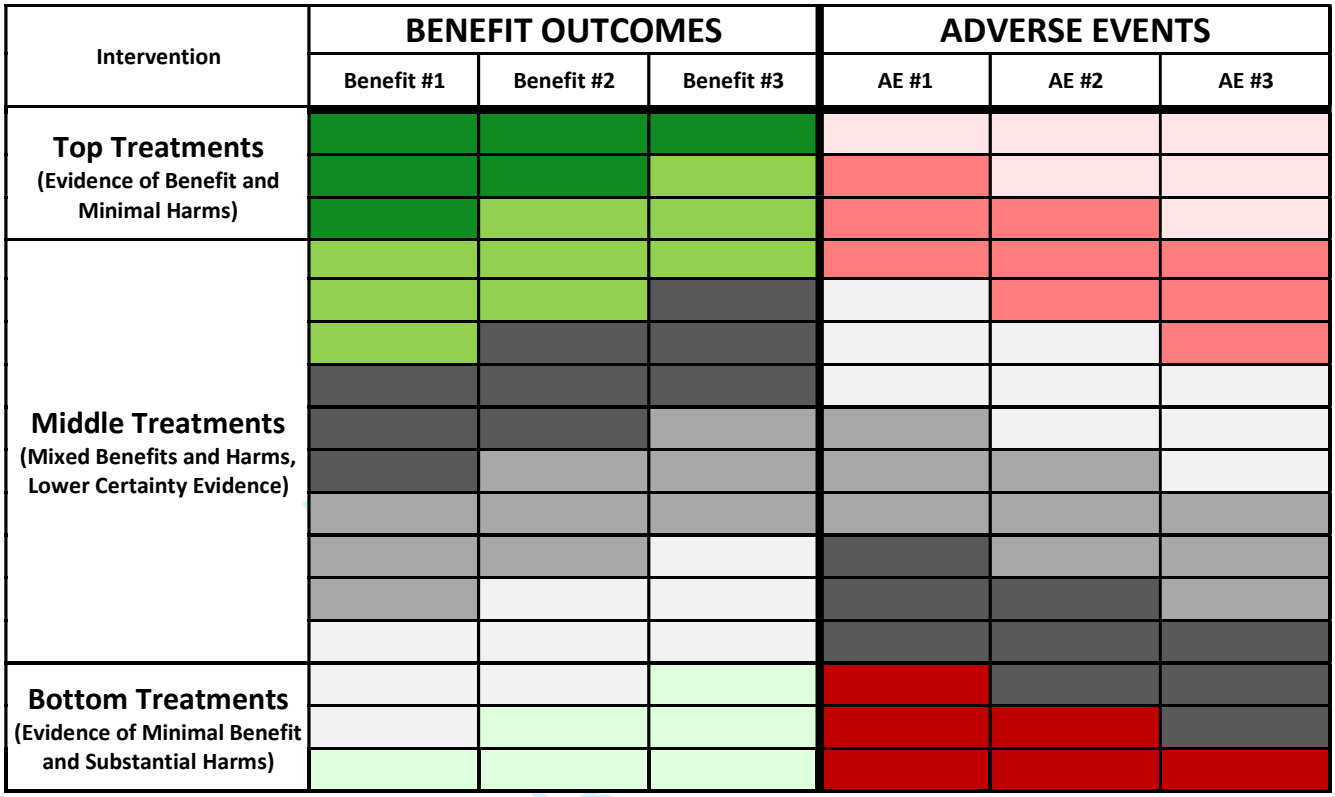
Figure 5: Gradient Colour-Coding Legend

Figure 6: Stoplight Colour-Coding Legend



Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	**-1.02 (-1.64,-0.39)**	**-1.08 (-1.40,-0.75)**	**1.66 (1.16,2.16)**	**5.20 (2.03,13.33)**	**6.39 (3.48,11.75)**	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	**-0.93 (-1.49,-0.37)**	**-0.99 (-1.46,-0.52)**	**0.73 (0.17,1.30)**	3.24 (0.43,24.70)	**3.10 (1.39,6.91)**	**1.77 (1.33,2.35)**	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	**-1.03 (-1.82,-0.24)**	**-1.07 (-1.89,-0.24)**	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	**-1.11 (-2.00,-0.21)**	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	**3.72 (1.02,13.52)**	-	-	-
Topical NSAID + Menthol Gel	**-1.68 (-0.27,-3.09)**	-0.89 (-2.33,0.54)	-	-	**13.34 (3.30,53.92)**	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	**-1.94 (-2.90,-0.98)**	**-1.18 (-2.09,-0.28)**	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	**-1.59 (-2.52,-0.66)**	**-2.09 (-3.86,-0.32)**	**1.51 (0.42,2.61)**	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	**-1.75 (-2.68,-0.81)**	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	**167.71 (6.67,4217.10)**	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	**-2.92 (-5.41,-0.43)**	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	**32.08 (4.93,208.60)**	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	**7.99 (1.29,49.41)**	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	**-1.71 (-2.97, -0.46)**	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	**5.63 (2.84,11.16)**	**3.53 (1.92,6.49)**	-
Acetaminophen, Ibuprofen + Codeine	**-1.36 (-2.49,-0.23)**	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	**4.91 (1.45,16.61)**	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	**-3.52 (-4.99,-2.04)**	-	-	-	-	**59.38 (6.21,567.71)**	**5.73 (1.20,27.47)**	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	**6.72 (1.24,36.39)**	-

Intervention	BENEFIT OUTCOMES					ADVERSE EVENTS		
	Pain ≤ 2 h post-tx	Pain 1 to 7 d post-tx	Physical function	Treatment satisfaction	Symptom relief	GI-related AE's	Neurologic AE's	Dermatologic AE's
	MD (95% CI)	MD (95% CI)	MD (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Topical NSAID	**-1.02 (-1.64,-0.39)**	**-1.08 (-1.40,-0.75)**	**1.66 (1.16,2.16)**	**5.20 (2.03,13.33)**	**6.39 (3.48,11.75)**	1.14 (0.65,2.01)	1.18 (0.51,2.74)	0.78 (0.52,1.15)
Oral NSAID	**-0.93 (-1.49,-0.37)**	**-0.99 (-1.46,-0.52)**	**0.73 (0.17,1.30)**	3.24 (0.43,24.70)	**3.10 (1.39,6.91)**	**1.77 (1.33,2.35)**	1.02 (0.65,1.59)	1.33 (0.43,4.09)
Acetaminophen	**-1.03 (-1.82,-0.24)**	**-1.07 (-1.89,-0.24)**	0.90 (-0.27,2.61)	2.43 (0.18,32.70)	2.73 (0.90,8.27)	0.50 (0.06,4.38)	-	-
Acetaminophen + Diclofenac	**-1.11 (-2.00,-0.21)**	-1.09 (-2.20,0.01)	-	3.45 (0.18,66.96)	**3.72 (1.02,13.52)**	-	-	-
Topical NSAID + Menthol Gel	**-1.68 (-0.27,-3.09)**	-0.89 (-2.33,0.54)	-	-	**13.34 (3.30,53.92)**	2.35 (0.04,124.85)	1.22 (0.02,69.98)	0.53 (0.05,6.29)
TENS	**-1.94 (-2.90,-0.98)**	**-1.18 (-2.09,-0.28)**	0.68 (-0.20,1.57)	-	6.00 (0.78,46.36)	1.25 (0.14,11.01)	1.12 (0.13,9.98)	1.18 (0.13,11.03)
Specific acupressure	**-1.59 (-2.52,-0.66)**	**-2.09 (-3.86,-0.32)**	**1.51 (0.42,2.61)**	0.50 (0.04,6.49)	2.54 (0.52,12.38)	0.80 (0.02,41.67)	0.80 (0.01,42.60)	0.80 (0.01,45.60)
Manipulation	**-1.75 (-2.68,-0.81)**	0.40 (-1.71,2.51)	0.09 (-1.06,0.87)	-	**167.71 (6.67,4217.10)**	0.50 (0.01,31.30)	1.41 (0.03,78.76)	-
Acetaminophen + Chlorzoxazone	-	**-2.92 (-5.41,-0.43)**	-	-	-	0.35 (0.01,10.59)	-	-
Laser therapy	-	-1.04 (-2.28,0.19)	-	-	**32.08 (4.93,208.60)**	0.49 (0.01,24.85)	0.49 (0.01,25.41)	0.49 (0.01,27.21)
Mobilization	-	3.40 (-0.05,6.85)	0.12 (-0.59,0.83)	2.07 (0.07,58.49)	**7.99 (1.29,49.41)**	0.93 (0.02,47.12)	0.93 (0.02,48.18)	0.93 (0.02,51.60)
Acetaminophen + Opioid	-0.52 (-1.47,0.43)	**-1.71 (-2.97, -0.46)**	-	2.50 (0.14,44.86)	1.47 (0.55,3.91)	**5.63 (2.84,11.16)**	**3.53 (1.92,6.49)**	-
Acetaminophen, Ibuprofen + Codeine	**-1.36 (-2.49,-0.23)**	-	-	-	-	-	-	-
Acetaminophen + Ibuprofen	-0.70 (-1.62,0.22)	-1.18 (-2.74,0.38)	-	-	3.62 (0.99,13.14)	-	-	-
Non-Specific Acupressure	-0.05 (-0.99,0.89)	-0.18 (-1.91,1.55)	-0.18 (-1.32,0.96)	0.44 (0.03,5.76)	1.80 (0.36,9.03)	0.85 (0.02,44.76)	0.85 (0.02,45.76)	0.85 (0.01,48.97)
Exercise	-	-0.81 (-2.64,1.02)	-0.43 (-1.00,0.14)	3.50 (0.21,59.42)	0.84 (0.31,2.29)	1.04 (0.06,17.06)	1.08 (0.07,17.95)	1.08 (0.06,18.84)
Cyclobenzaprine	-	-2.03 (-4.11,0.06)	-	-	-	0.64 (0.03,15.74)	1.95 (0.20,18.88)	-
Supervised Rehab	-	0.96 (-0.35,2.27)	0.24 (-0.59,1.07)	2.25 (0.15,34.07)	5.09 (0.84,30.78)	1.06 (0.02,54.49)	1.06 (0.02,55.71)	1.06 (0.02,59.65)
Ibuprofen + Cyclobenzaprine	-1.05 (-2.63,0.53)	-1.51 (-3.06,0.04)	-	5.52 (0.21,147.01)	-	1.10 (0.13,9.42)	**4.91 (1.45,16.61)**	-
Menthol Gel	-	-1.14 (-2.28,0.00)	0.70 (-0.61,2.02)	-	-	-	-	1.00 (0.11,8.91)
Ultrasound	-	-0.40 (-2.46,1.66)	-	-	-	-	-	-
Glucosamine	-	-0.10 (-1.89,1.69)	-	-	-	-	-	-
Phenyramidol	-	-	-	-	-	-	0.32 (0.01,8.45)	-
Massage therapy	-0.70 (-1.90,0.50)	-	-	-	-	-	-	-
Education	-	-	0.10 (-0.67,0.87)	-	0.93 (0.39,2.24)	-	-	-
Acetaminophen, Ibuprofen + Oxycodone	-0.94 (-2.27,0.38)	-	-	-	-	-	-	-
Fentanyl	**-3.52 (-4.99,-2.04)**	-	-	-	-	**59.38 (6.21,567.71)**	**5.73 (1.20,27.47)**	-
Tramadol	0.95 (-0.80,2.70)	-	-	-	-	5.98 (0.33,108.25)	**6.72 (1.24,36.39)**	-



	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**1**	**4**	**7**	**10**
INTERMEDIATE	**2**	**5**	**8**	**11**
AMONG THE WORST	**3**	**6**	**9**	**12**

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**1**	**4**	**7**	**10**
INTERMEDIATE	**2**	**5**	**8**	**11**
AMONG THE WORST	**3**	**6**	**9**	**12**

Appendix A: Open-Ended Interview Guide

Part 1: Introductions

1. Introductions

Part 2: NMA Familiarity

To begin, we would like to understand your current knowledge of NMA:

2. How familiar are you with NMA?
3. Have you ever been part of an NMA project?
    a. If so, what was your role in the NMA project?
4. Have you ever read an NMA?

Part 3: Review of the table format

The table I am showing you summarizes the results of an NMA that assessed acute pain management treatment options.

Please think aloud as you interpret this table

Prompts regarding the legend:

5. Do you find the language within the legend to be understandable? If not, what is confusing?
6. Do you have any feedback regarding the format of the legend?
    i. Do you have feedback regarding the coloring used?
    ii. Do you have feedback regarding the language used?
    iii. Do you have feedback regarding the indication of the certainty of evidence component of the legend?

Prompts regarding the results table:

7. Now that you have reviewed the legend in more detail, does the legend accurately and completely summarize the results table?
    a. If not, what could be changed?
8. Please provide any feedback you have regarding the results within the table
    a. Are the results easily understandable? If not, what is confusing or could be changed?
9. Do you have any feedback regarding the format of the table?
    a. Do you have feedback regarding the coloring used?
    b. Do you have feedback regarding the language used?
    c. Do you have feedback regarding the outcome reporting within the table?
    d. Do you have feedback regarding the indication of the certainty of evidence component of the results?
10. Please provide any other feedback that you may have regarding the table

Part 4: Assessing Participant Interpretation

Based on the results within the table, please describe how you interpret the findings?

Prompts regarding interpretation:

11. Based on both the benefits and the harms, which treatment(s) do you consider to be the optimal choice(s)?
12. Which treatment(s) do you believe are the least optimal choices? What information is important for you in deciding this?
13. How confident are you in your interpretation?
    a. Why are you/aren't you confident in your interpretation?
    b. What would aid in improving your interpretation?

Part 5: Closing Remarks

We would like to ask if you have any colleagues that may be interested in participating in this study. Following this interview, it would be great if we could connect with anyone who you believe may be able to provide valuable insights to this project.

Appendix B: Example Legend When Active Treatment is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than reference**	**May be better than reference**	**Less harmful than reference**	**May be less harmful than reference**
INTERMEDIATE	**No better than reference**	**May be no better than reference**	**No more harmful than reference**	**May be no more harmful than reference**
AMONG THE WORST	**Worse than reference**	**May be worse than reference**	**More harmful than reference**	**May be more harmful than reference**

Appendix C: Example Legend When Placebo (Or Any Sham/Null Treatment Effect) is Reference

	BENEFIT OUTCOMES		ADVERSE EVENTS	
	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence	High/Moderate Certainty Evidence	Low/Very Low Certainty Evidence
AMONG THE BEST	**Better than placebo and some other interventions**	**May be better than placebo and some alternatives**	**No more harmful than placebo**	**May be no more harmful than placebo**
INTERMEDIATE	**Better than placebo, but no better than any other interventions**	**May be better than placebo, but no better than other interventions**	**More harmful than placebo, but no worse than other interventions**	**May be more harmful than placebo, but no worse than other interventions**
AMONG THE WORST	**No better than placebo**	**May be no better than placebo**	**More harmful than placebo and some other interventions**	**May be more harmful than placebo and some alternatives**

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

	Page/line no(s).
Title and abstract	
Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	1/3
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	2/28
Introduction	
Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	4/74
Purpose or research question - Purpose of the study and specific objectives or questions	4/86
Methods	
Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	6/112
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	5/102
Context - Setting/site and salient contextual factors; rationale**	5/105, 6/123, 7/149
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	6/122
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	5/114
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	7/134

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	9/186
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	5/105, 6/111
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	7/134, 8/163
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	8/163
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	8/171, 9/185

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	10/202
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	10/210

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	15/320, 16/328, 17/358
Limitations - Trustworthiness and limitations of findings	16/336

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	20/426
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	20/428

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014
DOI: 10.1097/ACM.0000000000000388