A randomised trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement

Steven MacLennan[1*], Jamie Kirkham[2], Thomas B. L. Lam[1,3,] Paula R. Williamson[2]

1 Academic Urology Unit, University of Aberdeen

2 Department of Biostatistics, University of Liverpool

3 Department of Urology, Aberdeen Royal Infirmary

*Corresponding author Tel: +44 (0)1224 438123, Email: steven.maclennan@abdn.ac.uk

Abstract

Objective: To explore the impact of different feedback strategies on 1) subsequent agreement and 2) variability in Delphi studies.

Study Design and Setting: A two-round Delphi survey, with a list of outcomes generated from the results of a systematic review and interviews, was undertaken whilst developing a core outcomes set for prostate cancer including two stakeholder groups (health professionals and patients). Seventy-nine outcomes were scored on a scale of one (not important) to nine (critically important). Participants were randomised in round two to receive round one feedback from: peers only, multiple stakeholders separately, or multiple stakeholders combined.

Results: Agreement on outcomes retained for all feedback groups was high (peer: 92%, multiple separate: 90%, multiple combined: 84%). There were no statistically significant reduction in variability for peer versus multiple-separate (0.016 (-0.035, 0.067); p=0.529), or multiple-separate versus multiple-combined feedback (0.063 (-0.003, 0.129); p=0.062). Peer feedback statistically significantly reduced variability compared to multiple-combined feedback (0.079 (0.001, 0.157); p=0.046).

Conclusions: We found no evidence of a difference between different feedback strategies in terms of the number of outcomes retained or reduction in variability of opinion. However, this may be explained by the high level of existing agreement in round one. Further methodological studies nested within Delphi surveys will help clarify the best strategy.

Keywords: Consensus methods; Core Outcome Set development; Delphi study; RCT; stakeholders; Feedback strategies

What is new?

Key findings: We found no meaningful differences between the various feedback groups
What this adds to what is known: Despite this, the results are not inconsistent with other exploratory research showing that multiple separate feedback benefits agreement and reduces variability. This study added a new feedback strategy (multiple combined) which has not been studied in previous research.
What is the implication, what should change now: Further nested research using similar methods to investigate the influence of different types of feedback will help clarify the best strategy to be used in future Delphi studies in the context of COS development

1. Introduction

A core outcome set (COS) is an agreed minimum set of outcomes which should be reported in all effectiveness trials of an intervention or condition. [1] Reaching consensus on the most important outcomes to measure in clinical trials for stakeholders with potentially diverse opinions, such as patients and health professionals, is central to maximising the efficiency of clinical trials of effectiveness. [2] If there is consensus on what outcomes ought to be measured, then heterogeneity in the range of outcomes reported will reduce, selective outcome reporting will be reduced, evidence synthesis will be easier to perform, and the results are likely to be more informative. The COMET initiative have promoted methods to facilitate achieving consensus on COS. [3] A transparent way to incorporate diverse opinions and move toward consensus is to use Delphi surveys. Delphi surveys use more than one round of a questionnaire, with feedback after the first round, to elicit opinion on, for example, how important the participants think each outcome is. A strength of the Delphi method is that because the questionnaires are completed by participants anonymously and in isolation, they are not prone to social influences such as dominant personalities or pressure to conform to the majority, or to agree with perceived experts, [4, 5] yet still give participants an opportunity to consider and revise their own opinions in light of what other participants think. Additionally, using online surveys, Delphi techniques are not limited by geography. [5] Around 30% of core outcome set development projects listed in the COMET database incorporate Delphi methods. [6]

When asking participants to re-score outcomes in the subsequent Delphi round(s), in addition to reminding the participant of their own score, there are a few options available with regard to the type of feedback given. These include showing participants a summary of what their own stakeholder group's scores were (peer only), showing them a summary of the other stakeholder group's scores also (multiple separate), or showing a combined summary of all participants' scores regardless of stakeholder group (multiple combined). Furthermore, the type of data used to summarise the feedback could be a measure of central tendency, such as mean or median scores, or distributions of

the number of participants choosing each score for each outcomes, such as a histogram. Mean or median scores are proposed to be generally easier to understand but may also mask important divergences in opinion. Showing a distribution of scores is less succinct, and may be harder to assimilate, but gives more detailed information on diverse opinions if they exist. [1]

Whilst there is some evidence from social psychology that different presentation of feedback between rounds may influence subsequent scores differently, [4] there is no guidance on the optimal strategy as yet. Brookes et al recently published exploratory research from three COS Delphis comparing responses of participants randomised to receive peer-only or multiple separate stakeholder feedback using mean scores from the previous round to communicate the information. [7] Their results suggested consistently and statistically significantly that multiple separate feedback increased agreement on the number of outcomes retrained by both stakeholder groups after round two and reduced variability between rounds compared to peer-only feedback.

In this study we report the results of a nested randomised controlled trial (RCT) comparing peer-only, multiple separate and multiple combined feedback. Formally investigating feedback strategy is a new methodological concept in Delphi studies used in COS development. It is still unclear which strategies might reduce the number of retained or reduce the variability and therefore we do not state directionality in the hypotheses we used.

The exploratory hypotheses tested were:

1. There is a difference in the number of outcomes retained after Delphi round two between peer-only, multiple single and multiple combined feedback (agreement).
2. There is a difference in the variability of outcome scores after Delphi round two between peer-only, multiple single and multiple combined feedback (variability).

2. Materials and Methods

This nested RCT was undertaken within the development of a protocol-driven COS for localised prostate cancer [8, 9] and the study was approved by the National Research Ethics Service (NRES) – North of Scotland Committee (reference 12/NS0042). The list of outcomes for inclusion in the Delphi questionnaire were identified through a systematic review of the literature and semi-structured interviews with men who had been treated for localised prostate cancer. Verbatim outcomes were recoded to common names. This generated a list of 79 outcomes which were used in the Delphi questionnaire. [9] Methodological details about the design and analysis of the systematic review, and semi-structured interviews and the methods for the nested RCT are available in the study protocol [8] and the main COS report [9]. An overview of the Delphi creation process can be seen in Figure 1. A list

and description of all 79 outcomes and can be viewed in appendix 2 of the main COS development paper. [9] We reported the main study in accordance with the Core Outcome Set-STAndards for Reporting (COS-STAR) Statement. [10]

All outcomes where entered into a bespoke online Delphi tool, written in C# using WebForms and a MySQL backend. Two stakeholder groups were invited to participate via email: patients (men who received any treatment type for localised prostate cancer) and health professionals (urological surgeons, uro-oncologists, medical oncologists and urological cancer nurse specialists). Participants were asked "How important are the following outcomes in making decisions regarding treatments?" The Delphi questionnaire was completed online. Participants were directed to score the importance of each outcome on a 9-point scale. This scale was adapted from GRADE [11] (i.e. 1-3 = not important; 4-6 = important; 7-9 = critical; together with an 'unable to score' option). All Delphi outcomes were accompanied by a lay description and all participants saw the same questionnaire and descriptions. All participants completing round one were invited to complete round two. In round two participants were shown their score from the previous round and were centrally randomised on a 1:1 ratio, using variable block randomisation, to be shown one of three additional feedback options:

1. The 'peer' group received feedback from their own stakeholder group only
2. The 'multiple separate' group received feedback from their own stakeholder group and the other stakeholder group separately
3. The 'multiple combined' group received feedback from their own stakeholder group and the other stakeholder group combined

Central randomisation via the online Delphi software enabled allocations to be concealed from the research team. Participants could not be blinded to which feedback group they were allocated to precisely because of the nature of the study, but the research team were blinded to allocation until the analysis stage. An example of what the participants were shown in round two is depicted in Figure 2. All outcomes were retained after round one and all participants saw and had the opportunity to rescore all outcomes in round two.

2.1 Sample size

This was an opportunistic sample and the statistical testing was intended to be exploratory.

2.2 Analysis 1: Agreement between stakeholders and between participants randomised to different feedback groups

To investigate hypothesis one (there is a difference in the number of outcomes retained after Delphi round two between peer only, multiple single and multiple combined feedback) we analysed the

numbers of outcomes the stakeholder groups agreed on. To do this, after round two, outcomes scored as critical (i.e. 7-9) by ≥70% of patients and HCPs separately AND not important (i.e. 1-3) by <15% of patients and HCPs separately after round one were considered 'retained'. Retained outcomes were identified for patients and health professionals separately. We then calculated the number of outcomes retained for a) both stakeholder groups (agreement); b) neither group (agreement); and c) each stakeholder group independently (i.e. discordant) for each of the three feedback groups.

2.3 Analysis 2: Differences in variability between participants randomised to different feedback groups

To investigate hypothesis two (there is a difference in the variability of item scores after Delphi round two between peer only, multiple single and multiple combined feedback), for each of the 79 items, we computed the standard deviation of scores across all participants (ignoring stakeholder group) for each feedback group. This was done separately for round one and round two and the reduction in variability between rounds calculated. The mean reductions in standard deviations were compared, across all items, using paired t-tests for the peer versus multiple separate, peer versus multiple combined, and multiple separate versus multiple combined feedback groups.

3. Results

Of the 153 patients and 110 health professionals invited to participate in the Delphi, 118 patients (77%) and 56 health professionals (51%) completed round 1 indicating that initial response rates were lower for health professionals than patients. Both groups had good retention for round two at 92% (109/118) and 88% (49/56) for patients and health professionals respectively. The mean (SD) scores for 13 participants dropping out after round one (6 (1.3)) did not differ markedly from those remaining in the study (5.8 (1.6)). The numbers randomised to each feedback group are shown in Figure 3.

The numbers and percentages of participants changing at least one score between round one and round two can be seen in Figure 4. Regardless of stakeholder and randomised group, a high percentage of participants changed their scores from round one to round two (≥ 88%). In instances where a participant changed scores between rounds a free-text pop-up box asked them to give a reason why. This was for any score change regardless of how large or small, and giving a reason was optional. Of the 146 who changed their score, 110 (75%) filled in the free text box to explain their reason for changing score. Patients (86/102; 84%) commented more often than health professionals (24/44; 55%). The qualitative responses were coded. Some participant's comments were coded with more than one category. The most frequently given reasons for score change, regardless of stakeholder group or feedback, were that participants had 'time to reflect' between rounds (56/110;

51%), or that knowledge of 'other's scores' had influenced them (36/110; 33%). Although the sample sizes are very small, it is interesting to note that of the health professionals who commented, those who saw peer only feedback were more likely justify score change with the 'influence of other's' scores (7/10; 70%) than those who saw multiple separate (3/7; 42%) or multiple combined feedback (4/7; 57%). No such differences were seen in the patient group. Some patients, however, stated that their perception of the outcome had changed since the previous iteration of the questionnaire.

3.1 Agreement between stakeholders and between participants randomised to different feedback groups

The number of outcomes retained and the number of outcomes where there was agreement or discordance after applying the pre-specified criteria is shown in table 1a. Percent agreement was high across all feedback groups but the peer feedback group (92%) retained marginally more outcomes compared to 87% in the multiple combined group, and 90% in the multiple separate group. The absolute percent difference was very low and ranged from 2% in the peer compared to multiple separate to 5% in the peer compared to multiple combined groups.

To explore this finding further we repeated the analysis for round one. The results are shown in table 1b which shows that agreement between stakeholder groups was already very good at the outset in this Delphi: it was 84% in the multiple combined group, 90% in the multiple separate group and 91% in the peer feedback group.

3.2 Differences in variability between participants randomised to different feedback groups

As table 2 shows Peer feedback (0.22 (0.19)) was marginally better at reducing the average variability in scores compared to multiple separate (0.20 (0.12)) or multiple combined (0.14 (0.25)) feedback. However, the magnitude of effects was very small. The results of the pairwise comparisons are shown in table 3. Peer only feedback was statistically significantly better at reducing variability compared to multiple combined feedback (0.079 (0.001, 0.157); p=0.046) but this was not felt to be a meaningful difference. There were no statistically significant differences showing that peer only was better than multiple separate feedback for reducing variability (0.016 (-0.035, 0.067); p=0.529)). Likewise for multiple separate compared to multiple combined feedback (0.063 (-0.003, 0.129); p=0.062).

4. Discussion

The nested methods study shows no evidence of a difference between different feedback strategies in terms of the number of outcomes retained nor in reducing variability. However, this may be

explained by the good agreement in round one. There are some limitations to the study. For example the small sample size in the healthcare professional group resulted in relatively small numbers allocated to each of the 3 feedback groups (n ~15). Another limitation is the fact that all of the patient participants, and most of the health professional participants were from and were from the United Kingdom. A larger and more geographically diverse participant group may have shown greater divergence in opinion at the outset. If there had been greater variability in opinion at baseline then we may have seen clearer differences in the influence of different feedback strategies. Also, our study asked participants to score 79 outcomes which, although necessary for the aim of the main core outcome set development project, may have been burdensome to score on three separate occasions. However, it is not clear how this may have affected scoring.

4.1 Comparison with other studies

Brooke's et al [7] also compared peer with multiple separate (but not multiple combined) feedback and the similarity in methods allows some comparison of the results across the two studies. Brookes et al explored the effect of peer versus multiple separate feedback in three separate Delphis: breast reconstruction surgery, surgery for colorectal cancer and surgery for oesophageal cancer. They consistently found that agreement was better when participants were shown multiple separate feedback as shown in table 4. With regards to variability, Brookes et al found that multiple separate feedback consistently and statistically significantly reduced variability, whereas our prostate cancer Delphi showed no evidence of a difference as presented in the Forest plot in figure 5.

It is important to note that in the colorectal Delphi and the oesophageal Delphi, where there was a more pronounced effect in favour of multiple separate feedback, there was poorer initial agreement between stakeholder groups. Whereas in the breast reconstruction Delphi, albeit still finding a statistically significant result in favour of multiple separate feedback, the results are more in keeping with our findings. These differences then may be explained by the comparatively better initial agreement between stakeholder groups in the breast Delphi and prostate Delphi. In summary, the 95% CIs in our prostate cancer Delphi do not disagree with Brookes et al's results, but suggest that differences may not be as large as Brookes et al found. However, more research is needed to clarify these trends.

Other methodological differences between the prostate Delphi and the three Delphis reported by Brookes et al are 1) the prostate Delphi tested three feedback strategies instead of two and had a smaller number of participants in each feedback allocation group; 2) the prostate Delphi retained all outcomes throughout all rounds of the Delphi, whereas the other three Delphis dropped outcomes not meeting scoring thresholds after round 1 (i.e. only outcomes which scored between 7-9 by 50%

or more participants and between 1-3 by less than 15% (done separately for each stakeholder group) were retained after round one in Brookes et al's Delphis); and 3) the prostate Delphi presented feedback as the percentage of participants scoring each outcome at each possible score from 1-9, whereas the other three presented feedback as mean scores. It is currently unclear how any of these methodological differences may influence results and further research is required to investigate these.

For future Delphi studies, an important point to bear in mind is that if agreement is already good between stakeholder groups, then the type of feedback given may not make any difference to the results. Conversely, in instances where initial agreement is poor, multiple separate feedback may be a better strategy to reach consensus. Also, a priori, we planned three rounds of Delphi which was time consuming to complete. Future studies could propose to analyse results after each round to assess agreement and, according to predefined thresholds of agreement, could use these results to decide whether is necessary to have subsequent Delphi rounds.

5. Conclusion

In summary, our study found no meaningful differences between the various feedback groups. Despite this, the results are not inconsistent with Brookes et al's finding that multiple separate feedback benefits agreement and reduces variability. Further nested research using similar methods to investigate the influence of different types of feedback will help clarify the best strategy to be used in future Delphi studies in the context of COS development.

References

1. Williamson, P. R. Altman, D. G. Bagley, H. Barnes, K. L. Blazeby, J. M. Brookes, S. T. Clarke, M. Gargon, E. Gorst, S. Harman, N. Kirkham, J. J. McNair, A. Prinsen, C. A. C. Schmitt, J. Terwee, C. B. Young, B. *The COMET Handbook: version 1.0.* Trials, 2017. **18**(Suppl 3).
2. Williamson, P.R., et al., *Developing core outcome sets for clinical trials: issues to consider.* Trials, 2012. **13**(1): p. 1.
3. COMET. *Home: Core Outcome Measures in Effectiveness Trials Initiative (COMET)*. Available from: http://www.comet-initiative.org/.
4. Rowe, G., Wright, G., and McColl, A. *Judgment change during Delphi-like procedures:The role of majority influence, expertise, and confidence.* Technological Forecasting & Social Change 2005. **72** p. 377–399.
5. Sinha, I.P., R.L. Smyth, and P.R. Williamson, *Using the Delphi technique to determine which outcomes to measure in clinical trials: recommendations for the future based on a systematic review of existing studies.* PLoS Med, 2011. **8**(1): p. e1000393.
6. Gorst, S.L., et al., *Choosing Important Health Outcomes for Comparative Effectiveness Research: An Updated Review and User Survey.* PLoS One, 2016. **11**(1): p. e0146444.
7. Brookes, S.T., et al., *Three nested randomized controlled trials of peer-only or multiple stakeholder group feedback within Delphi surveys during core outcome and information set development.* Trials, 2016. **17**(1): p. 1.

8. MacLennan, S., et al., *A core outcome set for localised prostate cancer effectiveness trials: protocol for a systematic review of the literature and stakeholder involvement through interviews and a Delphi survey.* Trials, 2015. **16**(1): p. 76-76.
9. MacLennan, S., et al., *A core outcome set for localised prostate cancer effectiveness trials.* BJU Int, 2017.
10. Kirkham, J.J., et al., *Core Outcome Set-STAndards for Reporting: The COS-STAR Statement.* PLoS Med, 2016. **13**(10): p. e1002148.
11. Guyatt, G.H., et al., *What is "quality of evidence" and why is it important to clinicians?* 2008.