Consensus Building in OMERACT: Recommendations for Use of the Delphi for Core Outcome Set Development

Susan Humphrey-Murto, Richard Crew, Beverley Shea, Susan J. Bartlett, Lyn March, Peter Tugwell, Lara J. Maxwell, Dorcas Beaton, Shawna Grosskleg, and Maarten de Wit

ABSTRACT. ***Objective.*** Developing international consensus on outcome measures for clinical trials is challenging. The following paper will review consensus building in Outcome Measures in Rheumatology (OMERACT), with a focus on the Delphi.

Methods. Based on the literature and feedback from delegates at OMERACT 2018, a set of recommendations is provided in the form of the OMERACT Delphi Consensus Checklist.

Results. The OMERACT delegates generally supported the use of the checklist as a guide. The checklist provides guidance for clearly outlining the multiple aspects of the Delphi process.

Conclusion. OMERACT is deeply committed to consensus building and these recommendations should be considered a work in progress. (First Release February 15 2019; J Rheumatol 2019; 46:1041–6; doi:10.3899/jrheum.181094)

Key Indexing Terms:
OMERACT DELPHI TECHNIQUE CONSENSUS RESEARCH DESIGN

From the University of Ottawa, Departments of Medicine and Innovation in Medical Education; Center for Global Health, University of Ottawa; Division of Rheumatology, Department of Medicine, School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa; Clinical Epidemiology Program, Ottawa Hospital Research Institute; Centre for Practice-Changing Research, Ottawa Hospital Research Institute; OMERACT, Ottawa; Institute for Work and Health, Toronto, Ontario; Department of Medicine, McGill University, Montreal, Quebec, Canada; University of Liverpool, Department of Biostatistics, COMET, Liverpool, UK; Department of Medicine, Johns Hopkins University, Baltimore, Maryland, USA; Sydney Medical School, University of Sydney, Sydney; Institute of Bone and Joint Research, Kolling Institute, Northern Sydney Local Health; Department of Rheumatology, Royal North Shore Hospital, St Leonards, Australia; Amsterdam University Medical Centre, Department of Medical Humanities, Amsterdam Public Health, Amsterdam, the Netherlands.

S. Humphrey-Murto, Associate Professor of Medicine, MD, MEd, University of Ottawa, Departments of Medicine and Innovation in Medical Education; R. Crew, BSc (Hons), University of Liverpool, Department of Biostatistics, COMET; B. Shea, PhD, Center for Global Health, University of Ottawa; S.J. Bartlett, PhD, Professor, Department of Medicine, McGill University, and Adjunct Professor, Department of Medicine, Johns Hopkins University; L. March, MBBS, MSc, PhD, FRACP, FAFPHM, Sydney Medical School, University of Sydney, and Institute of Bone and Joint Research, Kolling Institute, Northern Sydney Local Health, and Department of Rheumatology, Royal North Shore Hospital; P. Tugwell, MD, Division of Rheumatology, Department of Medicine, School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, and Clinical Epidemiology Program, Ottawa Hospital Research Institute; L.J. Maxwell, PhD, Centre for Practice-Changing Research, Ottawa Hospital Research Institute, and University of Ottawa; D. Beaton, BscOT, PhD, Senior Scientist, Institute for Work and Health; S. Grosskleg, OMERACT, and University of Ottawa; M. de Wit, PhD, Amsterdam University Medical Centre, Department of Medical Humanities, Amsterdam Public Health.

Address correspondence by Dr. S. Humphrey-Murto, The Ottawa Hospital-Riverside Campus, 1967 Riverside Dr., Ottawa, Ontario, K1H 7W9, Canada. E-mail: shumphrey@toh.on.ca

Accepted for publication December 6, 2018.

Outcome Measures in Rheumatology (OMERACT) is an international collaboration devoted to developing consensus on outcome measures for trials involving rheumatic diseases[1]. Consensus building is a crucial component of the process but may be challenging. As a result, the OMERACT 2018 meeting put "consensus" front and center to bring attention to our current practices and provide an opportunity to reflect and improve on them. This was done through 2 plenaries, formal voting, informal discussions, and training sessions for nominal group technique facilitators. The following paper will review consensus building in OMERACT and reflect on feedback received during the formal sessions, as well as informal discussions throughout the meeting. We also outline plans for future activities to improve the process. To limit the scope of this paper, we will focus primarily on the Delphi, a formal consensus method used in many OMERACT initiatives.

Consensus is at the core of OMERACT. Building consensus on core outcome sets (COS) for clinical trials in rheumatic diseases has numerous benefits such as reducing biased reporting, a more comprehensive assessment of efficacy, and better opportunities for comparison and metaanalysis. The key principles of consensus building in OMERACT can be summarized as follows: consensus is not simply "majority wins"; consensus must be evidence-based; all relevant groups must be represented; a face-to-face interaction must occur at some point in the process; and there is a formal iterative process to work toward consensus[2].

A majority vote does not guarantee consensus. It simply reflects that a majority is in favor, but there could be groups

who consider the outcome unacceptable. For OMERACT, consensus indicates that although the result of the process may not be everyone's preferred choice, the aim is to reach an agreement that all participants can accept as a "working arrangement." It is also noteworthy that OMERACT does not seek to force consensus because this ultimately leads to poor acceptance in the long term. With OMERACT, the evidence must be considered sufficient to support decisions, and when not sufficient, questions are added to the research agenda[2].

In OMERACT, the process has evolved over time to include a wide variety of participants from multiple continents with strong representation from patients, carers, clinicians, researchers, regulators, payers, and industry. OMERACT firmly believes that these groups provide a wider range of knowledge and experience and that the interaction between participants stimulates consideration of a broader range of options. Although a recognized strength of the process, this can also present a challenge. For example, during the OMERACT 2016 meeting, some patients perceived themselves to be underrepresented in numbers (10%) in the final voting on core outcome domains. If patients play a major role in the phase of generating and prioritizing core outcome domains, they should be adequately represented in the final stages of decision making. Finally, the OMERACT meetings occur biennially and are an integral part of the process, bringing participants face to face for several days so that healthy discussion and debate can occur.

A formal consensus method: The Delphi. Among different strategies used to work toward consensus, the Delphi is frequently chosen[3]. The Delphi is one part of the entire consensus process and is defined as a systematic means to measure and facilitate consensus[4]. It is used when empiric evidence is limited or contradictory and is based on the premise that accurate and reliable decisions can best be achieved by consulting a panel of experts and accepting group consensus[5]. At OMERACT, the Delphi method is used to prioritize critically important domains from an initial list of candidate outcomes that should be included in a COS[6,7]. The OMERACT Rheumatoid Arthritis Flare Group, as an example, described in detail how the Delphi process was used to gain consensus among several hundred international patients, clinicians, and others on a COS for measuring rheumatoid arthritis flares[8]. The Delphi can also be used to obtain consensus on a list of candidate instruments that should subsequently be studied for their psychometric properties.

The Delphi method involves sending out surveys over several rounds. Participants, who are anonymous, rate potential items, or in the OMERACT context, candidate domains for a COS. In the first round they may also generate new items/domains. Then, in the next round, participants receive feedback comparing their own scores to the distribution of scores from other groups. Each participant is provided with an opportunity to re-rate domains. Although not part of the formal Delphi process, the final "round" may involve ranking items to ensure arriving at a reasonable number of core outcome domains is obtained.

The Delphi has several advantages, including the ability to reach many geographically dispersed participants, and it provides anonymity, thus reducing the potential for dominant individuals to sway the group. Disadvantages include the inability to discuss areas where there is lack of agreement directly with other participants, and it can be labor-intensive to collate scores and distribute feedback between rounds. Delphi software may facilitate the collation of scores and OMERACT currently mandates a face-to-face meeting to ensure that discussion can occur.

Despite extensive use of the Delphi in many contexts, several concerns have been raised in the literature. Studies using the Delphi for selecting performance indicators for healthcare, for medical and nursing education, or for determining outcomes to measure in clinical trials, often fail to adequately report sufficient methodological detail. Examples include poor reporting of background information provided to participants, response rates for all rounds, level of anonymity, formal feedback between rounds, and the definition of consensus[6,9,10,11,12,13].

MATERIALS AND METHODS

To improve the use and reporting of the Delphi within OMERACT, a preliminary OMERACT Delphi checklist (Figure 1) was developed based on previous recommendations and expert input[5,6,9,10,11,12,14]. The experts included all the authors (n = 10) of this article, which consisted of a patient, rheumatologists, and researchers with extensive experience in consensus methods. This was presented to the delegates in addition to specific recommendations for consideration. Feedback was solicited using voting keypads during the 2 plenary sessions, and through discussions throughout the meeting. The research team reviewed both the formal and informal feedback and adapted the recommendations after extensive discussion.

RESULTS

The total number of delegates attending OMERACT 2018 was 170 and included 106 clinicians/researchers (62%), 17 patients (10%), 11 pharmaceutical representatives (7%), 33 fellows (19%), and 3 regulation authorities (2%). Feedback on specific aspects of the Delphi process were sought. Regarding the number of items sent in Delphi surveys, based on previous experience, the consensus team (all authors) recommended including a maximum of 50 items (potential domains) in round 1. The delegates selected 70 as a more realistic number. There is no literature to support either recommendation. In the study by Boulkedid, *et al* that reported on 80 studies using the Delphi, the initial number of items ranged from 11 to 767, with a median of 59[9].

When considering what type of feedback should be provided to participants between rounds, a small majority (24/41, 58%) agreed that feedback between rounds should include individuals' scores for each item and the distribution of votes by participant group. Some, however, preferred to view aggregated feedback (11/41, 27%). The few studies that

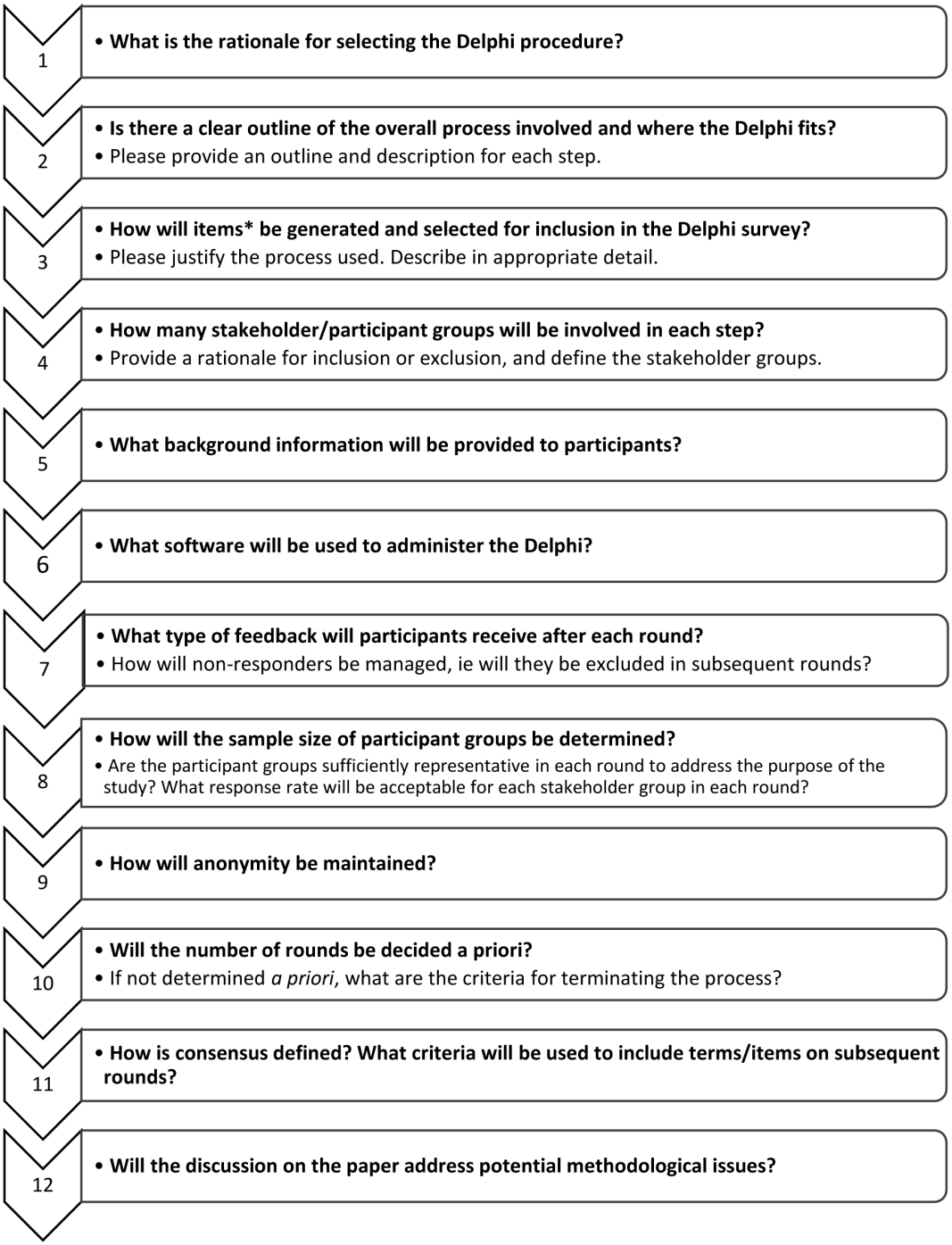

Figure 1. OMERACT Delphi consensus checklist. *The word "Item" refers to a domain or an instrument (outcome measure).

have formally assessed this have provided mixed results[15,16,17].

To provide a feasible minimum number of participant groups and to facilitate the incorporation of patient-relevant outcomes, the consensus team suggested a minimum of 2, including patients and clinicians. In fact, most delegates suggested that more than 3 be selected. After discussions it was felt that apart from patients and clinicians, trialists/researchers should always be involved. Involvement of others would depend on the context of use of the COS. Clearly, OMERACT participants value the involvement of many groups and consider that the selection be dependent on the context[7].

There was no consensus on how nonrespondents should

be handled. It could be argued they should be excluded from voting in future rounds because they may not be well informed. However, to ensure sufficient numbers of participants for the Delphi, informal discussions led to the conclusion that nonrespondents should be allowed to participate in future rounds at the discretion of the researcher.

Regarding the number of participants in each group, for logistical reasons the consensus team suggested a minimum of 50 participants for each of the 3 predominant groups (patients, clinicians, researchers) at the end of the final Delphi round. When delegates were surveyed, there was a wide distribution of opinions, demonstrating that participants preferred "as many as humanly possible." Informal discussions revealed delegates were concerned that for some groups, engaging 50 participants may not be a realistic goal, especially for rare diseases and may reduce anonymity. Delegates were surveyed regarding what should be the maximum number of rounds in the Delphi. Votes were divided; 3 rounds were selected by 25/58 (43%), 4 rounds by 16/58 (28%), 5 rounds by 15/58 (26%), and 2/58 (3%) selected 2 or fewer. The recommendation put forth suggests a minimum of 3 rounds. Greater attrition rates with an increasing number of rounds is a concern, but a recent publication demonstrated impressive retention (92%) after 5 rounds as the result of a strategy of tailored reminders by e-mail and telephone[18].

Through the Delphi surveys presented at OMERACT 2018, it became apparent that working groups used different ways to prioritize items. Participants were asked to (1) select how important the item was on a rating scale of 1–9; (2) select the top 10 from a long list; or (3) indicate, for instance in the final round, where the item should be [i.e., inner circle (mandatory), middle circle (important but not critical), outer circle (research agenda), or removed (not important)]. This supported the need for further discussion surrounding the definition of consensus. Two-thirds of the delegates agreed that consensus in a COS Delphi should be defined as ≥ 70% of participants in each group voting for the domain as critically important (rating 7–9 on a scale from 1–9). In this case, the domain will be included in the draft core domain set. The advantage is that the voice of various groups (e.g., patients) with fewer numbers can be adequately represented. One third of voting delegates were in favor of a combined definition of consensus, meaning ≥ 70% of all participants should vote the domain as very important, independent of the single groups' opinion. There is no "correct" definition of consensus, but determining the definition *a priori* in a manner acceptable to the key groups (i.e., OMERACT) is essential to prevent data mining[13].

DISCUSSION

The workshop has increased awareness surrounding the Delphi method and delegates agreed that more standardization is desirable. However, experienced delegates shared concerns that proposing standards that are too proscriptive may be problematic. We therefore suggest that the OMERACT Delphi consensus checklist be used as guidance to working group members. (More detailed recommendations regarding the use of the Delphi Consensus Checklist can be found in the supplementary material, available with the online version of this article.) Because of the lack of literature and empiric evidence regarding the methods themselves, more stringent guidelines are not justified[13,15,19]. During the meeting, an identified potential tool that may improve the use and reporting of the Delphi is a Delphi software package. Table 1[6,18] lists considerations when selecting a software package.

Our conclusions:

- Awareness regarding the Delphi has been increased;
- There is agreement that more standardization is desirable;
- OMERACT provides guidance, not absolute standards, because there is insufficient evidence to support decision making;
- Standardization may be improved by using software that provides structure and prompts decision making at each stage; and
- More research regarding the method itself is needed.

Suggestions moving forward. The updated recommendations accompanying the OMERACT Delphi Consensus Checklist (available with the online version of this article) will be available for OMERACT 2020. Most delegates (74/108, 69%) at the OMERACT 2018 meeting agreed to use this checklist, although some (27/108, 25%) were unsure and some (7/108, 6%) refused. A suggestion to improve uniformity is to use a software program that provides structure and help with reporting all relevant outcomes (e.g., DelphiManager, http://comet-initiative.org/delphimanager/). A majority [69/110 (63%)] of the delegates were willing to use it, 38/110 (34%) were unsure, and 3/110 (3%) refused. To further inform Delphi best practices, we will conduct an internal review of all Delphi surveys done in OMERACT since 2012 and compare them to the Delphi surveys done after the guiding document was made available. Finally, more research is required on the appropriate use of the Delphi method itself.

Limitations. Not all delegates voted at each session, but the response rates represent about 65% of delegates. Our paper is based on the opinions of the authors who have widespread experience with consensus methods, and after extensive consultation with the delegates at OMERACT 2018.

Our paper describes the ongoing strategies to improve processes and procedures surrounding consensus. This work should not be considered the final word, but a step forward as OMERACT continually strives to better itself.

ACKNOWLEDGMENT

Table 1. Considerations when selecting a software package to administer the Delphi.

Does the Software Package Provide the Following?	Considerations
Initial recruitment e-mail to participants to solicit willingness and consent to participate in all rounds of the Delphi	Response rates may be improved by sending an initial e-mail to potential participants outlining the purpose of the Delphi and the number of rounds planned. Example of text: "Thank you for agreeing to participate. It is very important that you complete the survey in each round, as the validity of the study could be compromised if participants drop out. If participants drop out because they feel their opinions are in the minority, the final results will overestimate how much agreement there is on the topic[6]."
Reminders (e.g., e-mail, other)	Only send to those who have not completed a round. Can these e-mails be personalized? A recent study suggests repeated reminders by e-mail, phone, and texts are acceptable to participants and produced high response rates[18]. Important to get all contact information at the recruitment stage.
Ability for participants to add domains in the first round	In many studies, participants should have an opportunity to add new domains in the first round.
Ability for the administrator to modify the list of domains between rounds based on results of the previous round	Does the software send all the domains for re-scoring or is there the option to provide a summary with a list of domains that achieved consensus (i.e., important based on *a priori* definition), domains that were removed (i.e., not important) and the list of domains to be re-scored? Is there an opportunity to clarify which domains were combined, reworded, or added? Summary lists, if provided, would be given in the final round, because it is important that participants have an opportunity to re-score domains after consideration of the feedback of others.
Feedback to individual participants	For each domain, how are individual participant scores and the distribution of scores from other participants displayed? Consider whether scores will be provided as aggregate or broken down by participant group (e.g., patients and other participants). Can participants include qualitative data (i.e., written comments)?
Ability to apply consensus *a priori* and allow for ranking	How much flexibility is there in defining consensus and can it vary between rounds? For example, in the initial rounds the participants would be asked to *score* domains, but in the final round they would be asked to *rank* domains in order of importance. Example for scoring: rating scale 1–9; if 70% of all participants select 7–9 (very important), the domains will be kept. If 70% of participants select 1-3 (not important), the domains will be dropped. Example for ranking: if you have a large number of domains that have achieved consensus, and want to reduce them to a manageable number, have participants select and rank their "top 10" in the final round.
Ability to extract pertinent data from the system	Is all the data entered by participants easily downloadable in a useful format? Is the data anonymized to maintain the spirit of the Delphi process?

The authors thank Dr. Paula Williamson for her thoughtful review of the manuscript.

ONLINE SUPPLEMENT

Supplementary material accompanies the online version of this article.

REFERENCES

1. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53.
2. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT Handbook. [Internet. Accessed May 17, 2017.] Available from: https://omeract.org/resources
3. Biggane AM, Brading L, Ravaud P, Young B, Williamson PR. Survey indicated that core outcome set development is increasingly including patients, being conducted internationally and using Delphi surveys. Trials 2018;19:113.
4. Murphy KR, Balzer WK, Lockhart MC, Eisenman EJ, Murphy K. Effects of previous performance on evaluations of present performance. J Appl Psychol 1985;70:72-84.
5. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995;311:376-80.
6. Sinha IP, Smyth RL, Williamson PR. 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:e1000393.
7. Maxwell LJ, Beaton DB, Shea BJ, Wells GA, Tugwell P, Boers M, et al. Core domain set selection according to filter 2.1: the OMERACT methodology. J Rheumatol 2019;46:1014-20.
8. Bartlett SJ, Hewlett S, Bingham CO 3rd, Woodworth TG, Alten R, Pohl C, et al; OMERACT RA Flare Working Group. Identifying core domains to assess flare in rheumatoid arthritis: an OMERACT international patient and provider combined Delphi consensus. Ann Rheum Dis 2012;71:1855-60.
9. Boulkedid R, Abdoul H, Loustau M, Sibony O, Alberti C. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. PLoS One 2011;6:e20476.
10. Humphrey-Murto S, Varpio L, Wood TJ, Gonsalves C, Ufholz LA, Mascioli K, et al. The use of the Delphi and other consensus group methods in medical education research: a review. Acad Med 2017;92:1491-8.
11. Foth T, Efstathiou N, Vanderspank-Wright B, Ufholz LA, Dütthorn N, Zimansky M, et al. The use of Delphi and Nominal group technique in nursing education: a review. Int J Nurs Stud

2016;60:112-20.
12. Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, et al. Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014;67:401-9.
13. Grant S, Booth M, Khodyakov D. Lack of preregistered analysis plans allows unacceptable data mining for and selective reporting of consensus in Delphi studies. J Clin Epidemiol 2018;99:96-105.
14. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. Med Teach 2017;39:14-9.
15. MacLennan S, Kirkham J, Lam TBL, Williamson PR. A randomized trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement. J Clin Epidemiol 2018;93:1-8.
16. Brookes ST, Macefield RC, Williamson PR, McNair AG, Potter S, Blencowe NS, 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;17:409.
17. Campbell SM, Hann M, Roland MO, Quayle JA, Shekelle PG. The effect of panel membership and feedback on ratings in a two-round Delphi survey: results of a randomized controlled trial. Med Care 1999;37:964-8.
18. Turnbull AE, Dinglas VD, Friedman LA, Chessare CM, Sepúlveda KA, Bingham CO 3rd, et al. A survey of Delphi panelists after core outcome set development revealed positive feedback and methods to facilitate panel member participation. J Clin Epidemiol 2018;102:99-106.
19. Hutchings A, Raine R. A systematic review of factors affecting the judgments produced by formal consensus development methods in health care. J Health Serv Res Policy 2006;11:172-9.