Journal of Clinical Epidemiology 106 (2019) 136–139

Journal of Clinical Epidemiology

COMMENTARY

The Delphi method—more research please

Susan Humphrey-Murto[a,*], Maarten de Wit[b]

[a]Department of Medicine and Innovation in Medical Education, University of Ottawa, The Ottawa Hospital-Riverside Campus, 1967 Riverside Drive, Ottawa, ON K1H 7W9, Canada

[b]Department of Medical Humanities, VU Medical Centre Amsterdam

Accepted 14 October 2018; Published online 21 October 2018

In this journal, Turnbull and colleagues have reported the results of a survey of Delphi panelists who had recently participated in the development of a Core Outcome Set (COS) and a Core Outcome Measurement Set (COMS) [1]. The COS and COMS define ''what to measure'' and ''how to measure'' outcomes in all clinical trials for the same condition. The purpose of their study was to help researchers navigate design decisions when conducting Delphi studies. The following commentary will situate the present study in the literature by reviewing the Delphi method and highlighting concerns regarding poor standardization and reporting, briefly touch upon the literature informing the method itself, and propose current recommendations and future directions.

The Delphi is a formal consensus method and is a systematic means for measuring and developing consensus among stakeholders [2]. It was developed in the 1950s to structure group communication to increase accuracy of forecasts by the RAND Air Force Corporation, which at the time was concerned with estimating key nuclear targets in America [3]. Most researchers are quite familiar with the stages of the Delphi that include identifying a research problem, selecting participants (also referred to as panelists), developing a questionnaire of items or statements, conducting anonymous iterative rounds to collect individual and group feedback, determining if consensus has been reached, and summarizing the findings. Arguably, researchers may find it challenging to describe best practices for each of those steps. Why? Because there is little to no empiric evidence to support best practices. This has led to a lack of standardized definitions (We challenge anyone to provide a clear definition of the ''modified'' Delphi), poor implementation of the method itself, and inadequate reporting of results.

Several recent reviews have highlighted concerns surrounding the Delphi. A systematic review of 80 studies using the Delphi to measure health care quality indicators found that there was a lack of consistency in providing details considered important for interpreting the results of the study; for example, only 39% of studies reported response rates for all rounds, 60% described the feedback provided, and 77% reported the method used to achieve consensus [4]. Sinha and colleagues published a systematic review of 15 studies using the Delphi to select outcome measures for clinical research and found that the following were poorly reported: information provided to the participants at the start of Delphi, the information fed back to participants after each round, the level of anonymity, attrition rates, and a list of outcomes after each round [5]. Studies from other academic disciplines such as resource management, agriculture, social policy, business, and medical education have found similar results [6,7].

Some researchers have attempted to study the actual method itself. A systematic review by Hutchings of consensus methods is over a decade old but provides some valuable information [8]. They examined 22 studies comparing the impact of the characteristics of individual participants within groups, and 30 studies comparing the results produced by two or more groups. They concluded that participant specialty influenced outcomes, but otherwise, there was little generalizable evidence for how characteristics of participants and groups influence judgments. More recently, studies in this journal have examined feedback strategies between rounds. After each round of the Delphi, responses for each item are summarized and fed back within the subsequent questionnaire enabling participants to consider the view of others before rerating. MacLennan and colleagues compared different types of feedback: feedback from peers only, from multiple stakeholders separately, or from multiple stakeholders combined [9]. They found no difference in number of outcomes retained or reduction in variability of opinion but do concede that a very high level of existing agreement in the first round may have accounted for the results. Other studies found that providing feedback from all stakeholder groups separately

DOI of original article: 10.1016/j.jclinepi.2018.06.007.
Conflict of interest: None.
Funding: None.

* Corresponding author. Tel.: 613-737-8899x81850; fax: 613-738-8228.
E-mail address: shumphrey@toh.on.ca (S. Humphrey-Murto).

https://doi.org/10.1016/j.jclinepi.2018.10.011

may influence the final core set and improve consensus between the groups. Differences in item scores and variability in scores between patients and professionals were smaller among those receiving feedback from both stakeholder groups rather than their peer group alone [10]. Similar findings were noted by Campbell; panel composition and type of feedback influenced the outcome. Not surprisingly, managers rated items differently than physicians, in this case higher. When manager and physician participant groups received collective feedback, results in the next rounds were moderated with clear evidence that each group was influenced by the other compared to those receiving only peer feedback [11].

One particular concern of the Delphi that has been studied is how to define consensus. Diamond et al. investigated how consensus was operationalized in 98 Delphi studies from multiple disciplines. They demonstrated that the definitions of consensus varied widely and were poorly reported and that very few studies described achievement of consensus as a decision to terminate the Delphi process [12]. Even more concerning is that many studies did not define consensus a priori. In another study, Grant et al. used the same data set but calculated final consensus based on several commonly used definitions, medians and means with and without measures of dispersion, and multiple levels of agreement with varying levels of stringency [13]. The percentage of items reaching consensus varied dramatically from 0% to 84% depending on the analysis procedure. The authors caution that this allows for the potential for unacceptable data mining. It is somewhat surprising that although the notion of consensus is fundamental to the Delphi, the definition and operationalization of what constitutes consensus appears inadequate in many studies.

In this issue, Turnbull et al. focuses their efforts on improving our understanding of the panelists in the Delphi [1]. This study examines how they perceive the burden of participation, how they use background information provided, how they consider and weighed feedback and voting from earlier rounds, and panelists' understanding of COS and COMS. The study itself strategically targets panelists who had recently completed a Delphi to develop COS/COMS for postdischarge clinical research studies evaluating acute respiratory failure survivors. Three months after completion of the Delphi, all panelists were sent a follow-up survey. The response rate was robust at 92% ($n = 70$) and included 32 clinical researchers, 19 clinical professional association representatives, four research funding representatives, and 15 survivors/caregivers (20%). The panel included stakeholders from over 16 countries.

The main findings of the study were as follows: 96% of participants agreed that participating in the Delphi was important, 91% agreed that the time required to participate in the process was appropriate, and 89% would participate again. Despite 94% of panelists receiving at least one e-mail reminder and approximately 40% receiving phone calls or text messages, only 3 (4%) panelists reported being bothered by these reminders. The impressive response rate of 90% over five rounds of the Delphi may have been related to what appears to be a highly motivated panel and multiple reminders. Previous research has suggested that the typical number of rounds should be 2–4 to improve completion rates, but the authors clearly admit the recommendation is based on very little evidence [4].

A commonly asked question regarding the Delphi is as follows: How many rounds are required? The easiest answer would be "till consensus has been reached." However, there are several issues for consideration. One concern that has been raised is increasing attrition over subsequent rounds. This could lead to "false" consensus as participants with dissenting views start dropping off. Many studies have instructions to participants emphasizing the importance of completing all rounds. A useful example is provided by Sinha [5]. It should also be clear to participants that they do not need to conform. It is our opinion that forcing consensus is not useful and continuing until "consensus is achieved" through multiple rounds may lead participants to agree just to make it end. Thus, some suggest that the number of rounds for each Delphi be determined a priori, whereas others suggest stopping the process when agreement has been reached or the variation in responses between rounds is reduced to a predetermined level [12]. It is not clear from the Turnbull study if the impressive response rate over five rounds was from intrinsic panelist motivation or, as we suspect, was more likely from the substantial efforts of the research team in increasing extrinsic motivation through reminders and phone calls.

Turnbull also explored how panelists considered and weighed feedback and voting from earlier rounds. It appears that most participants considered written results from other panel members. Interestingly, when specifically asked if they considered prior voting results from other stakeholder groups, patients and caregivers appeared less likely to consider (67% vs. all panel mean 83%). Although this did not reach statistical significance, arguably, this might be suggesting that patients do perceive their perspective as unique and may not be unduly influenced. It would be interesting to have actual data on changes in rating over the rounds for various stakeholder groups to verify this self-reported finding. For OMERACT, the organization for the development of COS and COMS for rheumatic conditions, it is no question that greater patient involvement is not only desirable but essential because they have a unique perspective [14].

Background information provided to participants is an important component of the Delphi, but often neglected [5,7]. The authors asked panelists if they reviewed the background information, in this case "Measure cards" for each measurement instrument. Only 56% of panelists reviewed over 75% of the information provided. It is not clear if clinicians and researchers were already familiar with the content and did not require the information, or if it was simply ignored by many panelists. Without further information,

Table 1. A research agenda for future Delphi method studies

	Topic	Research questions
1	Defining consensus	The key purpose of the Delphi is to determine if there is consensus on a set of items. Therefore, research should focus on how consensus is defined and operationalized and when to stop the Delphi process. This may include the study of statistical measures for reporting a move toward consensus, or stability of responses.
2	Stakeholder group participation	Patient involvement is considered essential to OMERACT (Outcome Measures in Rheumatology). Many questions remain regarding when and how to involve patients, percentage representation, required training, and potential influence from other participants.
3	Exploring the nature of consensus building	How does decision making occur in groups? How are participants influenced? Consider using qualitative research methods and borrowing from other disciplines such as psychology to explore the nature of consensus building in the Delphi.
4	Direct outcome versus the ultimate longer-term goals of the Delphi	Position the use of Delphi method in the broader context of improving health care. For example, if the goal is to develop a COS, the success of the Delphi is not determined by developing the set, it is determined by the uptake of the COS by stakeholders and the impact on improving patient outcomes.
5	Using multiple methods	The Delphi is not used in isolation. Studies of how and when to use the Delphi and combine with focus groups and other consensus methods such as Nominal Group Technique and consensus conferences are highly relevant.

Abbreviation: COS, Core Outcome Set.

it is difficult to interpret this finding. The authors then assessed the panelists' understanding of information provided about measurement instrument properties. Although a laudable effort, a single multiple-choice question is insufficient to assess knowledge [15]. Future studies may want to explore panelists' understanding with more rigorous assessment tools. Pragmatically, it might be a good reminder to consider providing background information at each round because the entire Delphi lasted 157 days in this instance and most of us can't recall what we ate for lunch yesterday.

This study would have been strengthened by examining responses to open-ended free-text questions. The authors state that the responses were not evaluated, but as a reader, we are left wondering why. Arguably, understanding how panelist considered and weighed feedback from different stakeholder groups in previous rounds might be best explored by evaluating these open-ended responses and better yet through semistructured interviews. This could certainly be a consideration for future work.

There are several practical take-home messages from this study. First, very high response rate over five rounds of Delphi voting is an achievable goal. Secondly, multiple reminders, even phone calls and text messages, appear acceptable, so be sure to acquire all that information at the beginning of the process and have enough research staff to support. Finally, do not assume your participants will recall the purpose of the study and background information so consider providing an easily acceptable summary during each round.

Why should we care about the results of this study? The Delphi is extensively used across multiple disciplines. In health care, it is often used for the selection of core outcome sets and core outcome measurement sets to be used in all effectiveness trials of an intervention, guideline development, and education of physicians. The input from diverse stakeholders is essential to ensure that the outcomes of these inquiries are relevant and acceptable. The methods need to be rigorous as the results of these studies have the potential to impact patient care.

Finally, the Delphi is still a challenge to all academic communities. With so many unanswered questions, all researchers using consensus methods should consider tackling one aspect of the Delphi method as part of their study. Ideally, this could be done in a systematic way, by critically examining each step and including both quantitative and qualitative research methodologies to help inform the process. Table 1 lists suggestions for future research. For the present time, in the absence of empiric evidence to guide choices, we would advocate for a minimum standard that includes clear reporting and justification for choices made for each stage of the Delphi. Practical recommendations can be found in several papers [5,6,12,16]. Reporting on these aspects is essential for readers to evaluate the validity and credibility of the study results.

References

[1] Turnbull AE, Dinglas VD, Aronson Friedman L, Sepulveda KA, Bingham CO, Needham DM. 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.

[2] Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995;311:376–80.

[3] Campbell SM, Cantrill JA. Consensus methods in prescribing research. J Clin Pharm Ther 2001;26:5–14.

[4] 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.

[5] 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.

[6] de Loë RC, Melnychuk N, Murray D, Plummer R. Advancing the state of policy Delphi practice: a systematic review evaluating methodological evolution, innovation, and opportunities. Technol Forecast Soc 2016;104:78−88.

[7] Humphrey-Murto S, Varpio L, Wood TJ, Gonsalves C, Ufholz L, Mascioli K, et al. The use of the delphi and other consensus group methods in medical education: a review. Acad Med 2017;92:1491−8.

[8] 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.

[9] MacLennan S, Kirkham J, Lam TBL, Williamson P. A randomized trial comparing three Delphi strategies found no evidence of a difference in a setting with high initial agreement. J Clin Epidemiol 2018;93:1−8.

[10] Brookes ST, Macefield RC, Williamson PR, McNair AG, Potter S, Blencowe NS, et al. Three nested randomized controlled trials of pee-only or multiple stakeholder group feedback within Delphi Surveys during core outcome and information set development. Trials 2016;17:409.

[11] Campbell SM, Hann M, Toland MO, Quayle JA, Shekelle PG. The effect of panel membership and feedback on ratings in a two-round Delphi survey. Med Care 1999;37:964−8.

[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] De Wit M, Kirwan JR, Tugwell P, Beaton D, Boers M, Brooks P, et al. Successful stepwise development of patient research partnership: 14 years' experience of actions and consequences in outcome measures in rheumatology (OMERACT). Patient 2017;10(2):141−52.

[15] Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York: Taylor and Francis; 2009.

[16] 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(1):14−9.