Journal of Clinical Epidemiology 99 (2018) 96−105

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Lack of preregistered analysis plans allows unacceptable data mining for and selective reporting of consensus in Delphi studies

Sean Grant*, Marika Booth, Dmitry Khodyakov

RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, USA

Accepted 10 March 2018; Published online 17 March 2018

Abstract

Objectives: To empirically demonstrate how undisclosed analytic flexibility provides substantial latitude for data mining and selective reporting of consensus in Delphi processes.

Study Design and Setting: Pooling data across eight online modified-Delphi panels, we first calculated the percentage of items reaching consensus according to descriptive analysis procedures commonly used in health research but selected post hoc in this article. We then examined the variability of items reaching consensus across panels.

Results: Pooling all panel data, the percentage of items reaching consensus ranged from 0% to 84%, depending on the analysis procedure. Comparing data across panels, variability in the percentage of items reaching consensus for each analysis procedure ranged from 0 (i.e., all panels had the same percentage of items reaching consensus for a given analysis procedure) to 83 (i.e., panels had a range of 11% to 94% of items reaching consensus for a given analysis procedure). Of 200 total panel-by-analysis-procedure configurations, four configurations (2%) had all items and 64 (32%) had no items reaching consensus.

Conclusion: Undisclosed analytic flexibility makes it unacceptably easy to data mine for and selectively report consensus in Delphi processes. As a solution, we recommend prospective, complete registration of preanalysis plans for consensus-oriented Delphi processes in health research.

Keywords: Consensus; Delphi process; Expert panel; Preanalysis plan; Preregistration; Open science framework

1. Introduction

1.1. The Delphi technique

The Delphi technique is a widely used approach to systematically seek expert and stakeholder opinion on topics important to future research, practice, and policy [1−3]. It involves an anonymous, iterative survey process in which participants respond to a series of written questionnaires interspersed with controlled feedback of the group's responses [4]. This structured process facilitates an efficient and effective way to solicit input from geographically dispersed stakeholders on important and potentially complex topics. Recent developments in the Delphi technique, such as online and Real-Time Delphi processes, are particularly promising advances for faster, more cost-effective Delphi processes that involve large groups of participants [5].

Although the general purpose of the Delphi technique is to facilitate and structure group communication processes [3,6], determining the existence of consensus among the participating stakeholders is often the primary outcome of

Funding: This research was supported by PCORI contract CDRN-1306-04819. The funder did not directly participate in the study design, collection, analysis, or interpretation of the data for this article; and the preparation, review, and approval of the article for publication. The authors are solely responsible for the content and the decision to submit it for publication.

Ethics approval: This study was determined not to be human subjects by IRBs at RAND and University of California Davis.

Availability of data and materials: Data from this article also appear in Khodyakov, D., Grant, S., Meeker, D., Booth, M., Pacheco-Santivanez, N., & Kim, K. K. (2016). Comparative analysis of stakeholder experience with an online approach to prioritizing patient-centered research topics. Journal of the American Medical Informatics Association. doi: https://doi.org/10.1093/jamia/ocw157.

Conflicts of interest: S.G.'s spouse is a salaried employee of Eli Lilly and Company, and owns stock. S.G. has accompanied his spouse on company-sponsored travel. All other authors declare that they have no conflicts of interest.

* Corresponding author. RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138, USA. Tel.: +1-310-393-0411×6438.

E-mail address: sgrant@rand.org (S. Grant).

https://doi.org/10.1016/j.jclinepi.2018.03.007

What is new?

Key findings

- Analyzing data from eight online modified-Delphi panels, we found the percentage of items reaching consensus to vary considerably depending on the analysis procedure used to determine consensus, ranging from no items to almost all items reaching consensus.

What this adds to what is known?

- Undisclosed analytic flexibility makes it unacceptably easy to data mine for and selectively report consensus in Delphi processes.

What is the implication and what should change now?

- Standards and norms for prospectively defining analysis plans are needed to improve the credibility of Delphi processes for informing health research, practice, and policy.
- Delphi researchers should prospectively and completely register their preanalysis plans in a publicly available, independently controlled platform that time-stamps entries. Journals and research sponsors should require prospective registration of Delphi processes as a precondition for publication and funding.

Delphi processes in the health research literature (the target audience for this article) [7,8]. Delphi processes typically ask participants to rate (nonleading and unambiguous) items in an initial questionnaire, review summaries of group responses once participants have finished rating items, and then revise their ratings of items in light of the results from previous rating rounds [9]. Delphi researchers continue this process until a stopping rule has been met or a predetermined number of rating rounds has been completed [10]. In consensus-oriented Delphi processes, the data collected during the Delphi process are then analyzed to determine which items reached consensus among the participating stakeholders. These consensus-oriented Delphi processes are increasingly influential in health research, as they are being used to define key terms for rapidly developing health research areas [11], decide core outcome sets to measure in clinical trials [12], develop reporting guidelines for research articles [13], estimate the prevalence of health-related conditions [14], and establish priorities for future research studies [15]. Specifying the analysis procedure for consensus is therefore a critical consideration when designing consensus-oriented Delphi processes in health research.

1.2. Analytic flexibility in consensus-oriented Delphi processes

The concept of consensus has been defined in numerous ways by Delphi researchers [6,16]. Many Delphi processes in health research specifically seek to determine—and therefore restrict their definitions of consensus to—the items that have been rated favorably by participants [1,6]. Negatively rated items, as well as those with responses in the middle of the rating scale, are consequently less relevant—other than their use to indicate the dispersion of responses. To illustrate, Delphi processes to develop reporting guidelines typically ask participants to rate how important they believe it is to include each item in the reporting guideline under development, with the ultimate goal of creating a checklist of the items that participants rated as most important [17,18]. In this context, a research team must choose an analysis procedure for determining whether an item has reached consensus (e.g., a median $\geq$7 on a nine-point Likert scale) that may or may not incorporate an indicator of the dispersion of responses as part of the criteria for consensus (e.g., the interquartile range of responses on the nine-point scale must be no greater than two) [19].

A multitude of different analysis procedures to determine consensus in Delphi processes exist, but there are no agreed-upon standards or guidelines for choosing one analysis procedure over another [6,8,16]. To classify this multitude of procedures, the most comprehensive overview of consensus measurement in Delphi processes to date lists three different types of analysis procedures: subjective decisions by the research team, descriptive statistics, and inferential statistics [6]. In health research, descriptive statistical measures are the most common [1]. These include measures of central tendency (e.g., median, mean), level of agreement (e.g., percentage of participants giving specific ratings to each item), dispersion of responses (e.g., width of inter-quartile range), and combinations of the above measures (e.g., a measure of central tendency in combination with a measure of dispersion) [8,9]. Thus, even when considering only descriptive statistical measures to determine which items have been rated favorably by participants, a multitude of analysis procedures are currently accepted by the health research community [1]. These analysis procedures differ in their sensitivity to different distributions of the data [8,20], and different procedures are specifically intended to involve more (or less) strict criteria to make it more difficult (or easy) for items to achieve consensus [21]. As a result, the set of items that reach consensus has the potential to vary considerably depending on the chosen analysis procedure.

The multitude of accepted procedures for determining consensus, and their differing sensitivities to different distributions of the data, leave Delphi processes susceptible to data mining for and selective reporting of consensus by health researchers. Without committing to an analysis procedure in advance of data collection, researchers have the flexibility to try different analysis procedures to obtain

a desired result, which jeopardizes the veracity of claims that the results of the study accurately reflect the consensus opinion of participating stakeholders [22]. That is, without prespecifying their analysis procedures in a study registry, health researchers conducting consensus-oriented Delphi processes can mine for and selectively report the most desirable set of items reaching consensus—and even present the reported analysis as the only one conducted. Undisclosed flexibility in data collection, analysis, and reporting is a growing concern in empirical research [23,24], and initiatives to address such concerns are increasing priority for the scientific community [25,26]. However, attention to these biases—and to the transparent, open, and reproducible research practices (such as study preregistration) [22] that aim to address these biases—is currently not normative in Delphi research. As a result, haphazard and even post hoc selection of analysis procedures for determining consensus can threaten the credibility of consensus-oriented Delphi processes in health research.

1.3. Study objective

Building on previous studies examining the measurement of consensus in Delphi processes [9,27], this study pools empirical data from an actual Delphi process to demonstrate the considerable variability in the percentage of items reaching consensus as a result of changing the analysis procedure used to determine consensus. The objective of these analyses is to highlight how undisclosed analytic flexibility allows data mining for and selective reporting of consensus, thereby posing a serious threat to the primary outcome of consensus-oriented Delphi processes.

2. Methods

Data for this article come from a project to develop research priorities for the Patient-centered SCAlable National Network for Effectiveness Research (pSCANNER): a stakeholder-driven, clinical data research network [28]. pSCANNER is a member of the Patient-Centered Outcomes Research Institute's national patient-centered clinical research network of learning health care systems created to harness the power of very large amounts of health data [29]. As part of pSCANNER stakeholder engagement activities, we conducted an online modified-Delphi process (involving eight stakeholder panels) to explore the existence of consensus on research topics to be prioritized for future studies conducted using pSCANNER data [30]. Namely, we aimed to determine priorities of patient-centered outcomes research for three specific health issues: three panels focused on weight management and obesity (WMO), three panels focused on congestive heart failure (CHF), and two panels focused on Kawasaki disease (KD). Substantive panel findings (including qualitative data analyses) will be published separately.

2.1. Study participants

We recruited researchers, clinicians, and patients or their caregivers relevant to the condition for each panel. Recruitment strategies involved email, messages to members-only social media communities, and in-person contact by members of the pSCANNER stakeholder advisory board and clinicians at pSCANNER clinical sites. Informed by previous research on online modified-Delphi processes, the intended sample size for each panel was approximately 40 participants [11]. To be eligible, patients had to be either overweight (body mass index $\geq$25 kg/m^2) for the WMO panels or diagnosed with heart failure for the CHF panels; eligible caregivers had to provide care for a child diagnosed with KD for the KD panels. Researchers and clinicians had to be involved in research or care for patients with the condition relevant to each panel. Participants also had to be 18 years of age or older, read and write in English, and be able to use a computer or similar device to access the online system. WMO and CHF each had a clinician-only panel (panels A and D), a patient-only panel (panels B and E), and a mixed panel of patients, clinicians, and researchers (panels C and F). KD had a patient-only (panel G) and a mixed panel (panel H) but no clinician-only panel.

2.2. Research design

We used RAND's ExpertLens (EL) for all eight panels [31]. EL is an online modified-Delphi platform that combines rounds of questions with interspersed rounds of statistical feedback and online discussion. EL offers an innovative approach to online modified-Delphi processes by allowing participants to answer questions and explain their responses, as well as discuss results using an asynchronous, anonymous, and moderated online discussion board. For efficiency of running online modified-Delphi processes, it automatically analyzes group responses and displays participants' previous responses in later rounds, eliminating the time between rounds needed to generate individualized reports that detail group responses in comparison to individual participants' responses. EL has been used in numerous health research studies that involved online modified-Delphi processes [28,32–35].

Procedures across rounds were consistent across panels; the only differences were the participants in a panel, the condition under study, and the research topics being rated. Participants considered patient-centered research topics for a given condition (WMO, CHF, or KD) with a particular aspirational goal in mind (e.g., "reduce unwarranted hospital readmissions for heart failure by 25% by 2020"); the research topics and aspirational goals were the same across stakeholder panels (patient-only, clinician-only, and mixed panels) for the same condition (WMO, CHF, or KD). Each study round occurred over approximately 1 to 2 weeks, with participants receiving periodic reminders and financial

incentives (a $300 gift card as a compensation for approximately 4 hours of their time) to maximize engagement.

During a brainstorming round 0, participants suggested research topics and rating criteria, and they provided feedback on a preliminary list of aspirational goals developed by the study team. In round 1, participants rated research topics on five different criteria—informed decision-making, collaboration, relevance, impact, and innovation (see Appendix on the journal's web site at www.elsevier.com)—on a nine-point Likert scale, in which higher scores indicated more "favorable" ratings (e.g., higher relevance, higher impact). Participants also had the ability to explain or comment on each Likert scale response using open text boxes. In round 2, the EL platform automatically calculated medians and interquartile ranges for each combination of research topic and rating criteria and displayed these on bar charts showing the frequencies of all participants' responses on the nine-point scale. Each participant also saw their own response highlighted as a red dot on each chart (see Fig. 1). Participants then engaged in an asynchronous and anonymous discussion of round 1 results using an online discussion board within EL. Members of the project team moderated discussions by posing questions and ensuring participants did not post comments containing personal health information. In round 3, participants revised their round 1 responses based on reviewing results from round 1 (bar charts and a list of explanatory comments for each rating question) as well as all of the round 2 discussion comments.

2.3. Measures

For the current article, we reanalyzed the data from the final rating round (i.e., round 3) of each panel according to five descriptive measures commonly used to determine consensus in Delphi processes reported in the health research literature [1], namely medians, medians plus dispersion of responses, means, means plus dispersion of responses, and levels of agreement. We used various specific thresholds within each of these definitions of consensus, yielding 25 analysis procedures in total (see Box 1). The reported analyses only included data for the participants who completed round 3 (see Table 1).

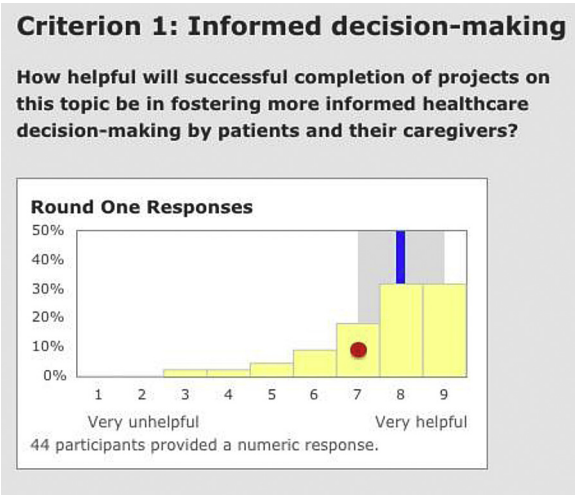

Fig. 1. Example screenshot of controlled, individualized feedback in round 2.

2.4. Statistical analysis

We first pooled data from all panels to calculate the percentage of items that reached consensus according to the 25

Box 1 Descriptive analysis procedures for determining consensus

Median alone
1. Median of 7.
2. Median of 8.
3. Median of 9.

Median and a measure of dispersion
4. Median of 7 and an interquartile range (IQR) less than or equal to 1.
5. Median of 7 and an IQR less than or equal to 2.
6. Median of 8 and an IQR less than or equal to 1.
7. Median of 8 and an IQR less than or equal to 2.
8. Median of 9 and an IQR less than or equal to 1.
9. Median of 9 and an IQR less than or equal to 2.
10. RAND-UCLA Appropriateness Method: median of 7 or higher and a disagreement index less than 1.21

Mean alone
11. Mean of 7.
12. Mean of 8.
13. Mean of 9.

Mean and a measure of dispersion
14. Mean of 7 and a standard deviation (SD) less than or equal to 1.
15. Mean of 7 and an SD less than or equal to 2.
16. Mean of 8 and an SD less than or equal to 1.
17. Mean of 8 and an SD less than or equal to 2.
18. Mean of 9 and an SD less than or equal to 1.
19. Mean of 9 and an SD less than or equal to 2.

Level of agreement (percentage of 7–9 ratings on a nine-point Likert scale)
20. Plurality of ratings (higher percentages of participants gave ratings of 7–9 than 1–3 or 4–6)
21. 50% of ratings.
22. 66% of ratings.
23. 75% of ratings.
24. 80% of ratings.
25. 100% of ratings.

Table 1. Study characteristics

Study information	Weight management and obesity	Congestive heart failure	Kawasaki disease
Dates	March–April 2015	May–June 2015	June–July 2015
Number of panels	3	3	2
Number of research topics rated by participants[a]	9	7	7
Number of rating criteria per topic[a]	5	5	5
Number we approached to participate	124	114	113
Number of invited panelists who participated (% of number we approached to participate)	99 (80%)	85 (75%)	103 (91%)
Number of panelists who completed round 3 (% of participating panelists)	82 (83%)	76 (89%)	99 (96%)

Notes: We defined ''number we approached to participate'' as those whom we sent an invitation and username/password to the ExpertLens system. We defined ''invited panelists who participated'' as those who accessed round 1. We defined ''panelists who completed round 3'' as those who completed the final rating round and therefore whose data we included in our analyses.

[a] The research topics and rating criteria can be found in Online Supplement 1.

different analysis procedures for determining consensus and rank-ordered these procedures from most lenient (i.e., highest percentage of items reaching consensus) to most stringent (i.e., lowest percentage of items reaching consensus). Then, we compared the percentage of items reaching consensus in each individual panel for each analysis procedure (i.e., eight panels and 25 analysis procedures, for a total of 200 panel-by-analysis-procedure configurations).

3. Results

Panels took place from March–July 2015 (see Table 1). Of 351 panelists, 287 (82%) answered questions in round 1. Of these 287 participants, 257 (90%) participated in round 3 and therefore provided data for the current analyses.

The percentage of items reaching consensus varied considerably by analysis procedure when pooling data from all panels (see Table 2). The most stringent analysis procedures all required a ''mean of 9'' to reach consensus; all of these analysis procedures yielded 0% of items reaching consensus. The least stringent analysis procedure was a ''plurality'' of respondents rating an item as favorable, for which 84% of items reached consensus.

The percentage of items reaching consensus also varied considerably within individual panels by analysis procedure and across panels when using the same analysis procedure (see Table 3). Similar to the analyses pooling data from all

Table 2. Percentage of items reaching consensus (all panels combined)

Measure information		Rank		Percentage reaching consensus
Descriptive measure	Consensus threshold	Measure rank	Threshold rank	Total
Level of agreement	Plurality	1	1	84%
Level of agreement	50%	2	2	79%
Median	Med 7	1	3	76%
RAND/UCLA	IPRAS	1	4	76%
Median + IQR 2	Med 7 + IQR 2	1	5	62%
Level of agreement	66%	3	6	60%
Mean	Mean 7	1	7	46%
Mean + SD 2	Mean 7 + SD 2	1	8	44%
Median	Med 8	2	9	39%
Level of agreement	75%	4	10	36%
Median + IQR 2	Med 8 + IQR 2	2	11	35%
Level of agreement	80%	5	12	28%
Median + IQR 1	Med 7 + IQR 1	1	13	24%
Median + IQR 1	Med 8 + IQR 1	2	14	16%
Mean	Mean 8	2	16	9%
Mean + SD 2	Mean 8 + SD 2	2	16	9%
Median	Med 9	3	18	6%
Median + IQR 2	Med 9 + IQR 2	3	18	6%
Mean + SD 1	Mean 7 + SD 1	1	19	5%
Median + IQR 1	Med 9 + IQR 1	3	20	4%
Mean + SD 1	Mean 8 + SD 1	2	21	3%
Level of agreement	100%	6	22	2%
Mean	Mean 9	3	25	0%
Mean + SD 1	Mean 9 + SD 1	3	25	0%
Mean + SD 2	Mean 9 + SD 2	3	25	0%

Abbreviations: IPRAS, Inter-Percentile Range Adjusted for Symmetry; IQR, interquartile range; SD, standard deviation.

Table 3. Percentage of items reaching consensus (individual panel data)

Measure information		Items reaching consensus across panels	Individual panel data							
Descriptive measure	**Consensus threshold**	**Range (lowest to highest)**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Level of agreement	Plurality	29% (71–100%)	73%	84%	93%	71%	100%	74%	86%	89%
Level of agreement	50%	36% (64–100%)	64%	82%	89%	69%	100%	69%	80%	83%
Median	Med 7	42% (58–100%)	58%	78%	87%	60%	100%	66%	80%	83%
RAND/UCLA	IPRAS	42% (58–100%)	58%	78%	87%	60%	100%	66%	80%	83%
Level of agreement	66%	62% (36–97%)	36%	53%	56%	49%	97%	51%	71%	74%
Mean	Mean 7	73% (24–97%)	24%	33%	31%	26%	97%	37%	66%	69%
Median + IQR 2	Med 7 + IQR 2	50% (44–94%)	44%	64%	69%	54%	94%	63%	57%	54%
Median	Med 8	83% (11–94%)	18%	13%	24%	11%	94%	40%	69%	60%
Level of agreement	75%	83% (11–94%)	11%	20%	40%	11%	94%	34%	49%	43%
Mean + SD 2	Mean 7 + SD 2	67% (24–91%)	24%	33%	31%	26%	91%	37%	63%	60%
Median + IQR 2	Med 8 + IQR 2	80% (11–91%)	16%	13%	24%	11%	91%	40%	54%	43%
Level of agreement	80%	79% (7–86%)	11%	7%	27%	11%	86%	26%	43%	29%
Median + IQR 1	Med 7 + IQR 1	59% (7–66%)	7%	9%	33%	17%	66%	34%	14%	14%
Median + IQR 1	Med 8 + IQR 1	63% (0–63%)	0%	2%	20%	0%	63%	26%	14%	14%
Mean	Mean 8	51% (0–51%)	0%	0%	0%	0%	51%	3%	14%	14%
Mean + SD 2	Mean 8 + SD 2	51% (0–51%)	0%	0%	0%	0%	51%	3%	14%	14%
Mean + SD 1	Mean 7 + SD 1	34% (0–34%)	0%	0%	0%	6%	34%	3%	0%	3%
Mean + SD 1	Mean 8 + SD 1	26% (0–26%)	0%	0%	0%	0%	26%	0%	0%	3%
Median	Med 9	20% (0–20%)	2%	0%	0%	0%	17%	0%	20%	17%
Median + IQR 2	Med 9 + IQR 2	20% (0–20%)	2%	0%	0%	0%	17%	0%	20%	17%
Level of agreement	100%	17% (0–17%)	0%	0%	0%	0%	17%	0%	0%	0%
Median + IQR 1	Med 9 + IQR 1	14% (0–14%)	0%	0%	0%	0%	14%	0%	11%	11%
Mean	Mean 9	0% (0–0%)	0%	0%	0%	0%	0%	0%	0%	0%
Mean + SD 1	Mean 9 + SD 1	0% (0–0%)	0%	0%	0%	0%	0%	0%	0%	0%
Mean + SD 2	Mean 9 + SD 2	0% (0–0%)	0%	0%	0%	0%	0%	0%	0%	0%

Abbreviations: IPRAS, Inter-Percentile Range Adjusted for Symmetry; IQR, interquartile range; SD, standard deviation.

panels, the three most stringent analysis procedures (i.e., requiring a "mean of 9") had no variability in the percentage of items reaching consensus (i.e., each individual panel had 0% of items reaching consensus). In contrast, the largest range in the percentage of items reaching consensus across panels for a given analysis procedure was 83: that is, the percentage of items reaching consensus ranged across panels from 11% to 94% when using a "median of 8" and "75% level of agreement" as the analysis procedure. Of the 200 total panel-by-procedure configurations, four configurations (2%) had all items reaching consensus, although 64 configurations (32%) yielded no items reaching consensus.

4. Discussion

4.1. Summary of findings

Deciding which analysis procedure to use is a critical aspect of designing consensus-oriented Delphi processes. In addition to often serving as the primary outcome of Delphi processes, reaching consensus is frequently the sole criterion for stopping a Delphi process in health research—although the stability of round results (i.e., consistency of responses between rounds) or a predetermined number of rounds are considered as more appropriate stopping criteria by experts in the Delphi technique [3,36,37]. Given the importance and influence of consensus measurement in Delphi processes in heath research, we built on previous studies examining how study design features can influence the results of Delphi processes to show how flexibility in the selection of the analysis procedure can considerably influence which items reach consensus [8–10,21,38,39]. Specifically, we empirically demonstrated that different analysis procedures for determining consensus lead to drastically different results, ranging from literally all items to no items reaching consensus. In other words, the set of items that reach consensus can change significantly based on the chosen analysis procedure.

In their seminal article on false-positive rates in psychology [23], Simmons et al demonstrated how undisclosed flexibility in data collection, analysis, and reporting allows psychology researchers the possibility to present nearly anything as statistically significant. We are concerned that undisclosed flexibility in data collection, analysis, and reporting of Delphi processes allows health researchers the similar possibility to present nearly any set of items as reaching consensus. Given that the analysis procedure used to determine consensus in Delphi processes is often arbitrarily or not rigorously decided [6], the possibility of health researchers trying different post hoc analysis procedures and reporting only the analysis procedure with the most desirable set of items reaching consensus is a legitimate concern. Without preregistering and reporting all of the attempted analysis procedures and when they were attempted, the extent and impact of researchers trying

different analysis procedures is nearly impossible for peer reviewers, editors, and consumers of Delphi research to assess [22].

4.2. A call to prospectively register analysis plans in consensus-oriented Delphi processes

For over a decade, health researchers have continually called for standards to improve the methodological rigor of Delphi processes [39,40]. To this end, and in light of our findings and concerns, health researchers should adopt practices used to address undisclosed analytic flexibility in other research methods to safeguard the credibility of consensus-oriented Delphi processes [22]. Otherwise, existing latitude in the selection and reporting of analysis procedures leaves items reaching consensus in Delphi processes from health research at risk of being (perceived as) arbitrary—or worse, an artifact of data mining combined with selective outcome and analysis reporting [41].

Health research using Delphi processes can meet this call by following developments from other areas of health research in the implementation of transparent, open, and reproducible research practices [1,12,42]. Of most relevance to the current article, pre-registration of research studies has been used to improve the credibility of health research for several decades—in particular, the credibility of reported outcomes in clinical trials [43−45]. One reason health researchers (namely clinical trialists) are expected to record important information about study design prospectively in public registries (e.g., ClinicalTrials.gov) is to allow others to compare the outcomes and analyses reported in articles with the outcomes and analyses they planned to use before the trial began [46]. Prospective, complete definitions of outcomes and how they will be analyzed can help prevent (or at least allow the detection of) post hoc specification searching: that is, conducting different analysis procedures on the same data set until the desired results are found and then selectively reporting these results [24,46,47]. Consequently, prospectively registering clinical trials has become an expected norm and standard by those involved in conducting, publishing, and sponsoring medical research [48−55].

Although many Delphi processes in health research provide operational definitions of consensus in final articles [1], there are fewer examples of Delphi researchers preregistering studies or publishing protocols that specify how they planned to determine consensus before collecting and analyzing the data [56,57]. Lack of prespecification allows for arbitrary, inappropriate, post hoc, selectively reported, and even purposely biased definitions of and procedures for analyzing consensus [6]. This norm is a great concern for the veracity of findings from Delphi processes, given that consensus measurement has been described by some as the "least-developed component of the … Delphi method," varying "from study to study" (p. 310) [58]. We believe preregistration of analysis plans would improve the credibility and utility of consensus-oriented Delphi processes in health research.

We recommend that researchers conducting consensus-oriented Delphi processes prospectively and completely register the intended procedure for identifying which items reach consensus. To be prospectively registered, the analysis procedure for determining consensus for Delphi processes should be chosen a priori—ideally before starting the first round but at the very latest before completing data collection—to improve the validity of findings. In other words, health researchers conducting consensus-oriented Delphi processes should commit themselves in advance to an analytic procedure for determining which items reach consensus before they see the actual data (or, ideally, before they even collect the data) [22]. To be completely registered, the preanalysis plan should precisely describe the essential elements of the analysis procedure for determining consensus (see Box 2). In addition, registrations should be in a publicly available and independently controlled platform that time-stamps entries. For instance, the Open Science Framework (https://osf.io/) provides a free, open-source system that Delphi researchers can use to preregister their analysis plans and store and share full study protocols, materials, code, and final raw data. Researchers can then compare the information about the analysis procedure reported in final article to the information about the analysis procedure registered in the preanalysis plan. Researchers should use existing guidance on reporting completed Delphi processes to provide sufficient information for comparing the final article to the registered preanalysis plan [1,12,42], with particular attention in the final article to any changes from the preanalysis plan in the items, rating criteria, analytic procedure (measure and threshold), and data and participants included in the analysis.

To facilitate the preregistration of consensus-oriented Delphi processes by researchers, other stakeholders in the

Box 2 Minimum set of items to include in prospectively registered preanalysis plans for consensus-oriented Delphi processes

1. All items to be rated (e.g., research projects to prioritize)
2. All criteria on which items will be rated (e.g., importance)
3. Type of measure for determining which items reach consensus (e.g., median plus a measure of variability in responses)
4. Specific threshold for the chosen measure (e.g., median of at least 7 on a nine-point scale and an interquartile range of less than 2)
5. The data and participants to be used for determining consensus (e.g., all data from all participants who responded in the final rating round).

scientific ecosystem also play an important role. Most notably, journal and research sponsor enforcement has been crucial in improving the rate of prospectively registered clinical trials [59–61]. Consequently, we also call on journals that publish and research funders who sponsor consensus-oriented Delphi processes to require prospective registration of these studies as a precondition for publication and funding.

4.3. Strengths, limitations, and future research

Several aspects of our analysis are worth discussing. First, numerous types of analysis procedures exist for determining consensus in Delphi processes [6]. This article focused on descriptive measures of consensus, given their prominence in health research; future research could further investigate the sensitivity of consensus to different subjective and inferential analysis procedures as well [6,9]. In addition, the online panels we used in this study focused on different health conditions; we thought it was appropriate to pool data across panels as they were part of the same study, had the same objective, had similar groups of stakeholders, and used the exact same rating criteria. Third, we only used data from the final round of the Delphi process, whereas other descriptive analysis procedures for determining consensus use data from multiple rounds—such as reductions in the dispersion in data across rounds [62–64]. It is also worth noting that this study analyzed data from an online, modified-Delphi process that focused on measuring consensus. Consensus-oriented Delphi processes taking place online have become particularly prominent as of late due to their increased efficiency and ability to engage stakeholders from disparate locations [65]. Nonetheless, this type of analysis could be replicated in other Delphi process formats. For example, as we focused exclusively on measures of consensus, future research could investigate the sensitivity of results to dissent-based analysis approaches in Delphi processes [20,66]. Moreover, data from this study (and similar research) [1,6,8,9,36–38] can assist researchers in prespecifying an analysis procedure that is likely to be strict (or lenient) enough for the aspirations of a given Delphi process [16]. Future empirical research on this topic can also provide the foundation for the development of a comprehensive list of consensus measures and corresponding analysis procedures [1], ideally combined with information on how other factors of study design (e.g., number of participants, participant characteristics, number of statements, rating scales) may influence results [6]. This future empirical research would complement existing research on other methodological features and biases that can influence the results in Delphi research [67,68].

5. Conclusion

This study demonstrates considerable variability in the items that reach consensus in Delphi processes when changing the analysis procedure used to determine consensus. These results highlight how undisclosed analytic flexibility makes it unacceptably easy to data mine for and selectively report consensus in Delphi processes. Consequently, a lack of standards and norms for defining and publicly registering analysis plans in advance threatens the credibility of using consensus-oriented Delphi processes to inform health research, practice, and policy [6]. We therefore call for the prospective, complete registration of pre-analysis plans for consensus-oriented Delphi processes.

Acknowledgments

The authors wish to acknowledge Nathaly Pacheco-Santivanez for assistance with panel administration.

Author contributions: S.G. and D.K. designed the study. S.G. and M.B. conducted statistical analyses. S.G. and D.K. wrote the first draft of the article, and all authors contributed to and have approved the final article.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.03.007.

References

[1] 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.

[2] Keeney S, McKenna H, Hasson F. The Delphi technique in nursing and health research. Oxford, UK: John Wiley & Sons; 2010.

[3] Linstone HA, Turoff M. Delphi: a brief look backward and forward. Technol Forecast Soc Change 2011;78(9):1712–9.

[4] de Meyrick J. The Delphi method and health research. Health Educ 2003;103(1):7–16.

[5] Aengenheyster S, Cuhls K, Gerhold L, Heiskanen-Schüttler M, Huck J, Muszynska M. Real-Time Delphi in practice—a comparative analysis of existing software-based tools. Technol Forecast Social Change 2017;118:15–27.

[6] von der Gracht HA. Consensus measurement in Delphi studies: review and implications for future quality assurance. Technol Forecast Social Change 2012;79:1525–36.

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

[8] Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess 1998;2:1–88.

[9] Holey EA, Feeley JL, Dixon J, Whittaker VJ. An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. BMC Med Res Methodol 2007;7:52.

[10] Black N, Murphy M, Lamping D, McKee M, Sanderson C, Askham J, et al. Consensus development methods: a review of best practice in creating clinical guidelines. J Health Serv Res Policy 1999;4:236–48.

[11] Khodyakov D, Hempel S, Rubenstein L, Shekelle P, Foy R, Salem-Schatz S, et al. Conducting online expert panels: a feasibility and

experimental replicability study. BMC Med Res Methodol 2011;11:174.
[12] 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(1):e1000393.
[13] Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7(2):e1000217.
[14] Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, et al. Global prevalence of dementia: a Delphi consensus study. Lancet 2006;366:2112−7.
[15] Cresswell KM, Panesar SS, Salvilla SA, Carson-Stevens A, Larizgoitia I, Donaldson LJ, et al. Global research priorities to better understand the burden of iatrogenic harm in primary care: an international Delphi exercise. PLoS Med 2013;10(11):e1001554.
[16] Mitchell VW. The Delphi technique: an exposition and application. Techno Anal Strateg Manage 1991;3(4):333−58.
[17] Tetzlaff JM, Moher D, Chan AW. Developing a guideline for clinical trial protocol content: Delphi consensus survey. Trial 2012;13(1):176.
[18] Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and Elaboration. PLoS Med 2008;5(1):e20.
[19] Williams PL, Webb C. The Delphi technique: a methodological discussion. J Adv Nurs 1994;19:180−6.
[20] Warth J, Heiko A, Darkow IL. A dissent-based approach for multi-stakeholder scenario development: the future of electric drive vehicles. Technol Forecast Soc Change 2013;80(4):566−83.
[21] Fitch K, Bernstein SJ, Aguilar MD, Burnand B, LaCalle JR. The RAND/UCLA appropriateness method user's manual (No. RAND/MR-1269-DG-XII/RE). Santa Monica, CA: RAND Corporation; 2001.
[22] Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. Perspect Psychol Sci 2012;7(6):632−8.
[23] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22(11):1359−66.
[24] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124.
[25] Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, et al. Promoting transparency in social science research. Science 2014;343:30−1.
[26] van Assen MA, van Aert RC, Nuijten MB, Wicherts JM. Why publishing everything is more effective than selective publishing of statistically significant results. PLoS One 2014;9(1):e84896.
[27] Naylor CD, Basinski A, Baigrie RS, Goldman BS, Lomas J. Placing patients in the queue for coronary revascularization: evidence for practice variations from an expert panel process. Am J Public Health 1990;80:1246−52.
[28] Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, et al. pSCANNER: patient-centered scalable national network for effectiveness research. J Am Med Inform Assoc 2014;21:621−6.
[29] Timbie JW, Rudin RS, Towe V, et al. National Patient-Centered Clinical Research Network (PCORnet) Phase I. Santa Monica, CA: RAND; 2015.
[30] Khodyakov D, Grant S, Meeker D, Booth M, Pacheco-Santivanez N, Kim KK. Comparative analysis of stakeholder experience with an online approach to prioritizing patient-centered research topics. J Am Med Inform Assoc 2017;24:537−43.
[31] Dalal S, Khodyakov D, Srinivasan R, Straus S, Adams J. ExpertLens: a system for eliciting opinions from a large pool of non-collocated experts with diverse knowledge. Technol Forecast Social Change 2011;78(8):1426−44.
[32] Jones MM, Pickett J, Chataway J, et al. Mapping pathways: Developing evidence-based, people-centred strategies for the use of antiretrovirals as prevention. Cambridge, UK: RAND Europe; 2013.
[33] Khodyakov D, Savitsky TD, Dalal S. Collaborative learning framework for online stakeholder engagement. Health Expect 2016;19:868−82.
[34] Rubenstein L, Khodyakov D, Hempel S, Danz M, Salem-Schatz S, Foy R, et al. How can we recognize continuous quality improvement? Int J Qual Health Care 2014;26:6−15.
[35] Claassen CA, Pearson JL, Khodyakov D, Satow PM, Gebbia R, Berman AL, et al. Reducing the burden of suicide in the U.S.: the aspirational research goals of the national action alliance for suicide prevention research prioritization task force. Am J Prev Med 2014;47(3):309−14.
[36] Dajani JS, Sincoff MZ, Talley WK. Stability and agreement criteria for the termination of Delphi studies. Technol Forecast Social Change 1979;13(1):83−90.
[37] Chaffin WW, Talley WK. Individual stability in Delphi studies. Technol Forecast Social Change 1980;16(1):67−73.
[38] Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. BMC Med Res Methodol 2005;5:37.
[39] Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000;32:1008−15.
[40] Hasson F, Keeney S. Enhancing rigour in the Delphi technique research. Technol Forecast Social Change 2011;78(9):1695−704.
[41] Humphreys M, Sanchez de la Sierra R, Van der Windt P. Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. Political Anal 2013;21(1):1−20.
[42] 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(6):e20476.
[43] Meinert CL. Toward prospective registration of clinical trials. Control Clin Trials 1988;9:1−5.
[44] Dickersin K. Why register clinical trials?−Revisited. Control Clin Trials 1992;13:170−7.
[45] Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol 1986;4:1529−41.
[46] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database - update and key issues. N Engl J Med 2011;364:852−60.
[47] Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009;302:977−84.
[48] De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. N Engl J Med 2005;26:1857−9.
[49] De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the international committee of medical journal editors. N Engl J Med 2004;351:1250−1.
[50] Council of Science Editors. ICMJE's statement on clinical trial registration 2006. Available at http://www.councilscienceeditors.org/resource-library/editorial-policies/cse-policies/other-supported-statements/icmjes-statement-on-clinical-trial-registration/. Accessed October 17, 2015.
[51] Wager E, Kleiniert S. Responsible research publication: international standards for authors 2010. Available at publicationethics.org/resources/international-standards-for-editors-and-authors. Accessed October 17, 2015.
[52] Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. N Engl J Med 2015;372:1031−9.
[53] Miller JE, Korn D, Ross JS. Clinical trial registration, reporting, publication and FDAAA compliance: a cross-sectional analysis

and ranking of new drugs approved by the FDA in 2012. BMJ Open 2015;5(11):e009758.

[54] Marin Dos Santos DH, Atallah AN. FDAAA legislation is working, but methodological flaws undermine the reliability of clinical trials: a cross-sectional study. PeerJ 2015;3:e1015.

[55] Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.

[56] Merlin JS, Young SR, Azari S, Becker WC, Liebschutz JM, Pomeranz J, et al. Management of problematic behaviours among individuals on long-term opioid therapy: protocol for a Delphi study. BMJ Open 2016;6(5):e011619.

[57] Schneider P, Evaniew N, Rendon JS, McKay P, Randall RL, Turcotte R, et al. Moving forward through consensus: protocol for a modified Delphi approach to determine the top research priorities in the field of orthopaedic oncology. BMJ Open 2016;6(5):e011780.

[58] Rayens MK, Hahn EJ. Building consensus using the policy Delphi method. Policy Politics Nurs Pract 2000;1(4):308−15.

[59] Gorman DM. 'Everything works': the need to address confirmation bias in evaluations of drug misuse prevention interventions for adolescents. Addiction 2015;110(10):1539−40.

[60] Thaler K, Kien C, Nussbaumer B, Van Noord MG, Griebler U, Klerings I, et al. Inadequate use and regulation of interventions against publication bias decreases their effectiveness: a systematic review. J Clin Epidemiol 2015;68:792−802.

[61] Pandis N, Shamseer L, Kokich VG, Fleming PS, Moher D. Active implementation strategy of CONSORT adherence by a dental specialty journal improved randomized clinical trial reporting. J Clin Epidemiol 2014;67:1044−8.

[62] Spinelli T. The Delphi decision-making process. J Psychol 1983; 113(1):73−80.

[63] Ray PK, Sahu S. Productivity management in India: a Delphi study. Int J Oper Prod Manage 1990;10(5):25−51.

[64] English JM, Kernan GL. The prediction of air travel and aircraft technology to the year 2000 using the Delphi method. Transp Res 1976;10(1):1−8.

[65] Donohoe H, Stellefson M, Tennant B. Advantages and limitations of the e-Delphi technique: implications for health education researchers. Am J Health Educ 2012;43(1):38−46.

[66] Steinert M. A dissensus based online Delphi approach: an explorative research tool. Technol Forecast Social Change 2009;76(3): 291−300.

[67] Ecken P, Pibernik R. Hit or miss: what leads experts to take advice for long-term judgments? Manage Sci 2015;62(7):2002−21.

[68] Winkler J, Moser R. Biases in future-oriented Delphi studies: a cognitive perspective. Technol Forecast Social Change 2016;105: 63−76.