Schifano and Niederberger
Systematic Reviews (2025) 14:14
https://doi.org/10.1186/s13643-024-02738-3

Systematic Reviews

RESEARCH Open Access

How Delphi studies in the health sciences find consensus: a scoping review

Julia Schifano[1*] and Marlen Niederberger[1]

Abstract

Background Delphi studies are primarily used in the health sciences to find consensus. They inform clinical practice and influence structures, processes, and framework conditions of healthcare. The practical research—how Delphi studies are conducted—has seldom been discussed methodologically or documented systematically. The aim of this scoping review is to fill this research gap and to identify shortcomings in the methodological presentation in the literature. On the basis of the analysis, we derive recommendations for the quality-assured implementation of Delphi studies.

Methods Forming the basis of this scoping review are publications on consensus Delphi studies in the health sciences between January 1, 2018, and April 21, 2021, in the databases Scopus, MEDLINE via PubMed, CINAHL, and Epistemonikos. Included were publications in German and English containing the words "Delphi" in the title and "health" and "consensus" in the title or abstract. The practical research was analyzed for the qualitative content of the publications according to three deductive main categories, to which an influence on the result of Delphi studies can be imputed (expert panel, questionnaire design, process and feedback design).

Results A total of 287 consensus Delphi studies were included in the review, whereby 43% reported having carried out a modified Delphi. In most cases, heterogeneous expert groups from research, clinical practice, health economics, and health policy were surveyed. In about a quarter of the Delphi studies, affected parties, such as patients, were part of the expert panel. In the Delphi questionnaires it was most common for standardized Likert scales to be combined with open-ended questions. Which method was used to analyze the open-ended responses was not reported in 62% of the Delphi studies. Consensus is largely (81%) defined as percentage agreement.

Conclusions The results show considerable differences in how Delphi studies are carried out, making assessments and comparisons between them difficult. Sometimes an approach points to unintended effects, or biases in the individual judgments of the respondents and, thus, in the overall results of Delphi studies. For this reason, we extrapolate suggestions for how certain comparability and quality assurance can be achieved for Delphi studies.

Keywords Expert survey, Agreement, Health, Conducting, Reporting, Bias

Background

Delphi studies are used in the health sciences with the primary goal of finding consensus [1–3]. The aim is "to obtain the most reliable consensus of opinion of a group of experts" [4]. The concept of consensus is often understood to be the majority of the participants agreeing on a standardized item [5]. In healthcare, consensus is most frequently measured using percentage agreement [6, 7]. Zarnowitz and Lambros [8] define consensus as "the

*Correspondence:
Julia Schifano
julia.schifano@ph-gmuend.de
[1] Department of Research Methods in Health Promotion and Prevention, Institute for Health Sciences, University of Education Schwäbisch Gmünd, Oberbettringer Straße 200, Schwäbisch Gmünd 73525, Germany

degree of agreement among point predictions aimed at the same target by different individuals" [8]. To reach a consensus, experts participating in a Delphi study evaluate concrete epistemic issues over multiple rounds [3, 9]. A crucial difference from one-time surveys is that the expert panel is not randomly selected and there is no claim of statistical representativity in the results [10].

In a classic Delphi study, the experts' judgments are typically collected anonymously using (online) questionnaires [11]. However, researchers will sometimes modify the classic Delphi so that the survey process fits the goals or the resources available for the project. For instance, this can involve having participants meet face-to-face or limiting the number of rounds from the start [12–14]. In a systematic review of Delphi studies that identify healthcare quality indicators ($n=80$), Boulkedid et al. [13] found that more than half of the Delphi studies reported following a modified Delphi. Yet, these modifications are not always described or justified [15, 16].

Now, alongside what often appear to outsiders to be nontransparent modifications are Delphi variants whose approaches are clearly articulated and justified. Among them are the policy Delphi [17] and real-time Delphi [18]. Some of the reasons for these developments are to better record the general context behind standardized judgments and to enable anonymous debates of the arguments in real time [17, 18]. However, only in individual cases have these different variants been reflected on or evaluated in terms of their methodology. Moreover, how modifications affect the overall results of Delph studies is still to the largest extent unclear. Evaluations of the real-time and classic Delphi found no significant difference in the overall result between the Delphi variants [19, 20]. However, the analysis by Quirke et al. [20] suggests that respondents are less likely to adjust their judgment in a real-time Delphi. When adjusted, it is more in the direction of the group mean than in the classic Delphi [20]. In our view, despite differences in study design, the following characteristics constitute a Delphi study [3, 4, 9]:

1) Survey of several people with specialized knowledge (known as experts) (e.g., operational knowledge, experiential knowledge, functional knowledge, contextual knowledge);
2) Carrying out at least two survey rounds or the option to respond at least two times;
3) Feedback and the (interim) results are presented to the respondents with a possibility to respond.

Typically, Delphi studies also share a focus on complex topics and questions that require a certain level of expertise and experience to answer [9]. The theoretical assumption underlying the Delphi technique is that multiple experts must be asked in order to cover different perspectives [21, 22]. To gather valid and practicable results, those conducting the Delphi must understand the response behavior of the experts, especially when the aim of the Delphi is to reach a consensus [23, 24]. This is because it has direct practical relevance, for instance, when consensus is sought to make concrete recommendations for routine clinical practice (e.g., [25]) or for medical curricula (e.g., [26]). Furthermore, Delphi studies are also very widespread in the field of health [27]. For these reasons, we want to shed light from a methodological perspective on how consensus is found in Delphi studies in the health sciences. However, first, we explain how judgments are formed in Delphi studies and which factors influence them.

Theoretical insights on judgment formation in consensus Delphi studies

In standardized surveys, response behavior is described as an ideal type of process consisting of four steps: (I) understanding the question, (II) retrieving information, (III) evaluating the question, and (IV) submitting the response. Making the cognitive effort in all steps is described by Krosnick [28] as *optimizing*. The intensity of the effort at each step of the response process model depends, among other things, on the motivation and ability of the respondent and the difficulty of the task [28]. Depending on what influences these factors exert, the response strategy of *satisficing* can come into play [28]. When satisficing, the respondent engages more superficially with the steps of the response process or skips over individual steps entirely to give an answer to the question asked [28, 29].

From a cognitive psychology perspective, judgment formation in consensus Delphi studies can be augmented by the idea of mental models [22, 30]. With Delphi studies, it must be assumed that participants go through the steps of the response process in a state of cognitive uncertainty because, typically, uncertain and incomplete knowledge exists regarding the topics, which sometimes extend beyond the experts' main area of expertise [21, 22]. Specialized knowledge on the part of the respondents is required to comprehend, contextualize (step I, Fig. 1), and evaluate questions [3, 22]. When forming judgments (steps II and III, Fig. 1), experts are often required to place the question in a larger context and generate transformation knowledge beyond the scope of their specialty [22, 31], e.g., in regard to the consequences of the judgment for affected groups [32], future generations [33] or other specialized areas in an organization [34]. Furthermore, Delphi studies sometimes integrate additional information which the respondents should consider when forming their judgments, e.g., a summary

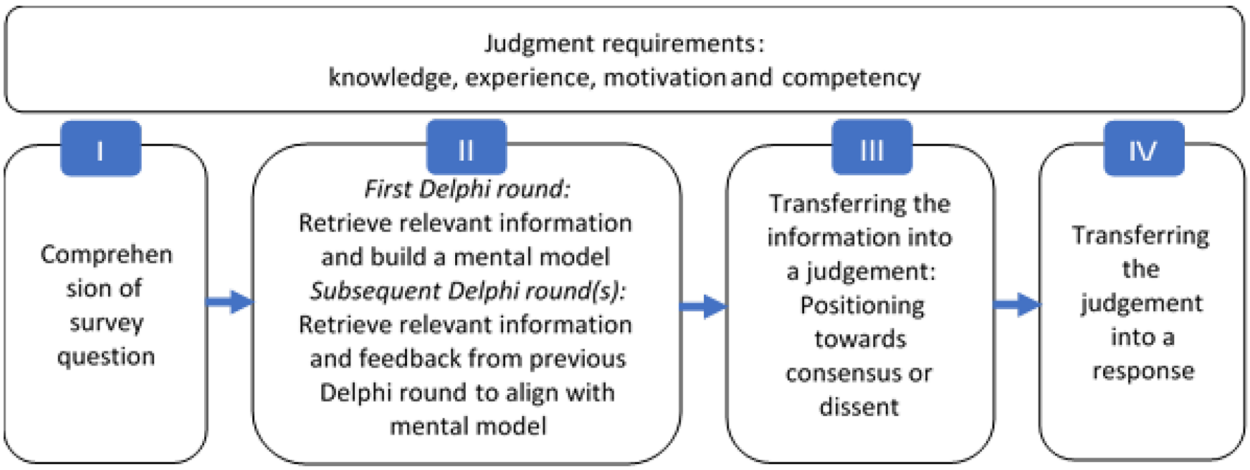

Fig. 1 Theoretical process for reaching optimal judgments in consensus Delphi studies based on Tourangeau et al. 2000 [35]

of the current state of research [25]. Ideally, all of the information is taken into consideration by the respondents when forming judgments (see step II, Fig. 1).

From the second round onward, feedback plays another central role in the process of forming judgments [4, 11]. The difference from the first Delphi round is that the experts have already formed a mental model and they receive the feedback as additional information (step II, Fig. 1), which can consist of arguments put forth by other experts, statistical data regarding the group response, the expert's response from the previous round, or a combination of these [7]. The feedback is meant to encourage the experts to include previously unconsidered aspects in their mental models in order to give a well-founded and carefully thought-out judgment according to the *optimizing* strategy [31, 36].

The individual process of coming to a judgment in a Delphi study is complex and can only function "optimally" under certain conditions (see "Judgment requirements", Fig. 1), and, to be specific, if the respondents [30, 37]:

- Have extensive knowledge of the topic,
- Are familiar with the topic and have experience with it, meaning they regularly engage with the topic under investigation (usually for professional reasons, but also because they are personally affected),
- Have certain cognitive abilities and the motivation to specify, structure, and evaluate information.

In Delphi studies, judgments can also be influenced by other factors, such as the stated aim of finding consensus, thus making optimizing more difficult [38]. There is still little reflection on the theoretical level about how response behavior in Delphi studies takes shape in practice, even though this seems to be highly relevant. If there is little success in getting the experts to optimize their response process, the results will be less precise and less reliable [29]. *Satisficing* [28, 29] in Delphi studies could take these forms:

- Experts avoid a clear judgment on a specific position, e.g., in that they tend toward the middle of the scale or consciously take a decision that differs from the majority.
- Experts respond arbitrarily, e.g., in that they select the first answer on offer.
- Experts do not form a judgment, e.g., in that they leave out questions, choose an evasive category (e.g., "don't know"), or discontinue the survey process.
- Experts form deliberate judgments but, consciously or unconsciously, do not consider all of the information equally, e.g., in that they only include the first response options or arguments in the qualitative feedback when forming their judgment.
- Experts respond such that the Delphi will be terminated, e.g., in that they more or less agree with the majority opinion as presented in the feedback in order to support a statistical consensus.

Contrary to the process outlined in the ideal model (Fig. 1), evidence shows that judgment formation is subject to suboptimal conditions that can make *optimizing* more difficult [29]. Hence, respondents' individual personal characteristics, the situation, or the questionnaire's content and visual presentation have effects on response behavior and thus on the overall results. In the following, we present an overview of methods studies that shed light on these aspects and show the effects on individual judgments.

Methodological findings on judgment formation in Delphi studies

Different methodological findings exist in regard to how respondents form their judgments in Delphi studies, namely:

- Systematic reviews based on publications of Delphi studies [1, 2, 6, 7, 14, 16, 39]
- Method experiments [24, 34, 38, 40–44]
- Evaluation studies [23, 45–48]
- Reports by Delphi practitioners [11, 49, 50]

According to these findings, three factors have a direct or indirect influence on the overall result of a Delphi study: the expert panel, the questionnaire design, and the process and feedback design (Fig. 2). How these three factors exert influence on individual response behavior and the overall result of a Delphi study is explained in the following.

1. The *expert panel* is a central feature of the Delphi technique. Empirical evidence demonstrates that five aspects affect the individual judgments of the respondents:
    - The subjective perception of the baseline (e.g., estimation of the topic's relevance, majority opinions in the field) [41, 42, 48]
    - The actual professional knowledge and experience (e.g., knowledge about current studies, position in an organization, lifeworld experiences) [23, 32, 51–53]
    - The intention to participate in the Delphi study (e.g., personal and/or institutional interests and objectives) [23, 48]
    - Personal characteristics (e.g., value systems) and sociodemographic profile (e.g., age) [23, 32, 43, 52]
    - The assessment of the Delphi study (e.g., relevance or clarity of the study's aims and the Delphi technique) [46, 47]

Although it must be assumed that, along with expertise, these diversity variables have an effect, they are generally not considered in the selection of experts for Delphi studies [52]. Selection is typically done on the basis of professional expertise, e.g., through professional associations [16, 25]. Furthermore, the nature and size of the expert panel composition are relevant to the process of finding consensus. To ensure sufficient professional heterogeneity, it is recommended that experts be recruited through purposive sampling and not snowballing [10, 11]. How large a Delphi panel should be is not determined in terms of method, usually, the number is in the low double-digits [7]. Statistical models demonstrate that groups of this size can deliver stable final results, which depend, of course, on the topic and composition of the expert panel [54, 55]. When different expert groups, e.g., different disciplines, are included and the numbers between them are unequal, biases can emerge in favor of the judgments from the expert group with the most members. Beiderbeck

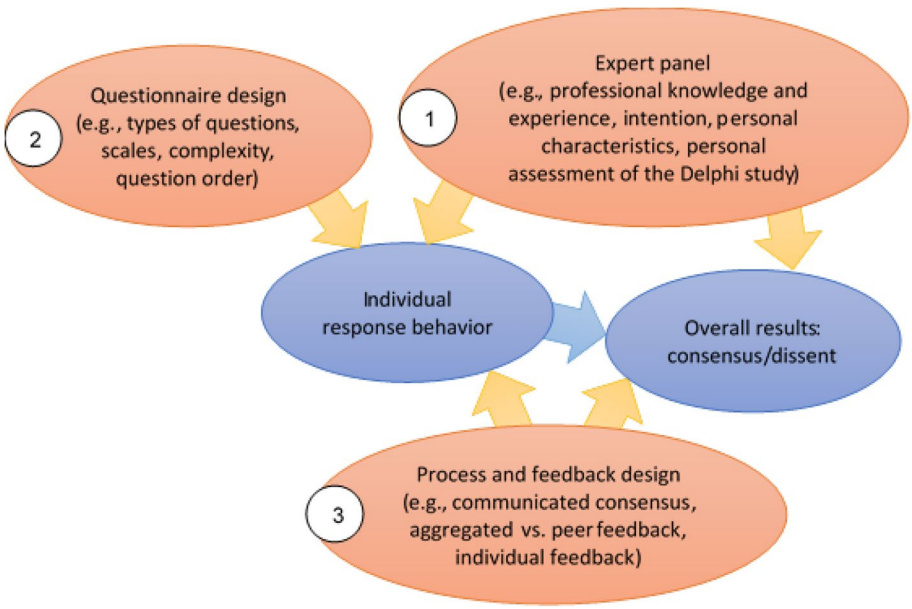

Fig. 2 Factors influencing the results of consensus Delphi studies in the health sciences

et al. [56] therefore advise performing subgroup analyses if there are 15 to 20 members per expert group.

Generally speaking, heterogeneous panels reach consensus overall for fewer aspects of a question than homogeneous panels [34, 53]. Still, the heterogeneity of the expert panel regarding professional knowledge counts as quality enhancing for Delphi studies, despite the partially unclear influences on the overall study results [14, 16, 49, 50]. Influences arising from the individual personalities of the respondents carry less weight as a result [49].

2. With *questionnaire design*, we mean how questions in Delphi studies are (a) presented visually and in terms of content and sequence and (b) designed methodologically, e.g., what types of questions (open/closed) and scales. Bias in the filling out of a questionnaire, as is seen with every standardized survey, can also be avoided in Delphi studies by adhering to the recommendations for designing standardized questionnaires [57].

a) Potential biases due to how questions and responses are formulated and presented can be mitigated in Delphi studies [49]. The background here appears to be that experts (including their mental models) analyze the questions more cognitively than citizens do with questions in survey polls [58].

Nevertheless, a questionnaire's complexity plays an important role in Delphi studies [59]. An item's length should not exceed 25 words according to a recommendation from futures research [60]. Markmann et al. [24] found that longer and more abstract statements make the formation of individual judgments in Delphi studies more difficult and lead to more moderate judgments.

Brookes et al. [61] investigated the effect of biases on judgments in Delphi studies arising from the order in which the questions are asked by presenting topic blocks to the participants in different sequences. By doing so, they were able to determine effects on the judgments of patients and healthcare professionals which were equally relevant to the overall result but could have different effects on it. Brookes et al. [61] and also Hallowell & Gambatese [62] therefore recommend randomizing the questions in Delphi studies, which has been reported by several studies (e.g., [63, 64]).

b) The relevance of open questions varies in regard to Delphi studies. Standardized items dominate mostly, and open comments are used only in individual cases or in an initial qualitative round [10, 65]. However, there are also Delphi studies that focus on the interchange of arguments from open comments [66]. The aim of free-text responses can be to supplement or specify details or to justify or appraise the judgments [65]. The problem, though, is that the handling and analysis of free-text responses is often not undertaken systematically [65]. In these cases, it is questionable if the increased cognitive effort required can be justified to the respondents [67].

The study findings are unclear on scale range and the design of rating scales in Delphi studies. Different reviews show that rating scales are typically used to measure consensus in Delphi procedures and often have five or more graduations [6, 7, 16, 39]. Based on the results of their review of Delphi studies in histopathology, Taze et al. [16] recommend the use of a "nine-point Likert scale with a 'no opinion' option and a free-text comment box" [16]. Initial analyses indicate that scales with different lengths lead to a different end result [68, 69]. In a comparison of three scale lengths (3-point, 5-point, and 9-point rating scales), Lange et al. [68] determined that the 5-point rating scale with a cut-off value of 75% achieved the least consensus and the 9-point scale the most. It must be noted that, while the scales had the same defined cut-off value, different numbers of scale points were included in the definition of consensus [68]. De Meyer et al. [69] found higher consensus with a longer scale, whereby consensus was also defined here using one (3-point scale) or more (9-point scale) scale points. Both studies concluded that recommendations for direct action in clinical practice can be derived with a 3-point scale (e.g., "main goal," "secondary goal" and "no goal") and the result is simpler to interpret than with longer scales [68, 69].

3. Another factor influencing individual judgments and the overall result is *process and feedback design***.** This influence is seen on three levels: a) the communicated consensus, b) the aggregated feedback and c) the individual feedback on a participant's response from the previous Delphi round.

a) Barrios et al. [38] differentially analyzed the influence of the level of agreement on the individual judgments and observed that if the consensus was over 75%, participants were more likely to converge with the opinion of the group than if

the group's aggregated agreement was below this value [38]. Signs of a conscious judgment against the consensus were observed by Barrios et al. [38] when the value in the feedback lay below the consensus level of the percent agreement. They speculate that experts consciously manipulate results as a result of revealing the consensus [38]. Although this influence is theoretically probable, we are not aware of other publications on the effect of disclosing the level consensus or dissent on individual judgments, e.g., in the communication of the Delphi study's aims or as part of the feedback.

b) Feedback involving the statistical group response dominates in Delphi studies, while peer feedback is less frequently given [7]. The extent to which the form and type of feedback (qualitative or quantitative) influences judgment behavior is controversial [12, 41, 49]. Some Delphi practitioners support the use of qualitative instead of quantitative feedback because then the experts do not prematurely side with the majority opinion (bandwagon effect) [49, 50]. However, this assumes that the open responses are not presented in an unfiltered form, but rather systematically analyzed, which, as already described, is not always the case [49, 65]. In addition, the effects of differentiating the feedback by expert group have already been shown [40]. Brookes et al. [40] have demonstrated that, if the feedback contains information on different groups of participants, the level of agreement between the expert groups increases compared to peer feedback. MacLennan et al. [70] have also carried out a randomized Delphi study with different feedback strategies but were unable to confirm the effects observed by Brookes et al. [40]. Fish et al. [43] assert the hypothesis that in comparison to healthcare professionals, patients less often integrate the feedback of other expert groups and hence do not reflect as much on judgments made from other perspectives [43]. Turnbull et al. [45] show similar findings.

c) A randomized experimental study on urban sustainability by Meijering & Tobi [44] demonstrated that experts less often adjust their judgment when they see their response from the previous round; however, an effect on the final consensus could not be determined [44].

We are not aware of analyses of other factors affecting the process and feedback design, e.g., how the termination criterion or the numbers of rounds influence individual judgments and the overall result. Having said this, though, the termination criterion and the number of rounds are relevant in order to ascertain whether the consensus is stable and valid [1, 2].

Aims of the scoping review

The methodical tests, experiments, and discussions presented here concerning Delphi studies in the health sciences ultimately identify three factors that can be alleged to exert an influence on the results of a Delphi and thus on the consensus: the expert panel, questionnaire design, and process and feedback design (Fig. 2). These three factors serve as the basis for the present scoping review. The aim is to highlight the range of methodological approaches in relation to these three factors and to identify the areas that have received little attention [71].

The following research question is answered in this scoping review:

How are the influential factors described here used in the practice of consensus Delphi studies in the health sciences?

In addition to these three proven factors, there are indications of other factors that influence individual judgments and the overall results of Delphi studies, e.g., the effect of the time between rounds [43], though such factors have not yet been fully examined explicitly for Delphi studies, e.g., the effect of sponsors or members of a supervisory group on the overall result. In general, a decrease can be seen in publications that examine Delphi studies in terms of methodology compared to Delphi primary studies [27]. Flostrand et al. [27] showed that the ratio of methodological studies to Delphi primary studies was 1:1 in 1975 and 1:19 in 2016. Other factors will not be considered in this scoping review due to a lack of evidence. In addition to the three proven factors, we identify general criteria for describing Delphi studies, including the Delphi variants and the definition of consensus. It must be noted that this review of research practice is based on publications of Delphi studies even though a lack of clarity and sometimes even errors have been repeatedly shown to exist in such publications [5, 7, 14]. Based on the analysis of research practice and taking into account theoretical and methodological findings on judgment building in Delphi studies, we draw conclusions for the quality-assured implementation of Delphi studies.

Method

The reporting in this review is based on the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [71]. This methodological approach has been discussed at different times with members of the German-speaking Delphi expert network (DEWISS), which is comprised of over

20 academics from various disciplines in the health and social sciences and epistemological approaches. DEWISS receives funding from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), an interdisciplinary institution promoting science and research in Germany (project number 429572724, more information is available at https://delphi.ph-gmuend.de/). No study protocol was published.

Search strategy and study selection

The search for publications was conducted on the basis of an already existing collection of 7,044 Delphi primary studies compiled by the Delphi expert network (available via https://www.zotero.org/groups/4396781/dewiss_datenbanken_delphi-studien/collections/25H44TFI). Table 1 summarizes the method used to compile the primary studies.

The authors are part of the Delphi network and collaborated in creating the collection of Delphi primary studies, which was compiled using central databases in the health and social sciences (Scopus, MEDLINE via PubMed, CINAHL, and Epistemonikos). Original works in German or English published between January 1, 2016, and April 21, 2021, were searched for in the four databases using the search term "title:(delphi*) OR abstract:(delphi*)." Following this, the titles and abstracts were screened by one researcher to include the English- and German-language Delphi studies in the health and social sciences. A total of six researchers were involved in the screening process and were in close contact with each other.

Inclusion criteria

Guidance on reporting Delphi studies in palliative care has been available since 2017 via the Equator Network [15]. It is possible that this has an influence on the reporting and practice of Delphi studies in the health sciences and differences are visible from previous systematic reviews of Delphi study reporting [6, 7, 16]. For this reason, publications starting in 2018 are included in this review.

To identify classic or modified Delphi procedures in the health sciences aimed at establishing consensus, the DEWISS data resource of 7044 Delphi primary studies were filtered using the search term "(Title fields: health OR Abstract: health) AND (Title fields: consensus OR Abstract: consensus) AND Title fields: Delphi." The inclusion criteria in the subsequent full-text screening entail the presence of a Delphi primary study on a health-related topic published in German or English. In regard to Delphi variants, both classic and modified Delphi studies were included when the aim was to find consensus and the consensus criteria were defined (Table 2). Studies that conducted a Delphi study in combination with another study, e.g., a previous systematic review to develop the Delphi questionnaire, were also included.

Any publications that did not meet one or more of the three constitutive characteristics (expert survey, iterative rounds, feedback) or did not report on them were excluded. According to our definition, studies that survey experts in multiple rounds but do not share the results of previous rounds with the experts during the survey process are not Delphi studies. Here it must be noted that this decision is made solely on the basis of the publication leaving the possibility that some studies were excluded even though they have all the characteristics of a Delphi. The reason for exclusion was documented (Fig. 3). The full texts were screened by one of the authors (JS). The authors conferred with each other when uncertainty arose about the inclusion of individual studies. The authors then discussed and agreed whether the study should be included or excluded.

Data extraction and analysis

A qualitative analysis strategy, namely qualitative content analysis, was selected to assess the publications [72]. This enables the analysis of large datasets and the inductive identification of categories based on the material and a picture of their range. Quantification of the results is

Table 1 Delphi expert network (DEWISS) data resource on Delphi primary studies

Search process	Description
Databases searched	• Scopus, MEDLINE, CINAHL, Epistemonikos in April 2021
Search strategy	• Keywords "Delphi*" in title or abstract from 2016 to April 2021
Selection criteria	• English or German language • Original papers with Delphi studies
Data extraction	• Title-abstract screening by six researchers and research fellows in the Delphi expert network without verification by a second reviewer

The * symbol allows you to include other terms that are used in connection with the word Delphi, e.g. Delphi study, Delphi-study, Delphi-Verfahren, Delphi-Studie

Table 2 Inclusion and exclusion criteria

	Inclusion criteria	Exclusion criteria
Article type	Original paper of a Delphi study	Systematic reviews, study protocols, and commentaries (on Delphi studies)
Language	German, English	A language other than English or German
Topic	Relevance to health	Studies on economic and technical topics not related to health
Delphi variants	Traditional or modified Delphi study that has the constitutive characteristics of a Delphi: 1) Survey of several people with specialized knowledge (known as experts) (e.g., operational knowledge, experiential knowledge, functional knowledge, contextual knowledge) 2) Carrying out at least two survey rounds or the option to respond at least two times 3) Feedback, the (interim) results are presented to the respondents with a possibility to respond	Study types other than Delphi studies, if these were not conducted in combination with a Delphi study
Study aim	Finding consensus, whereby criteria are defined according to which the consensus was determined	It is explicitly stated that reaching a consensus is not the aim of the Delphi study

also possible. Deductive main categories were formed as the basis for the qualitative analysis. These are the three factors that allegedly influence judgment formation: (1) the expert panel, (2) the questionnaire design, and (3) the process and feedback design. Using these deductive main categories, the research on the conduction of Delphi studies was documented, inductively filled in, and supplemented in terms of content. We have divided the inductive subcategories into two levels, with the second level further differentiating the first level. The formation of inductive subcategories was done by the first author using a random selection of approximately 10% of the included publications. This complies with current guidance for qualitative content analyses [72].

The category system was then discussed and revised with three of the seven members of the DEWISS network from the "Reporting Guideline" working group to assess the completeness and clarity of the definitions. The revised category system was verified anew using a random selection of 10% of the included publications. Some of the subcategories turned out to be difficult to analyze or did not yield much information due to a lack of uniformity in the reporting. For example, it was not possible to comparatively determine how much consensus was achieved in the Delphi studies because it was often unclear how many items had been analyzed in total. Also, the terms "question" and "item" were sometimes used synonymously or imprecisely.

Finally, the first author formulated explanations and added examples for all subcategories, and then discussed and finalized the category system again with the second author.

Final category system

The final category system entailed 4 deductive main categories, 22 subcategories for the first level, and 58 subcategories for the second level (Table 3). The analysis of all of the publications was carried out by one researcher (first author) using Microsoft Excel.

Results

The search for consensus Delphi studies in the health sciences in the database of Delphi primary studies yielded 538 hits. Forty-eight studies were excluded in the first step of full-text screening because they were not original papers on Delphi primary studies or did not meet other inclusion criteria (e.g., German or English language) (Fig. 3). For 40% ($n=194/490$) of the studies it was unclear if a Delphi study with the constitutive characteristics (Table 2) had been carried out or the criteria for determining consensus remained unclear (Fig. 3). A total of 287 studies satisfying the inclusion criteria were included in the analysis (Additional file 1).

General aspects

Area

Consensus Delphi studies in healthcare were assigned to four topical areas: clinical patient care (e.g., ID1, ID21, ID204, ID213),[1] healthcare/public health (e.g., ID6, ID18, ID93, ID168), medical education (e.g., ID83, ID128, ID162) and methodological health research (e.g., ID134, ID164) (Table 4). The aims for reaching consensus include the development of guidelines and recommendations for the diagnosis and therapy of disease (e.g., ID11, ID170), the forecasting of healthcare needs and research priorities (e.g., ID67, ID199), the development

[1] ID refers to the analyzed publication. An overview of the studies analyzed and the results is presented in Additional file 1.

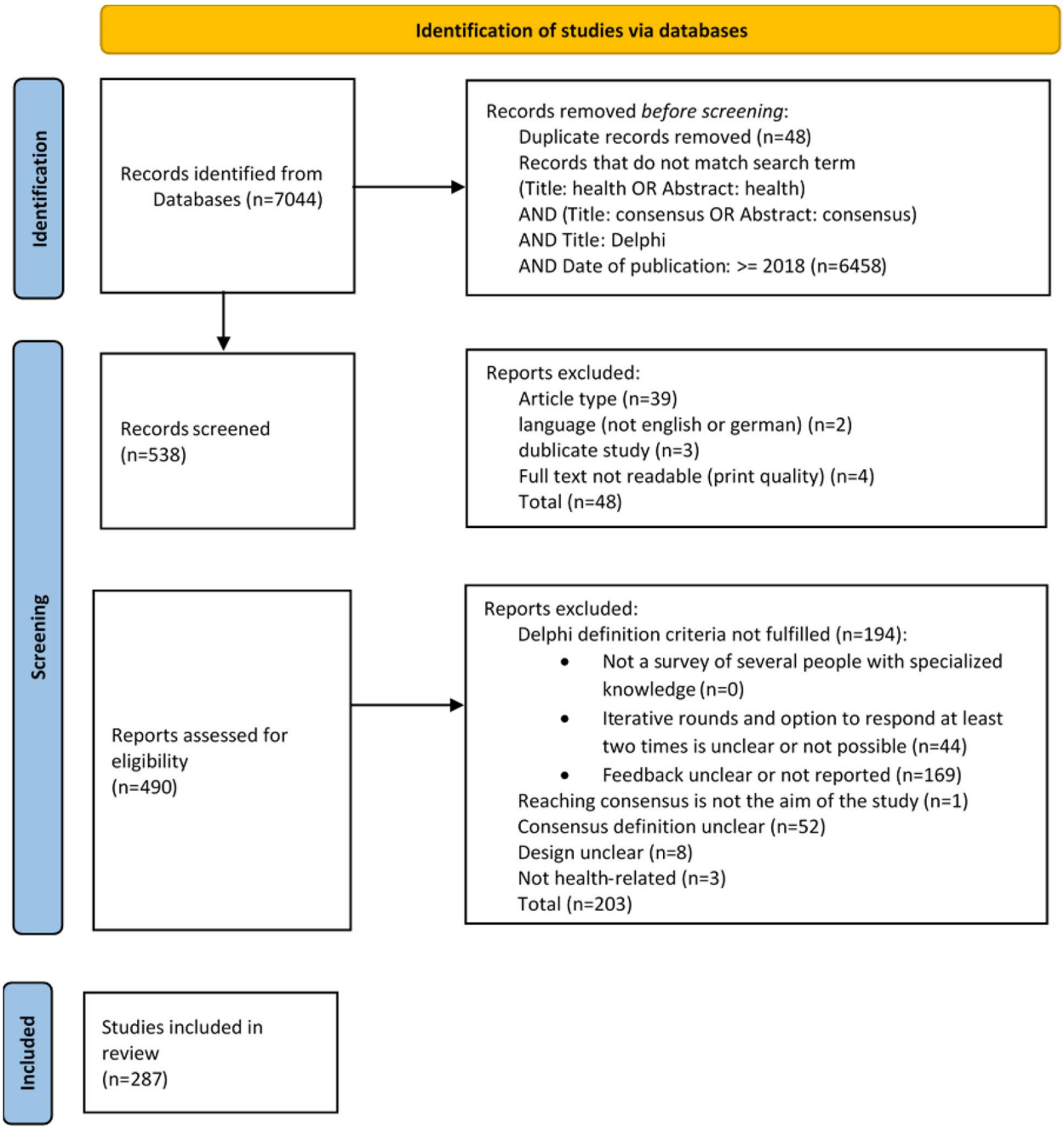

Fig. 3 Literature screening process

of measuring instruments and validation of findings (e.g., ID180, ID191), agreement on definitions and terminologies (e.g., ID1, ID146) and defining competency profiles and curricula in healthcare (e.g., ID176, ID267).

Delphi variant

The classic Delphi technique is most often selected in the studies analyzed, but in 43% ($n=122/287$) of the Delphi studies it is reported that a modified Delphi was carried out (Table 4). Among others, the modifications identified were an initial survey round with closed instead of open questions (e.g., ID5, ID71), face-to-face meetings of the participants, or a combination of anonymous survey rounds with group discussions (e.g., ID6, ID32, ID95), the integration of collecting and analyzing qualitative data (e.g., ID19, ID36), no meeting of the participants

Table 3 Category system to analyze the Delphi studies

No	Main category	Subcategory 1	Subcategory 2
1	General aspects	Area	1. Clinical patient care = diagnosis and therapy of diseases in inpatient settings, e.g., ID5[a] 2. Healthcare services/public health = management of diseases, availability of care, access to healthcare, policy implication, e.g., ID35 3. Medical education = teaching and studying in health science programs, competencies of healthcare professionals, e.g., ID83 4. Methodological health research = methods in healthcare, research on research, e.g., ID134
2	General aspects	Delphi variant	1. Classic = reported as a classic Delphi study or not reported as modified Delphi study 2. Modified = reported as modified Delphi study
3	General aspects	Consensus criterion for rating scales	1. Standardized measure of dispersion = e.g., coefficient of variation, interquartile range, standard deviation 2. Standardized measure of central tendency = e.g., median, mean 3. Percentage agreement (one scale point) = proportion of agreement with a value, e.g., 70% vote for 5 on a 5-point scale 4. Percentage agreement (adjacent scale points) = proportion of agreement with two adjacent values, e.g., 70% vote for 3 or 4 on a scale of 1–5 5. Percentage agreement (other conditions) = other criteria for measuring percentage agreement, e.g., less than 15% vote 1 or 2 and at least 70% vote 6 or 7 on a 7-point scale or proportion of agreement within specific subgroups 6. Percentage agreement (unclear) = unclear definition of consensus, e.g., unclear which scale items were used to measure percent agreement 7. Dependency analyses = e.g., Kendall's coefficient of concordance, Spearman's rho 8. Other criteria = e.g., number of outcomes predefined, content validity index, RAND/UCLA disagreement index, diversity of responses
4	General aspects	Percentage level consensus	Reported percentage level consensus in, e.g., 75%. Criteria may differ between Delphi rounds. In this case, all criteria were noted
3	Panel of experts	Sampling strategy	1. Snowball sampling = researcher relies on participant referrals to recruit new participants, e.g., recruiting colleagues from your own network 2. Purposive sampling = researcher seeks out participants with specific characteristics, e.g., recruiting researchers on the topic of artificial intelligence in clinical patient care 3. Purposive quota/random sampling = researcher randomly selects cases from within several different subgroups/quota, e.g., random selection of a number of the identified researchers on the topic of artificial intelligence in clinical patient care 4. Pool from a previous project or register = researchers select cases from a previous project or register, e.g., participants from a previous study 5. Convenience sampling = the authors reported to have selected according to convenience sampling, e.g., researcher gathers data from whatever cases happen to be convenient 6. Open calls = researchers recruit through open calls, e.g., through professional societies, regional networks, and advertisements on social media platforms
4	Panel of experts	Number of participants first round	Reported number of experts completing the first survey round
5	Panel of experts	Number of participants last round	Reported number of experts completing the last survey round
6	Panel of experts	Heterogeneity of expertise	1. Homogeneous = only one group of participants, e.g., nurses 2. Heterogeneous = the panel consists of participants from different disciplines and/or subject areas, e.g., nurses, care managers, nursing researchers 3. Heterogeneous including everyday life experts (e.g., patients) = the panel consists of participants from different disciplines and/or subject areas including affected persons, e.g., patients, patient representatives, affected persons

Table 3 (continued)

No	Main category	Subcategory 1	Subcategory 2
7	Panel of experts	Scope	1. national = one country, e.g., Germany 2. international = two or more countries without local scope, e.g., Germany and South Africa 3. Local = local cross-national focus, e.g., German-speaking region 4. Regional = specific region in one country, e.g., central Berlin
8	Questionnaire design	Survey software	Name of the digital platform for conducting the survey rounds, e.g. SurveyMonkey, LimeSurvey, Google forms, Microsoft Excel
9	Questionnaire design	Question types first Delphi round	1. Closed questions = questions with one or more answer options to choose from, e.g., rating, ranking, and multiple-choice questions 2. Closed questions with the possibility to comment = questions with one or more answer options to choose from and the possibility to comment on answers, e.g., including the option to reformulate or suggest new items 3. Open-ended questions = exclusively questions with free-text fields, e.g., in a first qualitative Delphi round
10	Questionnaire design	Number of items first Delphi round	Reported number of items or questions of the first survey round. Subdivision according to items and questions was not given in every case
11	Questionnaire design	Question types last Delphi round	1. Closed questions 2. Closed questions with the possibility to comment 3. Open-ended questions
12	Questionnaire design	Number of items last Delphi round	Number of items or questions of the last survey round. Subdivision according to items and questions was not given in every case
13	Questionnaire design	Width of rating scales	Width of rating scales, e.g., 4-point scale (1 = strongly agree, 4 = strongly disagree). If the response options or scale endpoints are not reported or are unclear, only the scale width is noted, e.g., 4-point scale
14	Questionnaire design	Rating scale, evasive category	Use of an evasive category, e.g., "unsure" or "don't know"—option to answer, option to answer "Absent" due to a lack of perceived expertise. Recorded as reported or not reported
15	Questionnaire design	Randomization of questionnaire content	Randomization of question blocks, questions in question blocks, answer options in questions, e.g., through the survey software randomly assigning respondents. Recorded as reported or not reported
18	Process and feedback design	Timing of consensus definition	1. A priori = determined before the Delphi round 2. A posteriori = determined after the Delphi round
19	Process and feedback design	Method or literature reference for the analysis of qualitative data	1. Content analysis 2. Thematic analysis 3. Inductive approach 1–3 = reported method as mentioned in the text, e.g., thematic analysis 4. Other = e.g., grounded theory, quantitative analysis
20	Process and feedback design	Feedback designed to reconsider the judgments	1. Group response or summary = summary of qualitative date, e.g., comments from open-ended questions, or summary of quantitative data, e.g., statistical feedback of results from closed-ended questions 2. Group response or summary of different groups of participants = peer feedback of one or different groups of participants, e.g., caregivers received feedback from patients or only from caregivers 3. Individual response = display the respondent's answer from the previous round

Table 3 (continued)

No	Main category	Subcategory 1	Subcategory 2
21	Process and feedback design	Termination criterion	1. Consensus reached = achieving consensus in the Delphi study on all or the majority of the issues 2. Number of rounds = terminate the Delphi study after a predefined number of rounds, e.g., after two rounds of voting 3. Stability of judgments = terminate the Delphi study if the judgments are stable, e.g., determined through interquartile range, changes in mean scores 4. Other criteria = other criteria to terminate the Delphi study, e.g., when no new items are proposed, the judgments align, the response rate dropped below a certain value
22	Process and feedback design	Number of rounds	Reported number of survey rounds/iterations, e.g., three Delphi rounds

[a] The ID refers to the analyzed publication. An overview of the analyzed studies and results is shown in Additional file 1

Table 4 Main category "General aspects" of the Delphi studies included in the scoping review

No	Subcategory 1	Subcategory 2: Frequency % (*n*/287) and statistics
1	Area	1. Clinical patient care: 38% (*n* = 110) 2. Healthcare services/public health: 39% (*n* = 112) 3. Medical education: 14% (*n* = 40) 4. Methodological health research: 9% (*n* = 25)
2	Delphi variant	1. Classic: 57% (*n* = 165) 2. Modified: 43% (*n* = 122)
3	Consensus criterion for rating scales	1. standardized measure of dispersion: 20% (*n* = 57) 2. standardized measure of central tendency: 21% (*n* = 60) 3. Percentage agreement (one scale point): 14% (*n* = 40) 4. Percentage agreement (adjacent scale points): 45% (*n* = 129) 5. Percentage agreement (other conditions): 13% (*n* = 37) 6. Percentage agreement (unclear): 12% (*n* = 35) 7. Dependency analyses: 2% (*n* = 6) 8. Other criteria: 3% (*n* = 10)
4	Percentage level consensus	Mean (standard deviation): 73,1% (8,3) Minimum/maximum: 40%/100% Median: 75%

during the process (e.g., ID14), conducting the study as an online survey/e-Delphi (e.g., ID28, ID76) or defining the number of rounds a priori (e.g., ID16, ID42). What is viewed as a modification in these studies varies and is, in part, even contradictory. Sometimes it is not reported at all (e.g., ID8, ID16, ID89, ID206, ID279). For 80% (*n* = 231/287) of the Delphi studies, it is reported that the Delphi took place online. Meetings between the participants took place in 16% (*n* = 45/287) of the Delphi studies as part of the process. Specific Delphi variants were identified in individual cases, for instance, the argumentative Delphi (e.g., ID214), policy Delphi (e.g., ID31, ID73, ID150, ID214), real-time Delphi (e.g., ID111, ID241), or fuzzy Delphi (e.g., ID284).

Definition of consensus

Reporting the definition of consensus was stipulated as an inclusion criterion for the scoping review. Consensus was defined in 30% of the Delphi studies (*n* = 87/287) using statistical measures. Most frequently, in 81% (*n* = 232/287) of the Delphi studies, the consensus was defined as percentage agreement, which is understood differently among the Delphi studies, e.g., whether one or several adjacent scale points are used (Table 4). The cut-off values are generally at 70% (*n* = 87), 75% (*n* = 49) or 80% (*n* = 67). Some Delphi studies divide the strength of consensus according to different levels (e.g., ID73, ID118, ID142, ID150, ID155, ID177), whereby these assignments entail three to five levels (ID73: perfect consensus, consensus, wide agreement, majority, and large minority; ID118: high consensus, low consensus, no consensus; ID142: very high, high, moderate, low; ID150, ID155 and ID177: high, moderate, low, no consensus).

Panel of experts

Sampling strategy

Ninety-four percent (*n* = 271/287) of the Delphi studies report how the participants were recruited, e.g., through

purposive or snowball sampling (Table 5). In 62% ($n=178/287$) of the Delphi studies the authors describe a concrete sampling strategy and 32% ($n=93/287$) apply a combination of different strategies.

Number of participants (first and last round)

Around 60 experts on average participated in the first Delphi round and 47 in the final Delphi round. The studies vary widely from 4 (ID39) to 1014 (ID51) participants in the first Delphi round and from 3 (ID234) to 713 (ID51) in the last round (Table 5). In 83% ($n=239/287$) of the Delphi studies the number of experts in the final Delphi round is the same size or smaller than in the first round. Often only the experts who completed the previous round are invited to participate in the subsequent rounds (e.g., ID1, ID24, ID49, ID73, ID147, ID207). However, there are also Delphi studies in which the panel is enlarged (e.g., ID16, ID70, ID71, ID95, ID111) or all of the experts are invited to participate, regardless of their participation in the separate Delphi rounds (e.g., ID2, ID27, ID28, ID62, ID124, ID205, ID233). Regarding the Delphi studies with the same or fewer number of participants in the final Delphi round as compared to the first, the expert panel in the final round is on average 19% smaller than in the first Delphi round.

Heterogeneity of the panel and scope

Recruitment was generally done according to criteria defining the expertise or the professional background of the experts (e.g., ID3, ID20, ID79, ID268). A homogeneous expert panel is reported by 6% ($n=18/287$) of the Delphi studies, meaning that it consisted of only one expert group (e.g., ID7, ID33). In the analyzed publications, the heterogeneity of the panel is reported as a quality criterion for the studies (e.g., ID81) but the shape this takes differs. These panels are, as a result, multidisciplinary in their composition, meaning that different disciplines are present in one area of expertise (e.g., ID5, ID38, ID130), or they are transdisciplinary, meaning different areas of expertise represented, such as theory and practice (e.g., ID92, ID100, ID181). Affected populations, e.g., patients, are included in 27% ($n=76/278$) of the Delphi studies. The scope of recruitment for the Delphi studies is 51% ($n=145/287$) at the national level and 36% ($n=102/287$) internationally (Table 5). For 3% ($n=8/287$) of the publications, the specific regional focus remained unclear.

Questionnaire design

Survey software

Approximately 30 different software programs were named as the online survey platform, whereby these are primarily designed for conventional population surveys

Table 5 Main category "Panel of experts" of the Delphi studies included in the scoping review

No	Subcategory 1	Subcategory 2: Frequency % (*n*/287) and statistics
3	Sampling strategy	1. Snowball sampling: 29% ($n=82$) 2. Purposive sampling: 78% ($n=225$) 3. Purposive quota/random sampling: 2% ($n=6$) 4. Pool from a previous project or register: 6% ($n=17$) 5. Convenience sampling: 3% ($n=10$) 6. open calls: 12% ($n=34$) *Unclear or not reported: 6% (n = 16)*
4	Number of participants first round	Mean (standard deviation): 60,2 (105,8) Minimum/maximum: 4/1014 Median: 31 *Unclear or not reported: 4%* ($n=12$)
5	Number of participants last round	Mean (standard deviation): 46,8 (72,7) Minimum/maximum: 3/713 Median: 26 *Unclear or not reported: 5% (n = 15)*
6	Heterogeneity of expertise	1. Homogeneous: 6% ($n=18$)[a] 2. Heterogeneous: 65% ($n=187$) 3. Heterogeneous including everyday life experts (e.g., patients): 25% ($n=73$) *Unclear or not reported:* 3% *(n = 9)*
7	Scope	1. National: 51% ($n=145$) 2. International: 36% ($n=102$) 3. Local: 6% ($n=17$) 4. Regional: 5% ($n=15$) *Unclear or not reported: 3% (n = 8)*

[a] Of these, three studies were only with patients

and not explicitly for carrying out Delphi studies, e.g., Qualtrics (https://www.qualtrics.com), SurveyMonkey (https://www.surveymonkey.com), and Google Forms (https://www.google.com/forms/about/). Qualtrics software was used most often making up 19% of the cases (Table 6). In 39% ($n=111/287$) of the Delphi studies the survey software was not identified or none was used if, for instance, the questionnaires were sent directly to the participants via email.

Question types and number of items (first and last Delphi round)

Standardized and open questions were used in the questionnaires, with the number of standardized items increasing with the number of rounds. A total of 10% ($n=28/287$) of the Delphi studies reported asking only closed questions or did not report on the options to comment openly in the first round; the same applied to 44% ($n=125/287$) of the Delphi studies in regard to the final round (Table 6). The number of standardized questions ranged from only a few (e.g., ID132, ID136) to several hundred (e.g., ID55, ID56, ID101) in the first Delphi round, with the observation that there is a tendency to decrease with each Delphi round.

Rating scale (width and evasive category)

In 79% of the Delphi studies ($n=226/287$) it was reported that an odd-numbered rating scale was used

Table 6 Main category "Questionnaire design" of the Delphi studies included in the scoping review

No	Subcategory 1	Subcategory 2: Frequency % (*n*/287) and statistics
8	Survey software	1. Qualtrics: 19% ($n=55$) 2. SurveyMonkey: 12% ($n=34$) 3. REDCap: 7% ($n=21$) 4. Google Forms: 5% ($n=15$) 5. LimeSurvey: 4% ($n=12$) 6. Other: 15% ($n=42$) *Unclear or not reported: 39% (n = 111)*
9	Question types first Delphi round	1. Closed questions: 10% ($n=28$) 2. Closed questions with a possibility to comment: 66% ($n=189$) 3. Open-ended questions: 22% ($n=63$) *Unclear or not reported: 2% (n = 7)*
10	Number of items first Delphi round*	*Closed questions (with a possibility to comment)* Mean (standard deviation): 60,6 (61,1) Minimum/maximum: 1/525 Median: 45 *Open-ended questions* Mean (standard deviation): 5,2 (6,1) Minimum/maximum: 1/26 Median: 3 *Unclear or not reported: 17% (n = 50)*
11	Question types last Delphi round	1. Closed questions: 44% ($n=125$) 2. Closed questions with a possibility to comment: 50% ($n=144$) 3. Open-ended questions: 0% ($n=0$) *Unclear or not reported: 6% (n = 18)*
12	Number of items last Delphi round*	Mean (standard deviation): 39,1 (43,5) Minimum/maximum: 1/289 Median: 24 *Unclear or not reported: 23% (n = 66)*
13	Width of rating scale	2-point scale: 11% ($n=31$) 3-point scale: 9% ($n=27$) 4-point scale: 10% ($n=28$) 5-point scale: 40% ($n=116$) 6-point scale: 2% ($n=5$) 7-point scale: 11% ($n=33$) 9-point scale: 20% ($n=56$) 10-point scale: 4% ($n=12$) 11-point scale: 1% ($n=4$) *Unclear or not reported/not applicable: 5% (n = 14)*
14	Rating scale, evasive category	Reported: 20% ($n=56$) *Unclear or not reported: 80% (n = 231)*
15	Randomization of questionnaire content	Reported: 2% ($n=7$) *Not reported or not applicable:* 98% ($n=280$)

*In one Delphi study the mean value was based on the given range of items because the study participants had received differing numbers of items

Table 7 Main category "Process and feedback design" of the Delphi studies included in the scoping review

No	Subcategory 1	Subcategory2: frequency % (*n*/287) and statistics
17	Timing of consensus definition	1. a priori: 36% (*n* = 103) 2. a posteriori: 3% (*n* = 8) *Unclear or not reported: 61% (n = 176)*
18	Method and/or literature reference for the analysis of qualitative data	1. Content analysis: 14% (*n* = 41) 2. Thematic analysis: 13% (*n* = 38) 3. Inductive approach: 1% (*n* = 3) 4. Other: 1% (*n* = 3) *Unclear or not reported: 62% (n = 178)* *Not applicable (n = 26): 9%*
19	Feedback design to reconsider the judgments	1. Group response or summary: 93% (*n* = 268) 2. Group response or summary of different groups of participants: 6% (*n* = 18) 3. Individual response: 46% (*n* = 132) *Unclear or not reported: 0% (n = 0)*
20	Termination criterion	1. Consensus reached: 23% (*n* = 66) 2. Number of rounds: 30% (*n* = 85) 3. Stability of judgments: 9% (*n* = 26) 4. Other criteria: 1% (*n* = 4) *Unclear or not reported: 46% (n = 132)*
21	Number of rounds	Mean (standard deviation): 2,8 (0,8) Minimum/maximum: 2/8 Median: 3 *Unclear or not reported: < 1% (n = 1)*

to capture the individual judgments. Five-point Likert scales were most commonly used in 40% (*n* = 116/287) of the Delphi studies (Table 6). Sometimes over the course of a Delphi study, other rating scales with different scale widths are used (e.g., ID13, ID73, ID114, ID124, ID143, ID251). No Delphi study reported more than 11 scale points. Differences can be seen in the identification of the scale gradations. The scales or response options are either offered verbally (e.g., "strongly agree," "strongly disagree" (ID67)) or verbally and numerically (e.g., 1 = disagree, 5 = agree (ID11)). Overall, in 20% (*n* = 56/278) of the Delphi studies it is reported that an evasive category could be selected. These are presented either as a separate response option (e.g., ID1, ID20, ID45, ID51, ID260) or a middle category (e.g., ID47, ID56, ID66, ID81, ID152).

Randomization of questionnaire content

The content of the Delphi study questionnaires was organized on the level of the questionnaire itself, e.g., in the sequencing of the content according to topic (e.g., ID24, ID34, ID52, ID56), and on the level of the questions, e.g., when response options were sorted according to frequency in the previous round (e.g., ID14, ID237, ID236). Only 2% (*n* = 7/287) of the Delphi studies stated that the content of the questionnaire—the topic segments or response options—was randomized.

Process and feedback design

Timing of consensus definition

The point in time at which consensus was defined is not reported in 61% (*n* = 176/287) of the Delphi studies. In 36% (*n* = 103/287) of the Delphi studies the authors state that consensus was defined in advance (Table 7).

Data analysis and feedback design to reconsider the judgments

In almost all of the Delphi studies (93%(*n* = 268/287)) it was reported that feedback was given to the participants in the form of information on the statistical group responses and/or a summary of the qualitative or quantitative data. It is also the case, where over half of the Delphi studies included free-text fields in the questionnaire (Table 6), and 62% (*n* = 178/287) of the Delphi studies do not report the analytical method applied to the qualitative data (e.g., ID2, ID20, ID71, ID123). For 46% (*n* = 132/287) of the Delphi studies, it was also indicated that the participants were able to see their responses from the previous round (e.g., ID106, ID110, ID243). Peer feedback from one or more participant groups was reported by 6% of the Delphi studies (e.g., ID58, ID247, ID255).

Termination criterion and number of rounds

In 30% (*n* = 85/287) of the Delphi studies the survey process was terminated after a previously defined number of

rounds. Three rounds were held in 54% ($n=156/287$) of the Delphi studies, and two rounds in 36% ($n=104/287$). The highest number of rounds was eight (ID71). The criterion for terminating the Delphi study remained unclear for around half of the Delphi studies (e.g., ID4, ID89, ID151).

Discussion

The scoping review analyzes the conduction of consensus Delphi studies in the health sciences based on publications. As documented in previous reviews [14–16], there were many diverse modifications such that it is impossible to refer to *the* Delphi technique. Because these modifications are seldom reflected on or justified, negative effects on the quality and ultimately on the acceptance and implementation of the results cannot be ruled out. This can be seen in the factors of expert panel, questionnaire design, and process and feedback design.

Expert panel

The findings suggest that it can be difficult to acquire experts for Delphi studies and to retain them over multiple rounds. Indications of this are the sampling strategies and decreasing numbers that have already been discussed in other reviews [7, 14]. Reasons for dropping out remain unclear and were usually not reflected in the Delphi studies analyzed here. However, this is an important but under-researched issue for the validity of the results. Gargon et al. [73] suspect that the "attrition of participants could mean that people with minority opinions drop out of the Delphi study, leading to an overestimation of consensus." High dropout rates can therefore have a negative impact on the credibility of the process and the validity of the consensus [11]. It also remained unclear for which reasons a Delphi study was terminated [14]. Ideally, termination occurs when consensus or agreement regarding the dissent has been achieved and/or the judgments are stable, but pragmatic reasons are also conceivable when, for instance, too many experts drop out or the resources for further rounds are lacking.

The goal of Delphi studies is to include different expert groups [25, 32]. Yet the numerical ratio between expert groups often remains unclear. Complicating this is that expert groups are defined and differentiated in varying ways. The wording is not always consistent either. For example, the participants in a Delphi are usually referred to as experts, but sometimes also as panelists [2]. In general, one of the greatest challenges in Delphi studies is the identification and recruitment of relevant experts [14, 56].

As a result, it is impossible to make any reliable claims about the influence of a panel's composition on the results. Especially when lifeworld experts are included, such as patients or their family members, it may be necessary, due to the educational backgrounds, to adapt the technical terminology or provide additional information on the current state of research so that these participants are not overwhelmed and to enable a well-considered judgment [39]. Otherwise, there is the risk that satisficing effects or drop-out rates [28, 29] will be more probable for this group. To reduce the probability that important perspectives are lost, Boel et al. [74] recommend, based on a methodological experiment, always inviting all of the experts to each Delphi round regardless of their participation or nonparticipation in the previous round. Their recommendation is justified by their study in which the response rate for the "all-rounds group" was higher (61%) compared to the "respondents-only group" (46%) [74]. Another strategy to reduce attrition is to send reminders [56]. It is also possible to contact experts personally and send personalized emails, which can have an additional positive effect on the response rate [73].

Questionnaire design

In contrast to a review from the field of histopathology [16], the results here show a clear preference for five-point rating scales to record consensus. Regardless of the exact number of scale points, the use of an odd-numbered rating scale means that respondents have an evasive option in the 'middle." The effect of the number of scale points on judgments in Delphi studies has not been conclusively clarified [14]. With the mental model in mind (see "Theoretical insights on judgment formation in consensus Delphi studies" section), it could be argued that the experts can avoid taking a clear position and engaging deeply with the questions by using middle categories. This could be compounded if, in the case of an online Delphi, questions are programmed such that experts cannot skip over them but rather must submit an answer. This could mean that in the worst-case scenario, the results of Delphi studies with a high degree of satisficing behavior by the participants reflect a consensus of "collective ignorance" or collective uncertainty [14], rather than the wisdom of the crowd [22].

To maintain the motivation of the experts despite this, Delphi practitioners appear to want to keep the time and cognitive effort as low as possible, thus also making satisficing effects less likely [28]. An indication of this is the tendency to shorten the questionnaire as the Delphi procedure progresses. However, it is impossible to verify the stability of the judgments when items for which there is consensus are taken out [5].

Process and feedback design

As evidenced in earlier reviews [2, 7, 16], the design of the feedback is reported unclearly or not at all. In particular,

the method used for qualitative analysis remains vague, as does whether or not all of the expert groups have participated equally in the arguments. After the writing of this review, the AQUA (Argument-based QUalitative Analysis strategy) was published, offering for the first time a strategy to analyze open-ended responses in Delphi studies [65]. It remains to be seen if proposals of this nature have an influence on how free-text responses are handled in Delphi studies.

Research practice and the quality of the results of Delphi studies

The analysis shows a wide variance in how consensus Delphi studies are conducted. This diversity has been a consistent topic of discussion for decades in critical assessments of Delphi studies [15, 75, 76]. At the same time, the possibilities offered by modern software [14] and new scientific standards, such as the inclusion of affected parties [77] or the combination of methods or triangulation [78], evoke the necessity for further methodological development of the Delphi technique. However, there is the risk that such developments are applied uncritically or without reflection because they are currently "state of the art." As a result, the combination of quantitative and qualitative data in Delphi studies has been done in a manner that is hardly systematic, whereby potentially significant information and insights are lost [65]. Furthermore, the participation of affected persons also goes without careful consideration in terms of methodology.

Overall, the impression is reinforced that the conduction of Delphi studies as a form of research is pragmatic in approach. This is suggested by the unthought-out number of Delphi rounds, the incomprehensible removal of items, and the lack of analyses regarding reasons for the absence or stability of judgments. Beiderbeck et al. [56] recommend considering a priori criteria for stopping the Delphi process (e.g., when judgments are stable). However, as shown in this review, the traditional measure of stability is rarely used. Another option could be to measure expert contribution [79], e.g., by the number of comments if this falls below a certain level. The example shows that it is essential for the applicability of quality criteria that they are feasible for the researchers and comprehensible for the reader of the Delphi study publications [79].

The Delphi technique's enduring popularity, which is even increasing in healthcare [27], could persist precisely because this method can be "pragmatically" adapted. That said, though, pragmatically conducted research cannot take place at the cost of systematic method, transparency, and quality [1, 2]. Otherwise, there is the danger that the consensus has no validity in practice [14].

A first step to make Delphi studies potentially more comparable with each other and to support their application would be interdisciplinary and cross-disciplinary agreement on a recommended definition of consensus. Valid definitions from medicine could serve as proposals allowing for consensus to be subdivided into three categories, e.g., high (70% agreement) moderate (60% agreement), and low consensus (50% agreement) [17] or strong consensus (95% agreement), consensus (75% agreement), and majority agreement (50% agreement) [80].

Advice for Delphi practitioners to ensure the quality Delphi studies

Based on the analysis of research practice and given the current state of research on Delphi studies, there are, in our opinion, various suggestions for Delphi practitioners in the health sciences regarding the three factors analyzed here to ensure and improve the quality of study results. We view these suggestions as a contribution to the discourse on the development of practical guidelines for transparent and critical conduction of Delphi studies.

1) Expert panel

- Define the relevant expert groups and see that there is a balanced ratio in order to avoid potential distortions due to uneven ratios.
- Critically consider the heterogeneity of the expert panel, particularly in regard to its relevance for the questionnaire and feedback design.
- Be sure that the number of participants in each expert group is a two-digit number. Depending on the research field, it may be necessary at first to invite five to ten times the number of experts needed for each expert group to retain sufficient numbers over multiple rounds. Consider how participants' motivation can be maintained across all expert groups.

2) Questionnaire design

- Develop the initial questionnaire based on the current state of research and, if applicable, other preceding empirical studies. Include not only representatives from the relevant expert groups during development to test the content and its comprehensibility but also methodological experts to ensure the Delphi study's practicability and the quality of its results.
- When designing the questionnaire, bear in mind that answering some of the questions is cognitively demanding so attention should be paid to the overall length and complexity of the questionnaire. Consider

randomizing the questions if the questionnaires are lengthy and there is a sufficient number of participants.

- Reflect on the personal backgrounds and competencies of the experts. If dissimilarities exist, support the experts so that all of them can participate fully. For example, provide clear information on the state of research.
- Try to motivate the experts to produce explicit judgments (e.g., for or against a specific action). Consider if a middle category on a rating scale really makes sense. An alternative would be to weigh the option of a separate evasive category (e.g., "don't know").
- We recommend only recording qualitative and quantitative data in a Delphi study if these are systematically analyzed and the findings are combined or triangulated.

3) Process and feedback design

- Avoid modifications to the Delphi technique that have not been critically considered or put to the test previously. However, if modifications are necessary, make them in a transparent manner and justify them.
- Define the criterion for consensus. A cut-off value of less than 70% is unusual in the health sciences.
- Consider when to end the Delphi study. If you set the number of Delphi rounds or a target corridor in advance, share this with the experts so that they can better estimate the time needed. In regard to participant motivation and securing results, we recommend holding a maximum of three Delphi rounds.
- Follow published guidance on reporting Delphi studies, e.g., DELPHISTAR (Delphi studies in social and health sciences—recommendations for an interdisciplinary standardized reporting) [81].
- Evaluate the Delphi study or integrate evaluative elements into the study, e.g., questions about which feedback information was considered or about the time needed to participate in the Delphi study.

Limitations

The scoping review examines research practices involving Delphi studies in the health sciences. Since this paper focuses on consensus-oriented Delphi studies, the results here may only be limitedly transferable to Delphi studies pursuing other aims. The publications of the Delphi studies were screened and analyzed by one person, whereby the ongoing discussions within the research team and with scientists from the Delphi expert network (DEWISS) should have reduced this bias. There could be biases because only the publications were analyzed and not the Delphi studies themselves and, as a consequence, the findings here depend on the quality of the reporting. A selection bias may also exist because only available German and English full texts were analyzed.

Conclusion

Research practice shows that consensus Delphi studies in healthcare are carried out in different ways, whereby the approach can sometimes be described as pragmatically based research. Due to unclear reporting and variance in study design, Delphi studies are currently only comparable and evaluable to a limited extent. Nevertheless, it is clear that Delphi studies play an important role in a variety of health science issues, such as curriculum or guideline development, definition and terminology, and prioritization of research needs, possibly because of their pragmatic approach. In our view, the challenge of uniform criteria for conducting and reporting lies in increasing the comparability of Delphi studies with each other and ensuring the quality of study results while at the same time allowing for the flexibility of the Delphi technique and innovations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-024-02738-3.

Additional file 1: Overview of the analyzed publications and results of the qualitative content analysis.

Acknowledgements
We would like to thank the members of the DEWISS network for their support of the study and critical comments.

Authors' contributions
Julia Schifano: Conceptualization, research, and analysis of the publications, writing (original draft), and corrections. Marlen Niederberger: conceptualization, writing (original draft), and corrections. Members of DEWISS: consulting.

Funding
Open Access funding enabled and organized by Projekt DEAL. The DEWISS Network is supported by the German research foundation (Project Number: 429572724). The article processing charge was funded by the University of Education Schwäbisch Gmünd in the funding programme Open Access Publishing.

Data availability
The dataset supporting the conclusions of this article is included within the article and its additional file.

Declarations

Ethics approval and consent to participate
The Ethics Committee of the University of Education in Schwäbisch Gmünd confirmed in a letter dated 10.07.2023 that no ethics vote is necessary.

Consent for publication
Not applicable.

Competing interests
The authors declare that there are no conflicts of interest.

Received: 7 August 2023 Accepted: 17 December 2024
Published online: 14 January 2025

References

1. Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol. 2014;67(4):401–9. https://doi.org/10.1016/j.jclinepi.2013.12.002.
2. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. WJM. 2021;11(4):116–29. https://doi.org/10.5662/wjm.v11.i4.116. published Online First: 20 July 2021.
3. Niederberger M, Köberich S, members of the DeWiss Network. Coming to consensus: the Delphi technique. EJCN. 2021;20(7):692–95. https://doi.org/10.1093/eurjcn/zvab059.
4. Dalkey N, Helmer O. An experimental application of the DELPHI method to the use of experts. Manage Sci. 1963;9(3):458–67.
5. von der Gracht HA. Consensus measurement in Delphi studies. Technol Forecast Soc Change. 2012;79(8):1525–36. https://doi.org/10.1016/j.techfore.2012.04.013.
6. Jaam M, Awaisu A, El-Awaisi A, et al. Use of Delphi technique in pharmacy practice research. Res Social Adm Pharm. 2021;18(1):2237–48. https://doi.org/10.1016/j.sapharm.2021.06.028. published Online First: 12 August 2021.
7. Niederberger M, Spranger J. Delphi technique in health sciences: a map. Front Public Health. 2020;8:457. https://doi.org/10.3389/fpubh.2020.00457. published Online First: 22 September 2020.
8. Zarnowitz V, Lambros LA. Consensus and uncertainty in economic prediction. J Polit Econ. 1987;95(3):591–621.
9. Linstone HA, Turoff M. The delphi method. MA: Addison-Wesley; 1975.
10. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Inf Manag. 2004;42(1):15–29.
11. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs. 2000;32(4):1008–15.
12. Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. Int J Forecast. 1999;15(4):353–75. https://doi.org/10.1016/S0169-2070(99)00018-7.
13. Boulkedid R, Abdoul H, Loustau M, et al. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. PLoS ONE. 2011;6(6):e20476.
14. Shang Z. Use of Delphi in health sciences research: a narrative review. Medicine (Baltimore). 2023;102(7):e32829. https://doi.org/10.1371/journal.pone.0020476.
15. Jünger S, Payne SA, Brine J, et al. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med. 2017;31(8):684–706.
16. Taze D, Hartley C, Morgan AW, et al. Developing consensus in Histopathology: the role of the Delphi method. Histopathology. 2022;81(2):159–67. https://doi.org/10.1111/his.14650. published Online First: 24 April 2022.
17. Meskell P, Murphy K, Shaw DG, et al. Insights into the use and complexities of the Policy Delphi technique. Nurse Res. 2014;21(3):32–9. https://doi.org/10.7748/nr2014.01.21.3.32.e342.
18. Aengenheyster S, Cuhls K, Gerhold L, et al. Real-time delphi in practice — a comparative analysis of existing software-based tools. Technol Forecast Soc Change. 2017;118:15–27. https://doi.org/10.1016/j.techfore.2017.01.023.
19. Gnatzy T, Warth J, von der Gracht H, et al. Validating an innovative real-time Delphi approach - a methodological comparison between real-time and conventional Delphi studies. Technol Forecast Soc Change. 2011;78(9):1681–94. https://doi.org/10.1016/j.techfore.2011.04.006.
20. Quirke FA, Battin MR, Bernard C, et al. Multi-Round versus Real-Time Delphi survey ap-proach for achieving consensus in the COHESION core outcome set: a randomised trial. Trials. 2023;24:461. https://doi.org/10.1186/s13063-023-07388-9.
21. Dalkey NC. The Delphi Method: an experimental study of group opinion. Santa Monica, CA: RAND Corporation; 1969.
22. Jorm AF. Using the Delphi expert consensus method in mental health research. Aust N Z J Psychiatry. 2015;49(10):887–97. https://doi.org/10.1177/0004867415600891.
23. Campbell SM. How do stakeholder groups vary in a Delphi technique about primary mental health care and what factors influence their ratings? Qual Saf Health Care. 2004;13(6):428–34. https://doi.org/10.1136/qshc.2003.007815.
24. Markmann C, Spickermann A, von der Gracht HA, et al. Improving the question formulation in Delphi-like surveys: analysis of the effects of abstract language and amount of information on response behavior. Futures Foresight Sci. 2021;3(1):e56. https://doi.org/10.1002/ffo2.56.
25. Lüke C, Kauschke C, Dohmen A, et al. Definition and terminology of developmental language disorders-Interdisciplinary consensus across German-speaking countries. PLoS ONE. 2023;18(11):e0293736. https://doi.org/10.1371/journal.pone.0293736.
26. Homberg A, Krug K, Klafke N, et al. Consensus views on competencies and teaching methods for an interprofessional curriculum on complementary and integrative medicine: a Delphi study. J Integr Med. 2021;19(3):282–90. https://doi.org/10.1016/j.joim.2021.03.001.
27. Flostrand A, Pitt L, Bridson S. The Delphi technique in forecasting– a 42-year bibliographic analysis (1975–2017). Technol Forecast Soc Change. 2020;150:119773. https://doi.org/10.1016/j.techfore.2019.119773.
28. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cognit Psychol. 1991;5(3):213–36. https://doi.org/10.1002/acp.2350050305.
29. Krosnick JA. Survey research. Annu Rev Psychol. 1999;50:537–67. https://doi.org/10.1146/annurev.psych.50.1.537.
30. Häder M, Häder S. Delphi und Kognitionspsychologie: ein Zugang zur theoretischen Fundierung der Delphi-Methode. ZUMA Nachrichten. 1995;19(37):8–34 https://www.ssoar.info/ssoar/bitstream/document/20888/1/ssoar-zuma-1995-37-hader_et_al-delphi_und_kognitionspsychologie.pdf . Accessed 03 Aug 2023. Accessed 03 Aug 2023.
31. Gigerenzer G, Hoffrage U, Kleinbölting H. Probabilistic mental models: a Brunswikian theory of confidence. Psychol Rev. 1991;98(4):506–28. https://doi.org/10.1037/0033-295x.98.4.506.
32. Sautenet B, Tong A, Manera KE, et al. Developing consensus-based priority outcome domains for trials in kidney transplantation: a multinational Delphi survey with patients, caregivers, and health professionals. Transplantation. 2017;101(8):1875–86. https://doi.org/10.1097/TP.0000000000001776.
33. Ravensbergen WM, Drewes YM, Hilderink HBM, et al. Combined impact of future trends on healthcare utilisation of older people: a Delphi study. Health Policy. 2019;123(10):947–54. https://doi.org/10.1016/j.healthpol.2019.07.002. published Online First: 17 July 2019.
34. Förster B, von der Gracht H. Assessing Delphi panel composition for strategic foresight — A comparison of panels based on company-internal and external participants. Technol Forecast Soc Change. 2014;84:215–29. https://doi.org/10.1016/j.techfore.2013.07.012.
35. Tourangeau R, Rips LJ, Rasinski KA. The psychology of survey response. Cambridge: Cambridge Univ. Press; 2000.
36. Goldstein WM. Social judgment theory: applying and extending Brunswik's probabilistic functionalism. In: Koehler DJ, Harvey N, editors. Blackwell handbook of judgment and decision making. 1st ed. Oxford, UK, Malden, MA: Blackwell Pub; 2004. p. 37–61.
37. Mauksch S, von der Gracht HA, Gordon TJ. Who is an expert for foresight? A review of identification methods. Technol Forecast Soc Change. 2020;154:119982. https://doi.org/10.1016/j.techfore.2020.119982.
38. Barrios M, Guilera G, Nuño L, et al. Consensus in the delphi method: what makes a decision change? Technol Forecast Soc Change. 2021;163. https://doi.org/10.1016/j.techfore.2020.120484.
39. Barrington H, Bridget Y, Paula R. Williamson. Patient participation in Delphi surveys to develop core outcome sets: systematic review. BMJ Open. 2021;11(9). https://doi.org/10.1136/bmjopen-2021-051066.
40. Brookes ST, Macefield RC, Williamson PR, et al. Three nested randomized controlled trials of peer-only or multiple stakeholder group feedback within Delphi surveys during core outcome and information set development. Trials. 2016;17(1):1–14. https://doi.org/10.1186/s13063-016-1479-x. published Online First: 17 August 2016.
41. Rowe G, Wright G, McColl A. Judgment change during Delphi-like procedures: the role of majority influence, expertise, and confidence. Technol

Forecast Soc Change. 2005;72(4):377–99. https://doi.org/10.1016/j.techfore.2004.03.004.
42. Bolger F, Stranieri A, Wright G, et al. Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? Technol Forecast Soc Change. 2011;78(9):1671–80. https://doi.org/10.1016/j.techfore.2011.06.002.
43. Fish R, MacLennan S, Alkhaffaf B, et al. "Vicarious thinking" was a key driver of score change in Delphi surveys for COS development and is facilitated by feedback of results. J Clin Epidemiol. 2020;2020(128):118–29. https://doi.org/10.1016/j.jclinepi.2020.09.028. publishedOnlineFirst:1October.
44. Meijering JV, Tobi H. The effects of feeding back experts' own initial ratings in Delphi studies: a randomized trial. Int J Forecast. 2018;34(2):216–24. https://doi.org/10.1016/j.ijforecast.2017.11.010.
45. Turnbull AE, Dinglas VD, Friedman LA, 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;2018(102):99–106. https://doi.org/10.1016/j.jclinepi.2018.06.007. publishedOnlineFirst:30June.
46. Biggane AM, Williamson PR, Ravaud P, et al. Participating in core outcome set development via Delphi surveys: qualitative interviews provide pointers to inform guidance. BMJ Open. 2019;9(11):e032338. https://doi.org/10.1136/bmjopen-2019-032338. published Online First: 14 November 2019.
47. Khodyakov D, Chen C. Nature and predictors of response changes in modified-delphi panels. Value Health. 2020;23(12):1630–8. https://doi.org/10.1016/j.jval.2020.08.2093.
48. Makkonen M, Hujala T, Uusivuori J. Policy experts' propensity to change their opinion along Delphi rounds. Technol Forecast Soc Change. 2016;109:61–8. https://doi.org/10.1016/j.techfore.2016.05.020.
49. Winkler J, Moser R. Biases in future-oriented Delphi studies: a cognitive perspective. Technol Forecast Soc Change. 2016;105:63–76. https://doi.org/10.1016/j.techfore.2016.01.021.
50. Bolger F, Wright G. Improving the Delphi process: Lessons from social psychological research. Technol Forecast Soc Change. 2011;78(9):1500–13. https://doi.org/10.1016/j.techfore.2011.07.007.
51. Hussler C, Muller P, Rondé P. Is diversity in Delphi panelist groups useful? Evidence from a French forecasting exercise on the future of nuclear energy. Technol Forecast Soc Change. 2011;78(9):1642–53. https://doi.org/10.1016/j.techfore.2011.07.008.
52. Spickermann A, Zimmermann M, von der Gracht HA. Surface- and deep-level diversity in panel selection — exploring diversity effects on response behaviour in foresight. Technol Forecast Soc Change. 2014;85:105–20. https://doi.org/10.1016/j.techfore.2013.04.009.
53. Fraser GM, Pilpel D, Kosecoff J, et al. Effect of panel composition on appropriateness ratings. Int J Qual Health Care. 1994;6(3):251–5. https://doi.org/10.1093/intqhc/6.3.251.
54. 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. https://doi.org/10.1186/1471-2288-5-37. published Online First: 1 December 2005.
55. Alizadeh S, Maroufi SS, Sohrabi Z, et al. Large or small panel in the delphi study? Application of bootstrap technique. Jemds. 2020;9(15):1267–71. https://doi.org/10.14260/jemds/2020/275.
56. Beiderbeck D, Frevel N, von der Gracht HA, et al. Preparing, conducting, and analyzing Delphi surveys: cross-disciplinary practices, new directions, and advancements. MethodsX. 2021;8:101401. https://doi.org/10.1016/j.mex.2021.101401. published Online First: 28 May 2021.
57. Choi BCK, Pak AWP. A catalog of biases in questionnaires. Prev Chronic Dis. 2005;2(1):A13 https://pmc.ncbi.nlm.nih.gov/articles/PMC1323316.
58. Bassili JN, Krosnick JA. Do strength-related attitude properties determine susceptibility to response effects? New evidence from response latency, attitude extremity, and aggregate indices. Polit Psychol. 2000;21(1):107–32. https://doi.org/10.1111/0162-895X.00179.
59. Andersen PD. Constructing Delphi statements for technology foresight. Futures Foresight Sci. 2022;5(2): e144. https://doi.org/10.1002/ffo2.144.
60. Salancik JR, Wenger W, Helfer E. The construction of Delphi event statements. Technol Forecast Soc Change. 1971;3:65–73. https://doi.org/10.1016/S0040-1625(71)80004-5.
61. Brookes ST, Chalmers KA, Avery KNL, et al. Impact of question order on prioritisation of outcomes in the development of a core outcome set: a randomised controlled trial. Trials. 2018;19(1):66. https://doi.org/10.1186/s13063-017-2405-6. published Online First: 25 January 2018.
62. Hallowell MR, Gambatese JA. Qualitative research: application of the Delphi method to CEM Research. J Constr Eng Manage. 2010;136(1):99–107. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000137.
63. Stennett A, de Souza L, Norris M. Physical activity and exercise priorities in community dwelling people with multiple sclerosis: a Delphi study. Disabil Rehabil. 2018;40(14):1686–93. https://doi.org/10.1080/09638288.2017.1309464. published Online First: 10 April 2017.
64. Millward CP, Armstrong TS, Barrington H, et al. Development of 'Core Outcome Sets' for Meningioma in Clinical Studies (The COSMIC Project): protocol for two systematic literature reviews, eDelphi surveys and online consensus meetings. BMJ Open. 2022;12(5):e057384. https://doi.org/10.1136/bmjopen-2021-057384. published Online First: 9 May 2022.
65. Niederberger M, Homberg A. Argument-based QUalitative Analysis strategy (AQUA) for analyzing free-text responses in health sciences Delphi studies. MethodsX. 2023;10. https://doi.org/10.1016/j.mex.2023.102156.
66. Cuhls K, Dragomir B, Gheorghiu R, et al. Probability and desirability of future developments – Results of a large-scale argumentative Delphi in support of Horizon Europe preparation. Futures. 2022;138:102918. https://doi.org/10.1016/j.futures.2022.102918.
67. Holland JL, Christian LM. The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. Soc Sci Comput Rev. 2009;27(2):196–212. https://doi.org/10.1177/0894439308327481.
68. Lange T, Kopkow C, Lützner J, et al. Comparison of different rating scales for the use in Delphi studies: different scales lead to different consensus and show different test-retest reliability. BMC Med Res Methodol. 2020;20(28):1–11. https://doi.org/10.1186/s12874-020-0912-8. published Online First: 10 February 2020.
69. De Meyer D, Kottner J, Beele H, et al. Delphi procedure in core outcome set development: rating scale and consensus criteria determined outcome selection. J Clin Epidemiol. 2019;2019(111):23–31. https://doi.org/10.1016/j.jclinepi.2019.03.011. publishedOnlineFirst:25March
70. MacLennan S, Kirkham J, Lam TBL, et al. A randomized trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement. J Clin Epidemiol. 2017;2018(93):1–8. https://doi.org/10.1016/j.jclinepi.2017.09.024. publishedOnlineFirst:7October.
71. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73. https://doi.org/10.7326/M18-0850. published Online First: 4 September 2018.
72. Mayring P. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt 2014. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173. Accessed 03 Aug 2023.
73. Gargon E, Crew R, Burnside G, et al. Higher number of items associated with significantly lower response rates in COS Delphi surveys. J Clin Epidemiol. 2018;2019(108):110–20. https://doi.org/10.1016/j.jclinepi.2018.12.010. publishedOnlineFirst:15December
74. Boel A, Navarro-Compán V, Landewé R, et al. Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. J Clin Epidemiol. 2020;2021(129):31–9. https://doi.org/10.1016/j.jclinepi.2020.09.034[publishedOnlineFirst:28September. published OnlineFirst:28September.
75. Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud. 2001;38(2):195–200. https://doi.org/10.1016/S0020-7489(00)00044-4.
76. Sackman H. Delphi Assessment: Expert opinion, forecasting and group assessment. Santa Monica, California 1974. https://www.rand.org/pubs/reports/R1283.html. Accessed 28 Dec 2024.
77. Price A, Clarke M, Staniszewska S, et al. Patient and public involvement in research: a journey to co-production. Patient Educ Couns. 2022;105(4):1041–7. https://doi.org/10.1016/j.pec.2021.07.021. published Online First: 19 July 2021.

78. Lee SYD, Iott B, Banaszak-Holl J, et al. Application of mixed methods in health services management research: a systematic review. Med Care Res Rev. 2022;79(3):331–44. https://doi.org/10.1177/10775587211030393. published Online First: 12 July 2021.
79. Landeta J, Lertxundi A. Quality indicators for Delphi studies. Futures Foresight Sci. 2023;6(1): e172. https://doi.org/10.1002/ffo2.172.
80. German Association of the Scientific Medical Societies (AWMF). AWMF guidance manual and rules for guideline development 2013. https://www.awmf.org/en/clinical-practice-guidelines/awmfguidance.html. Accessed 03 Aug 2023.
81. Niederberger M, Schifano J, Deckert S, et al. Delphi studies in social and health sciences-recommendations for an interdisciplinary standardized reporting (DELPHISTAR). Results of a Delphi study. PloS ONE. 2024;19(8):e0304651. https://doi.org/10.1371/journal.pone.0304651.

Publisher's Note