Journal of Clinical Epidemiology 111 (2019) 23–31

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Delphi procedure in core outcome set development: rating scale and consensus criteria determined outcome selection

Dorien De Meyer[a], Jan Kottner[a,b], Hilde Beele[c], Jochen Schmitt[d], Toni Lange[d], Ann Van Hecke[a,e], Sofie Verhaeghe[a,f], Dimitri Beeckman[a,g,h,*]

[a]*Department of Public Health and Primary Care, Skin Integrity Research Group (SKINT), University Centre for Nursing and Midwifery, Ghent University, Ghent, Belgium*

[b]*Department of Dermatology and Allergy, Clinical Research Center for Hair and Skin Science, Charité-Universitätsmedizin, Berlin, Germany*

[c]*Department of Dermatology, Ghent University Hospital, Ghent, Belgium*

[d]*Center for Evidence-Based Healthcare, Medizinische Fakultät Carl Gustav Carus TU Dresden, Dresden, Germany*

[e]*Nursing Department, Ghent University Hospital, Ghent, Belgium*

[f]*Department Health Care, VIVES University College, Roeselare, Belgium*

[g]*Faculty of Medicine & Health Sciences, School of Nursing and Midwifery, Royal College of Surgeons in Ireland (RCSI), Dublin, Ireland*

[h]*School of Health Sciences, Örebro University, Örebro, Sweden*

Accepted 19 March 2019; Published online 25 March 2019

Abstract

Objective: The objective of this study was to compare two different rating scales within one Delphi study for defining consensus in core outcome set development and to explore the influence of consensus criteria on the outcome selection.

Study Design and Setting: Randomized controlled parallel group trial with 1:1 allocation within the first Delphi round of the Core Outcome Set in the Incontinence-Associated Dermatitis project. Outcomes were rated on a three-point or nine-point Likert scale. Decisions about which outcomes to retain were determined by commonly used consensus criteria (i.e., [combinations of] proportions with restricted ranges, central tendency within a specific range, and decrease in variance).

Results: Fifty-seven participants (group 1 = 28, group 2 = 29) rated 58 outcomes. The use of the nine-point scale resulted in almost twice as many outcomes being rated as ''critical'' compared with the three-point scale (24 vs. 13). Stricter criteria and combining criteria led to less outcomes being identified as ''critical''.

Conclusion: The format of rating scales in Delphi studies for core outcome set development and the definition of the consensus criteria influence outcome selection. The use of the nine-point scale might be recommended to inform the consensus process for a subsequent rating or face-to-face meeting. The three-point scale might be preferred when determining final consensus.

Keywords: Consensus; Core outcome set; Criteria; Delphi-procedure; Dermatology; Incontinence-associated dermatitis

1. Introduction

Clinical trials aim to evaluate the effects of interventions based on predefined outcomes. Results of clinical trials are an important source of information for evidence-based clinical decision-making [1]. Therefore, the selection of appropriate and useful outcomes is crucial. However, during the last years, the multitude of noncomparable outcomes in clinical trials has been identified as a major limitation for evidence-based practice [2].

The development of core outcome sets (COSs) is one approach to tackle this methodological challenge [3,4]. A COS is defined as an agreed standardized set of outcomes that should at least be measured and reported in clinical trials within a specific health area [2]. COS development is a standardized process which includes (1) making decisions on the specific health condition, population, intervention, and setting; (2) gaining agreement on what outcomes should be measured through the involvement of stakeholders, the identification of potential relevant outcomes and a consensus process [5], and (3) gaining agreement on how each outcome should be measured [6].

Conflict of interest: none.

* Corresponding author. Skin Integrity Research Group (SKINT), University Centre for Nursing and Midwifery, Ghent University, UZ Gent, 5K3, Corneel Heymanslaan 10, 9000 Ghent, Belgium. Tel.: +32 9 332 83 48.

E-mail address: Dimitri.Beeckman@UGent.be (D. Beeckman).

https://doi.org/10.1016/j.jclinepi.2019.03.011

What is new?

- Three- and nine-point rating scales in Delphi studies to evaluate the importance of an outcome for inclusion in a core outcome set lead to different outcome selection.
- The definition of the consensus criteria determines outcome selection.

What this adds to what was known?

- Nine-point scales may lead to a broader set of outcomes to be included into a (preliminary) core outcome set.

What is the implication and what should change now?

- In order to be able to discriminate between response options, simple cutoffs should not be used when using nine-point scales.

The Delphi technique is a frequently applied method to achieve consensus on core outcome domains [2,7]. The advantage of the Delphi method is that many participants have the opportunity to rate the importance of outcomes independently and anonymously and that large numbers of geographically divergent participants can be involved [8]. Methodological guidance and research of how to conduct Delphi studies within COS development is emerging [9−11], but a number of methodological challenges remain. Uncertainties exist regarding the most appropriate format of outcome importance rating and regarding the definition of consensus [2,12]. Nine-item rating scales as originally proposed by the RAND appropriateness method [13] and the Grading of Recommendations, Assessment, Development and Evaluations scale (GRADE) initiative to prioritize outcomes in evidence summaries [14] are widely used in the COS field. However, a number of other scales (such as three-, five-, and seven-item rating scales) and also simple yes/no classifications have been used as well [7]. It is unclear whether and how these different rating scales influence conclusions and decisions in outcome selection in Delphi studies and research on this topic is recommended [15].

A systematic review revealed that consensus criteria in Delphi studies vary widely [12]. Different procedures to define consensus can be applied such as formal measures of agreement, degrees of uncertainty around point estimates, decreases in variance of group responses, and the proportion of participants agreeing on a particular point of view [16−19]. However, the selection of consensus criteria is rarely justified [12] and it has been argued that the aim of Delphi studies in COS development is not to reach consensus but to decide which outcomes are core [2].

More research focusing on concerns regarding the Delphi method is needed [20]. To our knowledge, no studies exist that have investigated the way in which rating scales and consensus criteria affect the final agreement on outcomes of critical importance. Therefore, the aim of this study was to compare two different rating scales within one Delphi study and to explore the influence of consensus criteria on the outcome selection.

2. Methods

2.1. Design

This study was part of the Core Outcome Set in Incontinence-Associated Dermatitis project in 2017 [21,22]. Incontinence-associated dermatitis (IAD) is an irritant contact dermatitis, caused by the prolonged and repeated exposure of the skin to urine and/or feces. It is characterized by the presence of erythema and edema, sometimes accompanied by bullae, erosion, or secondary cutaneous infection [23]. After the design of a long list of possible outcomes, three Delphi rounds were conducted between April and September 2017 [24]. In the first Delphi round, a randomized controlled parallel group trial with 1:1 allocation was conducted.

2.2. Participants

An international group of health care providers, researchers, and industry product experts with established experience in IAD assessment, prevention, and management was invited by email to participate. They were authors of studies identified based on a literature review conducted previously [24] and using the professional network of the authors. All possible panel members were also asked to suggest additional experts to be invited. An online survey, hosted by the Cochrane Skin Core Outcome Set Initiative, was developed.

2.3. Intervention

Before the first Delphi round, participants were randomly assigned to two groups, rating the outcomes using two different rating scales. Group 1 rated the importance of outcomes on a three-point scale: (1) ‘‘not important enough to be considered in the COS’’, (2) ‘‘important but not critical to be considered in the COS’’, and (3) ‘‘critical, should be included in the COS’’. This scale was selected because the importance of outcomes is actually often assigned to three categories only even if GRADE and other methodologies use a nine-point scale initially [14]. Group 2 rated the importance of outcomes on a nine-point scale with following anchors: (1) ‘‘not important for inclusion in the COS’’ and (9) ‘‘critical, should be included in the COS’’. This scale was used because it is currently widely used by many COS groups [2,25,26]. The option ‘‘I can’t rate the importance of the

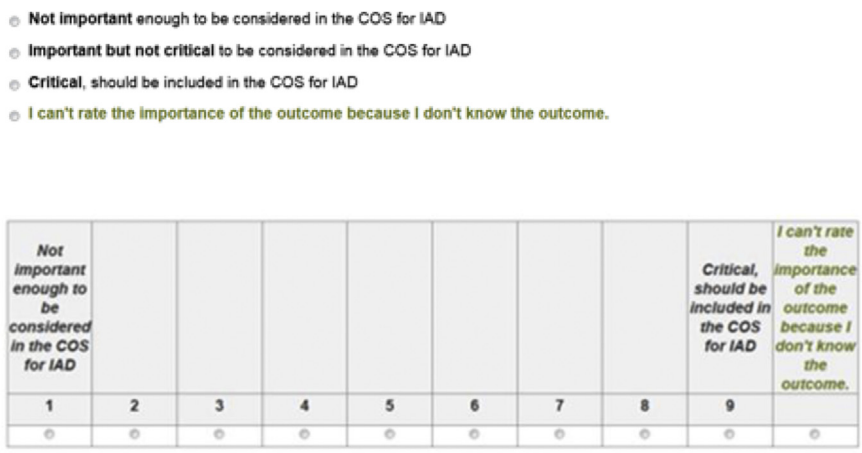

Fig. 1. Three-point scale and nine-point scale as used in the first round of the Delphi study. *Abbreviations*: COS, core outcome set; IAD, incontinence-associated dermatitis.

outcome because I don't know the outcome'' was provided for both groups (see Fig. 1). Outcomes were listed alphabetically within the core areas life impact, economical impact, and pathophysiological manifestations [27] to avoid weighting because of the order. Core areas refer to large ''containers'' encompassing all key aspects of interest (i.e., the outcomes) [27]. Other core areas such as death or adverse events were not presented because no outcome domains have been identified previously by the patients and the review [24].

2.4. Outcome

The distribution of outcome priorization was the primary outcome of this study.

2.5. Randomization

Simple randomization using allocation concealment before the invitation to participate was applied (Excel; Microsoft, NY, USA). Panelists were assigned a random number between 0 and 1,000 and were consequently sorted from smallest to largest and allocated to group 1 or 2.

2.6. Ethical considerations

The study was approved by the Ethics review Committee (April 2016–B670201628231). Participants were informed about the development of a COS. Return of a completed questionnaire was taken as consent to participate in the Delphi procedure. Information was treated anonymous and confidential.

2.7. Data analysis

Statistical analyses were performed using the software package SPSS statistics 24 (SPSS, Inc. Chicago, IL, USA). Demographic data and responses to the questionnaires were described using frequency distributions. Ratings of the nine-point scale were assigned to three categories: ''not important'' (scoring 1 to 3), ''important but not critical'' (scoring 4 to 6), and ''critical'' (scoring 7 to 9). In addition to the a priori defined consensus criterion of at least 70% rating the outcome as ''critical'' on the three-point and nine-point scale [21], other commonly used consensus criteria [2,12] were applied as well (see Table 1). The association between measures of central tendency (mean and median) and distribution (SD and IQR) was also investigated.

3. Results

3.1. Participants

A sample of 151 potential panelists was invited. Fifty-seven of the panelists participated in the first Delphi round (group 1 = 28, group 2 = 29) (see Fig. 2).

Table 1. Applied consensus criteria to the data set of the nine-point scale (Delphi round 1)

No	Criterion
Proportions with restricted ranges	
1	≥60% scoring 7 to 9
2	≥70% scoring 7 to 9
3	≥75% scoring 7 to 9
4	≥90% scoring 7 to 9
Combinations of restricted ranges	
5	≥60% scoring 7 to 9 and ≤ 15% 1 to 3
6	≥70% scoring 7 to 9 and ≤ 15% 1 to 3
7	≥75% scoring 7 to 9 and ≤ 15% 1 to 3
8	≥90% scoring 7 to 9 and ≤ 15% 1 to 3
Central tendency within a specific range	
9	Mean greater than 7
10	Median 7 to 9
Decrease in variance	
11	Median 7 to 9 and IQR less than 3

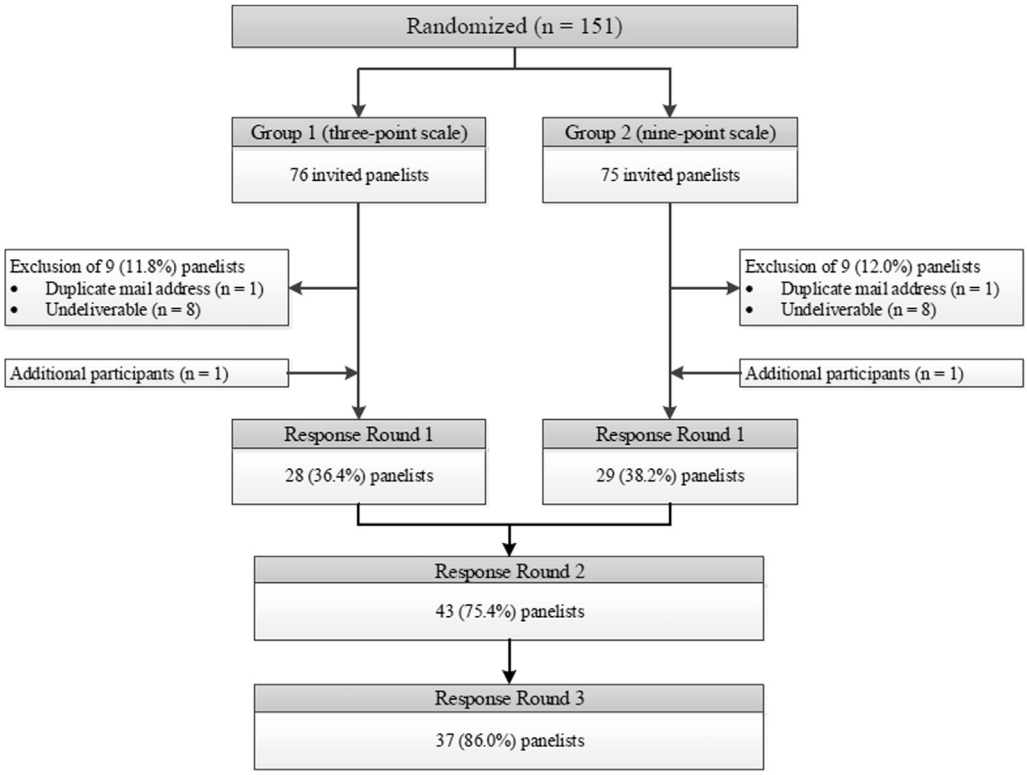

Fig. 2. Participant flow.

Demographic characteristics were compared. Most of the participants was female (group 1: 75%; group 2: 72.4%) and had a doctoral degree. Participants assigned to group 1 had an average work experience of 26.1 (SD = 10.2) years, vs. 25.3 years (SD = 9.6) in group 2. Most of the participants were based in Europe, followed by the United States. Detailed demographic characteristics of the participants can be found in Appendix 1.

3.2. Three-point vs. nine-point scales

Results and comparisons of ratings between groups are shown in Table 2 and Appendix 2. Based on the criterion of ≥70% rating the outcome as "critical" for inclusion, group 1 selected 13 in comparison with 24 of 58 outcomes in group 2. All outcomes defined as "critical" by group 1, were also rated as "critical" by group 2 (see Table 2).

3.3. Consensus criteria

Applying different consensus criteria to the nine-point scale (group 2), stricter criteria and combining criteria led to less outcomes being defined as "critical". For example, the criterion "≥60% scoring 7 to 9" resulted in 40 outcomes and "≥90% scoring 7 to 9 and ≤ 15% 1 to 3" resulted in two outcomes (see Appendix 3 and Fig. 3). The criterion of central tendency within a specific range, that is, median between 7 and 9 led to the highest number of outcomes being defined as "critical" for inclusion (47/58 = 81.0%), when the IQR was not taken into account. In case the IQR less than 3 was combined with a median between 7 and 9 (i.e., decrease in variance), 21 outcomes were defined as "critical". The criterion of central tendency within a specific range, mean greater than 7 resulted in 25 outcomes (43.1%).

Comparing the outcomes defined as "critical" based on the ratings of the participants using a three-point scale with the ratings using different consensus criteria showed that for 2 of 13 outcomes defined as "critical", these were also defined as "critical" by each of the consensus criteria (i.e., "Pain" and "Self-reported symptoms"). This was also the case for 10 of 13 outcomes, except when the criterion ≥90% scoring 7+ (and ≤15% scoring 1 to 3) was used. The outcome "Erosion" was not defined as being "critical" by 6 of 11 consensus criteria. For 11 of 45 outcomes which were not defined as "critical" based on the three-point scale, none of the consensus criteria defined the outcome as "critical".

When analyzing the final list of core outcome domains (obtained after three Delphi rounds using the consensus criterion of at least 70% rating the outcome as "critical") [24] in relation to the different consensus criteria, the outcome "IAD related pain" was included in all applied consensus criteria (see Appendix 3). The outcomes "Satisfaction with intervention from patient perspective", "Erythema," and

Table 2. Selection of outcomes on a three-point scale ($n = 28$) and nine-point scale ($n = 29$) based on the criterion of ≥70% rating the outcome as ''critical'' for inclusion (Delphi round 1)

Outcome	Group 1 (three-point scale) ($n = 28$)	Group 2 (nine-point scale) ($n = 29$)
Proportion within a range (restricted): Minimum 70% scoring the outcome critical for inclusion		
Outcomes per core area		
Life impact		
1. Burden of care from caregiver's perspective	Out	In
2. Burden of care from patient perspective	In	In
3. Health-related quality of life	Out	In
4. Independence (IAD related)	Out	In
5. Pain	In	In
6. Physical comfort	In	In
7. Physical functioning	Out	Out
8. Physical well-being	Out	Out
9. Psychological impact of the disease	Out	Out
10. Quality of life (general)	Out	Out
11. Quality of life (IAD related)	In	In
12. Satisfaction with intervention from caregiver's perspective	Out	In
13. Satisfaction with intervention from patient perspective	In	In
14. Self-reported symptoms	In	In
15. Sleep (IAD related)	Out	Out
Resource use/economical impact		
16. Caregivers' work productivity	Out	Out
17. Cost-effectiveness	In	In
18. Costs	Out	Out
Pathophysiological manifestations		
19. Bleeding	Out	Out
20. Bullae	Out	In
21. Clinical characteristics of skin surrounding IAD area assessed by caregiver	Out	In
22. Clinical signs of inflammation/colonisation/infection of IAD area assessed by caregiver	In	In
23. Cracking	Out	Out
24. Crusting	Out	Out
25. Denudation	In	In
26. Desquamation	Out	Out
27. Discoloration	Out	In
28. Dryness	Out	Out
29. Erosion	In	In
30. Erythema	In	In
31. Excoriation	Out	Out
32. Exudate	Out	In
33. Glossy/shiny appearance	Out	Out
34. Infection confirmed by culture	Out	Out
35. Lichenification	Out	Out
36. Maceration	In	In
37. Macules	Out	Out
38. Maculopapular rash	Out	Out
39. Necrosis	Out	Out
40. Nodules	Out	Out
41. Edema	Out	Out
42. Oozing	Out	Out

(*Continued*)

Table 2. Continued

Outcome	Group 1 (three-point scale) (n = 28)	Group 2 (nine-point scale) (n = 29)
43. Papules	Out	Out
44. Pigmentation	Out	Out
45. Purulent exudate	Out	Out
46. Pustules	Out	In
47. Roughness	Out	Out
48. Satellite lesions	Out	In
49. Scabbing	Out	Out
50. Scaling	Out	Out
51. Scratch marks	Out	Out
52. Shiny appearance	Out	Out
53. Skin barrier properties	Out	In
54. Skin loss	In	In
55. Slough present in the wound bed (yellow/brown/greyish)	Out	Out
56. Swelling	Out	Out
57. Vesicles	Out	Out
58. White scaling	Out	Out
TOTAL IN *N* (%)	13 (22.4)	24 (41.4)

Abbreviation: IAD, incontinence-associated dermatitis.

"Maceration" were defined to be "critical", except when the ≥90% scoring 7+ (+≤15% scoring 1 to 3) criterion was applied. Only one outcome, that is, "Erosion" was not rated "critical" based on several consensus criteria (≥90% scoring 7+ [+≤15% scoring 1 to 3]; ≥60%, ≥70%, ≥75% scoring 7 + and ≤15% scoring 1 to 3; median >7 and IQR <3).

Results indicate that the higher the SD, the more often the outcome is not rated as "critical" to be included in the COS (i.e., mean scoring 7+). This is similar for the association between the IQR and the median (see Fig. 4).

4. Discussion

4.1. Interpretation

This is the first study comparing two different rating scales and several commonly used consensus criteria and their impact on priorization of outcomes to be included in a COS.

The use of the nine-point scale in combination with the commonly used threshold of ≥70% to rate the outcome as "critical" resulted in almost twice as many outcomes selected as "critical" compared with using the three-point scale. This difference indicates that the format of the scales influences the Delphi study results, indicating that the choice of the scoring method is important. Using a scale with few response options might limit respondents to make full use of their capacity to discriminate; however, rating scales with a broad range of response options might contribute to measurement error because the respondent's capacity to discriminate is exceeded [28]. To reduce measurement error, it is therefore suggested to use verbal labels to anchor all scale points next to numeric labels in rating scales [28]. In addition, the possibly better discrimination when using more response options is ignored when using simple cutoffs such as 70%. In research, it is generally recommended not to arbitrarily cut (quasi)continues scales into categories [7,29]. In addition, the thresholds used to

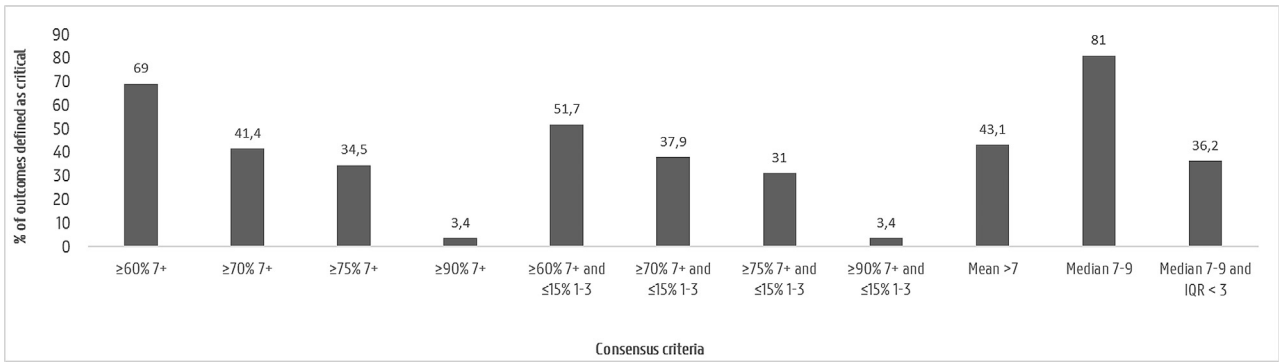

Fig. 3. Outcomes defined as "critical" on the nine-point scale based on different consensus criteria (Delphi round 1).

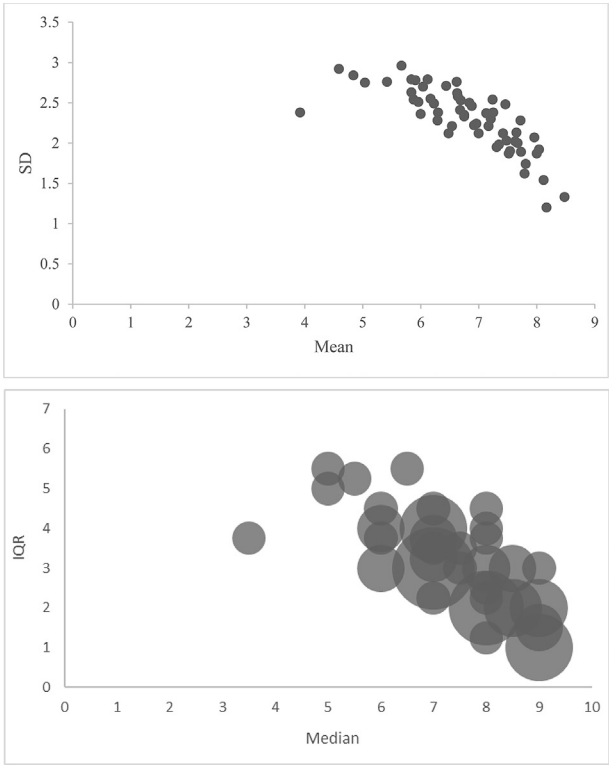

Fig. 4. Scatter plot SD vs. mean rating and bubble plot IQR vs. median rating on the nine-point scale (Delphi round 1). The size of the bubbles is the number of pairs with equal medians and IQRs.

categorize outcomes into "critical" (scoring 7 to 9 on the scale), "important but not critical" (scoring 4 to 6 on the scale), and "limited important" (scoring 1 to 3 on the scale) seem to be arbitrary [19,30]. Therefore, it is recommended to use the full range of information from the rating scales [28,31].

Because the main objective of Delphi studies in the COS field is to prioritize outcomes [2], direct questions about the outcomes seem to be an attractive option. Therefore, it might be recommended to use the nine-point scale, in accordance with the RAND appropriateness method [32], to inform the consensus process for a subsequent rating or face-to-face consensus meeting. The three-point scale is suggested to determine final consensus.

When using thresholds, stricter criteria and combining criteria led to less outcomes being identified as "critical". These results were expected but they confirm the arbitrariness and open the discussion about the most appropriate thresholds. Our results indicate that ≥60% rating the outcome "critical" might be not strict enough but ≥90% rating the outcome "critical" might be too strict. Therefore, thresholds of ≥70% or slightly higher might be a compromise. Our results support the recommendation by Williamson and Altman [2] to use a combination of a proportion to rate an outcome as "critical" and a proportion to rate an outcome as "not important".

The criterion of central tendency within a specific range, that is, median between 7 and 9 led to the highest number of outcomes being defined as "critical" for inclusion. Similar to the definition of proportions of restricted ranges, this consensus criterion is arbitrary and probably not strict enough. If measures of central tendency are used, they should be combined with measures of distribution, such as IQRs. Our results clearly show, that the wider the distribution, the less respondents rated outcomes as "critical". Reduction in the distribution of ratings between several Delphi rounds can be seen as an indicator of increasing agreement between participants [2]. However, according to Crisp and Pelletier [33], the stability of the ratings through a series of Delphi rounds is a more reliable indicator of agreement.

To reduce bias, researchers should always consider and publicly register in advance the consensus criterion that will be used to determine agreement within Delphi studies [2,15]. If thresholds are used and outcomes believed to be critical end up just below the threshold, it might be useful to reconsider the threshold a posteriori [12]. However, this somehow questions the robustness of the Delphi process itself. It might be argued that the results of Delphi studies are just the basis for subsequent face-to-face consensus meetings. If this would be the case, strict cutoffs should not be used at all.

The choice of consensus definitions and the way of the outcome selection finally also depends on the desired number core outcomes. There is no guidance about the optimal number of core outcomes in a set. It would be interesting to include this in future studies.

4.2. Generalizability

The generalizability of the study results might be limited because of the following reasons: the sample size was small and this study was conducted within one COS development project only. We only compared two scales: a three- and a nine-point scale, although a number of other scales and scale formats are used by COS developers as well. Therefore, the reproduction of this study in other fields of COS development, focusing on other topics and including larger sample sizes and different scales, is needed. On the other hand, it is highly likely that systematic differences regarding outcome ratings between different scale types would also be observed in studies with other scales because the reliability and the rating behavior is a function of the number of scale categories, the labeling, and the general way of scale presentation [31,34].

4.3. Limitations

This study had several limitations. Only two different rating scales were used to study the impact on outcome selection. We tried to contact as many experts as possible worldwide to participate in the Delphi study. However, the response rate of 37.7% (57/151 potential panelists) for the first Delphi round was low. The main reason is the small number of qualified experts with a deep understanding of clinical IAD research. This is also the case in

other areas in COS development [2]. Because of the low number of participants in the first Delphi round, it was decided to combine both groups in the second round to have a sufficient number for meaningful analysis. Therefore, the impact of receiving feedback on the subsequent rating of outcomes and the likelihood of respondents to change scores could not be investigated. Patients were involved in the search for possible outcomes but patients or patient representatives were not included in the Delphi study. The acute and mostly self-limiting nature of this skin condition in often care-dependent older people and intensive care patients makes it very challenging if not impossible to involve patients with IAD. Active involvement of patients is, however, recommended by the Patient-Centered Outcomes Research Institute [35] and in COS development [2]. As we did not involve patients, it could not be investigated whether the findings would have been different if patients would have been included as a separate stakeholder group in the Delphi procedure.

5. Conclusions

The format of rating scales in Delphi studies for COS development and the definition of the consensus criteria to decide about inclusion of outcomes influence outcome selection. This challenges the current methodology to achieve consensus on core outcome domains. The use of the nine-point scale might be recommended to inform the consensus process for a subsequent rating or face-to-face meeting. The three-point scale might be preferred when determining final consensus on the COS.

CRediT authorship contribution statement

Dorien De Meyer: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Jan Kottner:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Supervision. **Hilde Beele:** Conceptualization, Writing - review & editing. **Jochen Schmitt:** Conceptualization, Writing - review & editing. **Toni Lange:** Conceptualization, Software, Writing - review & editing. **Ann Van Hecke:** Conceptualization, Writing - review & editing. **Sofie Verhaeghe:** Conceptualization, Writing - review & editing. **Dimitri Beeckman:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Supervision.

Acknowledgments

The authors would like to acknowledge all panelists for their contribution to the Delphi procedure.

This study is partly supported through a doctoral fellowship from the Ghent University Special Research Fund.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.011.

References

[1] Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71−2.
[2] Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET handbook: version 1.0. Trials 2017;18(3):280.
[3] Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012;13(1):132.
[4] Clarke M. Standardising outcomes for clinical trials and systematic reviews. Trials 2007;8(1):39.
[5] Kirkham JJ, Davis K, Altman DG, Blazeby JM, Clarke M, Tunis S, et al. Core outcome set-STAndards for development: the COS-STAD recommendations. PLoS Med 2017;14(11):e1002447.
[6] Prinsen C, Mokkink L, Bouter L, Alonso J, Patrick D, de Vet H, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147.
[7] Kottner J, Jacobi L, Hahnel E, Alam M, Balzer K, Beeckman D, et al. Core outcome sets in dermatology: report from the second meeting of the international Cochrane skin group core outcome set initiative. Br J Dermatol 2018;178(4):e279−85.
[8] 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.
[9] Turnbull AE, Dinglas VD, Friedman LA, Chessare CM, Sepúlveda KA, Bingham CO III, et al. A survey of Delphi panelists after core outcome set development revealed positive feedback and methods to facilitate panel member participation. J Clin Epidemiol 2018;102:99−106.
[10] MacLennan S, Kirkham J, Lam TB, Williamson PR. A randomized trial comparing three Delphi feedback strategies found no evidence of a difference in a setting with high initial agreement. J Clin Epidemiol 2018;93:1−8.
[11] Brookes ST, Macefield RC, Williamson PR, McNair AG, Potter S, Blencowe NS, et al. Three nested randomized controlled trials of peer-only or multiple stakeholder group feedback within Delphi surveys during core outcome and information set development. Trials 2016;17(1):409.
[12] Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014;67:401−9.
[13] Campbell S, Cantrill J. Consensus methods in prescribing research. J Clin Pharm Ther 2001;26(1):5−14.
[14] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[15] Grant S, Booth M, Khodyakov D. Lack of preregistered analysis plans allows unacceptable data mining for and selective reporting of consensus in Delphi studies. J Clin Epidemiol 2018;99:96−105.
[16] Barrett S, Kristjanson LJ, Sinclair T, Hyde S. Priorities for adult cancer nursing research: a West Australian replication. Cancer Nurs 2001;24(2):88−98.

[17] Basson R, Berman J, Burnett A, Derogatis L, Ferguson D, Fourcroy J, et al. Report of the international consensus development conference on female sexual dysfunction: definitions and classifications. J Urol 2000;163(3):888−93.
[18] Broder MS, Landow WJ, Goodwin SC, Brook RH, Sherbourne CD, Harris K. An agenda for research into uterine artery embolization: results of an expert panel conference. J Vasc Interv Radiol 2000;11(4):509−15.
[19] Broomfield D, Humphris G. Using the Delphi technique to identify the cancer education requirements of general practitioners. Med Educ 2001;35(10):928−37.
[20] Humphrey-Murto S, de Wit M. The Delphi method-more research please. J Clin Epidemiol 2018;106:136−9.
[21] Van den Bussche K, De Meyer D, Van Damme N, Kottner J, Beeckman D. CONSIDER - core outcome set in IAD research: study protocol for establishing a core set of outcomes and measurements in incontinence-associated dermatitis research. J Adv Nurs 2017;73:2473−83.
[22] Kottner J, Beeckman D. Core Outcome Sets (cos) for clinical trials in health-and nursing science: the case of Incontinence-associated Dermatitis (iad). J Adv Nurs 2017;73:2268−9.
[23] Gray M, Beeckman D, Bliss DZ, Fader M, Logan S, Junkin J, et al. Incontinence-associated dermatitis: a comprehensive review and update. J Wound Ostomy Continence Nurs 2012;39(1):61−74.
[24] Van den Bussche K, Kottner J, Beele H, De Meyer D, Dunk AM, Ersser S, et al. Core outcome domains in incontinence-associated dermatitis research. J Adv Nurs 2018;74:1605−17.
[25] Thorlacius L, Ingram JR, Villumsen B, Esmann S, Kirby JS, Gottlieb AB, et al. A core domain set for hidradenitis suppurativa trial outcomes: an international Delphi process. Br J Dermatol 2018;179(3):642−50.
[26] Sahnan K, Tozer PJ, Adegbola SO, Lee MJ, Heywood N, McNair AG, et al. Developing a core outcome set for fistulising perianal Crohn's disease. Gut 2019;68(2):226−38.
[27] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino M-A, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745−53.
[28] Beckstead JW. On measurements and their quality. Paper 4: verbal anchors and the number of response options in rating scales. Int J Nurs Stud 2014;51(5):807−14.
[29] Kottner J, Streiner DL. Binary outcomes are not better than continuous variables in randomized controlled trials. J Invest Dermatol 2014;134(1):267−8.
[30] Thorlacius L, Ingram JR, Garg A, Villumsen B, Esmann S, Kirby JS, et al. Protocol for the development of a core domain set for hidradenitis suppurativa trial outcomes. BMJ Open 2017;7(2):e014733.
[31] Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. New York, NY: Oxford University Press; 2015.
[32] Fitch K, Bernstein SJ, Aguilar MD, Burnand B, LaCalle JR. The RAND/UCLA appropriateness method user's manual. Santa Monica, CA: RAND CORP; 2001.
[33] Crisp J, Pelletier D, Duffield C, Adams A, Nagy S. The Delphi method? Nurs Res 1997;46(2):116−8.
[34] Crocker L, Algina J. Introduction to classical and modern test theory Mason. OH: Cengage Learning; 2008.
[35] Frank L, Forsythe L, Ellis L, Schrandt S, Sheridan S, Gerson J, et al. Conceptual and practical foundations of patient engagement in research at the patient-centered outcomes research institute. Qual Life Res 2015;24:1033−41.