# Consensus development methods: Considerations for national and global frameworks and policy development

## Introduction

In this ever-changing world, key decisions have to be made in healthcare systems and policies often under uncertain conditions or without complete objective evidence.[1-3] It is not uncommon for key decisions in healthcare systems and policies to be made based on informal group decision making.[2] However, various drawbacks of such an approach have been identified, including: particular individuals are dominating discussion and decision making, pressures from powerful individuals and the power imbalance of individuals in the group, more extreme decisions with individuals with strong opinions, and complex issues remaining forgotten due to unstructured process.[2, 5]

When decisions are made at national or international levels, the impact of such decisions become even wider often affecting more stakeholders and sometimes entire populations. Decisions-making in health systems and health policies at national and international levels would affect national and global health, which might lead to inequality in health of population. In order to achieve transparent and accountable decision-making process for optimising national and global health and achieving equitable health access, consensus development methods (CDMs) can be adapted.

CDMs are commonly used for developing clinical guidelines for quality improvement of health care.[5, 6] A complex process is involved in its consensus-based decision making.[5] CDMs assist such complex processes by offering a systematic approach to synthesising information and expert views.[1, 7]

CDMs use a quantitative approach for synthesising qualitative data, aiming to achieve general agreement, convergence of opinion, or resolving inconsistencies in scientific information around a particular topic.[2, 8, 9] The methods are widely used in the field of health, social care and wellbeing,[10] including pharmacy practice,[8] and are regularly applied for the development of clinical guidelines. CDMs are officially used by the World Health Organisation for guideline development.[11]

CDMs usually involve repetitive interactions with a group of participants to reach a general agreement in the group. There are different types of CDMs available. The methods commonly consist of the Nominal Group Technique (NGT), Delphi technique (DT), Consensus Development Conference (CDC), RAND/UCLA appropriateness method (RAM). These differ in terms of anonymity, the number of participants, the use of face-to-face meeting, and time frame.

### Nominal Group Technique

The Nominal Group Technique (NGT) is a highly structured and controlled small group process involving face-to-face group interaction.[3, 8] The NGT was developed in the 1960s by Delbecq and Van de Ven, originally developed in psychological studies, management science studies, and social work studies.[12] The NGT has four main stages: silent generation, round robin, clarification and voting/ranking.[8]

Silent generation is the first stage of NGT, where participants list ideas or issues related to the particular topic silently and individually in a written format.[12] Round robin is the second stage, when participants in turn state a single idea to the group one at a time in a 'round robin' fashion, then continue to record one item from participants in sequence, until all group members have no more

ideas to state.[8, 12] Third stage of the NGT is clarification where group members clarify, elaborate, defend or dispute the ideas listed, using face-to-face focus group meeting.[12] During ranking, the last stage, participants rank the preferences, agreement, or appropriateness depending on what the NGT's aims are; ranking uses a structured questionnaire including Likert scales.[2, 8, 12] If the pre-defined level of consensus is not achieved, then the third and fourth stages are repeated until is it achieved. McMillan et al.[8] proposes that the first two stages (i.e. silent generation and round robin) can be replaced by literature reviews or exploratory surveys.

Throughout the stages, controlled interactions with small group of participants take place, with normally 5 – 12 people, depending on the aims and topics of the study.[1, 2] Cantril et al.[3] suggested that if a larger number of participants are required (i.e. more than 9 or 10 people), the participants may require dividing into two or more parallel or sequential groups. Participants of the NGT are 'experts' of an investigating issue. This has an impact on the validity, credibility, reliability and acceptability of the findings of the method.[13, 14] Careful consideration should be given to group composition, as different stakeholders tend to produce different ratings. However, within defined specialist or professional categories, selecting particular individuals have little impact on the rating.[15]

The main advantages of the NGT include relatively quick outcomes to be obtained,[8] a large number of ideas to be generated,[1] and greater 'ownership' of the decision and tools developed by participants,[2] which can affect how the decisions or tools implemented after the study. On the other hand, some limitations include relatively small number of participants,[2] and the difficulty of setting up face-to-face meetings.[8]

## Delphi technique

The Delphi technique (DT) is an iterative survey technique for group decision making without face-to-face interactions.[3] The DT was introduced by the RAND Corporation in US in the 1950s, as a means to forecasting the effects of atomic warfare in US.[5] Since the DT was first introduced, the method has been used in many academic fields, including science, technology, health, business, communication, education and policy analysis.[1, 7] The DT is applied within the pharmacy practice field. McMillan et al.[8] showcase examples which include scoping exercise of future practice and education, developing criteria, indicators and definitions, and the more wide utilisation for clinical guidelines development.

The DT process includes: collecting opinions/views on a particular issue, rating the agreement with each item/statement, and re-rating the agreement with updated items/statements.[7] All stages are carried out using an anonymous questionnaire. This means participants never meet or interact in the DT.[5] Originally, the questionnaires are to be distributed by mail, but recently this has been replaced largely by email or online questionnaire.[8] The re-rating stage can be repeated as often as needed depending on the degree of agreement reached from the second stage, or until a predetermined number of rounds are completed.[1] In the first stage of the DT, opinions or views from experts are usually collected by open ended questions,[8] and analysed using content analysis approach.[3] The results of the first stage is summarised and converted into a list of statements/items in a questionnaire to be distributed in the second phase.[8] The degree of agreement is commonly measured by Likert Scale,[8] often with written feedback especially when the participant does not agree with the item.[1] After reconsidering a group median of each item, participants will re-rate the updated items again by Likert Scale.[8] Some studies modified the DT process, particularly the first stage, by replacing it with a literature review or qualitative study to explore more in-depth opinions.[3]

The DT allows for repetitive interactions with a large number of participants,[7] which can range from 4 to 3000,[13] and presents no geographical barriers for conducting the study since the techniques are

conducted remotely.[2] This flexibility, the option to have a large number of participants, and the relatively low resources needed are significant advantages of the DT. Potential disadvantage may be found in the fact that the participants do not engage with others nor do are they exposed to their opinions and disagreements. Participants therefore do not change their opinions often; and it is argued that agreement or consensus achieved is lower in level than that in NGT.[2]

### Consensus Development Conference (CDC)

The consensus development conference (CDC) approach is thoroughly a face-to-face interactive method to develop consensus among panel members at a public forum. CDCs were developed by the National Institutes of Health (NIH) in US in 1977.[5] CDCs are commonly used to evaluate and disseminate health care technologies for clinical practices.[16]

The CDC involves a series of intensive face-to-face interactions of panel members (about 10 people). Its process is more flexible compared to previous two methods, using iterative face-to-face meetings of experts. A group of panel members are provided with evidence on a particular issue by a small group of experts on the particular issue, and who are not involved in the decision-making process.[17] The panel members will then ask questions to presenters. After clarifying all issues, the panel group members will deliberate on the issue, directed by their chairperson to attempt to reach consensus.[5]

The main advantage of the CDC technique is to foster dialogue, debate and discussion.[17, 18] Furthermore, CDC method embeds a dissemination process of guidelines by holding a form of press conference at the end of each round.[16] However, significant drawback of the CDC includes expensive study cost holding the public forum, which usually lasts more than two days. The time limitation to hear the evidence from experts also is also at the same time a disadvantage of the method, as presenters may not be able to cover all evidence related to the issue.[17]

### RAND/UCLA appropriateness method (RAM)

The RAND/UCLA Appropriateness Method (RAM) is a hybrid method combining the elements of both NGT and DT,[15] developed by the RAND corporation and University of California Los Angeles (UCLA) School of Medicine in the 1980s for a health services utilisation study, measuring the overuse and underuse of medical and surgical procedures.[19] RAM is commonly used in clinical practice for developing guidelines. However, the method is also used in policy and organisational interventions.[20]

RAM commonly involves 7-15 members of experts in the process.[19] The RAM consists of 5 stages, including (1) literature review to create a list of indicators for an intervention, (2) an expert panel to rate the appropriateness of the intervention for each item measuring the degree of agreement by a 9-point Likert scale using an anonymous questionnaire, (3) face-to-face meeting of the expert panel, (4) re-rating the appropriateness of updated items by Likert scale in an anonymous questionnaire, and (5) categorising indicators as appropriate, inappropriate, or uncertain, based on the group median rating.[3, 15] Basic data analysis to determine agreement and disagreement is similar to NGT and DT. However, the RAM uses more sophisticated analysis using the concept of interpercentile Range (IPR) and Interpercentile Range Adjusted for Symmetry (IPRAS). This enables to avoid the effect of different size of expert panel and to provide an in-depth analysis of 'disagreement' in searching the appropriateness of items in a tool.[19]

These CDMs can be used to develop health and pharmacy policy tools and frameworks with a variety of modifications and adaptations made over the decades. However, to the best of the authors' knowledge, there is no review yet on the use of CDMs particularly in health systems and policy at national and international levels. Understanding the different approaches, utilisations can support

researchers with ensuring rigour, validity, and transparency in the use of CDMs.. Therefore, this paper aims to identify and review papers which have used CDMs in order to develop national or international policy tools or framework in health field. The review identifies CDMs used in current research, identify modifications and adaptations to traditional methods, and distil considerations required for high level policy and framework development.

# Methods

A narrative systematic review was conducted, following the PRISMA guidelines.

## *Inclusion and exclusion criteria*

Eligible studies included those which used any of CDMs commonly used for developing a policy or framework, or equivalent to framework (eg. tools or indicators) at a country or international level, aiming to improve health system(s) or policy in health. Papers aimed to develop clinical guidelines were excluded, as these are often developed following country specific developmental guidelines (eg. National Institute for Health and Care Excellence in UK) or global WHO guideline,[11] and there are intensive literature reviews on clinical guideline development (e.g. Black et al.[6]). Books, editorials, and other sources reporting non-original research were excluded. Papers were excluded if they were written in non-English.

## *Search strategy*

Articles were searched in the electronic databases, including the EMBASE, PsycINFO, and PubMed as of 3rd November 2020. The search did not limit the publication years. The search terms 'consensus development' and 'framework', which were crosschecked with 'health system' OR 'health policy'. The exact strategy used was [(consensus development) AND (framework) AND ((health system) OR (health policy))]. All retrieved citation data were entered into EndNote X9 (Clarivate Analytics, Philadelphia, US), and cleaned to remove duplicated papers.

The authors applied the inclusion and exclusion criteria to independently examine the titles and abstracts of the original articles. A list of eligible articles was discussed including any divergence occurred between both authors until the final list was agreed.

## *Data extraction and synthesis*

A data extraction table was prepared in Microsoft Excel sheet and agreed by both authors. The primary investigator (NA) extracted the data, which were reviewed by LRB. Collected data were narratively synthesised using content analysis in terms of the use of CDMs and the way the methods were modified according to the study settings.

# Results

## *Screening and selection of studies*

The screening process following the PRISMA guidelines is depicted in Figure 1. The searches in EMBASE, PsycINFO, and PubMed identified 70 different titles. The titles and abstracts were reviewed for inclusion and exclusion criteria, which excluded 34 papers. The remaining 37 papers were assessed for their eligibility by reviewing full-text articles. This identified an additional eligible article from reference lists of the reviewed papers, and further excluded 11 articles in alignment with the inclusion and exclusion criteria. In total, 26 articles were included in the final analysis.

Table 1 summarises collected data from each study. Thirteen studies developed frameworks at international levels or engaged with multiple countries.[21-33] Other remaining studies were conducted at a national level, including India,[34] Iran,[35, 36] The Netherlands,[37] Canada,[38] Libya,[39] US,[40-43] and Spain.[44] Out of 13 international studies, 4 papers[28, 30-32] related to the same academic society for rheumatology, carrying out at the Outcome Measures in Rheumatology (OMERACT) activities. The review identified that a wide range of frameworks or tools were developed using CDMs. The purposes of frameworks or tools include improving health services, health systems, health policies, and workforce development.

The papers related to consensus development in health policy and system field published more recently in a last decade. Nineteen papers (73.1%) were published between 2010 and 2020,[21-26, 28, 30-36, 38, 39, 41-43] 6 papers (23.1%) in 2000s,[27, 29, 37, 40, 44, 45] and 1 (3.8%)in 1998,[46] although no restriction was made for publication years in the literature search. Considering the CDMs developed in 1950s to 1960s, it has taken around 50 years to populate the use of CDMs in health policy and system research field.

The search identified that over 60% of studies did not apply typical consensus development methods, but were stated as consensus meetings. Sixteen studies (61.5%)[21-24, 26-28, 30-35, 37, 39, 42] claimed the use of consensus meeting as a main or part of their process, 12 studies (46.2%)[21, 24, 25, 29, 30, 35, 36, 38, 40, 41, 45, 46] used the DT as a main or part of their process, and 4 studies (15.4%)[36, 40, 44, 46] applied NGT as a main or part of their methods, and 1 study (3.8%)[43] claimed iterative feedbacks as their consensus method. The review illustrated a wide range of modification of methods, which are either non-use of typical CDM or combining different methods to compensate weakness of used methods. The most common combination of different methods was a combined use of DT and some form of face-to-face meetings (CM or NGT).[21, 24, 30, 36, 40, 46] No paper described reasons for non-use of formal CDM or combination of methods. However, this review indicates the formal CDMs may not fitting in developing national or international level framework or tools in health system and health policy areas.

Many studies did not state the numbers of consensus panel members (9 studies, 34.6%)[22, 23, 27, 28, 32, 34, 36, 42, 43] nor consensus thresholds (15 studies, 57.7%).[21-23, 26-28, 30-34, 36, 37, 39, 44] Among the studies claimed the number of panel members, there was a varying number of members engaged from 4 to 500, depending on the purpose of the tool developed and audience whether national or international levels. Reporting the number of consensus panel members and consensus threshold have an impact on the quality of the report and affect the rigour and value of decisions made.

## Discussion

This review is first of its kind focusing the use of CDMs for developing frameworks and tools in health systems and policy at country or international levels. A number of variations of consensus development methods have been developed since its first development in 1950s and 1960s, and the use of methods over the last a half century. CDMs have been adapted for use in healthcare field, often for clinical guideline development or health technology assessment. This present review focused on the use of consensus development approach in health policy and system improvement outside of clinical guideline development. The review identified a wide range of modifications of CDMs, and many did not follow any of existing consensus development methods. This is not consistent with the use of CDMs in creating clinical guidelines where intensive reviews and guidelines for formal steps of CDMs were developed. Humphrey-Murto et al.[1] criticise that a lack of standardization in definitions of consensus, use of methods, and reporting of these methods would result in less certain the level of rigour or value of decisions made.

However, high-level consensus in national and international settings often encounter challenges of geographical issues of panel members and time commitment of members with different countries when it comes to international framework development. Considering these challenges, it is understandable to have consensus development processes in academic conference settings and combine some methods to utilise opportunities for meetings at conferences and other processes conducted at online platforms as the review identified. The authors experienced consensus development processes at national and international levels, in order to develop national and international policy, framework and tools in pharmacy field.

For example, the International Pharmaceutical Federation (FIP) has recently launched a global list of FIP Development Goals as a key resource for transforming the pharmacy profession over the next decade globally, regionally and nationally. For the goals to be globally relevant and to increase uptake, it was essential that the development approach was based on consensus; but engaging global experts and stakeholders required a practice approach to consensus developed that combined elements of the traditional CDMs. An extensive process of consultation built on the methodology used in the development of the Pharmaceutical Workforce Development Goals (PWDGs) and adapted to develop the practice and science elements in the new goals. FIP experts, members, partners and stakeholders have all taken part in this work to ensure that the goals are relevant, measurable and achievable. The development process included a consultation with the FIP Council in 2019; the Council is FIP's highest organ which includes all national pharmaceutical associations (member organisations) and national pharmaceutical scientific associations (predominantly scientific member organisations). This element of the process adopted and adapted the CDC method by conducting a face to face meeting during a global conference. Afterwards, a cross-FIP Internal Reference Group was commissioned to provide feedback and input into the draft Goals; the Group included representatives from FIP Boards of practice and science, FIP Education as well as the FIP Young Pharmacists Group. The Internal Reference Group was engaged in the consultation using Delphi technique approaches.

A lack of consensus criteria reporting in papers was identified throughout the review. Scott and Black[48] showed that differences in consensus definitions lead to the nature and level of consensus that panel would reach. For achieving transparent and rigour consensus decisions in health, reporting the consensus threshold and criteria is essential. This indicates the need of rigorous planning of the consensus development process at national and international levels. Especially any decisions made in health system and policy affect national and global health. The researchers and policy makers need to be aware of the consensus development methods for better decision making process. For that, creating a guidance of CDMs to develop national/international frameworks or tools in health system and policy would be warranted.

It is important to note a limitation of the review. A literature search in the review was only carried out with limited literature database (i.e., PubMed and EMBASE) due to the resource limited to authors. A narrative systematic review limits evaluation of selected articles for validity. Considering the nature of the methods developed outside of health field, there may be more literature using consensus development methods in health system and policy development. Having identified the variety of modifications in consensus development methods in health system and policy development, the development of practical standards for CDMs would be warranted.

## Conclusion

This review summarised the use of consensus development methods in health system and policy development. The review identified a wide range of variations in the selection, use and application of the methods in the field. Some elements of consensus development process (i.e., the number of panel members and consensus threshold) were not reported in all papers, which addressed the issues in the quality of reporting with the use of consensus development methods. For better utilisation and application of the consensus development methods in health system and policy development, some standardisation of the methods and reporting would be warranted while allowing for the necessary pragmatic adaptations needed by researchers.

## Declarations

### *Ethics*

A present literature review does not require ethical clearance.

### *Declaration of interest*

None

### *Author contributions*

NA: concept development, data collection, analysis and original draft writing

LRB: review for data collection and original draft, and edit

## References

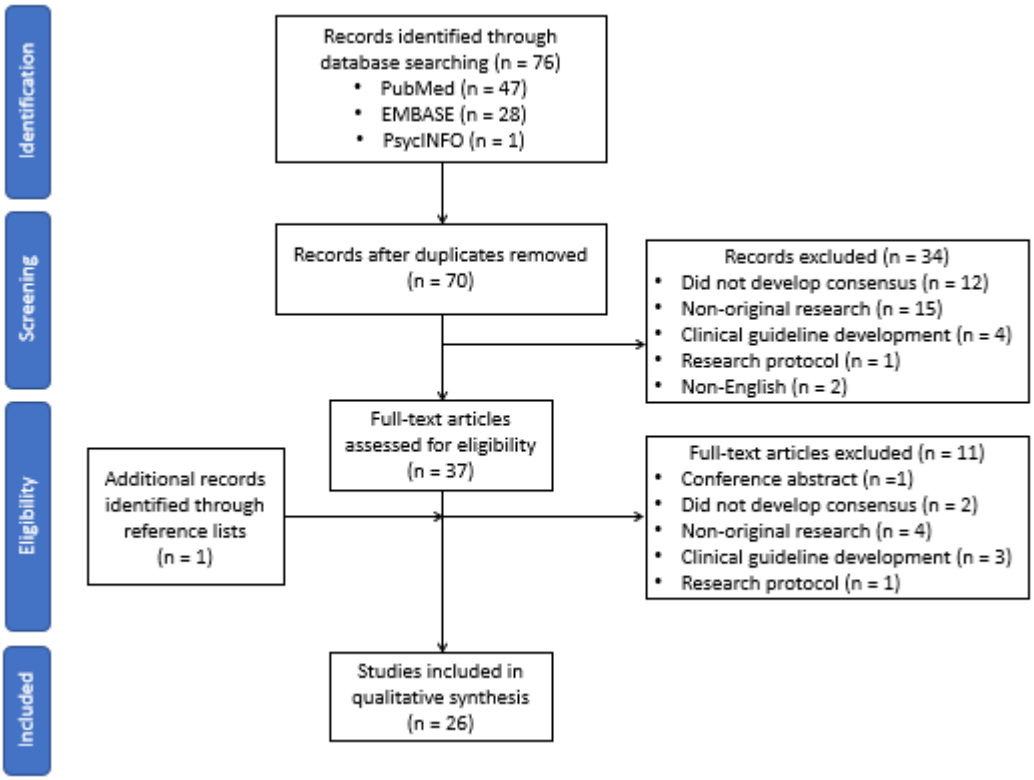

1. Humphrey-Murto, S., Varpio, L., Gonsalves, C., and Wood, T.J., Using consensus group methods such as Delphi and Nominal Group in medical education research. *Medical Teacher*, 2017;39:14-19.
2. Black, N., *Chapter 12: Consensus development methods*, in *Qualitative Research in Health Care - 3rd ed.*, C. Pope and N. Mays, Editors. Blackwell Publishing Ltd: Oxford; 2006.
3. Cantrill, J.A., Sibbald, B., and Buetow, S., The Delphi and nominal group techniques in health services research. *International Journal of Pharmacy Practice*, 1996;4:67-74.
4. Jennings, B., *Counsel and Consensus: Norms of Argument in Health Policy*, in *The argumentative turn in policy analysis and planning*, F. Fisher and J. Forester, Editors. Duke University Press: London; 1993.
5. Murphy, M.K., Black, N.A., Lamping, D.L., McKee, C.M., Sanderson, C.F., Askham, J., and Marteau, T., Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment*, 1998;2:i-iv, 1-88.
6. Black, N., Murphy, D.M., Lamping, D., Mckee, M., Sanderson, C., Askham, J., and Marteau, T., Consensus development methods: a review of best practice in creating clinical guidelines. *Journal of Health Services Research & Policy*, 1999;4:236-48.
7. Jones, J. and Hunter, D., Qualitative Research: Consensus methods for medical and health services research. *BMJ*, 1995;311:376.
8. McMillan, S.S., King, M., and Tully, M.P., How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy*, 2016;38:655-662.
9. Tammela, O., Applications of consensus methods in the improvement of care of paediatric patients: a step forward from a 'good guess'. *Acta Pediatrica*, 2013;102:111-115.

10. Guzys, D., Dickson-Swift, V., Kenny, A., and Threlkeld, G., Gadamerian philosophical hermeneutics as a useful methodological framework for the Delphi technique. *International Journal of Qualitative Studies on Health and Well-being*, 2015;10:26291.
11. World Health Organization, *WHO Handbook for Guideline Development - 2nd edition*. Geneva: World Health Organization; 2014.
12. Van de Ven, A.H. and Delbecq, A.L., The nominal group as a research instrument for exploratory health studies. *American journal of public health*, 1972;62:337-342.
13. Campbell, S.M. and Cantrill, J.A., Consensus methods in prescribing research. *Journal of Clinical Pharmacy and Therapeutics*, 2001;26:5-14.
14. Grimshaw, J. and Russell, I., Achieving health gain through clinical guidelines. I: Developing scientifically valid guidelines. *Quality in health care : QHC*, 1993;2:243-248.
15. Hutchings, A. and Raine, R., A systematic review of factors affecting the judgments produced by formal consensus development methods in health care. *Journal of Health Services Research & Policy*, 2006;11:172-9.
16. Fink, A., Kosecoff, J., Chassin, M., and Brook, R.H., Consensus Methods: Characteristics and Guidelines for Use. *American journal of public health*, 1984;74:979-983.
17. James, D. and Warren-Forward, H., Research methods for formal consensus development. *Nurse Researcher*, 2015;22:35-40.
18. Halcomb, E., Davidson, P., and Hardaker, L., Using the consensus development conference method in healthcare research. *Nurse Researcher*, 2008;16:56-71.
19. Fitch, K., Bernstein, S.J., Aguilar, M.D., Burnand, B., LaCalle, J.R., Lazaro, P., van het Loo, M., McDonnell, J., Vader, J.P., and Kahan, J.P., *The RAND/UCLA appropriateness method user's manual*. Santa Monica, CA: RAND; 2001.
20. Campbell, J.L., Fletcher, E., Abel, G., Anderson, R., Chilvers, R., Dean, S.G., Richards, S.H., Sansom, A., Terry, R., Aylward, A., Fitzner, G., Gomez-Cano, M., Long, L., Mustafee, N., Robinson, S., Smart, P.A., Warren, F.C., Welsman, J., and Salisbury, C., Policies and strategies to retain and support the return of experienced GPs in direct patient care: the ReGROUP mixed-methods study. *Hearth Services and Delivery Research*, 2019;7.
21. Calvert, M., Kyte, D., Mercieca-Bebber, R., Slade, A., Chan, A.-W., King, M.T., and the SPIRIT-PRO Group, Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. *JAMA*, 2018;319:483-494.
22. Cornel, M.C., Rigter, T., Weinreich, S.S., Burgard, P., Hoffmann, G.F., Lindner, M., Gerard Loeber, J., Rupp, K., Taruscio, D., and Vittozzi, L., A framework to start the debate on neonatal screening policies in the EU: an Expert Opinion Document. *European Journal of Human Genetics*, 2014;22:12-17.
23. Salvador-Carulla, L., Reed, G.M., Vaez-Azizi, L.M., Cooper, S.-A., Leal, R.M., Bertelli, M., Adnams, C., Cooray, S., Deb, S., Dirani, L.A., Girimaji, S.C., Katz, G., Kwok, H., Luckasson, R., Simeonsson, R., Walsh, C., Munir, K., and Saxena, S., Intellectual developmental disorders: towards a new name, definition and framework for "mental retardation/intellectual disability" in ICD-11. *World Psychiatry*, 2011;10:175-180.
24. Forrest, C.B., Chesley Jr, F.D., Tregear, M.L., and Mistry, K.B., Development of the Learning Health System Researcher Core Competencies. *Health Services Research*, 2018;53:2615-2632.
25. Balakrishnan, K., Sidell, D.R., Bauman, N.M., Bellia-Munzon, G.F., Boesch, R.P., Bromwich, M., Cofer, S.A., Daines, C., de Alarcon, A., Garabedian, N., Hart, C.K., Ida, J.B., Leboulanger, N., Manning, P.B., Mehta, D.K., Monnier, P., Myer, C.M., Prager, J.D., Preciado, D., Propst, E.J., Rahbar, R., Russell, J., Rutter, M.J., Thierry, B., Thompson, D.M., Torre, M., Varela, P., Vijayasekaran, S., White, D.R., Wineland, A.M., Wood, R.E., Wootten, C.T., Zur, K., and Cotton, R.T., Outcome measures for pediatric laryngotracheal reconstruction: International consensus statement. *Laryngoscope*, 2018;129:244-255.

26. Dreesens, D., Stiggelbout, A., Agoritsas, T., Elwyn, G., Flottorp, S., Grimshaw, J., Kremer, L., Santesso, N., Stacey, D., Treweek, S., Armstrong, M., Gagliardi, A., Hill, S., Légaré, F., Ryan, R., Vandvik, P., and van der Weijden, T., A conceptual framework for patient-directed knowledge tools to support patient-centred care: Results from an evidence-informed consensus meeting. *Patient Education and Counseling*, 2019;102:1898-1904.
27. Garrison, L.P.J., Neumann, P.J., Erickson, P., Marshall, D., and Mullins, C.D., Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. *Value in Health*, 2007;10:326-335.
28. Kirwan, J.R., Boers, M., Hewlett, S., Beaton, D., Bingham, C.O., Choy, E., Conaghan, P.G., D'Agostino, M.-A., Dougados, M., Furst, D.E., Guillemin, F., Gossec, L., van der Heijde, D.M., Kloppenburg, M., Kvien, T.K., Landewé, R.B.M., Mackie, S.L., Matteson, E.L., Mease, P.J., Merkel, P.A., Ostergaard, M., Saketkoo, L.A., Simon, L., Singh, J.A., Strand, V., and Tugwell, P., Updating the OMERACT Filter: Core Areas as a Basis for Defining Core Outcome Sets. *The Journal of Rheumatology*, 2014;41:994.
29. Mattke, S., Epstein, A.M., and Leatherman, S., The OECD Health Care Quality Indicators Project: history and background. *International Journal for Quality in Health Care*, 2006;18:1-4.
30. Boers, M., Idzerda, L., Kirwan, J.R., Beaton, D., Escorpizo, R., Boonen, A., Magasi, S., Sinha, I., Stucki, G., and Tugwell, P., Toward a Generalized Framework of Core Measurement Areas in Clinical Trials: A Position Paper for OMERACT 11. *The Journal of Rheumatology*, 2014;41:978.
31. Kirwan, J.R., Bartlett, S.J., Beaton, D.E., Boers, M., Bosworth, A., Brooks, P.M., Choy, E., de Wit, M., Guillemin, F., Hewlett, S., Kvien, T.K., Landewé, R.B., Leong, A.L., Lyddiatt, A., March, L., May, J., Montie, P.L., Nikaï, E., Richards, P., Voshaar, M.M.J.H., Smeets, W., Strand, V., Tugwell, P., and Gossec, L., Updating the OMERACT Filter: Implications for Patient-reported Outcomes. *The Journal of Rheumatology*, 2014;41:1011.
32. D'Agostino, M.A., Boers, M., Kirwan, J., van der Heijde, D., Østergaard, M., Schett, G., Landewé, R.B., Maksymowych, W.P., Naredo, E., Dougados, M., Iagnocco, A., Bingham, C.O.r., Brooks, P.M., Beaton, D.E., Gandjbakhch, F., Gossec, L., Guillemin, F., Hewlett, S.E., Kloppenburg, M., March, L., Mease, P.J., Moller, I., Simon, L.S., Singh, J.A., Strand, V., Wakefield, R.J., Wells, G.A., Tugwell, P., and Conaghan, P.G., Updating the OMERACT filter: implications for imaging and soluble biomarkers. *Journal of Reumatology*, 2014;41:1016-24.
33. Kumanyika, S., Taylor, W.C., Grier, S.A., Lassiter, V., Lancaster, K.J., Morssink, C.B., and Renzaho, A.M.N., Community energy balance: A framework for contextualizing cultural influences on high risk of obesity in ethnic minority populations. *Preventive Medicine*, 2012;55:371-381.
34. Lalitha, A.V., Fassl, B., Gist, R.E., Shah, B.R., Chawla, N., Singh, A., Baranawal, A., Shamarao, S., Vanaki, R., Mahajan, P., Patel, R., Chauhan, V., Batra, P., Saha, A., Galwankar, S., and Soans, S., 2019 WACEM - Academic College of Emergency Experts Consensus Recommendations on Admission Criteria to Pediatric Intensive Care Unit from the Emergency Departments in India. *Journal of Emergencies, Trauma, and Shock*, 2019;12:155-162.
35. Fazaeli, S., Ahmadi, M., Rashidian, A., and Sadoughi, F., A framework of a health system responsiveness assessment information system for Iran. *Iranian Red Crescent Medical Journal*, 2014;16:e17820.
36. Ardalan, A., Rajaei, M.H., Masoumi, G., Azin, A., Zonoobi, V., Sarvar, M., Vaskoei Eshkevari, K., Ahmadnezhad, E., and Jafari, G., 2012-2025 Roadmap of I.R.Iran's Disaster Health Management. *PLoS currents*, 2012;4:e4f93005fbcb34-e4f93005fbcb34.
37. ten Asbroek, A.H.A., Arah, O.A., Geelhoed, J., Custers, T., Delnoij, D.M., and Klazinga, N.S., Developing a national performance indicator framework for the Dutch health system. *International Journal for Quality in Health Care*, 2004;16:i65-i71.

38. Rae, C.S., Pole, J.D., Gupta, S., Digout, C., Szwajcer, D., Flanders, A., Srikanthan, A., Hammond, C., Schacter, B., Barr, R.D., and Rogers, P.C., Development of System Performance Indicators for Adolescent and Young Adult Cancer Care and Control in Canada. *Value in Health*, 2020;23:74-88.
39. El Oakley, R.M., Ghrew, M.H., Aboutwerat, A.A., Alageli, N.A., Neami, K.A., Kerwat, R.M., Elfituri, A.A., Ziglam, H.M., Saifenasser, A.M., Bahron, A.M., Aburawi, E.H., Sagar, S.A., Tajoury, A.E., and Benamer, H.T.S., Consultation on the Libyan health systems: towards patient-centred services. *Libyan Journal of Medicine*, 2013;8:20233.
40. Hermann, R.C., Palmer, H., Leff, S., Shwartz, M., Provost, S., Chan, J., Chiu, W.T., and Lagodmos, G., Achieving Consensus Across Diverse Stakeholders on Quality Measures for Mental Healthcare. *Medical Care*, 2004;42.
41. Spyropoulos, A.C., Viscusi, A., Singhal, N., Gilleylen, J., Kouides, P., Howard, M., Rudd, K., Ansell, J., and Triller, D.M., Features of electronic health records necessary for the delivery of optimized anticoagulant therapy: consensus of the EHR Task Force of the New York State Anticoagulation Coalition. *Annals of Pharmacotherapy*, 2015;49:113-124.
42. Leone, F.T., Carlsen, K.-H., Chooljian, D., Crotty Alexander, L.E., Detterbeck, F.C., Eakin, M.N., Evers-Casey, S., Farber, H.J., Folan, P., Kathuria, H., Latzka, K., McDermott, S., McGrath-Morrow, S., Moazed, F., Munzer, A., Neptune, E., Pakhale, S., Sachs, D.P.L., Samet, J., Sufian, B., and Upson, D., Recommendations for the Appropriate Structure, Communication, and Investigation of Tobacco Harm Reduction Claims. An Official American Thoracic Society Policy Statement. *American Journal of Respiratory and Critical Care Medicine*, 2018;198:e90-e105.
43. Mazzone, P., Powell, C.A., Arenberg, D., Bach, P., Detterbeck, F., Gould, M.K., Jaklitsch, M.T., Jett, J., Naidich, D., Vachani, A., Wiener, R.S., and Silvestri, G., Components Necessary for High-Quality Lung Cancer Screening: American College of Chest Physicians and American Thoracic Society Policy Statement. *CHEST*, 2015;147:295-303.
44. Grandes, G., Sanchez, A., Cortada, J.M., Balague, L., Calderon, C., Arrazola, A., Vergara, I., Millan, E.S., and Prescribe Vida Saludable group, Is integration of healthy lifestyle promotion into primary care feasible? Discussion and consensus sessions between clinicians and researchers. *BMC Health Services Research*, 2008;8:213.
45. Hughes, R., Competencies for effective public health nutrition practice: a developing consensus. *Public Health Nutrition*, 2004;7:683-691.
46. Nijs, H.G., Bleeker, J.K., van Der Waal, M.A., and Casparie, A.F., Need for consensus development in prehospital emergency medicine: effect of an expert panel approach. *European Journal of Emergency Medicine*, 1998;5:329-34.
47. Schünemann, H.J., Wiercioch, W., Etxeandia, I., Falavigna, M., Santesso, N., Mustafa, R., Ventresca, M., Brignardello-Petersen, R., Laisaar, K.-T., Kowalski, S., Baldeh, T., Zhang, Y., Raid, U., Neumann, I., Norris, S.L., Thornton, J., Harbour, R., Treweek, S., Guyatt, G., Alonso-Coello, P., Reinap, M., Brožek, J., Oxman, A., and Akl, E.A., Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *Canadian Medical Association Journal*, 2014;186:E123.
48. Scott, E.A. and Black, N., When does consensus exist in expert panels? *Journal of Public Health Medicine*, 1991;13:35-39.

Figure 1: PRISMA flowchart




1

2

Table 1: Consensus development studies in health policy and system field

| Authors (year) | Country | Framework/policy that consensus developed upon | Consensus development method | No. of panel members | Consensus threshold | Method modification |
|---|---|---|---|---|---|---|
| Lalitha et al. (2019)[34] | India | Admission criteria on paediatric intensive care unit in India | CM | Not stated | Not stated | • Consensus was not developed in the CM.<br>• Full taskforce team discussed, and consensus developed on the refined draft framework from CM |
| Calvert et al. (2018)[21] | International | SPIRIT-Patient Reported Outcomes (PRO) Extension | DT and CM | DT: n=99<br>CM: n=29 | Not stated | • DT – 2 rounds of online surveys<br>• CM – 2-day face-to-face meeting<br>• Final consultation through 3-weel period with all attendees from DT and CM<br>• Consensus/final agreement by the SPIRIT-PRO group |
| Fazaeli et al. (2014)[35] | Iran | A framework of a health system responsiveness assessment information system | DT and CM | DT: n=25 | Components with ≥75% agreement were used for the secondary framework. Components with 50 to 75% agreement were entered into the second round of DT. Items with <50% agreement were excluded. | • DT – 2 rounds of surveys with 3-point Likert scale (no information on distribution method)<br>• CM – face-to-face meeting by panel members to discuss results to reach consensus |
| Cornel et al. (2014)[22] | International (EU) | A framework on newborn screening implementation | CM | Not stated | Not stated | • Consensus was not developed in the CM.<br>• Consultation by email and face-to-face meeting with panel members<br>• Final document was endorsed by the Boards of the International Society. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Salvador-Carulla et al. (2011)[23] | International | Classification of intellectual disabilities in ICD-11 | CM | Not stated | Not stated | • A series of meetings – 3 face-to-face meetings, 7 teleconferences, and electronic exchanges |
| Ten Asbroek et al. (2004)[37] | The Netherlands | A national performance indicator framework for the Dutch health system | CM | 38 | Not stated | • A series of meetings – not stated the frequency and delivery of the meetings |
| Rae et al. (2020)[38] | Canada | System performance indicators for adolescent and young adult cancer care and control | DT | 10 | Survey 1 - indicator eliminated if <70% of group scored the indicator as a 3<br>Survey 2 - indicator eliminated if <60% of group scored the indicator as a 3<br>Final consensus - endorsed by ≥ 80% of group in Round 1 and 3 | • DT – 3 rounds of surveys (no information on distribution method) |
| Forrest et al. (2018)[24] | International | A list of core competencies for learning health system researchers | DT and CM | 19 | CM consensus: Not stated<br><br>DT consensus: Competencies with a median of at least 7 and ≥ 75% of panel members rating between 7 and 9 were evaluated for retention | • CM – 2 face-to-face expert panel meeting, rating competencies with 5-point Likert scale<br>• DT – 1 round of a modified process with 9-point Likert scale (no information on distribution method)<br>• CM – final face-to-face expert panel meeting |
| Balakrishnan et al. (2018)[25] | International | Outcome measures for paediatric laryngotracheal reconstruction | DT | 40 | Consensus for mean rating ≥ 7, with ≤ 1 response ≥2 points away from mean | • 2 different rounds of surveys at different levels of important to include (with 9-point Likert scale) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | Near consensus for mean rating ≥ 6.5, with ≤ 2 response ≥2 points away from mean | |
| Dreesens et al. (2019)[26] | International | A conceptual framework for patient-directed knowledge tools to support patient-centred care | CM | 15 (9 attended in person, 4 attended through teleconference, and 2 participated via email) | Not stated | • 2-day face-to-face consensus meeting with international experts |
| Garrison et al. (2007)[27] | International | A framework to assist health-care decision-makers in dealing with Real-world data | CM | Not stated | Not stated | • Face-to-face meetings and electronic exchange via email with Taskforce members and public consultation from the membership of the ISPOR |
| Kirwan et al. (2014)[28] | | Outcome Measures in Rheumatology (OMERACT) Filters: core outcome sets | CM | Not stated | Not stated | • Face-to-face meetings using breakout small group discussion at conference got for discussion and a final plenary voting |
| Mattke et al. (2006)[29] | International (OECD) | Health care quality indicators | DT | 4 expert panels (the number of panel members of each panel was not stated) | Retain measures with high ratings (≥7) on both relevance and soundness, and also measures with intermediate scores on those dimensions if data collection considered to be feasible | • A modified DT – evaluate the relevance and scientific soundness of identified indicators with 9-point Likert scale (no statement on the frequency of rounds) |
| El Oakley et al. (2013)[39] | Libya | Recommendations mapping a modern health systems for the 21st century | CM | 500 | Not stated | • A half-day conference session, in which potential solutions and priorities for change were identified – where consensus |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | achieved, included in proposed recommendations |
| Boers et al. (2014)[30] | International | A framework of core measurement areas in clinical trials for OMERACT 11 | DT and CM | CM – 18<br>DT - 262 | Not stated | • CM – held at conference for structured discussion<br>• DT – invited 2293 people – a total of 262 returned, indicating broad agreement |
| Kirwan et al. (2014)[31] | International | A framework for identifying the relevant core outcomes universal to all studies of the effects of intervention effects | CM | >115 | Not stated | • Group discussions with small breakout groups at conference |
| Hermann et al. (2004)[40] | US | A core set of quality measures for mental health and substance-related care | DT and NGT | 12 | Meaningfulness of 3 or less with dispersion of ≤0.8<br>Feasibility of ≤6 with dispersion ≤0.9<br>The highest-rated measure in that category was added to the core set. | • A modified DT – 2 rounds of surveys, assessing with 3 scales of meaningfulness, 3 scales of feasibility and 1 scale for overall agreement of inclusion, using 9-point Likert scale<br>• Face-to-face meeting – discussing the results of 1st round of DT on measures with overall scores of 6 or less and significant dispersion |
| Grandes et al. (2008)[44] | Spain | Factors in the Basque Health System hindering/facilitating the integration of healthy lifestyle promotion in primary health care setting | NGT | 12 | Not stated | • 5 structured meetings for discussion<br>• A draft circulation to verify validity of document |
| D'Agostino et al. (2014)[32] | International | An imaging or biochemical measurement | CM | Not stated | Not stated | • 3 disease-related groups further divided into small breakout |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | instrument in OMERACT Filter | | | | groups to discuss and propose new hierarchical structure<br>• Each discussion group reported to its main plenary session of all participants |
| Spyropoulos et al. (2015)[41] | US | A list of electronic health records features clinically necessary to delivery optimized anticoagulation management | DT | 12 | No more than one reviewer expressing opposition to any recommended feature | • Electronic circulation of draft logic and technical guidance among taskforce members (n = 150)<br>• 1 round of DT – assessing a clinical necessity using 5-poing Liker Scale |
| Leone et al. (2018)[42] | US | Standards of practice in making tobacco harm reduction claims | CM | Not stated | Unanimous approval | • Online and face-to-face meetings, and electronic circulation of successive drafts |
| Kumanyika et al. (2012)[33] | International | A visual and narrative framework for community-level interventions for obesity | CM | 6 | Not stated | • A 2-day session and literature review for preliminary framework<br>• 2-year iterative consultation |
| Nijs et al. (1998)[46] | The Netherlands and Belgium | Agreement on prehospital emergency medicine care | DT and NGT | 7 | Multiple rater kappa values (a kappa value of 1 denotes perfect agreement, corrected for chance-expected agreement. A value of 0 denotes that observed agreement equals chance-expected agreement. A value below 0 denotes that observed agreement is even less than chance-expected agreement) | • A modified DT – 2 rounds of posting process assessing the need of cares<br>• A modified NGT – discussing the lowest agreement case and re-rating |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mazzone et al. (2015)[43] | US | Policy statements on the components necessary for high-quality lung cancer screening | Iterative feedback | Not stated | Unanimous approval | • Iterative written and verbal feedback of committee members on 2 quality metrics assessing validity, feasibility and relevance |
| Ardalan et al. (2012)[36] | Iran | 2012-2025 roadmap of Disaster Health Management | NGT and DT | Not stated | Not stated | • NGT – discussing and assessing necessity of items (no information of specific process)<br>• DT – No information of the frequency and distribution method |
| Hughes (2004)[45] | International (EU, US, AUS) | Competencies for effective public health nutrition practice | DT | 20 | Consensus cut-offs of ≥80% agreement rated essential in Rounds 1 and 3. | • A modified DT – 3 rounds of surveys (via email) with 3-point rating scales assessing the relevance |

*Abbreviation for consensus development method: Consensus meeting (CM), Delphi Technique (DT), Nominal Group Technique (NGT), Consensus Development Conference (CDC).