Developing a Clinician Friendly Tool to Identify Useful Clinical Practice Guidelines: G-TRUST

Allen F. Shaughnessy, PharmD, MMedEd[1]

Akansha Vaswani, MS[2]

Bonnie K. Andrews, MPH[2]

Deborah R. Erlich, MD, MMedEd[1]

Frank D'Amico, PhD[3]

Joel Lexchin, MD[4]

Lisa Cosgrove, PhD[2]

[1]Tufts University School of Medicine, Boston, Massachusetts

[2]Department of Counseling and School Psychology, University of Massachusetts, Boston, Massachusetts

[3]McAnulty College and Graduate School of Liberal Arts, Duquesne University, Pittsburgh, Pennsylvania

[4]School of Health Policy and Management, York University, Toronto, Canada

MORE ONLINE
www.annfammed.org

CORRESPONDING AUTHOR

Allen F. Shaughnessy, PharmD, MMedEd
Family Medicine Residency at Cambridge
Health Alliance
Tufts University
195 Canal St
Malden, MA 02148
Allen.Shaughnessy@tufts.edu

ABSTRACT

BACKGROUND Clinicians are faced with a plethora of guidelines. To rate guidelines, they can select from a number of evaluation tools, most of which are long and difficult to apply. The goal of this project was to develop a simple, easy-to-use checklist for clinicians to use to identify trustworthy, relevant, and useful practice guidelines, the Guideline Trustworthiness, Relevance, and Utility Scoring Tool (G-TRUST).

METHODS A modified Delphi process was used to obtain consensus of experts and guideline developers regarding a checklist of items and their relative impact on guideline quality. We conducted 4 rounds of sampling to refine wording, add and subtract items, and develop a scoring system. Multiple attribute utility analysis was used to develop a weighted utility score for each item to determine scoring.

RESULTS Twenty-two experts in evidence-based medicine, 17 developers of high-quality guidelines, and 1 consumer representative participated. In rounds 1 and 2, items were rewritten or dropped, and 2 items were added. In round 3, weighted scores were calculated from rankings and relative weights assigned by the expert panel. In the last round, more than 75% of experts indicated 3 of the 8 checklist items to be major indicators of guideline usefulness and, using the AGREE tool as a reference standard, a scoring system was developed to identify guidelines as useful, may not be useful, and not useful.

CONCLUSION The 8-item G-TRUST is potentially helpful as a tool for clinicians to identify useful guidelines. Further research will focus on its reliability when used by clinicians.

Ann Fam Med 2017;15:413-418. https://doi.org/10.1370/afm.2119.

INTRODUCTION

Clinicians expect clinical practice guidelines to have 3 characteristics. Good guidelines should (1) be trustworthy, in that the recommendations are based on the best available evidence; (2) be relevant, meaning the recommendations are pertinent to one's practice population and focus on affecting outcomes of importance to patients; and, (3) have a high degree of utility, in that the recommendations are clear and actionable.

Numerous researchers have documented issues with the guideline development process. Guidelines vary in their relevance to specific clinical practice,[1-5] their use of evidence,[6-14] and the role of other factors on the process of drafting recommendations.[15-32]

Tools are available to evaluate the quality of clinical practice guidelines.[33-38] These tools, however, are designed in part to guide guideline development and are difficult to use by nonresearchers without extensive training. None of these tools considers the need for a focus on patient-oriented outcomes, and none allows users to conclude whether a guideline should be followed. The aim of this study was to develop the Guideline Trustworthiness, Relevance, and Utility Scoring Tool (G-TRUST) for clinicians to easily identify useful clinical practice guidelines.

METHODS

The study design used a Delphi approach[39] to obtain expert consensus on items for inclusion, to hone the wording of the items, and to develop a ranking system. The Delphi approach is designed to gather the wisdom of the group without succumbing to issues of group process, such as social pressure (groupthink[40]), personality influence,[15] and individual dominance[16,41] It can be applied to generate consensus within groups of individuals who hold different views. Its main attributes include anonymity of participants, structured information flow to participants, and regular feedback to the group on the progress of the decision making.

Because we have already developed and piloted a set of items to be used to evaluate the validity of guidelines, we used a modified Delphi technique, which can be used when basic information is already available.[42] The source of checklist items is outlined in Supplemental Appendix 1 at http://www.annfammed. org/content/15/5/413/suppl/DC1.

Selection of Experts

We selected a representative group of volunteer experts from 2 populations: producers of practice guidelines known to be of high quality in several clinical areas;[7,13,14,43-46] and self-identified and recognized experts in evidence-based medicine. Physicians in family medicine and primary care internal medicine made up some of both groups. The sources of these experts are further described in the supplementary material (Supplemental Appendix 2, http://www.annfammed. org/content/15/5/413/suppl/DC1).

Initial Items on the Tool

The initial 8 items of the instrument were derived from several sources, including the National Academy of Medicine's (formerly Institute of Medicine) "Clinical Practice Guidelines We Can Trust,"[47] the AGREE II instrument,[48,49] and the research of ours[50] and others on guideline validity.[34] The items hew most closely to the National Academy of Medicine standards[47] and are critical for evaluating and recognizing flaws in the evidence development process, the relevance of recommendations to clinical practice, and the threats to the judgment process of creating recommendations from the evidence.

Delphi Process

The modified Delphi process consisted of 4 rounds and was conducted using an online survey instrument (http://www.surveymonkey.com). Participants were not told the number or identity of other participants. At each stage we analyzed results from each subgroup (evidence-based medicine and guideline experts) separate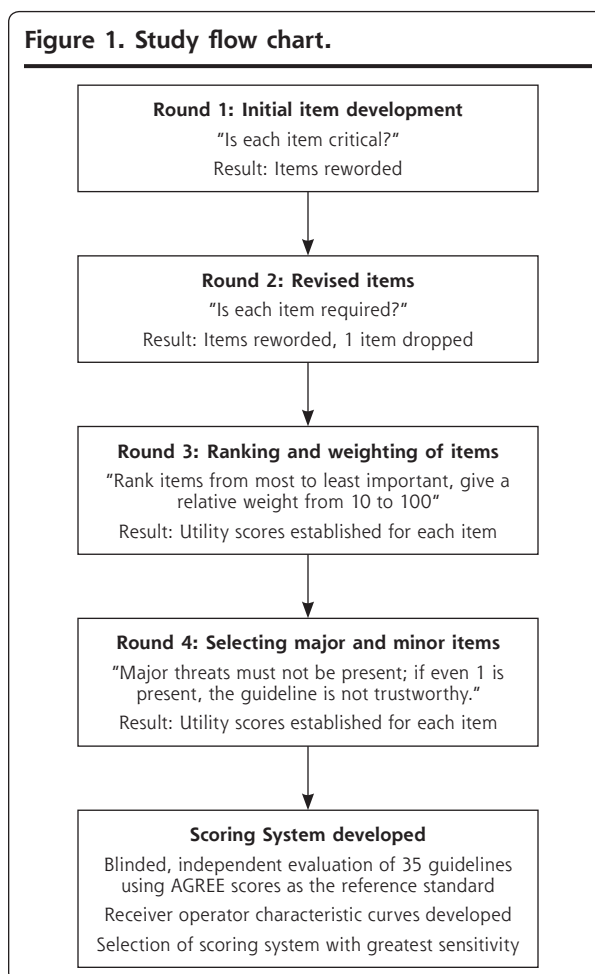ly to identify any discrepancies in opinion. The complete process is outlined in Figure 1 and explained in Supplemental Appendix 3, http://www.annfammed. org/content/15/5/413/suppl/DC1.

The goal of the first round was to develop the wording of the items and identify additional items to be added to the tool. For the second round, participants were asked whether the revised items were "required to identify guidelines that present both relevant and trustworthy recommendations." For the third round, participants were given aggregate responses from the second round and asked to rank, weight, and order the items. Based on these rankings and weights, we used multiple attribute utility analysis[51,52] to obtain utility scores for each item on a scale from 0 to 100. During the fourth round, participants determined whether each item was a major or minor threat to the usefulness of a practice guideline.

Scoring System

To determine concurrent validity and to develop a scoring system, the final items were used to assess the quality of 26 (74.3%) low-quality and 9 high-

Figure 1. Study flow chart.

Round 1: Initial item development
"Is each item critical?"
Result: Items reworded

Round 2: Revised items
"Is each item required?"
Result: Items reworded, 1 item dropped

Round 3: Ranking and weighting of items
"Rank items from most to least important, give a relative weight from 10 to 100"
Result: Utility scores established for each item

Round 4: Selecting major and minor items
"Major threats must not be present; if even 1 is present, the guideline is not trustworthy."
Result: Utility scores established for each item

Scoring System developed
Blinded, independent evaluation of 35 guidelines using AGREE scores as the reference standard
Receiver operator characteristic curves developed
Selection of scoring system with greatest sensitivity

quality guidelines previously assessed by others using the AGREE instrument.[13,14] Two authors (L.C. and A.F.S.), independently and masked to the AGREE quality scores, assessed each guideline using 7 of the 8 G-TRUST items, excluding the item evaluating the clinical relevance of the recommendations because this criterion is not considered in AGREE or AGREE II and will vary based on user. For each item they determined whether the criterion was met, was not met, or could not be determined from the guideline description. Results from each investigator were compared and discrepancies resolved via discussion.

Analysis

For each round, except for utility calculations after round 3, we calculated average responses. For round 4, we used either the Fisher exact test or the χ^2 test with Yates correction to determine whether designation of major and minor flaws was different between evidence-based medicine experts and guideline writers.

RESULTS

Expert Consensus Panel

The group (Table 1) comprised 40 members representing expertise in evidence-based medicine (n = 22) and in guideline development (n = 17); 1 consumer representative had expertise in risk communication and health policy. All panel members participated in rounds 1 and 2, 95% of members (n = 38) participated in the round 3, and 85% (n = 34) participated in round 4.

Item Selection and Refinement

Responses from the first Delphi round resulted in changed wording and explanations of several items. A general statement (and not a specific item) was added that the guideline should have been written or updated within the past 5 years, which is similar to the requirement for inclusion in the National Guideline Clearinghouse.[53]

In the second round, 6 of the 8 items were deemed to be critical to assess the relevance and validity of practice guideline recommendations. The item, "The guidelines are the official stance or policy of a professional society," was deemed to be critical by only 7.5% and was dropped from the instrument. Based on written comments, wording for 1 item was slightly changed, and 1 compound item was split into 2 items.

For the third round, Table 2 contains median utility scores calculated from respondents' ranking and weighting. Utility varied widely, and items evaluating evidence quality (systematic review, evidence grading) had the highest utility, followed by items evaluating relevance.

The last round produced a cutoff between major and minor threats that corresponded to a utility of

greater than 16 for the weighted scores (Table 2). Two evidence validity items were considered major items: 1 pertained to identifying systematic review (100%), and the other related to the use of graded evidence (85.3%). One relevance item, "recommendations focus on improving patient-oriented outcomes, explicitly comparing benefits versus harms to support clinical decision making," was also considered to be a major threat to the usefulness of guidelines by most participants (82.4%). The rest of the items were considered major threats by fewer than one-half the participants. There was no difference in these designations between evidence-based medicine experts and guideline developers. Results from each Delphi round are outlined in detail in Supplemental Appendix 4, http://www.annfammed.org/content/15/5/413/suppl/DC1.

G-TRUST Scoring System

Using AGREE scores as our reference standard, we evaluated various combinations of item responses from

Table 1. Demographic Composition of the Expert Panel

Characteristic	No. (%)
Background	
Evidence-based medicine expert	22 (55)
Guideline developer	17 (42.5)
Consumer representative	1 (2.5)
Sex	
Male	26 (65)
Female	14 (35)
Self-identified background	
Clinician	16 (40)
Methodologist	17 (42.5)
Methodologist/clinician	6 (15)
Consumer representative	1 (2.5)
Geography	
Europe	8 (20)
United Kingdom	4 (10)
Canada	4 (10)
United States	21 (52.5)
South America	1 (2.5)
Africa	1 (2.5)
Asia	1 (2.5)
Organization	
Academia/university	20 (50)
Government	8 (20)
Nongovernmental, nonacademic	4 (10)
Industry	3 (7.5)
Professional society	5 (12.5)
Identification method	
Evidence-based medicine listserve	15 (38)
Guideline developers	9 (23)
Author contacts	16 (40)

Table 2. Final Item Wording With Utility Scores and Ratings

Final Item Wording	Median Utility Score[a]	Rating Item as Major Threat %
Relevance threats		
1. The patient populations and conditions are relevant to my clinical setting.	15.7	29.4
2. The recommendations are clear and actionable.	11.2	35.3
3. The recommendations focus on improving patient-oriented outcomes, explicitly comparing benefits versus harms to support clinical decision making.	18.0	82.4
Evidence threats		
4. The guidelines are based on a systematic review of the research data.	22.5	100.0
5. The recommendation statements important to you are based on graded evidence and include a description of the quality (e,g, strong, weak) of the evidence.	20.2	85.3
6. The guideline development includes a research analyst, such as a statistician or epidemiologist.	2.3	26.5
Interpretation threats		
7. The Chair of the guideline development committee and a majority of the rest of the committee are free of declared financial conflicts of interest, and the guideline development group did not receive industry funding for developing the guideline.	3.4	47.1
8. The guideline development includes members from the most relevant specialties and includes other key stakeholders, such as patients, payer organizations, and public health entities, when applicable.	6.7	41.2

[a] Ranked from 0 to 100, with higher scores indicating greater utility.

the consensus panel to develop a scale aimed at fewer lower quality guidelines identified as trustworthy:

Useful: no major items answered "can't tell" or "no", and 0-1 minor items answered "no".

May not be useful: no major items answered "can't tell" or "no", but 2 minor items answered "no".

Not useful: any major item answered "can't tell" or "no", or more than 2 minor items answered "no".

Applying these cutoffs, 3 of 26 (11%) low-quality guidelines were identified as being useful. These guidelines were downgraded by the AGREE instrument because of scores of 0 for the domain "editorial independence," comprising editorial independence from the funding body and recording of conflicts of interest. A lack of clarity of the guideline's description of conflicts of interest resulted in a score of "cannot be determined" using the G-TRUST instrument. After the analysis, we added the following wording (in italics) to this item to add clarity: "The Chair of the guideline development committee and a majority of the rest of the committee are free of declared financial conflicts of interest, *and the guideline development group did not receive industry funding for developing the guideline.*"

The G-TRUST instrument identified 5 guidelines (55%) as either may not be useful or not useful of the 9 determined to be of high quality by AGREE. Using AGREE, 1 guideline received a high score (81%) for rigor even though it was not based on a systematic review of the literature. All the other guidelines were graded as may not be useful because they did not include members from most of the relevant specialties or were not substantially free of conflicts of interest.

DISCUSSION

Through expert consensus we developed an 8-item checklist designed to help clinicians quickly identify useful guidelines to follow in practice. Using AGREE as our reference standard, our checklist identified almost all (92%) of the low-quality guidelines and disqualified many high-quality guidelines because of a stricter definition of trustworthiness. The items in the G-TRUST (tool available from the authors) address issues and concerns voiced by the National Academy of Medicine report[47] and the AGREE II instrument[33] and add additional issues of relevance not considered by either.[54,55] G-TRUST is more stringent than AGREE II in that it stipulates an independent (ie, nonconflicted) research analyst or methodologist be part of the process, based on recent research findings that including independent methodological experts may better ensure evidence-based and conservative recommendations.[56,57] The tool is also more stringent than AGREE II in its handling of conflicts of interest (barring them rather than simply addressing them) and in broad representation on the guideline development group. Using the stricter requirement for conflicts of interest reflected in the G-TRUST led to many guidelines being rated as may not be useful that would be rated as of high quality by AGREE.

A major advantage of the G-TRUST is that it gives different importance to individual items (eg, major, minor) and arrives at a determination of overall guideline quality (useful, may not be useful, not useful).

In the development of the scoring system, we produced a conservative cutoff score that, while preventing false positives (eg, falsely identifying guidelines as high quality), will exclude some high-quality guidelines. Given the large number of guidelines, this emphasis is needed to ensure that fewer low-quality guidelines will be incorrectly identified as useful.

A second limitation is that it may be difficult for users to determine conflicts of interest and the presence of a research analyst on the guideline development group. Despite extensive searching, we could not determine the answer to these items for almost one-half the studied guidelines. In a previous study we found that more than one-half (57%) of the guidelines for the treatment of major depressive disorder did not include a conflicts of interest policy or disclosure statement.[8]

The evidence supporting clinical practice guideline development is very preliminary; the Institute of Medicine report, upon which we based our initial development, is widely seen as the best we have. Still, much of the evidence supporting what constitutes a reliable and valid practice guideline is expert opinion.

Further research should determine the reliability of G-TRUST by comparing scores obtained by single users. Also, the use of technology, such as smartphone applications that integrate with the National Guideline Clearinghouse, could be explored to determine the usefulness of the tool. In addition, neither the G-TRUST nor any of the other guideline evaluation tools evaluate whether guidelines provide enough information to support shared decision making.

Considering the proliferation of guidelines in all areas of medicine and the well-documented concerns about their validity and trustworthiness, clinicians need an easy-to-use screening tool to enhance evidence-based care. The 8-item G-TRUST instrument is a potentially helpful tool for clinicians to identify clinical practice guidelines that are trustworthy in their development, reliable in their application to patient care, and have high utility in clinical practice.

To read or post commentaries in response to this article, see it online at http://www.annfamed.org/content/15/5/413.

References

1. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance Group. Evidence based medicine: a movement in crisis? *BMJ*. 2014;348:g3725.

2. Steel N, Abdelhamid A, Stokes T, et al. A review of clinical practice guidelines found that they were often based on evidence of uncertain relevance to primary care patients. *J Clin Epidemiol*. 2014;67(11): 1251-1257.

3. Scullard P, Abdelhamid A, Steel N, Qureshi N. Does the evidence referenced in NICE guidelines reflect a primary care population? *Br J Gen Pract*. 2011;61(584):e112-e117.

4. Hegarty K, Gunn J, Blashki G, Griffiths F, Dowell T, Kendrick T. How could depression guidelines be made more relevant and applicable to primary care? A quantitative and qualitative review of national guidelines. *Br J Gen Pract*. 2009;59(562):e149-e156.

5. McCartney M, Treadwell J, Maskrey N, Lehman R. Making evidence based medicine work for individual patients. *BMJ*. 2016;353:i2452.

6. Feuerstein JD, Pelsis JR, Lloyd S, Cheifetz AS, Stone KR. Systematic analysis of the quality of the scientific evidence and conflicts of interest in osteoarthritis of the hip and knee practice guidelines. *Semin Arthritis Rheum*. 2016;45(4):379-385.

7. Feuerstein JD, Akbari M, Gifford AE, et al. Systematic review: the quality of the scientific evidence and conflicts of interest in international inflammatory bowel disease practice guidelines. *Aliment Pharmacol Ther*. 2013;37(10):937-946.

8. Cosgrove L, Bursztajn HJ, Erlich DR, Wheeler EE, Shaughnessy AF. Conflicts of interest and the quality of recommendations in clinical guidelines. *J Eval Clin Pract*. 2013;19(4):674-681.

9. Lee DH, Vielemeyer O. Analysis of overall level of evidence behind Infectious Diseases Society of America practice guidelines. *Arch Intern Med*. 2011;171(1):18-22.

10. McAlister FA, van Diepen S, Padwal RS, Johnson JA, Majumdar SR. How evidence-based are the recommendations in evidence-based guidelines? *PLoS Med*. 2007;4(8):e250.

11. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA*. 2009;301(8):831-841.

12. Rodríguez-Gutiérrez R, Montori VM. Glycemic control for patients with type 2 diabetes mellitus: Our evolving faith in the face of evidence. *Circ Cardiovasc Qual Outcomes*. 2016;9(5):504-512.

13. Burda BU, Norris SL, Holmer HK, Ogden LA, Smith ME. Quality varies across clinical practice guidelines for mammography screening in women aged 40-49 years as assessed by AGREE and AMSTAR instruments. *J Clin Epidemiol*. 2011;64(9):968-976.

14. Bancos I, Cheng T, Prokop LJ, Montori VM, Murad MH. Endocrine clinical practice guidelines in North America. A systematic assessment of quality. *J Clin Epidemiol*. 2012;65(5):520-525.

15. Berg AO. Clinical practice guideline panels: personal experience. *J Am Board Fam Pract*. 1996;9(5):366-370.

16. Raine R, Sanderson C, Hutchings A, Carter S, Larkin K, Black N. An experimental study of determinants of group judgments in clinical guideline development. *Lancet*. 2004;364(9432):429-437.

17. Levinsky NG. Nonfinancial conflicts of interest in research. *N Engl J Med*. 2002;347(10):759-761.

18. Thagard P. The moral psychology of conflicts of interest: Insights from affective neuroscience. *J Appl Philos*. 2007;24(4):367-380.

19. Guyatt G, Akl EA, Hirsh J, et al. The vexing problem of guidelines and conflict of interest: a potential solution. *Ann Intern Med*. 2010; 152(11):738-741.

20. Tibau A, Bedard PL, Srikanthan A, et al. Author financial conflicts of interest, industry funding, and clinical practice guidelines for anticancer drugs. *J Clin Oncol*. 2015;33(1):100-106.

21. Ballini L, Grilli R. Practice guidelines by specialist societies: a jump back into the past? *Int J Clin Pract*. 2007;61(7):1075-1076.

22. Sinuff T, Patel RV, Adhikari NK, Meade MO, Schünemann HJ, Cook DJ. Quality of professional society guidelines and consensus conference statements in critical care. *Crit Care Med.* 2008;36(4): 1049-1058.

23. Choudhry NK, Stelfox HT, Detsky AS. Relationships between authors of clinical practice guidelines and the pharmaceutical industry. *JAMA.* 2002;287(5):612-617.

24. Abramson J, Starfield B. The effect of conflict of interest on biomedical research and clinical practice guidelines: can we trust the evidence in evidence-based medicine? *J Am Board Fam Pract.* 2005; 18(5):414-418.

25. Cosgrove L, Bursztajn HJ, Krimsky S. Developing unbiased diagnostic and treatment guidelines in psychiatry. *N Engl J Med.* 2009; 360(19):2035-2036.

26. *Conflict of Interest in Medical Research, Education, and Practice.* Washington, DC: National Academies Press; 2009.

27. Norris SL, Burda BU, Holmer HK, et al. Author's specialty and conflicts of interest contribute to conflicting guidelines for screening mammography. *J Clin Epidemiol.* 2012;65(7):725-733.

28. Norris SL, Holmer HK, Ogden LA, Burda BU. Conflict of interest in clinical practice guideline development: a systematic review. *PLoS One.* 2011;6(10):e25153.

29. Johnson L, Stricker RB. The Infectious Diseases Society of America Lyme guidelines: a cautionary tale about the development of clinical practice guidelines. *Philos Ethics Humanit Med.* 2010;5:9.

30. Taylor R, Giles J; R T. Cash interests taint drug advice. *Nature.* 2005; 437(7062):1070-1071.

31. Neumann J, Korenstein D, Ross JS, Keyhani S. Prevalence of financial conflicts of interest among panel members producing clinical practice guidelines in Canada and United States: cross sectional study. *BMJ.* 2011;343:d5621.

32. Newman TB, Pletcher MJ, Hulley SB. Overly aggressive new guidelines for lipid screening in children: evidence of a broken process. *Pediatrics.* 2012;130(2):349-352.

33. Brouwers MC, Kho ME, Browman GP, et al; AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting and evaluation in health care. *J Clin Epidemiol.* 2010;63(12):1308-1311.

34. Siering U, Eikermann M, Hausner E, Hoffmann-Eßer W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. *PLoS One.* 2013;8(12):e82915.

35. Brouwers MC, Kho ME, Browman GP, et al; AGREE Next Steps Consortium. The Global Rating Scale complements the AGREE II in advancing the quality of practice guidelines. *J Clin Epidemiol.* 2012; 65(5):526-534.

36. Coroneos CJ, Voineskos SH, Cornacchi SD, Goldsmith CH, Ignacy TA, Thoma A. Users' guide to the surgical literature: how to evaluate clinical practice guidelines. *Can J Surg.* 2014;57(4):280-286.

37. Semlitsch T, Jeitler K, Kopp IB, Siebenhofer A. [Development of a workable mini checklist to assess guideline quality]. *Z Evid Fortbild Qual Gesundhwes.* 2014;108(5-6):299-312.

38. Grimmer K, Dizon JM, Milanese S, et al. Efficient clinical evaluation of guideline quality: development and testing of a new tool. *BMC Med Res Methodol.* 2014;14:63.

39. Dalkey NC. *The Delphi Method: An Experimental Study of Group Opinion.* Santa Monica, CA: RAND Corp; 1969: http://www.rand.org/pubs/research_memoranda/RM5888.html.

40. Madigosky W, van Schaik S. Context matters: groupthink and outcomes of health care teams. *Med Educ.* 2016;50(4):387-389.

41. Callaway MR, Marriott RG, Esser JK. Effects of dominance on group decision making: toward a stress-reduction explanation of groupthink. *J Pers Soc Psychol.* 1985;49(4):949-952.

42. Kerlinger FN. *Foundations of Behavioral Research.* New York, NY: Holt, Rinehart, and Winston, Inc; 1973.

43. Gaebel W, Riesbeck M, Wobrock T. Schizophrenia guidelines across the world: a selective review and comparison. *Int Rev Psychiatry.* 2011;23(4):379-387.

44. Hogeveen SE, Han D, Trudeau-Tavara S, et al. Comparison of international breast cancer guidelines: are we globally consistent? cancer guideline AGREEment. *Curr Oncol.* 2012;19(3):e184-e190.

45. Kredo T, Gerritsen A, van Heerden J, Conway S, Siegfried N. Clinical practice guidelines within the southern african development community: A descriptive study of the quality of guideline development and concordance with best evidence for five priority diseases. *Health Res Policy Syst.* 2012;10:1.

46. Knai C, Brusamento S, Legido-Quigley H, et al. Systematic review of the methodological quality of clinical guideline development for the management of chronic disease in Europe. *Health Policy.* 2012; 107(2-3):157-167.

47. Institute of Medicine, ed. *Clinical Practice Guidelines We Can Trust.* Washington, DC: The National Academies Press; 2011.

48. Brouwers MC, Kho ME, Browman GP, et al; AGREE Next Steps Consortium. Development of the AGREE II, part 1: performance, usefulness and areas for improvement. *CMAJ.* 2010;182(10):1045-1052.

49. Brouwers MC, Kho ME, Browman GP, et al; AGREE Next Steps Consortium. Development of the AGREE II, part 2: assessment of validity of items and tools to support application. *CMAJ.* 2010;182(10): E472-E478.

50. Cosgrove L, Shaughnessy AF, Wheeler EE, Austad KE, Kirsch I, Bursztajn HJ. The American Psychiatric Association's guideline for major depressive disorder: a commentary. *Psychother Psychosom.* 2012;81(3):186-188.

51. Butler J, Morrice DJ, Mullarkey PW. A multiple attribute utility theory approach to ranking and selection. *Manage Sci.* 2001;47(6): 800-816.

52. Weirich P. *Decision Space : Multidimensional Utility Analysis.* Cambridge, UK; New York: Cambridge University Press; 2001.

53. National Guideline Clearinghouse. Inclusion criteria. 2014; http://www.guideline.gov/about/inclusion-criteria.aspx. Accessed Oct 24, 2015.

54. Slawson DC, Shaughnessy AF, Bennett JH. Becoming a medical information master: feeling good about not knowing everything. *J Fam Pract.* 1994;38(5):505-513.

55. Psaty BM, Weiss NS, Furberg CD, et al. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. *JAMA.* 1999;282(8): 786-790.

56. Cosgrove L, Shaughnessy AF, Peters SM, Lexchin JR, Bursztajn HJ, Bero LA. Conflicts of interest and the presence of methodologists on guideline development panels: A cross sectional study of clinical practice guidelines for major depressive disorder. *Psychother Psychosom.* 2017;86(3):168-170.

57. Eccles MP, Grimshaw JM, Shekelle P, Schünemann HJ, Woolf S. Developing clinical practice guidelines: target audiences, identifying topics for guidelines, guideline group composition and functioning and conflicts of interest. *Implement Sci.* 2012;7:60.

58. Lenzer J, Hoffman JR, Furberg CD, Ioannidis JP; Guideline Panel Review Working Group. Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. *BMJ.* 2013;347:f5535.

59. Aczél J, Saaty TL. Procedures for synthesizing ratio judgements. *J Math Psychol.* 1983;27(1):93-102.

60. Kemp S, Grace RC. How do people order stimuli? *Atten Percept Psychophys.* 2014;76(6):1847-1859.