Development and preliminary verification of the evaluation system for clinical practice guidelines in China

Ji-Yao Wang a,b,* , Qiang Wang c,d, Xiao-Qin Wang b,e, Xue-Juan Jin b,d, Bo-Heng Zhang b,f, Shi-Yao Chen a,b, Xue-Cheng Gao c,**

a Department of Gastroenterology, Zhongshan Hospital, Fudan University, Shanghai 200032, China
b Evidence-based Medicine Center, Fudan University, Shanghai 200032, China
c Medical Management Center of the National Health and Family Planning Commission, Beijing 100191, China
d Institute of Cardiovascular Diseases, Zhongshan Hospital, Fudan University, Shanghai 200032, China
e Department of Hematology, Huashan Hospital, Fudan University, Shanghai 200040, China
f Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China

Original Article

Abstract

Objective: Clinical practice guidelines can improve healthcare processes and patient outcomes; however, the quality of these guidelines varies greatly in China. The aim of this study was to construct a comprehensive instrument for the appraisal of clinical practice guidelines in China (AGREE-CHINA), and to validate its reliability as a tool for helping potential guideline users in assessing guideline quality.

Methods: First, an interdisciplinary working group was established for developing the methods. They also created a checklist as a tool according to the Appraisal of Guidelines, Research and Evaluation II (AGREE II) standards, considering the particularity of Chinese clinical practice. Next, the first draft of AGREE-China was developed by vote, modification, preliminary trial, and cross-verification. To ensure the objectivity, credibility, and reproducibility of the draft assessment, all of the checklists and standards were cross-reviewed fairly widely. Finally, AGREE-CHINA and AGREE II were used to assess the Chinese guidelines published in the past five years, and the results were compared.

Results: The presented AGREE-CHINA covered five main checkpoints (science and rigor, effectiveness and safety, economy, usability and feasibility, and conflicts of interest) with each point divided into several more specific checkpoints. Definitions and rationales for each main checkpoint appear in the Appendix. The quality ratings based on the total scores of AGREE-China and AGREE II were consistent ($r = 0.508, P = 0.020$). Compared with AGREE II, the study showed a higher level of interrater-reliability for AGREE-CHINA overall ($ICC = 0.957, P < 0.001$). The mean time required for AGREE-CHINA was less than that for AGREE II; this was approximately 30 minutes for every assessment. User satisfaction was generally high.

* This paper was first published in the National Medical Journal of China. 2018; 98:1544-1548 [in Chinese].
* Corresponding author. Department of Gastroenterology, Zhongshan Hospital, Fudan University, Shanghai, 200032, China.
** Corresponding author.
E-mail addresses: wang.jiyao@zs-hospital.sh.cn (J.-Y. Wang), gaoxch@live.cn (X.-C. Gao).
Peer review under responsibility of Chinese Medical Association.

https://doi.org/10.1016/j.cdtm.2019.08.007
© 2019 Chinese Medical Association. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Conclusions: This paper has presented the first edition of the AGREE-CHINA appraisal tool for clinical guidelines. It is quick and easy to use; it assesses and performs well in comparison to AGREE II. This first version of AGREE-CHINA will require further development and validation.

© 2019 Chinese Medical Association. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Appraisal of guidelines research and evaluation; China; Development; Verification

In the past two decades, as physicians and patients have come to recognize the importance of clinical practice guidelines, the number of guidelines has increased year by year. From 1993 to 2016, a total of 664 guidelines were published in Chinese journals. However, most of them were based on expert consensus. There are few evidence-based guidelines based on systematic review. The average number of references in Chinese guidelines is 36, while that of international guidelines is 400. The issue of conflicts of interest is serious, and 88% of Chinese guidelines have no statement regarding conflicts of interest. Thus, the quality of Chinese guidelines is generally low. Moreover, there are multiple guidelines for the same problem, and the recommendations from different guidelines can differ, leaving physicians unsure which to follow. Therefore, it is essential to establish an evaluation system that aids physicians in selecting high-quality guidelines for clinical practice and guideline developers in standardizing their recommendations.

In 2003, the International Working Group on Clinical Guidelines Research and Evaluation, established by researchers from 13 countries including Canada and the United Kingdom, published the Appraisal of Guidelines Research and Evaluation in Europe (AGREE), which consisted of six major domains, 23 entries, and two overall evaluation entries. In 2009, to further improve the reliability and effectiveness of the AGREE tool and better meet users’ needs, the AGREE working group revised the first version and launched AGREE II, which had the same number of domains and entries and clearer, more specific content. At present, AGREE II has become the internationally recognized “gold standard” for guideline evaluation.

We evaluated the quality of 30 Chinese guidelines using AGREE II and found that most clinicians had some difficulty applying AGREE II for evaluation. First, the scores given by different reviewers varied greatly. Reasonably applying the AGREE tool proved quite demanding for reviewers. It not only requires the reviewers to have professional knowledge of the diseases but also systematic training and a deep understanding of the standardized development process of guidelines and the basic concepts of evidence-based medicine (such as the evidence grading system and evidence retrieval methods). Otherwise, reviewers may make very different decisions based on the scoring results. Second, a full mark of seven points is assigned to each entry in AGREE II, resulting in the same weight for all entries; in reality, each entry is of different importance. Third, some entries are not mentioned in the Chinese guidelines, such as audit tools. Fourth, it is time-consuming. It takes, on average, 50 minutes to evaluate a single guideline. To meet the current needs of Chinese guideline evaluation, it is necessary to develop an “appraisal of guidelines research and evaluation in China (AGREE-China)” with a substantial equivalence of AGREE II in its internationally recognized framework.

Development processes and methods

Establishment of a working group for developing guideline evaluation criteria

After the standardization department of the Medical Management Center of the National Health Commission of the People's Republic of China (from here on referred to as the Commission) approved the project, the criteria development process was officially launched. First, a working group for developing the guideline evaluation criteria was established. The members came from the personnel of the Evidence-based Medicine Center of Fudan University and included multidisciplinary experts from all over the country. The working group consisted of national policymakers, guideline methodologists, experts in clinical epidemiology and evidence-based medicine, clinicians, nursing experts, and journal editors for a total of 20 people.

Establishing the initial entry list

The group members studied the establishment of domestic and international clinical guidelines and the system of Grading of Recommendations Assessment,
Development, and Evaluation (GRADE). Then, they evaluated the published evaluation criteria and systems of clinical guidelines at home and abroad, including the content, advantages, and disadvantages of each evaluation system, and compared them to the current situation in China. Because AGREE II is currently an internationally accepted evaluation criteria, our study focused on analyzing its entries and application status. Also, two guidelines from each of the different research fields were evaluated using AGREE II to collect data on the application experience. Each group member proposed a written version of the initial evaluation entries. Then, a group meeting was held to select and form a unified, initial list of entries.

Developing the first draft of AGREE-China by vote, modification, preliminary trial, and cross-verification

Experts from various fields across the country, including clinicians of various disciplines and experts in clinical epidemiology, evidence-based medicine, information retrieval, and nursing, were invited to the expert consensus meetings to discuss, revise, and anonymously vote on the initial list of entries. Based on the opinions gathered, the working group revised the initial list of entries to form a first draft of the guideline evaluation criteria list that consisted of 15 entries.

Chinese clinical practice guidelines published in the past five years, which were representative of the disciplines of pediatrics, rheumatology, nursing, gastroenterology, cardiology, hematology, ophthalmology and oncology, were retrieved and selected for this study. Two independent reviewers from relevant professional fields with senior professional titles evaluated the selected guidelines using AGREE II and AGREE-China, respectively. The reviewers evaluated the advantages and disadvantages of the two criteria in the practical application of the different disciplines, spotted existing problems, and proposed suggestions for improvement. They cross-verified the criteria to evaluate the consistency (external consistency) between AGREE II and AGREE-China, the scoring consistency (internal consistency) between different experts when evaluating the same guideline using AGREE-China, and its repeatability.

Developing the revised draft of AGREE-China based on revisions of the first draft, according to a wide range of expert opinions and suggestions

The researchers consulted experts from various disciplines, including clinical epidemiology, evidence-based medicine, neurosurgery, general surgery, hepatobiliary surgery, imaging, nursing, gastroenterology, hematology, oncology, and pediatrics. A face-to-face discussion was held, opinions and suggestions were collected, and the first round of revisions was made to the first draft of AGREE-China. According to the list of external reviewers offered by the Medical Management Center of the National Health Commission of the People's Republic of China, a written review was conducted by more than 150 external individuals from different disciplines from the Chinese Medical Association, the Chinese Nursing Association, the Shanghai Medical Association, and the Shaanxi Provincial Medical Association.

During that period, the Evidence-based Medicine Center of Fudan University organized several expert consultations and seminars to summarize the external reviewers’ opinions and suggestions. They also formulated a 2017 version that would be in keeping with domestic clinical practice, and studied its evaluation criteria, the interpretation of the evaluation criteria, and the implementation instructions. Thus, the revised draft of AGREE-China was completed.

Developing the formal version of AGREE-China after a cross-review of the revised draft

The independent reviewers applied AGREE II and AGREE-China to evaluate the quality of the same guideline. They analyzed the correlation coefficient of the total scores of AGREE-China and AGREE II and the inter-reviewer reliability of AGREE-China to verify and determine AGREE-China's reliability as a guideline assessment tool. Then the Standing Committee of the Clinical Epidemiology and Evidence-Based Medicine Branch of the Chinese Medical Association voted independently on the revised manuscripts to form a formal draft. The expert group researched and developed the inclusion criteria for the national clinical practice guidelines. Finally, the Medical Management Center of the National Health Commission of the People's Republic of China organized experts to perform an acceptance inspection before the project's closure.

The flow chart of the development process for AGREE-China is shown in Fig. 1. After two rounds of independent expert voting, three expert consensus meetings, two trial evaluations of domestic guidelines, an external expert review, and ten periods of revisions, the formal list of entries, scores, and weights, and the specifications of the evaluation criteria were finalized.
Evaluation criteria and specifications of AGREE-China

The content and scoring criteria of AGREE-China

The guideline evaluation criteria include five domains (scientificity/preciseness, effectiveness/safety, economic efficiency, usability/feasibility, conflicts of interest), a total of 15 entries, and an overall evaluation, which is the “overall impression of the guideline: strongly recommendable, weakly recommendable, not recommendable” (Table 1). The criteria emphasize the domain of “scientificity/preciseness,” which consists of a total of eight entries. The score for each entry is assigned by the Likert rating scale method (0—5 points), with different weights depending on the importance of the entry. Both the total scores of different domains and the total score of the entire scale can be calculated. The higher the score is, the higher the quality is.

The specifications and instructions for the evaluation criteria are shown in Table S1 in the Supplementary Appendix. For example, Article 1, “The guideline development group consists of experts from relevant disciplines,” is divided into six situations. If experts from only one discipline are involved, zero points are assigned; one point is assigned when two to five disciplines are involved; two points are assigned when five or more disciplines are involved; three points are assigned when the guideline development group consists of experts from multiple disciplines; four points are assigned when the guideline development group includes methodologists. Furthermore, five points are assigned when the role and contribution of the methodologists are well-defined. The scoring criteria are straightforward.

Comparison between AGREE-China and AGREE II

AGREE II consists of six domains: the scope and purpose of the guideline, guideline developers, the preciseness of the guideline development, the clarity of expression, the applicability, and the independence of the guideline development. AGREE-China includes
five significant domains: scientificity/preciseness, effectiveness/safety, economic efficiency, usability/feasibility, and conflicts of interest. The scope and purpose of the guideline, guideline developers, and the preciseness of the guideline development in AGREE are combined into one, and some entries have been deleted. Considering that clinicians are more concerned with effectiveness and safety, those concerns are listed as a separate domain of evaluation. Some items, such as “The guideline retrieves and evaluates evidence from Chinese studies,” have been added to emphasize that Chinese guideline should take into account evidence from Chinese studies. The quality ratings based on the total scores of AGREE-China and AGREE II were consistent ($r = 0.508, P = 0.020$).

After two trials, it was found that AGREE-China was easier to use than AGREE II, and the difference between reviewers decreased. The quality scores from different reviewers using AGREE-China were highly consistent [intra-group correlation coefficient (ICC) = 0.957, $P < 0.001$]. The entries of AGREE-China showed greater ICCs than AGREE II. The evaluation time was shortened. When evaluating the same guideline, reviewers spent 45–60 minutes using AGREE II depending on their evaluation experience. However, it only took 30 minutes on average when they applied AGREE-China. When AGREE-China is used for evaluation, the quality ratings from experienced clinicians and inexperienced clinicians are highly consistent, and both can decide whether to recommend the concerned guideline. Meanwhile, using AGREE II is very time-consuming and requires well-trained personnel, which is not suitable for general clinicians.
Conclusion

AGREE-China has been modified based on the framework of AGREE II, including changing the scoring of each entry from a 7-point to a 5-point system, reducing 23 entries to 15, and deleting the entries that are not currently available in Chinese guidelines. The experts emphasized that Chinese guidelines should include evidence from Chinese studies. With detailed scoring criteria, AGREE-China is simpler and more efficient to apply, and it is suitable for domestic clinical practice.

Establishing China's guideline evaluation criteria is a significant achievement. It is suitable for China's current situation and is highly applicable. It provides a reference standard for the development of Chinese guidelines so that developers can understand what a good guideline would include. It provides inclusion criteria for establishing a Chinese guideline library in the future. In most Chinese clinical guidelines nowadays, the conflicts of interest for guideline developers and participation of patients are not fully considered. AGREE-China has assigned appropriate weights to these aspects, and it will pay more attention to them when it is used to develop and improve guidelines in the future.

It is the first attempt in China to establish evaluation criteria for Chinese guidelines and their interpretation. According to international standardized methods, the evaluation criteria were formulated accurately and transparently. The trial verification in different clinical disciplines also shows that the evaluation criteria are highly effective and reliable. However, there are still some shortcomings that need to be improved by further verification in future studies, such as determining which entries and quality dimensions are necessary when assessing the quality of a guideline and how to weigh entries more precisely. AGREE II is by far the most widely validated tool, and AGREE-China is not independent of AGREE II. It is expected that AGREE-China will be continuously revised and improved in future practice, and it is expected that it will be revised every one to two years.

Conflicts of interest

None declared.

Acknowledgment

This work was supported by the Entrusted Project of the Medical Management Center of the National Health and Family Planning Commission (2109901); Evidence-based public health and health economics of the fourth-round public health three-year action plan of Shanghai (15GWZK0901).

Sincere gratitude goes to the following organizations. The standardization department of the Medical Management Center of the National Health Commission of the People's Republic of China funded and convened the external review. The Evidence-based Medicine Center of Fudan University provided technical, personnel, and financial support. The Clinical Epidemiology and Evidence-based Medicine Branch of the Chinese Medical Association greatly supported the drafting and voting process in the development of the evaluation criteria.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cdtm.2019.08.007.

References

1. Chen Y, Wang C, Shang H, Yang K, Norris SL. Clinical practice guidelines in China. BMJ. 2018;360:j5158.
2. Brouwers MC, Kho ME, Browman GP, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. J Clin Epidemiol. 2010;63:1308–1311.
3. Brouwers MC, Kho ME, Browman GP, et al. The Global Rating Scale complements the AGREE II in advancing the quality of practice guidelines. J Clin Epidemiol. 2012;65:526–534.
4. Wang XQ, Wang JY. Development and Implementation of Evidence-Based Clinical Practice Guidelines. Beijing: People's Medical Publishing House; 2016:85–100 [in Chinese].
5. Jiang ZM, Zhan SY, Jia XW, et al. The basic methods and procedures for developing/revising the Clinical Practice Guideline. Natl Med J China. 2016;96:250–253 [in Chinese].
6. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924–926.

Edited by Pei-Fang Wei