ORIGINAL ARTICLE
Medicine General & Social Medicine

JKMS

http://dx.doi.org/10.3346/jkms.2014.29.6.771 • J Korean Med Sci 2014; 29: 771-775

Improving the Reliability of Clinical Practice Guideline Appraisals: Effects of the Korean AGREE II Scoring Guide

Moo-Kyung Oh,[1] Heuisug Jo,[1,2,3] and You Kyoung Lee[3,4,5]

[1]Department of Preventive Medicine, Kangwon National University Hospital, Chuncheon; [2]Department of Health Management and Policy, Kangwon National University School of Medicine, Chuncheon; [3]The Executive Committee for Clinical Practice Guideline, The Korean Academy of Medical Sciences, Seoul; [4]Department of Laboratory Medicine and Genetics, Soonchunhyang University Bucheon Hospital, Bucheon; [5]Department of Laboratory Medicine and Genetics, Soonchunhyang University College of Medicine, Cheonan, Korea

Received: 28 November 2013
Accepted: 28 March 2014

Address for Correspondence:
You Kyoung Lee, MD
Department of Laboratory Medicine and Genetics, Soonchunhyang University Bucheon Hospital, 170 Jomaru-ro, Wonmi-gu, Bucheon 420-767, Korea
Tel: +82.32-621-5941, Fax: +82.32-621-5944
E-mail: cecilia@schmc.ac.kr

This study was supported by a 2012 research grant for health policy (2012-0708-017) from the Ministry of Health and Welfare, Republic of Korea.

The Korean translated Appraisal of Guidelines for Research and Evaluation II (Korean AGREE II) instrument was distributed into Korean medical societies in 2011. However, inter-rater disagreement issues still exist. The Korean AGREE II scoring guide was therefore developed to reduce inter-rater differences. This study examines the effects of the Korean AGREE II scoring guide to reduce inter-rater differences. Appraisers were randomly assigned to two groups (Scoring Guide group and Non-Scoring Guide group). The Korean AGREE II instrument was provided to both groups. However, the scoring guide was offered to Scoring Guide group only. Total 14 appraisers were participated and each guideline was assessed by 8 appraisers. To evaluate the reliability of the Korean AGREE II scoring guide, correlation of scores among appraisers and domain-specific intra-class correlation (ICC) were compared. Most scores of two groups were comparable. Scoring Guide group showed higher reliability at all guidelines. They showed higher correlation among appraisers and higher ICC values at almost all domains. The scoring guide reduces the inter-rater disagreement and improves the overall reliability of the Korean-AGREE II instrument.

Keywords: Clinical Practice Guideline; Observer Variation; Reproducibility of Results

INTRODUCTION

Knowledge translation strategies for evidence-based clinical decision-making are embodied in Clinical practice guidelines (CPGs). In Korea, more than 100 CPGs have been developed in the last decade, and the types and development of CPGs are increasing (1-3). Nevertheless, there have been no discussions on the scientific methodology of guideline development or an appraisal for the developed CPGs. Moreover, the quality management of CPGs such as accreditation by prestigious institutions differs according to the organization developing the CPGs.

The purpose of an appraisal was to enhance the quality of information that CPGs provide to decision-makers and recommendations that have been developed using critically evaluated high-quality processes (4). For this purpose, in Europe and North America, where CPGs are actively applied, quality management of CPGs is performed using standardized tools or outlining the requirements developers must follow during the development process (5). In Korea, the Korean Academy of Medical Sciences (KAMS), the federation of professional societies of medical sciences in Korea, established a center for appraisal of clinical practice guideline in 2013 and took a major role in quality management of CPGs through the peer assessment by using the Appraisal of Guidelines for Research & Evaluation (AGREE) instrument .

AGREE instrument is a tool that assesses the methodological rigor and transparency by which a CPG is developed (6). The original AGREE instrument was developed in 2003 in collaboration with researchers from 13 countries. In 2009, AGREE II was produced, improving the reliability, validity, and performance of appraisal. In Korea, a translation of the original AGREE instrument was introduced in 2006. In 2010, AGREE II was translated into Korean and distributed to Korean medical societies. The usefulness of AGREE II have been verified through various quality assessment studies of specific diseases (8-11), international comparison of the level of guideline development (12, 13), and overall quality assessment of CPGs developed in specific countries (14-17).

However, differences in the developmental environments, health care systems, and medical cultures across the countries make it difficult to apply AGREE II uniformly to all assessments. Consequently, AGREE II is simply a comprehensive reference,

pISSN 1011-8934
eISSN 1598-6357

not any details of the proposed standards (7). In addition, Korean medical societies did not have enough experience developing CPGs and applying AGREE II. These limitations have led to obstacles appraising CPGs, including a lack of consensus among appraisers and a significant score variability. That is, there is an inter-rater disagreement.

Therefore, the Korean AGREE II Scoring Guide (hereafter, Scoring Guide) was developed to reduce these inter-rater differences (7). By reflecting the characteristics of the developmental environment of Korea, it provides detailed evaluation criteria for each item. It is expected to reduce the variation in scores among appraisers and presents a desirable target level for the development of CPGs.

This study examined the effects of the Scoring Guide, with an emphasis on reducing inter-rater differences and improving assessment reliability.

MATERIALS AND METHODS

AGREE II

AGREE II consists of 23 key items organized within six domains followed by two global rating items ("Overall assessment"). Each item is rated on a 7-point scale (1 = strongly disagree to 7 = strongly agree). Each domain captures a unique dimension of guideline quality. AGREE II recommends that each guideline be assessed by at least two appraisers, and preferably four, to increase the reliability of the assessment.

Scoring guide

The Scoring Guide was developed in accordance with Korean AGREE II. The first draft established requirements for anchor points 1, 3, 5, and 7 for each of the Korean AGREE II items and presented a specific checklist for each anchor point. Final agreement was derived through a modified *Delphi* consensus process. Thirteen specialists participated in the process and the modified *Delphi* was conducted twice.

Assessment and appraisal

Appraisers were randomly assigned to two groups (Scoring Guide group and Non-Scoring Guide group). The Korean AGREE II instrument was provided to both groups. However, the scoring guide was offered to Scoring Guide group only. Total 14 appraisers were participated and each guideline was assessed by 8 appraisers.

Clinical practice guidelines

Two CPGs that were officially submitted for appraisal to the Korean medical guideline information center (KoMGI) were appraised (Table 1). The KoMGI recognizes Korean AGREE II as the only official tool for appraisal, but the Scoring Guide was applied for this study with the consent of the developers.

Statistical analysis

A descriptive analysis of the distribution of domain scores according to the group was performed. To evaluate the reliability of the Korean AGREE II Scoring Guide, domain-specific intraclass correlations (ICCs) were calculated. The consistency of scores among appraisers was assessed by analyzing the correlation between the scores within the group. The statistical program SPSS for Windows 18.0 (SPSS, Chicago, IL, USA) was used. *P* values of 0.05 or less were considered significant.

RESULTS

Distribution of scores

Compared to the Non-Scoring Guide group, the distributions of the domain-specific scores in Scoring Guide group were characterized by low scores and low deviation for both CPGs. The differences were notable for domain 2 (Stakeholder involvement), domain 3 (Rigor of development), and domain 5 (Applicability) (Fig. 1). However, the difference was not statistically significant.

Reliability

Inter-rater reliability was analyzed by comparing the ICCs between the groups. The Scoring Guide groups had higher ICCs for both CPGs and this was common in all domains, except domain 4 (Clarity of presentation) and domain 6 (Editorial independence). Domain 3 and domain 5 were particularly significant. The overall scores were significantly higher in the Scoring Guide groups for both CPGs (Tables 2, 3).

Consistency

The Scoring Guide groups showed higher correlations among appraisers for both CPGs (Tables 4, 5).

DISCUSSION

Although quality management policy regarding CPG development varies widely across the countries, a strict quality assess-

Table 1. Clinical practice guidelines appraised in this study

Guideline	Publish year	Edition	Developer	Developmental method
Cancer pain management guideline	2012	5th	National Cancer Center	Adaptation
Guideline 2012 for Gastroesophageal reflux disease (GERD)	2012	3rd	The Korean Society of Neurogastroenterology and Motility	Adaptation

http://dx.doi.org/10.3346/jkms.2014.29.6.771

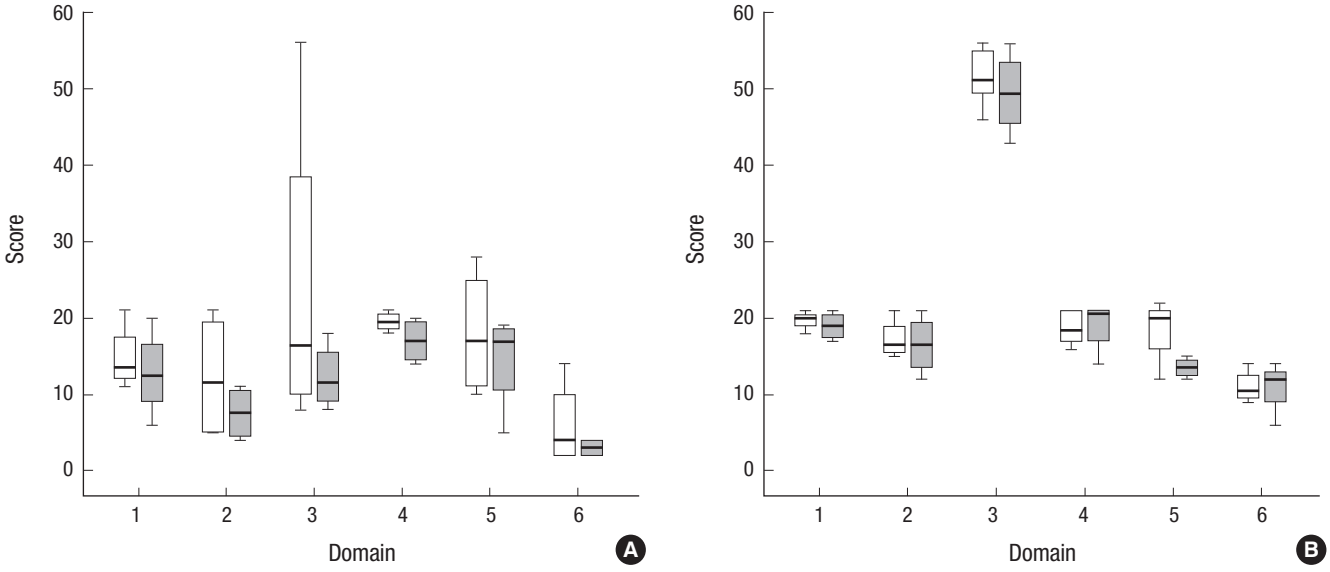

Fig. 1. Distribution of Korean AGREE II domain scores according to the use of Scoring Guide for Cancer pain management guideline (**A**) and GERD guideline (**B**). Black boxs are Scoring guide group and white boxs are Non-Scoring Guide group. The top and bottom of the box indicates the 75th (Q3) and 25th percentile (Q1), respectively, and the horizontal line in the box means the 50th percentile (the median). The upper and lower ends of the whisker represent Q3+1.5×(interquartile range), and Q1-1.5×(interquartile range), respectively.

Table 2. Inter-rater reliability of Korean AGREE II instrument domain scores for Cancer pain management guideline

Domain	Scoring guide group			Non-scoring guide group		
	ICC	(95% CI)	*P* value	ICC	(95% CI)	*P* value
Domain 1	0.815	(-0.344-0.995)	0.046	0.682	(-1.307-0.992)	0.116
Domain 2	0.430	(-3.137-0.986)	0.251	-0.762	(-11.791-0.955)	0.595
Domain 3	0.722	(0.175-0.938)	0.011	0.473	(-0.565-0.882)	0.121
Domain 4	0.718	(-1.048-0.993)	0.096	-0.296	(-8.411-0.967)	0.503
Domain 5	0.424	(-1.925-0.960)	0.229	0.273	(-2.693-0.950)	0.312
Domain 6	0.000	(-16.443-0.999)	0.391	0.000	(-16.443-0.999)	0.391
Overall	0.826	(0.671-0.918)	< 0.001	0.680	(0.395-0.850)	< 0.001

ICC, Intra-class correlation; CI, Confidence interval.

Table 3. Inter-rater reliability of Korean AGREE II instrument domain scores for GERD guideline

Domain	Scoring guide group			Non-scoring guide group		
	ICC	(95% CI)	*P* value	ICC	(95% CI)	*P* value
Domain 1	0.821	(-0.303-0.995)	0.043	-0.333	(-0.806-0.966)	0.512
Domain 2	0.769	(-0.675-0.994)	0.068	0.769	(-0.679-0.994)	0.069
Domain 3	0.796	(0.394-0.954)	0.002	0.424	(-0.710-0.871)	0.155
Domain 4	-1.333	(-15.940-0.941)	0.670	0.000	(-6.260-0.975)	0.422
Domain 5	0.888	(0.431-0.992)	0.005	0.272	(-2.696-0.950)	0.312
Domain 6	0.667	(-4.814-1.000)	0.182	0.792	(-2.634-1.000)	0.116
Overall	0.869	(0.753-0.939)	< 0.001	0.662	(0.362-0.841)	< 0.001

ICC, Intra-class correlation; CI, Confidence interval.

ment based on the AGREE II instrument is a common feature. This means that a thorough understanding and proper applications of AGREE II are essential for quality management of CPGs. In Korea, AGREE II was translated and distributed to medical societies in 2011. Nevertheless, inter-rater disagreement issues still exist. So, the Scoring Guide was developed to reduce inter-rater differences. This study examined the effects of the Scoring Guide on the reduction of inter-rater differences.

Most scores for the two groups were comparable, and the Scoring Guide groups showed higher reliability for both CPGs. They showed a stronger correlation among appraisers and higher ICC values for most domains, especially domain 2 (Stakeholder involvement), domain 3 (Rigor of development), and domain 5 (Applicability).

Table 4. Association among the appraisers according to use of Scoring Guide with Cancer pain management guideline

Group	Scoring guide group				Non-scoring guide group			
Appraiser	1	2	3	4	1	2	3	4
1	1	**0.622**	0.348	**0.481**	1	-0.225	**-0.434**	**-0.459**
2	**0.622**	1	0.393	**0.596**	-0.225	1	0.373	**0.502**
3	0.348	0.393	1	-0.052	**-0.434**	0.373	1	**0.833**
4	**0.481**	**0.596**	-0.052	1	**-0.459**	**0.502**	**0.833**	1

Boldface are statistically significant at the $P < 0.05$ level.

Table 5. Association among the appraisers according to use of Scoring Guide with GERD guideline

Group	Scoring guide group				Non-scoring guide group			
Appraiser	1	2	3	4	1	2	3	4
1	1	**0.853**	**0.453**	**0.491**	1	**0.556**	**0.441**	0.127
2	**0.853**	1	**0.651**	**0.749**	**0.556**	1	**0.479**	0.128
3	**0.453**	**0.651**	1	**0.641**	**0.441**	**0.479**	1	**0.372**
4	**0.491**	**0.749**	**0.641**	1	0.127	0.128	0.372	1

Boldface are statistically significant at the $P < 0.05$ level.

To better understand the results, the characteristics of the appraisers' environment, which affect their decisions, must be considered (11, 18-20). In the Korean healthcare environment, stakeholder involvement is an unfamiliar concept. Participation of the patients and citizen is seldom guaranteed, and even if they participated, their power and rights are generally weak (21, 22). As a result, only experts are recognized as stakeholders. This lack of stakeholder involvement experience in health policy and various definitions of stakeholder among appraisers are thought to have influenced the results. In this study, we found that the Scoring Guide reduced the gap in experience and understanding among appraisers by providing clear standards regarding the stakeholder and level of participation.

Another distinct domain is applicability. Applicability evaluates whether facilitators and barriers to its implementation and the potential resource implications of applying the recommendations have been considered. Strategies used to promote the implementation of CPGs are diverse, and the effect of application differs depending on the user's environment (23-25). Since there are few or no implementation strategies or efforts to promote the implementation level in Korea, differences in awareness and environments across appraisers are thought to affect inter-rater differences. In this respect, the Scoring Guide that provides specific criteria related to resources and methodology for measuring the level of implementation could complement the gaps in experience and awareness among appraisers.

This study suggests that the low reliability across appraisers arises from the effects of the healthcare environment and characteristics of the appraisers, rather than the validity of the Korean AGREE II instrument itself. Using these findings, we can find ways to overcome those limitations, and expand the use of evidence-based CPGs in Korea.

Low reliability was noticeable for low-quality CPG development. This means that enhancing CPG developmental competency is the first step in CPG quality management, to improve the reliability of appraisers. Therefore, priority should be placed on the development of high-quality CPGs, and developers should be provided with tools and programs for developing CPGs. Two guidebooks related to *de novo* and adaptation methods have been developed and disseminated, but few societies are aware of these manuals and active implementation strategies are insufficient (26). Guidebooks and programs to aid guideline development using scientific methodology and covering comprehensive developmental processes could be an effective strategy.

ACKNOWLEDGMENT

We thank for the National Cancer Center and the Korean Society of Neurogastroenterology and Motility for letting us appraise their CPGs.

DISCLOSURE

The authors have no conflicts of interest regarding the material presented here.

ORCID

Moo-Kyung Oh *http://orcid.org/0000-0002-2011-5708*
Heuisug Jo *http://orcid.org/0000-0003-0245-3583*
You Kyoung Lee *http://orcid.org/0000-0003-1835-2007*

REFERENCES

1. Shin YS, Kim YI. *Health policy and management. Seoul: Seoul National University Press, 2013.*
2. The Korean Society for Preventive Medicine. *Preventive medicine and public health. Seoul: Gyechuk munwhasa, 2010.*

 http://dx.doi.org/10.3346/jkms.2014.29.6.771

3. Korean Medical Guideline Information Center. *Available at http://www.guideline.or.kr/contents/index.php?code=015 [accessed on 9 July 2013].*
4. Darling G. *The impact of clinical practice guidelines and clinical trials on treatment decisions. Surg Oncol 2002; 11: 255-62.*
5. Legido-Quigley H, Panteli D, Brusamento S, Knai C, Saliba V, Turk E, Solé M, Augustin U, Car J, McKee M, et al. *Clinical guidelines in the European Union: mapping the regulatory basis, development, quality control, implementation and evaluation across member states. Health Policy 2012; 107: 146-56.*
6. AGREE Collaboration. *Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care 2003; 12: 18-23.*
7. Lee YK, Shin ES, Shim JY, Min KJ, Kim JM, Lee SH; the Executive Committee for CPGs; the Korean Academy of Medical Sciences. *Developing a scoring guide for the Appraisal of Guidelines for Research and Evaluation II instrument in Korea: a modified Delphi consensus process. J Korean Med Sci 2013; 28: 190-4.*
8. Sabharwal S, Patel V, Nijjer SS, Kirresh A, Darzi A, Chambers JC, Malik I, Kooner JS, Athanasiou T. *Guidelines in cardiac clinical practice: evaluation of their methodological quality using the AGREE II instrument. J R Soc Med 2013; 106: 315-22.*
9. Wijkstra J, Schubart CD, Nolen WA. *Treatment of unipolar psychotic depression: the use of evidence in practice guidelines. World J Biol Psychiatry 2009; 10: 409-15.*
10. Nagy E, Watine J, Bunting PS, Onody R, Oosterhuis WP, Rogic D, Sandberg S, Boda K, Horvath AR; IFCC Task Force on the Global Campaign for Diabetes Mellitus. *Do guidelines for the diagnosis and monitoring of diabetes mellitus fulfill the criteria of evidence-based guideline development? Clin Chem 2008; 54: 1872-82.*
11. Holmer HK, Ogden LA, Burda BU, Norris SL. *Quality of clinical practice guidelines for glycemic control in type 2 diabetes mellitus. PloS One 2013; 8: e58625.*
12. Van der Wees PJ, Hendriks EJ, Custers JW, Burgers JS, Dekker J, de Bie RA. *Comparison of international guideline programs to evaluate and update the Dutch program for clinical guideline development in physical therapy. BMC Health Serv Res 2007; 7: 191.*
13. Ansari S, Rashidian A. *Guidelines for guidelines: are they up to the task? a comparative assessment of clinical practice guideline development handbooks. PloS One 2012; 7: e49864.*
14. Jo MW, Lee JY, Kim NS, Kim SY, Sheen S, Kim SH, Lee SI. *Assessment of the quality of clinical practice guidelines in Korea using the AGREE Instrument. J Korean Med Sci 2013; 28: 357-65.*
15. Esandi ME, Ortiz Z, Chapman E, Dieguez MG, Mejía R, Bernztein R. *Production and quality of clinical practice guidelines in Argentina (1994-2004): a cross-sectional study. Implement Sci 2008; 3: 43.*
16. Tremblay MS, Warburton DE, Janssen I, Paterson DH, Latimer AE, Rhodes RE, Kho ME, Hicks A, LeBlanc AG, Zehr L, et al. *New Canadian physical activity guidelines. Appl Physiol Nutr Metab 2011; 36: 36-46.*
17. Rossignol M, Poitras S, Dionne C, Tousignant M, Truchon M, Arsenault B, Allard P, Coté M, Neveu A. *An interdisciplinary guideline development process: the Clinic on Low-back pain in Interdisciplinary Practice (CLIP) low-back pain guidelines. Implement Sci 2007; 2: 36.*
18. Yan J, Min J, Zhou B. *Diagnosis of pheochromocytoma: a clinical practice guideline appraisal using AGREE II instrument. J Eval Clin Pract 2013; 19: 626-32.*
19. MacDermid JC, Brooks D, Solway S, Switzer-McIntyre S, Brosseau L, Graham ID. *Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. BMC Health Serv Res 2005; 5: 18.*
20. Burgers JS, Cluzeau FA, Hanna SE, Hunt C, Grol R. *Characteristics of high-quality guidelines: evaluation of 86 clinical guidelines developed in ten European countries and Canada. Int J Technol Assess Health Care 2003; 19: 148-57.*
21. Lee WY. *Review on the patient and public involvement in health technology appraisals at NICE. J Crit Soc Welfare 2012; 34: 47-75.*
22. Kwon SM, You MS, Oh JH, Kim SJ, Jeon BY. *Public participation in healthcare decision making: experience of citizen council for health insurance. Korean J Health Policy Adm 2012; 22: 467-96.*
23. Kim YK, Lee SH, Seo JH, Kim JH, Kim SD, Kim GK. *A comprehensive model of factors affecting adoption of clinical practice guidelines in Korea. J Korean Med Sci 2010; 25: 1568-73.*
24. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, Robinson N. *Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006; 26: 13-24.*
25. Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L, Whitty P, Eccles MP, Matowe L, Shirran L, et al. *Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technol Assess 2004; 8: iii-iv, 1-72.*
26. Ahn HS, Kim HJ. *Development and implementation of clinical practice guidelines: current status in Korea. J Korean Med Sci 2012; 27: S55-60.*