Post-print of a Comparative Study on Quality Assessment and Recommendations of Guidelines for Degenerative Lumbar Spinal Stenosis

Authors: An Yi, Chen Hong, Zhou Yanji, Liu Luping, Chen Qianji, Lei Yuan, Sun Yanyan, Wang Xiyou, Liu Changxin, Zhang Yang, Yu Changhe

Date: 2022-09-05T00:00:00+00:00

Abstract

AGREE and RIGHT were employed to evaluate the methodological and reporting quality of guidelines for degenerative lumbar spinal stenosis (DLSS), providing a reference basis for the development and reporting of DLSS guidelines. Methods: Computerized searches were conducted in PubMed, CBM, CNKI, VIP, and Wanfang databases, with supplementary searches in the Medlive database, WHO, National Institute for Health and Care Excellence (NICE), Guidelines International Network (GIN), National Guideline Clearinghouse (NGC), and Scottish Intercollegiate Guidelines Network (SIGN) databases, with the search period from January 1, 2010 to January 1, 2022. Two reviewers independently assessed the methodological and reporting quality of included studies and compared recommendation grades after standardization. Results: A total of 6 relevant publications were included, comprising 4 guidelines and 2 consensuses; one guideline was a traditional Chinese medicine guideline, and three were evidence-based guidelines. AGREE evaluation showed that the ratio of actual total score to actual maximum score for the 6 included documents, in descending order, was 78.2%, 53.6%, 45.7%, 37.7%, 28.2%, and 15.9%. RIGHT evaluation showed that the ratio of actual total score to maximum possible score for the 4 included guidelines, in descending order, was 72.9%, 72.9%, 62.9%, and 34.3%. The 6 documents collectively formulated 46 treatment-related and 11 non-treatment-related recommendations. Conclusion: The methodological quality and reporting standards of current DLSS guidelines and consensuses need further improvement, with treatment recommendations tending toward surgical intervention for patients with moderate to severe DLSS.

Full Text

Quality Evaluation and Recommendations Analysis of Guidelines for Degenerative Lumbar Spinal Stenosis

Authors: An Yi[1], Chen Hong[2], Zhou Yanji[1], Liu Luping[1], Chen Qianji[1], Lei Yuan[2], Sun Yanyan[2], Wang Xiyou[2], Liu Changxin[2], Zhang Yang[2], *Yu Changhe*[2]

Affiliations: 1. The First Clinical Medical College, Beijing University of Chinese Medicine, Beijing 2. Dongzhimen Hospital, Beijing University of Chinese Medicine

Corresponding Author: Yu Changhe, MD, PhD. Research interests: methodology of traditional Chinese medicine efficacy evaluation. Email: yakno2@163.com, Phone: 010-84013296

Funding: - 2020 Capital Health Development Research Special Project: Standardization Research on Integrated Traditional Chinese Medicine Diagnosis and Treatment Protocol for Degenerative Lumbar Spinal Stenosis (First Issue 2020-4-4195) - National Natural Science Foundation Youth Project: Construction and Optimization of Core Outcome Sets for Traditional Chinese Medicine Clinical Research on Chronic Low Back Pain (81803956) - Beijing Municipal Science & Technology Commission Golden Bridge Engineering Seed Fund: Construction of Integrated Traditional Chinese Medicine Intervention Protocol for Degenerative Lumbar Spinal Stenosis Based on Evidence-Based Concepts (No. ZZ21053)

Abstract

Objective: To evaluate the methodological and reporting quality of guidelines for degenerative lumbar spinal stenosis (DLSS) using AGREE II and RIGHT, and to provide references for future DLSS guideline development and reporting.

Methods: We systematically searched PubMed, CBM, CNKI, VIP, and Wanfang databases, supplemented by searches of Medlive, WHO, NICE, GIN, NGC, and SIGN databases from January 1, 2010 to January 1, 2022. Two reviewers independently assessed the methodological and reporting quality of included studies and compared recommendations after standardizing their grading systems.

Results: Six relevant documents were included, comprising four guidelines and two consensus statements. One guideline was TCM-based, while three were evidence-based guidelines. AGREE II evaluation showed the ratio of actual total score to maximum possible score for the six documents ranged from 78.2% to 15.9% in descending order. RIGHT evaluation of the four guidelines showed total reporting rates from 72.9% to 34.3%. The documents generated 46 treatment-related and 11 non-treatment recommendations.

Conclusion: Current DLSS guidelines and consensus statements require further improvement in methodological quality and reporting standards. Recommendations for moderate-to-severe DLSS patients tend to favor surgical treatment.

Keywords: AGREE II; RIGHT; Guidelines; Consensus; Degenerative lumbar spinal stenosis

Introduction

Degenerative lumbar spinal stenosis (DLSS) is a degenerative disease characterized by leg pain, low back pain, and neurogenic intermittent claudication. In the United States, over 200,000 patients currently suffer from DLSS, making it a leading cause of pain and disability and the primary indication for spinal surgery in patients over 65 years old. Global data show that 102 million people (1.4%) are diagnosed with spinal stenosis annually. The Framingham study indicates that 19-47% of Americans over 60 have radiographic evidence of spinal stenosis, with prevalence expected to rise alongside aging populations and advances in diagnostic technology. High medical costs impose a substantial healthcare burden worldwide.

Clinical practice guidelines play an essential and positive role in medical care. However, only guidelines with sound methodological design and reporting standards can provide decision-making evidence that serves the interests of both patients and physicians, effectively guide clinical practice, improve patient experience, and reduce societal healthcare costs. The most widely used tool for evaluating guideline methodological quality is AGREE II (Appraisal of Guidelines for Research and Evaluation II), while the most common reporting standard is RIGHT (Reporting Items for Practice Guidelines in Healthcare). First published in 2003 and updated to AGREE II in 2008, this instrument has been extensively applied to guideline evaluation. Reporting quality of Chinese guidelines remains suboptimal, with insufficient attention from developers. Since RIGHT' s release in 2016, both guideline quality and their dissemination among target populations have improved. A good guideline must not only adhere to methodological standards during development but also ensure transparency and completeness through open reporting. This study employs AGREE II and RIGHT to evaluate DLSS guidelines, aiming to provide references for improving guideline quality and standardizing reporting.

Methods

1.1 Search Strategy

We searched CBM, CNKI, Wanfang Data, VIP, and PubMed databases, supplemented by guideline repositories including Medlive, WHO, NICE, GIN, NGC, and SIGN from January 1, 2010 to January 1, 2022. Chinese search terms included lumbar spinal stenosis, spinal stenosis, spinal degenerative disease, intermittent claudication, neurogenic claudication, guideline, consensus, and standard. English search terms included lumbar spinal stenosis, spinal stenosis, spinal osteophytosis, neurogenic claudication, guideline, consensus, and clinical practice guideline.

1.2 Inclusion and Exclusion Criteria

Inclusion criteria: (1) Clinical practice guidelines, consensus statements, or standards; (2) Target population: DLSS patients; (3) When multiple versions of the same guideline existed, only the most recent version was included.

Exclusion criteria: (1) Guideline interpretations or translations; (2) Guideline protocols, drafts, trial versions, abstracts, or meeting minutes; (3) Non-Chinese or non-English documents; (4) Documents lacking methodological support.

1.3 Literature Screening and Data Extraction

Two researchers (AY, LLP) independently screened literature and cross-checked results. Disagreements were resolved through consultation with YCH. When information was incomplete, corresponding authors were contacted for supplementary data. We designed a data extraction form based on the evaluation tools, collecting publication year, country, study population, evidence quality and recommendation grading methods, transparency issues, and other relevant information.

1.4 Quality Assessment

We used AGREE II and RIGHT tools to evaluate methodological and reporting quality respectively. AGREE II comprises 23 items across 6 domains, with each item scored from 1 to 7. The standardized domain score was calculated as: (actual score - minimum possible score) / (maximum possible score - minimum possible score). We also calculated the ratio of actual total score to maximum possible total score for overall methodological quality comparison.

RIGHT comprises 22 items across 7 domains, with each item qualitatively scored as "completely reported (Y)," "partially reported (P)," "not reported (N)," or "not applicable (I)." For items with multiple components, failure to report all required information was scored as "partially reported." We summarized reporting frequency and percentage by domain, assigning 2 points for "completely reported," 1 point for "partially reported," and 0 points for "not reported" or

"not applicable." The overall reporting rate was calculated as actual total score / maximum possible total score × 100%.

1.5 Recommendation Strength Grading

Due to inconsistent recommendation grading systems across guidelines, two reviewers (AY, LLP) independently classified interventions into four categories: strong recommendation, weak recommendation, uncertain, and not recommended based on textual descriptions and original grading systems. Strong and non-recommendations were classified according to explicit wording. Weak recommendations were assigned when benefits outweighed harms but evidence was insufficient for strong recommendation. Uncertain classifications were applied when evidence was insufficient to recommend an intervention or when benefit-harm balance was unclear. Disagreements were adjudicated by a third party (YCH).

1.6 Quality Control

The intraclass correlation coefficient (ICC) measures inter-rater reliability. We used ICC to test the reliability of AGREE II evaluations. For RIGHT assessments, disagreements between the two reviewers (AY, LLP) were resolved by a third party (YCH).

1.7 Statistical Analysis

We used Excel 2019 and SPSS 26.0 for data analysis and ICC testing of consistency between the two reviewers' scores. ICC > 0.80 indicated good inter-rater agreement.

Results

2.1 Literature Search Results

The search retrieved 1,402 documents. After deduplication using NoteExpress 3.2, 1,282 remained. Title and abstract screening left 22 documents, and full-text review ultimately included 6 documents [Figure 1: see original paper].

2.2 Basic Characteristics of Included Literature

Among the six included DLSS documents, four were guidelines and two were consensus statements. One was a TCM guideline, while the other five were modern medicine guidelines. Of the four guidelines, three were evidence-based. Three guidelines covered both diagnosis and treatment, while one addressed only treatment. The two consensus statements focused on diagnosis and diagnosis/treatment respectively. Three documents used GRADE (Grading of Recommendations Assessment, Development and Evaluation) for evidence quality

assessment, one used USPSTF (U.S. Preventive Services Task Force) recommendation grades, and two did not specify their grading method. Detailed baseline information is presented in Table 1 .

2.3 Quality Evaluation of Included Literature

AGREE II Evaluation Results ICC testing showed high consistency across domains (all ICC values > 0.92). AGREE II results indicated generally low methodological quality. The ratio of actual total score to maximum possible score for the six documents ranged from 78.2% to 15.9% in descending order. Across AGREE II domains, substantial variation existed between guidelines. Only "Scope and Purpose" and "Clarity of Presentation" domains exceeded 50% average scores (52.8% and 65.8% respectively). "Stakeholder Involvement," "Rigour of Development," "Applicability," and "Editorial Independence" domains all scored below 50%, with "Applicability" scoring lowest at 31.2%. The NASS guideline and Canadian Orthopaedic Association guideline scored highest across domains, with the latter achieving at least 66.7% in its lowest domain. Detailed AGREE II domain scores are shown in Table 2 .

RIGHT Evaluation Results Since many RIGHT items were not applicable to the two consensus statements, reporting quality ratings were conducted only for the four guidelines. RIGHT evaluation revealed suboptimal overall reporting quality for current DLSS guidelines. Total reporting rates for the four guidelines ranged from 72.9% to 34.3% in descending order, with the TCM-based guideline scoring lowest. Across RIGHT' s seven domains, "Other Aspects" (accessibility, recommendations for future research, and guideline limitations) showed best compliance with an average score of 79.1%. For evidence reporting, although the non-evidence-based guideline scored only 10%, the remaining three evidence-based guidelines achieved 100% compliance, yielding a 77.5% average. "Basic Information" and "Background" domains both exceeded 60% (73.2% and 66.1% respectively). However, "Review and Quality Assurance," "Recommendations," and "Funding and Conflicts of Interest" domains scored lowest at 50.0%, 44.7%, and 28.1% respectively. Domain-specific scores and averages are presented in Table 3 , with detailed reporting results in Table 4 .

2.4 Analysis of Clinical Questions and Recommendations

Among the six included guidelines/consensus statements, five presented clinical questions, recommendations, or consensus results, while one used an international Delphi method to establish consensus on seven specific DLSS questions. A total of 46 treatment-related and 11 non-treatment recommendations were generated.

The World Federation of Chinese Medicine Societies guideline proposed eight TCM treatment recommendations, including herbal medicine, external TCM therapies, and exercise regimens. The Danish national clinical guideline presented ten treatment-related clinical questions (seven on conservative therapy,

three on surgical therapy). The NASS guideline proposed 16 clinical questions (six on definition, disease course, diagnosis, and outcome measures; ten on treatment, with a 7:3 ratio of conservative to surgical questions). The Canadian Orthopaedic Association guideline included 12 non-surgical treatment recommendations for neurogenic claudication. The West Virginia Interventional Society consensus contained 11 items: five on DLSS concepts/diagnosis, four on surgical treatment, and two on non-surgical treatment. Specific therapy recommendations are detailed in Table 5 .

Discussion

3.1 Basic Characteristics of DLSS Guidelines/Consensus

The correlation between DLSS severity and clinical presentation is weak, as noted in included guidelines, though neurogenic intermittent claudication remains the characteristic clinical manifestation. The Canadian guideline focused clinical questions on addressing this primary symptom, while other documents provided recommendations on diagnosis, imaging, treatment, and outcome assessment. Based on AGREE II and RIGHT evaluations, guidelines demonstrated substantially higher quality and reporting standards than consensus statements, and evidence-based guidelines outperformed non-evidence-based ones. Modern medicine guidelines also surpassed TCM guidelines in quality and reporting. To improve clinical decision-making among Chinese healthcare providers, high-quality, evidence-based guidelines—particularly TCM guidelines—are urgently needed.

3.2 Methodological Quality Analysis of DLSS Guidelines

Chinese guidelines published in 2019 showed over 5% improvement in AGREE II and RIGHT scores compared to 2014-2018, yet significant gaps remain with international standards. Guideline evaluation drives developers toward stricter quality control and promotes advancement in Chinese medical guideline development. AGREE II assesses whether guideline development methods and content are reliable, whether processes meet standard requirements, and whether recommendations are based on current best evidence and warrant clinical promotion.

AGREE II results showed that the three evidence-based guidelines scored higher in total and domain scores, particularly in "Scope and Purpose," "Rigour of Development," and "Clarity of Presentation." This reflects the current trend toward evidence-based guideline development. The Institute of Medicine' s 2011 definition of clinical guidelines—as recommendations based on systematic review evidence evaluating benefits and harms of alternatives—emphasizes the evidence-based foundation. Evidence-based guidelines employ rigorous, systematic development processes superior to traditional expert consensus. However, scores were low (<40%) in "Stakeholder Involvement," "Applicability," and "Editorial Independence," indicating gaps in describing contributor expertise, collecting target

population preferences, identifying applicability groups, addressing implementation factors, and transparently reporting funding and conflicts. Only the TCM guideline considered patient preferences when formulating recommendations—concerning given that patient values represent one of evidence-based medicine' s three pillars. Most guidelines provided minimal information on promotion, application, and updates, with only two high-quality guidelines addressing these aspects.

3.3 Reporting Standards Analysis of DLSS Guidelines

RIGHT evaluation revealed that most items were not applicable to consensus statements, which focused extensively on consensus-building processes. Therefore, analysis was limited to the four guidelines. As the most widely used international reporting standard, RIGHT comprises seven domains: basic information, background, recommendations, evidence, review and quality assurance, funding and conflicts of interest, and other aspects.

Three guidelines achieved >60% total reporting rates, while the TCM guideline scored only 34.3%. The "Funding and Conflicts of Interest" and "Recommendations" domains scored poorest at 28.1% and 44.7% respectively. RIGHT requires transparent reporting of funding sources across all development stages, yet guidelines only briefly mentioned sponsors without detailing stage-specific funding usage. In the "Recommendations" domain, guidelines lost most points on rationale and explanation, as standards require consideration of user/target population preferences plus cost, resource utilization, equity, feasibility, and acceptability. Current DLSS guideline reporting falls short of international standards, particularly in item-specific descriptions. While length constraints may limit information, the highest score of 72.9% indicates room for improvement in reporting standards.

3.4 Recommendations Analysis of DLSS Guidelines

Five documents providing recommendations generated 57 total recommendations: 11 on definition, diagnosis, and outcome assessment; 46 on interventions. Three evidence-based guidelines provided recommendations based on systematic reviews or meta-analyses using GRADE for evidence quality and recommendation strength.

Thirty-six conservative intervention recommendations were identified. Epidural steroid injection received the most evidence, with moderate recommendation from NASS and West Virginia consensus. NASS limited this recommendation to within 24 months of onset, downgrading to uncertain after 36 months. Only the TCM guideline specified acupuncture and manipulation prescriptions; other guidelines gave weak or uncertain recommendations. The Canadian guideline limited acupuncture to early-stage disease and categorized manipulation as rehabilitation. Two guidelines gave weak recommendations for functional exercise, one was uncertain, and the Danish guideline described exercise as "preferred"

while noting insufficient evidence for neurogenic pain but gave weak recommendation for postoperative exercise. For pharmacological interventions, the TCM guideline weakly recommended herbal medicine. Western medications were generally not recommended, except for uncertain recommendation of neurotrophic drugs in NASS and weak recommendation for neurogenic pain analgesics in early-stage disease per the Danish guideline.

For surgical interventions, the Danish guideline weakly recommended laminectomy and fusion for patients with poor conservative treatment response. NASS weakly recommended surgery for moderate-to-severe patients. The West Virginia consensus strongly recommended percutaneous image-guided lumbar decompression using USPSTF criteria.

Given DLSS' s prolonged course and delayed conservative treatment effects, surgery offers rapid symptom relief with higher recommendation grades. However, neurological complications, high recurrence and reoperation rates, high costs, and low acceptability make conservative-first strategies preferable. NASS also proposes non-surgical treatment as the first strategy, making surgery a backup option aligned with patient preferences and current evidence. Greater emphasis on conservative treatments in guidelines is needed to help patients make optimal clinical decisions.

Conclusion

DLSS guidelines and consensus statements require improvement in methodological quality and reporting standards. Recommendations for moderate-to-severe DLSS tend toward surgical treatment. Few domestic DLSS guidelines exist. We urge developers to strictly follow AGREE II and RIGHT criteria to produce higher-quality, evidence-based guidelines that better serve clinicians and patients.

References

1. Jon L, Christy T. Management of lumbar spinal stenosis. BMJ. 2016;4(352):h6234.
2. Deyo RA, Mirza SK, Martin BI, et al. Trends, Major Medical Complications, and Charges Associated With Surgery for Lumbar Spinal Stenosis in Older Adults. JAMA. 2010;303(13):1259-1265.
3. Ravindra VM, Senglaub SS, Rattani A, et al. Degenerative Lumbar Spine Disease: Estimating Global Incidence and Worldwide Volume. Global Spine J. 2018;8(8):784-794.
4. Chad DA. Lumbar Spinal Stenosis. Neurologic Clinics. 2007;25(2):407-418.

5. Zhou Yanji, Liu Changxin, Liu Yangang, et al. Systematic Evaluation of the Effectiveness Research Trial for Lumbar Spinal Stenosis by the American Spine Patient Outcomes Research Trial. Chinese General Practice. 2022;25(05):535-541.
6. Yaolong C, Chen W, Hongcai S, et al. Clinical practice guidelines in China. BMJ. 2018;5(360):j5158.
7. Ma Y, Che G, Lian R, et al. Evaluation of Methodological and Reporting Quality of Guidelines for Henoch-Schönlein Purpura Using AGREE II and RIGHT. Journal of PLA Medicine. 2020;45(06):639-645.
8. K NT, Laura A, Ioana P, et al. Opioid prescribing: a systematic review and critical appraisal of guidelines for chronic pain. Annals of Internal Medicine. 2014;160(1):28-47.
9. Chen Yaolong, Yang Kehu. Correct Understanding, Development, and Application of Clinical Practice Guidelines. Peking Union Medical College Journal. 2018;9(04):367-373.
10. CBM, EKM, PBG, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ. 2010;182(18):E839-E842.
11. Liu Yunlan, Zhang Jingyi, Shi Qianling, et al. Investigation and Evaluation of Chinese Clinical Practice Guidelines Published in 2019—Methodological and Reporting Quality. Peking Union Medical College Journal. 13(02):324-331.
12. Yaolong C, Kehu Y, Ana M, et al. A Reporting Tool for Practice Guidelines in Health Care: The RIGHT Statement. Annals of Internal Medicine. 2017;166(2):128-132.
13. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. PLoS Medicine. 2010;7(2):e1000217.
14. Zhenwei X, Xiaoling W, Lin S, et al. Appraisal of clinical practice guidelines on community-acquired pneumonia in children with AGREE II instrument. BMC Pediatrics. 2016;Aug 2(16):119.
15. Lu Shuya, Zhao Siya, Wu Shouyuan, et al. Investigation and Evaluation of Chinese Clinical Practice Guidelines Published in 2019—Evidence Quality and Recommendation Strength. Peking Union Medical College Journal. 2022;13(01):130-137.
16. Ng JY, Mohiuddin U. Quality of complementary and alternative medicine recommendations in low back pain guidelines: a systematic review. Eur Spine J. 2020;29(8):1833-1844.
17. Zhu Qianqing, Zeng Manjie. International Clinical Practice Guideline for Traditional Chinese Medicine—Degenerative Lumbar Spinal Stenosis (2019-10-10). World Chinese Medicine. 2021;16(16):2371-2376.
18. Rikke R, Krüger JR, Søren F, et al. Danish national clinical guidelines for surgical and nonsurgical treatment of patients with lumbar spinal stenosis. European Spine Journal. 2019;28(6):1386-1396.
19. Kreiner DS, Shaffer WO, Baisden JL, et al. An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update). The Spine Journal. 2013;13(7):734-743.
20. André B, Carolina C, Carlo A, et al. Non-Surgical Interventions for Lum-

bar Spinal Stenosis Leading To Neurogenic Claudication: A Clinical Practice Guideline. The Journal of Pain. 2021;22(9):1015-1039.
21. Tomkins-Lane C, Melloh M, Lurie J, et al. Consensus on the Clinical Diagnosis of Lumbar Spinal Stenosis: Results of an International Delphi Study. SPINE. 2016;41(15):1239-1246.
22. RDT, SGJ, EPJ, et al. The MIST Guidelines: The Lumbar Spinal Stenosis Consensus Group Guidelines for Minimally Invasive Spine Treatment. Pain Practice. 2018;19(3):250-274.
23. Qi Z, Zijun W, Qianling S, et al. Clinical Epidemiology in China series. Paper 4: The reporting and methodological quality of Chinese clinical practice guidelines published between 2014 and 2018: A Systematic Review. Journal of Clinical Epidemiology. 2021;140:189-199.
24. Wang Q. Quality Evaluation of Chinese Clinical Practice Guidelines. Lanzhou University; 2017.
25. Liu Ming, Yang Jie, Wang Yiping. New Thoughts on Evidence-Based Guideline Development Methods and Clinical Application. Chinese Journal of Evidence-Based Medicine. 2009;9(02):127-128.
26. RG, MM, DMW, et al. Clinical Practice Guidelines We Can Trust. National Academies Press (US); 2011.
27. Xie Xuewan, Yang Wendeng. Review of Evidence-Based Treatment Guidelines for Attention Deficit Hyperactivity Disorder. Chinese Journal of Clinical Psychology. 2021;29(03):661-664.

Note: Figure translations are in progress. See original paper for figures.