PLOS ONE

RESEARCH ARTICLE

Guideline appraisal with AGREE II: Systematic review of the current evidence on how users handle the 2 overall assessments

Wiebke Hoffmann-Eßer[1,2]*, Ulrich Siering[1], Edmund A. M. Neugebauer[3], Anne Catharina Brockhaus[1], Ulrike Lampert[1], Michaela Eikermann[4]

1 Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany, **2** Institute for Research in Operative Medicine (IFOM), University of Witten/Herdecke, Campus Cologne, Cologne, Germany, **3** Brandenburg Medical School – Theodor Fontane Neuruppin, Germany & University of Witten/Herdecke, Witten/Herdecke, Germany, **4** Medical Advisory Service of the German Social Health Insurance (MDS), Essen, Germany

* wiebke.hoffmann-esser@iqwig.de

Citation: Hoffmann-Eßer W, Siering U, Neugebauer EAM, Brockhaus AC, Lampert U, Eikermann M (2017) Guideline appraisal with AGREE II: Systematic review of the current evidence on how users handle the 2 overall assessments. PLoS ONE 12(3): e0174831. https://doi.org/10.1371/journal.pone.0174831

Editor: Hong-Liang Zhang, National Natural Science Foundation of China, CHINA

Received: December 12, 2016

Accepted: March 15, 2017

Published: March 30, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Non-monetary support for this research was provided by the Institute for Quality and Efficiency in Health Care (IQWiG). No external funding was received.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Introduction

The Appraisal of Guidelines for Research & Evaluation (AGREE) II instrument is the most commonly used guideline appraisal tool. It includes 23 appraisal criteria (items) organized within 6 domains and 2 overall assessments (1. overall guideline quality; 2. recommendation for use). The aim of this systematic review was twofold. Firstly, to investigate how often AGREE II users conduct the 2 overall assessments. Secondly, to investigate the influence of the 6 domain scores on each of the 2 overall assessments.

Materials and methods

A systematic bibliographic search was conducted for publications reporting guideline appraisals with AGREE II. The impact of the 6 domain scores on the overall assessment of guideline quality was examined using a multiple linear regression model. Their impact on the recommendation for use (possible answers: "yes", "yes, with modifications", "no") was examined using a multinomial regression model.

Results

118 relevant publications including 1453 guidelines were identified. 77.1% of the publications reported results for at least one overall assessment, but only 32.2% reported results for both overall assessments. The results of the regression analyses showed a statistically significant influence of all domains on overall guideline quality, with Domain 3 (rigour of development) having the strongest influence. For the recommendation for use, the results showed a significant influence of Domains 3 to 5 ("yes" vs. "no") and Domains 3 and 5 ("yes, with modifications" vs. "no").

Conclusions

The 2 overall assessments of AGREE II are underreported by guideline assessors. Domains 3 and 5 have the strongest influence on the results of the 2 overall assessments, while the other domains have a varying influence. Within a normative approach, our findings could be used as guidance for weighting individual domains in AGREE II to make the overall assessments more objective. Alternatively, a stronger content analysis of the individual domains could clarify their importance in terms of guideline quality. Moreover, AGREE II should require users to transparently present how they conducted the assessments.

Introduction

According to the definition of the US Institute of Medicine, "clinical practice guidelines are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options." [1].

Various studies have shown that guidelines can improve health care [2–14]. However, their quality is variable and therefore their recommendations are often inconsistent [15–24].

In order to be able to use guidelines as a reliable basis for decision-making, their quality, i.e. their methodological rigour and transparency, needs to be ensured. Guideline appraisal tools are applied for this purpose. Forty such tools covering varying dimensions of guideline quality were identified in a systematic review published in 2013 [25], of which 6 contain a quantitative assessment of overall guideline quality.

In 2003, an international group of guideline developers and researchers developed the Appraisal of Guidelines for Research & Evaluation (AGREE) instrument [15]. The revised version, AGREE II [26], was published in 2009 and is currently the most commonly applied and comprehensively validated guideline appraisal tool worldwide [17–19]. It consists of 23 appraisal criteria (items) organized into 6 domains (Table 1), each of which "captures a unique dimension of guideline quality" [16]. The items within each domain are rated on a 7-point scale ("strongly disagree" to "strongly agree").

In addition, AGREE II includes 2 global rating items (overall assessments). In the first overall assessment, the overall guideline quality is rated on a 7-point scale ("lowest possible quality" to "highest possible quality"). In the second overall assessment, a recommendation is provided on whether to use the guideline in practice or not (recommendation for use: "yes", "yes with modifications", "no"). Both assessments should consider the 23 items evaluated beforehand and the resulting domain scores, but should not be calculated from them.

It has not yet been investigated in the literature how often AGREE II users conduct the 2 overall assessments. For this reason, it is unclear whether these assessments actually represent separate assessments (as specified by AGREE II) or whether users simply calculate the overall scores directly from the domain scores.

On the basis of recent publications on guideline appraisals, the aim of this systematic review was twofold. Firstly, to investigate how AGREE II users handle the 2 overall assessments, that is, how often they conduct them. Secondly, to investigate the influence of the 6 domain scores on each of the 2 overall assessments (1. overall guideline quality; 2. recommendation for use).

Table 1. Items and domains of the AGREE II instrument[a].

Item	Content	Domain
1	The overall objective(s) of the guideline is (are) specifically described.	Scope and Purpose
2	The health question(s) covered by the guideline is (are) specifically described.	
3	The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described.	
4	The guideline development group includes individuals from all relevant professional groups.	Stakeholder Involvement
5	The views and preferences of the target population (patients, public, etc.) have been sought.	
6	The target users of the guideline are clearly defined.	
7	Systematic methods were used to search for evidence.	Rigour of Development
8	The criteria for selecting the evidence are clearly described.	
9	The strengths and limitations of the body of evidence are clearly described.	
10	The methods for formulating the recommendations are clearly described.	
11	The health benefits, side effects, and risks have been considered in formulating the recommendations.	
12	There is an explicit link between the recommendations and the supporting evidence.	
13	The guideline has been externally reviewed by experts prior to its publication.	
14	A procedure for updating the guideline is provided.	
15	The recommendations are specific and unambiguous.	Clarity of Presentation
16	The different options for management of the condition or health issue are clearly presented.	
17	Key recommendations are easily identifiable.	
18	The guideline describes facilitators and barriers to its application.	Applicability
19	The guideline provides advice and/or tools on how the recommendations can be put into practice.	
20	The potential resource implications of applying the recommendations have been considered.	
21	The guideline presents monitoring and/or auditing criteria.	
22	The views of the funding body have not influenced the content of the guideline.	Editorial Independence
23	Competing interests of guideline development group members have been recorded and addressed.	

[a]: Extracted from the AGREE II instrument

https://doi.org/10.1371/journal.pone.0174831.t001

Materials and methods

A systematic search for relevant (primary and secondary) publications was conducted in MEDLINE, EMBASE, the Database of Abstracts of Reviews of Effects (Other Reviews), and the Health Technology Assessment Database (Technology Assessments).

Amongst others, the following search terms were used: "practice guidelines as topic", "AGREE instrument" and "methodological guideline appraisal". The full list of search terms is included in the search strategy (see supporting information S1 File), which was developed by an information specialist. The search was conducted in January 2016.

German- and English-language publications reporting results of at least one guideline appraisal with AGREE II were considered. These results had to include all 6 standardized domain scores of each guideline appraised.

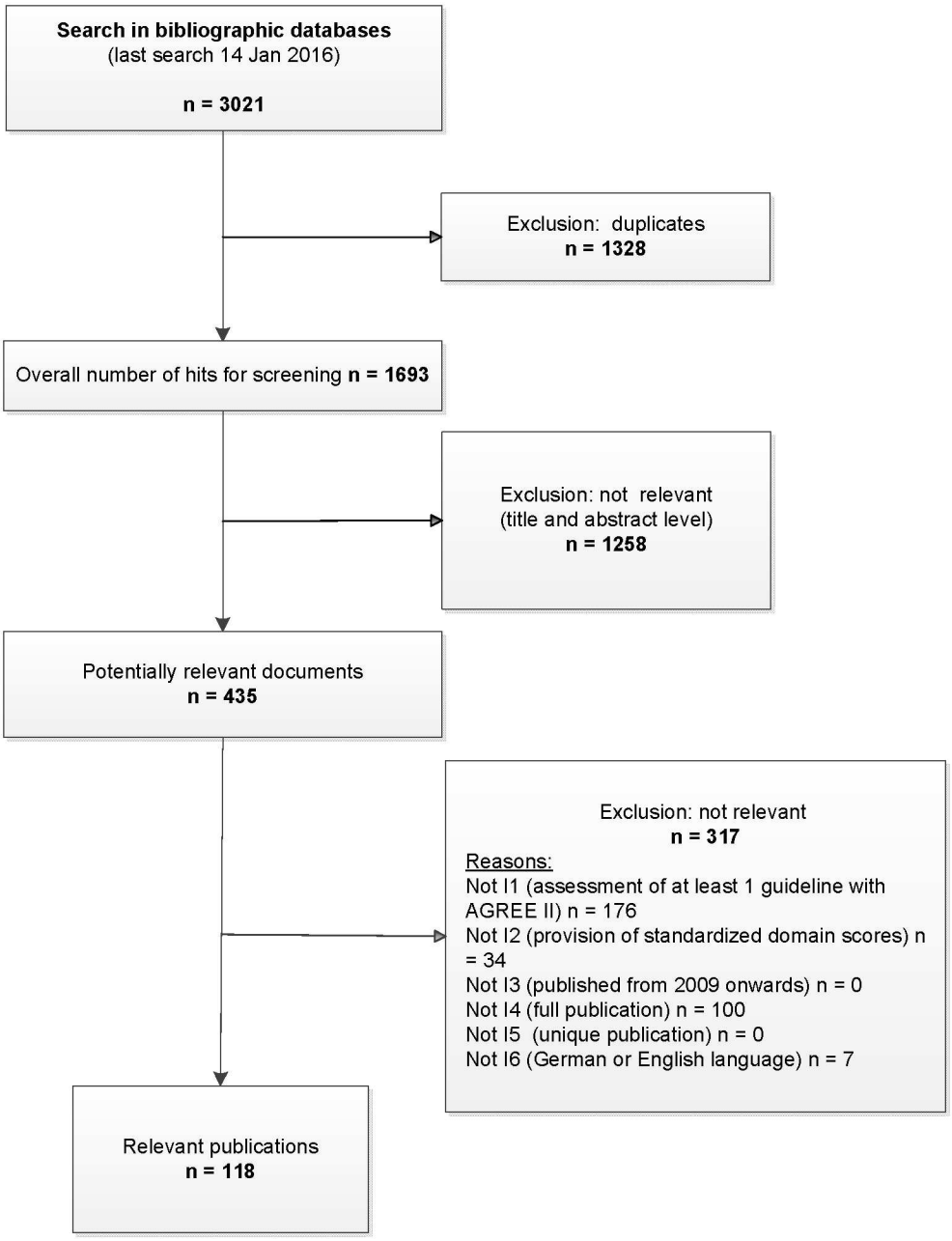

Fig 1. Results of the systematic literature search.

https://doi.org/10.1371/journal.pone.0174831.g001

The screening of titles and abstracts and subsequently of full texts was performed by 2 authors independently of one another. 96 discrepancies in the screening of title and abstracts and 128 discrepancies in the screening of full texts were resolved by discussion between both authors (see Fig 1).

Data extraction and analysis

The results of the AGREE II appraisals (standardized domain scores, and, if available, results of the overall assessments) were extracted from the publications included. In addition, the main

characteristics of these publications were extracted, namely, the aim of the publication, the number of assessors, the number of guidelines appraised with AGREE II, the publication dates of the guidelines included in the relevant publications, as well as the guideline topics (S4 File).

Further information was also extracted from the publications (S5 File). This referred to whether overall assessment 1 (overall guideline quality) and/or overall assessment 2 (recommendations for use) had been conducted or not. If yes, it was also examined whether the requirements of AGREE II had been followed.

Data extraction and analysis were performed by one reviewer and checked by another. Any discrepancies were resolved by discussion between them. It was then checked how often the overall assessments had been performed in the guideline appraisals.

The impact of the 6 standardized domain scores (independent variables) on the overall assessment of guideline quality (dependent variable) was examined using a multiple linear regression model. Guideline appraisals were excluded from the multiple linear regression analysis if a standardized domain score was not available for all 6 domains. Similarly, guidelines were excluded whose overall guideline quality had been calculated from the standardized domain scores using the mean values, as this approach is not recommended by AGREE II. The inclusion of such guideline appraisals could have biased our results concerning the influence of the 6 domains on the 2 overall assessments, as this influence would have been determined by calculation, not by evaluation.

In a second analysis, the impact of the 6 standardized domain scores (independent variables) on the recommendation for use (dependent variable) was examined using a multinomial regression model. Guideline appraisals were excluded from the multinomial regression if they did not contain data on standardized domain scores for all 6 domains.

It is possible to receive inconsistent information on the recommendations for use due to independent evaluations by several assessors (e.g. both "yes, with modifications" and "no" or both "yes" and "yes, with modifications"). In these cases, the recommendation for use was allocated to the category "yes, with modifications". In addition, guideline appraisals were excluded from the analysis if no allocation of the recommendation for use to one of the 3 categories ("yes", "yes, with modifications", "no") was meaningful. This could be the case if inconsistent recommendations for use were provided for the same guideline, such as both "yes" and "no", or all 3 categories ("yes", "yes, with modifications", "no").

Due to the multiple comparisons performed, we also present adjusted p-values for each regression analysis according to Benjamini and Hochberg [27] to control for the false discovery rate and maintain an overall significance level of 5%. The decision on whether a domain had a significant influence on the overall assessments or not was based on this adjusted p-value. The data were analysed with SPSS Statistics 18 and SAS 9.3.

Results

Selection of relevant publications

The systematic search in bibliographic databases identified a total of 3021 publications, of which 435 were screened in full text; 118 fulfilled the inclusion criteria (Fig 1). The supporting information contains the list of publications included (S2 File) and excluded (S3 File), with the reasons for exclusion, as well as the main characteristics of the guidelines appraised in the publications (S4 File).

Results for the first research question

Conduct of overall assessments. 91 (77.1%) of the 118 eligible publications reported results for at least one overall assessment of which 38 (32.2%) reported both overall

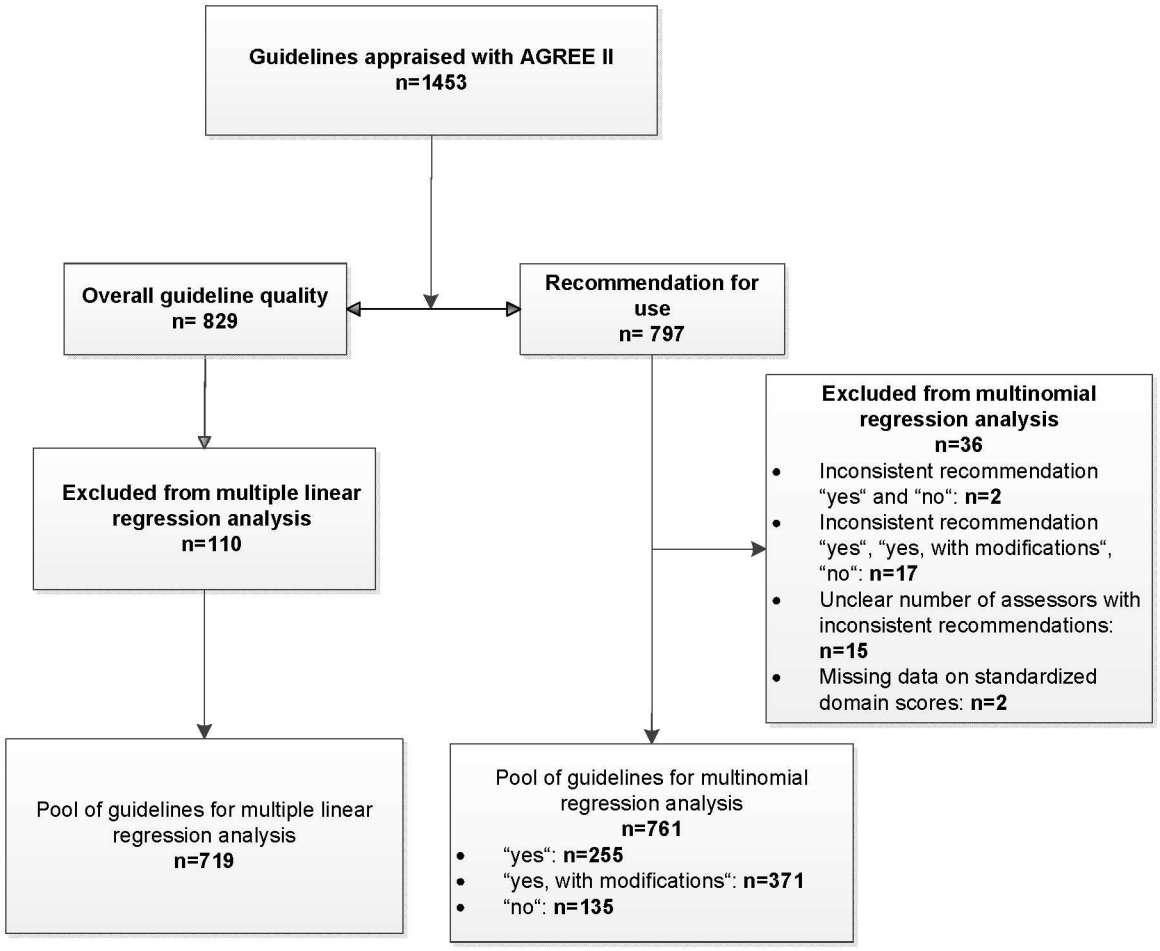

Fig 2. Guideline pool for the multiple linear and multinomial regression analyses.

https://doi.org/10.1371/journal.pone.0174831.g002

assessments, 32 (27.1%) reported only overall assessment 1 (overall guideline quality), and 21 (17.8%) reported only overall assessment 2 (recommendation for use); see S5 File. The 91 publications included 1453 guidelines appraised with AGREE II (Fig 2).

70 publications (38 + 32) therefore included at least one result on the assessment of overall guideline quality, while 59 publications (38 + 21) included at least one result on the assessment of the recommendation for use.

Overall assessment 1 (overall guideline quality). The overall guideline quality had been assessed for 829 (57.1%) of the 1453 guidelines.

In 10 (14.3%) of the 70 publications reporting overall guideline quality, the authors apparently calculated the overall score from the mean scores of the 6 standardized domain scores [28–37]; see S5 File. The data from these 10 publications, which contained 110 guidelines, were not considered in the multiple regression analysis.

719 (49.5%) guidelines thus formed the total pool for the analysis of the association between standardized domain scores and overall guideline quality (Fig 2).

Overall assessment 2 (recommendation for use). A recommendation for use was provided by the assessors for 797 (54.9%) of the 1453 guidelines. All guideline appraisals (n = 797) were performed by between 2 and 11 assessors independently of one another; different recommendations for use were therefore provided for the same guideline (e.g. both "yes, with

modifications" and "no" or both "yes" and "yes, with modifications"). In such cases (n = 53), the assessment was allocated to the category "yes, with modifications".

In addition, further inconsistent information on the recommendations for use was provided for the same guideline by the different assessors: both "yes" and "no" (n = 2) as well as all 3 categories ("yes", "yes, with modifications", "no"; n = 17). Moreover, in one publication the number of assessors was not clear for the guidelines with inconsistent recommendations (n = 15); these results could not be allocated to any of the 3 categories above and were thus not included in the multinomial regression analysis. Likewise, 2 guideline appraisals were excluded, since they did not contain data on standardized domain scores for all 6 domains. A total of 36 (4.5%) guidelines were thus excluded from the multinomial regression analysis (Fig 2).

Overall, consistent recommendations for use were provided for 708 (88.8%) of the 797 guidelines with a recommendation for use. Ultimately, 761 (52.4%) guidelines formed the pool for the multinomial regression analysis: 255 (33.5%), 371 (48.8%), and 135 (17.7%) were allocated to the categories "yes", "yes, with modifications", and "no" respectively (Fig 2).

Results for the second research question

Evaluation of model assumptions and correlations between independent variables: Multiple regression analysis. No major violations were shown in the evaluation of the model assumptions of the multiple linear regression analysis (S6 File). Weak ($r < 0.5$) to moderate ($0.5 \leq r < 0.8$) correlations were shown for all pairs of independent variables considered (domains 1 to 6).

Evaluation of model assumptions and correlations between independent variables: Multinomial regression analysis. Weak ($r < 0.5$) to moderate ($0.5 \leq r < 0.8$) correlations were shown for all pairs of independent variables considered in the multinomial regression analysis; S7 File.

Influence of the 6 domains on the results of overall assessment 1 (overall guideline quality). All domains had a statistically significant influence (adjusted p-value < 0.05) on overall guideline quality (Table 2). Domain 3 had the strongest influence (ß = 0.300; adjusted p-value < 0.001), followed by Domain 4 (ß = 0.203; adjusted p-value < 0.001) and Domain 1 (ß = 0.175; adjusted p-value < 0.001), as well as Domain 5 (ß = 0.163; adjusted p-value < 0.001), Domain 6 (ß = 0.065; adjusted p-value < 0.001), and Domain 2 (ß = 0.062; adjusted p-value = 0.018).

Table 2. Results of the multiple regression analysis (independent variable: Overall guideline quality).

Predictors	Unstandardized coefficients		95% confidence interval for B		t	P-value	Adjusted P-value (sig. < 0.05)
	B	Standard error	Lower bound	Upper bound			
Intercept	5.591	1.753			3.19	0.001	
Domain 1 (scope and purpose)	.175	0.026	.125	.226	6.784	< 0.001	< 0.001
Domain 2 (stakeholder involvement)	.062	0.026	.011	.114	2.381	0.018	0.018
Domain 3 (rigour of development)	.300	0.025	.250	.350	11.796	< 0.001	< 0.001
Domain 4 (clarity of presentation)	.203	0.027	.150	.255	7.583	< 0.001	< 0.001
Domain 5 (applicability)	.163	0.021	.123	.204	7.913	< 0.001	< 0.001
Domain 6 (editorial independence)	.065	0.017	.032	.099	3.841	< 0.001	< 0.001

Dependent variable: overall guideline quality; adjusted R2: 0.732

https://doi.org/10.1371/journal.pone.0174831.t002

Table 3. Results of the multinomial regression analysis (independent variable: Recommendation for use for the categories "yes" vs. "no").

Parameter	Estimate	Standard error	Wald chi-square	P-value	Adjusted p-value (sig. < 0.05)	OR[a]	95% confidence interval for OR[a]	
							Lower bound	Upper bound
Intercept (recommended)	-9.744	0.856	129.729	< 0.001				
Domain 1 (scope and purpose)	0.013	0.009	2.059	0.151	0.227	1.140	0.954	1.367
Domain 2 (stakeholder involvement)	0.013	0.010	1.603	0.206	0.247	1.135	0.933	1.381
Domain 3 (rigour of development)	0.109	0.011	93.824	< 0.001	< 0.001	2.963	2.395	3.719
Domain 4 (clarity of presentation)	0.046	0.010	20.521	< 0.001	< 0.001	1.581	1.301	1.934
Domain 5 (applicability)	0.022	0.009	6.026	0.014	0.028	1.250	1.048	1.498
Domain 6 (editorial independence)	0.003	0.006	0.200	0.657	0.657	1.029	0.909	1.166

[a]: The OR corresponds to the change in the respective domain score by 10 percentage points.

Dependent variable: recommendation for use; Reference category: "no".

https://doi.org/10.1371/journal.pone.0174831.t003

Influence of the 6 domains on the results of overall assessment 2 (recommendation for use). According to AGREE II, there are 3 categories for the recommendation for use ("yes", "yes, with modifications", "no"). For the multinomial regression analysis, the category "no" of the recommendation for use was chosen as the reference category, resulting in 2 comparisons: "yes" vs. "no" and "yes, with modifications" vs. "no".

The comparison of the categories "yes" and "no" showed that Domains 3 to 5 had a significant influence (adjusted p-value < 0.05) on whether the use of a guideline was recommended (Table 3). Domain 3 had the strongest influence (ß = 0.109; adjusted p-value < 0.001) followed by Domains 4 (ß = 0.046; adjusted p-value < 0.001), and 5 (ß = 0.022; adjusted p-value = 0.028). With an increase in the standardized domain score of Domain 3 by 10 percentage points, there was almost a 3-fold increase in the ratio of guidelines recommended for use versus those not recommended (if all other variables remained unchanged). With the same increase in Domains 4 and 5, there was almost a 1.6 and 1.3-fold increase in this ratio, respectively.

The comparison of the categories "yes, with modifications" and "no" showed that Domains 3 and 5 had a significant influence on whether the use of a guideline was recommended with modifications (Table 4). Domain 3 had the strongest influence (ß = 0.061; adjusted p-value $<$ 0.001), followed by Domain 5 (ß = 0.022; adjusted p-value = 0.019). With an increase in the standardized domain score of Domain 3 by 10 percentage points, there was about a 1.8-fold increase in the ratio of guidelines recommended for use with modifications versus those not recommended (if all other variables remained unchanged). With the same increase in Domain 5, there was a 1.2-fold increase in this ratio.

Discussion

Main findings

The aim of this systematic review was twofold. Firstly, to investigate how AGREE II users handle the 2 overall assessments (1. overall guideline quality, 2. recommendation for use), that is, how often they conduct them. Secondly, to investigate the influence of the 6 domain scores on each of the 2 overall assessments.

Table 4. Results of the multinomial regression analysis (independent variable: Recommendation for use for the categories; "yes, with modifications" vs. "no").

Parameter	Estimate	Standard error	Wald chi-square	P-value	Adjusted p-value (sig. < 0.05)	OR[a]	95% confidence interval for OR[a]	
							Lower bound	Upper bound
Intercept (recommended, with modifications)	-3.224	0.472	46.584	< 0.001				
Domain 1 (scope and purpose)	0.014	0.006	4.765	0.029	0.058	1.146	1.014	1.297
Domain 2 (stakeholder involvement)	0.005	0.008	0.303	0.582	0.699	1.047	0.889	1.233
Domain 3 (rigour of development)	0.061	0.009	43.945	< 0.001	< 0.001	1.843	1.549	2.226
Domain 4 (clarity of presentation)	0.012	0.007	3.438	0.064	0.096	1.132	0.994	1.293
Domain 5 (applicability)	0.022	0.008	7.497	0.006	0.019	1.246	1.068	1.465
Domain 6 (editorial independence)	0.000	0.005	0.009	0.926	0.926	1.005	0.908	1.114

[a]: The OR corresponds to the change in the respective domain score by 10 percentage points.

Dependent variable: recommendation for use; Reference category: "no".

https://doi.org/10.1371/journal.pone.0174831.t004

Even though the assessment of overall guideline quality and the recommendation for use are standard components of AGREE II, they are underreported: 77.1% of the eligible publications reported results for at least one overall assessment, but only 32.2% reported results for both overall assessments.

Regarding the influence of domains, both regression analyses showed that Domain 3 (rigour of development) had the strongest influence on the 2 overall assessments. Furthermore, all analyses showed a statistically significant influence of Domain 5 (applicability) on both overall assessments. For Domain 4 (clarity of presentation), the results were statistically significant for the multiple linear regression analysis (overall guideline quality), as well as for part of the multinomial regression analysis (recommendation for use: "yes" vs. "no"); in both of these analyses this domain showed the second strongest influence.

Relation to other studies

The strong influence of Domain 3 on the 2 overall assessments is not surprising, as previous research suggests that this domain is a stronger indicator of guideline quality than the other domains [16, 38], a high score indicating minimum bias and evidence-based guideline development [38]. On the other hand, a low score indicates serious methodological problems, for instance, a lack of methodological expertise in guideline developing teams or an inadequate systematic search due to a lack of resources [16].

The results for Domain 5 can be explained by what Gagliardi et al. note in their systematic review of guideline applicability. The items of Domain 5 refer to facilitators and barriers of guideline implementation, monitoring or audit criteria, and implementation instructions / tools. According to Gagliardi et al., for the latter there is evidence of association with guideline use: If a guideline is insufficiently implemented in clinical practice because of inadequate implementation instructions, this "contributes to omission of beneficial therapies, preventable harm, suboptimal patient outcomes or experiences, or waste of resources" [39].

However, the items in Domain 5 need to be distinguished from the specific applicability of a guideline in clinical practice (e.g. in the care of a certain patient population); a guideline may not be applicable to a specific context, but may still receive a high score in Domain 5 [40].

The strong influence of Domain 4 is not surprising either, as "[t]he main advantage of a well-reported guideline is that flaws in the methodology are more easily detected, so that inherent biases can be considered more explicitly and scrutinized by the potential users" [41].

Calculation of overall guideline quality

In contrast to the requirements of AGREE II, in 10 publications the overall assessment 1 (overall guideline quality) was calculated as the mean of the 6 standardized domain scores [28–37]; it is thus highly likely that no separate assessment of overall guideline quality independent of the domain scores was performed. According to the AGREE II requirements the assessment of the overall quality of a guideline should be subjective. This is not possible if the assessment is based on a calculation, as each of the 6 domains would have a similar influence on the overall quality. Such an approach would thus not have answered our second research question. As previously stated, this approach is not recommended by AGREE II and these 10 publications were excluded from further analysis.

Strengths and limitations

Language restrictions. The search for relevant publications was limited to German and English-language publications. Since AGREE II is an internationally recognized and validated instrument that has been translated into several languages not considered here, potentially relevant publications in other languages were not taken into account in our analysis, which may have led to language bias.

Choice of regression model. The data on the overall assessment 2 (recommendation for use) were ordinally scaled. Initially we attempted to adjust an ordinally scaled regression model to answer our objective. However, an evaluation of the model assumptions showed that the assumption of proportional odds was not fulfilled; this model was thus inappropriate and could have led to misleading results [42].

Independent variables. The adjusted determination coefficient (R2) for the multiple linear regression analysis shows that 73.2% of the variance in the overall guideline quality can be explained by the independent variables (standardized domain scores of the 6 domains) considered in the analysis (Table 2 and S6 File). The determination coefficient (R2) in the multinomial regression analysis was 58.2% (S7 File).

Besides the standardized domain scores, no further independent variables were considered in the 2 regression models. Further research would need to investigate to what extent further guideline characteristics affect the results of the regression models (e.g. guideline topic, country of origin, publisher, publication date, profession and level of experience of AGREE II users, and whether or not a consensus procedure was conducted in the event of deviating appraisals). In addition, factors affecting the internal validity of a guideline (e.g. consistency of recommendations) that go beyond the methodological aspects of an AGREE appraisal could play a role.

Sample size for regression analyses. According to Schneider et al. 2010, at least 20 observations should be available for each independent variable [43]. The multiple linear regression model included 6 independent variables (domains 1 to 6); therefore the minimum sample size of 6x20 = 120 observations (guidelines) had to be available in the multiple linear regression analysis. This analysis was based on 719 guidelines appraised with AGREE II. An insufficient sample size could falsely create strong associations between variables. However, the above estimate applies only to the multiple linear regression analysis; it is not directly applicable to the multinomial regression analysis.

Correlation of the independent variables in the regression analyses. The correlation of the independent variables should be considered in the overall interpretation of the results of the regression analyses. A statistically non-significant result for an independent variable (standardized domain score) does not necessarily mean a lack of association with the dependent variable (overall guideline quality or recommendation for use). If 2 independent variables correlate with each other, this can lead to a situation where one of these variables does not contribute additional information to the regression analysis [43].

Strengths and limitations of AGREE II. Due to the wide range of domain items, the AGREE II instrument offers the opportunity to systematically, specifically and objectively evaluate the quality of guidelines from all specialties [44]. However, as stated above and also noted by several other researchers [44–49] AGREE II lacks detailed information on how to perform the 2 overall assessments. In addition, several researchers emphasize that these assessments are subjective [31, 44, 48, 50, 51]; some regard it as a weakness of AGREE II that items or domains are not weighted, but are all considered equally [45–47, 50, 52]. Before appraising a guideline with AGREE II, Lytras et al. thus propose to weight domain items [47].

The results of an AGREE II appraisal should be viewed with caution, as different guideline assessors may interpret the items and scoring system differently [53].

Due to the problems described above, some researchers criticize that AGREE II allows no clear distinction between high- and low-quality guidelines [33, 47, 48, 54]. Several researchers use cut-offs to distinguish between high and low quality [32, 44, 50, 55–60]. This shows that AGREE II users would welcome such a clear distinction, but the instrument currently does not fulfil this requirement.

Implications for further research

Besides the standardized domain scores, no further independent variables were considered in the 2 regression models. Future research would need to investigate to what extent additional guideline characteristics potentially affect the results of the regression models (e.g. guideline topic, country of origin, publisher, publication date, profession and level of experience of AGREE II users, and whether or not a consensus procedure was conducted in the event of deviating appraisals). In addition, factors affecting the internal validity of a guideline (e.g. consistency of recommendations) that go beyond the methodological aspects of an AGREE appraisal might play a role.

Conclusion

The 2 overall assessments of the AGREE II instrument are underreported by guideline assessors. Domains 3 and 5 have the strongest influence on the results of the 2 overall assessments, while the other domains have a varying influence.

As a normative approach, the results of our study could be used as guidance for weighting individual domains in AGREE II, an approach already proposed by authors of guideline appraisals. Consequently, the 2 overall assessments would be performed in a more objective manner. Alternatively, a stronger content analysis of the individual domains or their items could be carried out to clarify their importance in terms of the quality of a guideline.

In addition, AGREE II should require users to transparently present how they performed the 2 overall assessments. This particularly refers to the recommendation for use; guideline assessors should explain on which criteria their recommendation is based, allowing readers to form their own judgement on whether they would have provided the same recommendation in the same healthcare setting.

Supporting information

S1 File. Search strategy.
(PDF)

S2 File. Publications included.
(PDF)

S3 File. Publications excluded (organized by reasons for exclusion).
(PDF)

S4 File. Characteristics of guidelines included in the publications.
(PDF)

S5 File. Information on the conduct of the overall assessments according to AGREE II.
(PDF)

S6 File. Statistics for the multiple regression analysis.
(PDF)

S7 File. Assessment of model quality of the multinomial regression analysis.
(PDF)

S8 File. PRISMA-checklist.
(PDF)

Acknowledgments

We thank Verena Wekemann for checking the format of the citations and Natalie McGauran for medical writing support.

Author Contributions

Conceptualization: WHE ME EN US.

Data curation: WHE US ACB UL.

Formal analysis: WHE ACB.

Methodology: WHE ACB EN ME.

Project administration: WHE.

Resources: WHE ACB ME.

Supervision: EN ME.

Visualization: WHE ACB.

Writing – original draft: WHE US EN ACB UL ME.

References

1. Graham RM, Mancher M, Miller-Wolman D, Greenfield S, Steinberg E, editors. Clinical practice guidelines we can trust. Washington: National Academies Press; 2011.
2. Baker R, Camosso-Stefinovic J, Gillies C, Shaw EJ, Cheater F, Flottorp S, et al. Tailored interventions to address determinants of practice. Cochrane Database Syst Rev. 2015;(4):CD005470. https://doi.org/10.1002/14651858.CD005470.pub3 PMID: 25923419
3. Baskerville NB, Liddy C, Hogg W. Systematic review and meta-analysis of practice facilitation within primary care settings. Ann Fam Med. 2012; 10(1):63–74. http://dx.doi.org/10.1370/afm.1312.

4. Brusamento S, Legido-Quigley H, Panteli D, Turk E, Knai C, Saliba V, et al. Assessing the effectiveness of strategies to implement clinical guidelines for the management of chronic diseases at primary care level in EU Member States: a systematic review. Health Policy. 2012; 107(2–3):168–83. http://dx.doi.org/10.1016/j.healthpol.2012.08.005.
5. Flodgren G, Parmelli E, Doumit G, Gattellari M, O'Brien MA, Grimshaw J, et al. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2011;(8): CD000125. http://dx.doi.org/10.1002/14651858.CD000125.pub4.
6. Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technol Assess. 2004; 8(6):iii–iv, 1–72. PMID: 14960256
7. Hakkennes S, Dodd K. Guideline implementation in allied health professions: a systematic review of the literature. Qual Saf Health Care. 2008; 17(4):296–300. https://doi.org/10.1136/qshc.2007.023804 PMID: 18678729
8. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2012;(6): CD000259. https://doi.org/10.1002/14651858.CD000259.pub3 PMID: 22696318
9. Medves J, Godfrey C, Turner C, Paterson M, Harrison M, MacKenzie L, et al. Systematic review of practice guideline dissemination and implementation strategies for healthcare teams and team-based practice. Int J Evid Based Healthc. 2010; 8(2):79–89. https://doi.org/10.1111/j.1744-1609.2010.00166.x PMID: 20923511
10. Okelo SO, Butz AM, Sharma R, Diette GB, Pitts SI, King TM, et al. Interventions to modify health care provider adherence to asthma guidelines 2013 [updated 05.2013; cited 2016 22.02.2016]. 95:[Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/372/1493/asthma-provider-adherence-report-130626.pdf.
11. Ray-Coquard I, Philip T, De Laroche G, Froger X, Suchaud JP, Voloch A, et al. A controlled "before-after" study: impact of a clinical guidelines programme and regional cancer network organization on medical practice. Br J Cancer. 2002; 86(3):313–21. https://doi.org/10.1038/sj.bjc.6600057 PMID: 11875690
12. Ray-Coquard I, Philip T, Lehmann M, Fervers B, Farsi F, Chauvin F. Impact of a clinical guidelines program for breast and colon cancer in a French cancer center. JAMA. 1997; 278(19):1591–5. PMID: 9370505
13. Rotter T, Kinsman L, James E, Machotta A, Gothe H, Willis J, et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database Syst Rev. 2010;(3):CD006632. http://dx.doi.org/10.1002/14651858.CD006632.pub2.
14. Smith TJ, Hillner BE. Ensuring quality cancer care by the use of clinical practice guidelines and critical pathways. J Clin Oncol. 2001; 19(11):2886–97. https://doi.org/10.1200/JCO.2001.19.11.2886 PMID: 11387362
15. Abdelsattar ZM, Reames BN, Regenbogen SE, Hendren S, Wong SL. Critical evaluation of the scientific content in clinical practice guidelines. Cancer. 2015; 121(5):783–9. https://doi.org/10.1002/cncr.29124 PMID: 25376967
16. Alonso-Coello P, Irfan A, Sola I, Gich I, Delgado-Noguera M, Rigau D, et al. The quality of clinical practice guidelines over the last two decades: a systematic review of guideline appraisal studies. Qual Saf Health Care. 2010; 19(6):e58. https://doi.org/10.1136/qshc.2010.042077 PMID: 21127089
17. Altman RD, Schemitsch E, Bedi A. Assessment of clinical practice guideline methodology for the treatment of knee osteoarthritis with intra-articular hyaluronic acid. Semin Arthritis Rheum. 2015; 45(2):132–9. https://doi.org/10.1016/j.semarthrit.2015.04.013 PMID: 26142320
18. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet. 2000; 355(9198):103–6. https://doi.org/10.1016/S0140-6736(99)02171-6 PMID: 10675167
19. Guyatt G, Vandvik PO. Creating clinical practice guidelines: problems and solutions. Chest. 2013; 144 (2):365–7. https://doi.org/10.1378/chest.13-0463 PMID: 23918097
20. Koh C, Zhao X, Samala N, Sakiani S, Liang TJ, Talwalkar JA. AASLD clinical practice guidelines: a critical review of scientific evidence and evolving recommendations. Hepatology. 2013; 58(6):2142–52. https://doi.org/10.1002/hep.26578 PMID: 23775835
21. Kryworuchko J, Stacey D, Bai N, Graham ID. Twelve years of clinical practice guideline development, dissemination and evaluation in Canada (1994 to 2005). Implement Sci. 2009; 4:49. https://doi.org/10.1186/1748-5908-4-49 PMID: 19656384
22. Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. Arch Intern Med. 2012; 172(21):1628–33. https://doi.org/10.1001/2013.jamainternmed.56 PMID: 23089902

23. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA. 1999; 281 (20):1900–5. PMID: 10349893

24. van den Berg J, Berger MY. Guidelines on acute gastroenteritis in children: a critical appraisal of their quality and applicability in primary care. BMC Fam Pract. 2011; 12:134. https://doi.org/10.1186/1471-2296-12-134 PMID: 22136388

25. Siering U, Hoffmann-Eßer W, Neugebauer EA, Eikermann M. Is there a cut-off for high-quality guidelines? A systematic analysis of current guideline appraisals using the AGREE-II instrument [poster]. G-I-N Conference; 7th - 10th October 2015; Amsterdam, The Netherlands.

26. AGREE Next Steps Consortium. Appraisal of guidelines for research and evaluation II: AGREE II instrument [Internet]. 2013 Sep [cited 2015 Nov 10]. http://www.agreetrust.org/wp-content/uploads/2013/10/AGREE-II-Users-Manual-and-23-item-Instrument_2009_UPDATE_2013.pdf.

27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57(1):289–300.

28. Burnett HF, Tanoshima R, Chandranipapongse W, Madadi P, Ito S, Ungar WJ. Testing for thiopurine methyltransferase status for safe and effective thiopurine administration: a systematic review of clinical guidance documents. Pharmacogenomics J. 2014; 14(6):493–502. https://doi.org/10.1038/tpj.2014.47 PMID: 25156214

29. Haran C, Van Driel M, Mitchell BL, Brodribb WE. Clinical guidelines for postpartum women and infants in primary care: a systematic review. BMC Pregnancy Childbirth. 2014; 14:51. https://doi.org/10.1186/1471-2393-14-51 PMID: 24475888

30. Rios E, Seron P, Lanas F, Bonfill X, Quigley EMM, Alonso-Coello P. Evaluation of the quality of clinical practice guidelines for the management of esophageal or gastric variceal bleeding. Eur J Gastroenterol Hepatol. 2014; 26(4):422–31. https://doi.org/10.1097/MEG.0000000000000033 PMID: 24535595

31. Sabharwal S, Patel NK, Gauher S, Holloway I, Athanasiou T. High methodologic quality but poor applicability: assessment of the AAOS guidelines using the AGREE II instrument. Clin Orthop. 2014; 472 (6):1982–8. https://doi.org/10.1007/s11999-014-3530-0 PMID: 24566890

32. Schoenmaker NJ, Tromp WF, Van der Lee JH, Offringa M, Craig JC, Groothoff JW. Quality and consistency of clinical practice guidelines for the management of children on chronic dialysis. Nephrology Dialysis Transplantation. 2013; 28(12):3052–61.

33. White PE, Shee AW, Finch CF. Independent appraiser assessment of the quality, methodological rigour and transparency of the development of the 2008 international consensus statement on concussion in sport. Br J Sports Med. 2014; 48(2):130–4. https://doi.org/10.1136/bjsports-2013-092720 PMID: 24128756

34. Chua ME, Mendoza J, See M, Esmena E, Aguila D, Silangcruz JM, et al. A critical review of recent clinical practice guidelines on the diagnosis and treatment of non-neurogenic male lower urinary tract symptoms. Can Urol Assoc J. 2015; 9(7–8):E463–E70. https://doi.org/10.5489/cuaj.2424 PMID: 26279717

35. Haddadi M, Muhammadnejad S, Sadeghi-Fazel F, Zandieh Z, Rahimi G, Sadighi S, et al. Systematic review of available guidelines on fertility preservation of young patients with breast cancer. Asian Pac J Cancer Prev. 2015; 16(3):1057–62. PMID: 25735331

36. Tunnicliffe DJ, Singh-Grewal D, Kim S, Craig JC, Tong A. Diagnosis, monitoring, and treatment of systemic lupus erythematosus: a systematic review of clinical practice guidelines. Arthritis Care Res. 2015; 67(10):1440–52.

37. Yaman ME, Gudeloglu A, Senturk S, Yaman ND, Tolunay T, Ozturk Y, et al. A critical appraisal of the North American Spine Society guidelines with the Appraisal of Guidelines for Research and Evaluation II instrument. Spine J. 2015; 15(4):777–81. https://doi.org/10.1016/j.spinee.2015.01.012 PMID: 25614152

38. Brosseau L, Rahman P, Poitras S, Toupin-April K, Paterson G, Smith C, et al. A systematic critical appraisal of non-pharmacological management of rheumatoid arthritis with Appraisal of Guidelines for Research and Evaluation II. PLoS One. 2014; 9(5):e95369. https://doi.org/10.1371/journal.pone.0095369 PMID: 24840205

39. Gagliardi AR, Brouwers MC. Do guidelines offer implementation advice to target users? A systematic review of guideline applicability. BMJ Open. 2015; 5(2):e007047. http://dx.doi.org/10.1136/bmjopen-2014-007047.

40. Damiani G, Silvestrini G, Trozzi L, Maci D, Iodice L, Ricciardi W. Quality of dementia clinical guidelines and relevance to the care of older people with comorbidity: evidence from the literature. Clin Interv Aging. 2014; 9:1399–407. Epub 2014/08/30. https://doi.org/10.2147/CIA.S65046 PMID: 25170263

41. Fervers B, Burgers JS, Haugh MC, Brouwers M, Browman G, Cluzeau F, et al. Predictors of high quality clinical practice guidelines: examples in oncology. Int J Qual Health Care. 2005; 17(2):123–32. https://doi.org/10.1093/intqhc/mzi011 PMID: 15665068

42. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond. 1997; 31(5):546–51. PMID: 9429194

43. Schneider A, Hommel G, Blettner M. Lineare Regressionsanalyse: Teil 14 der Serie zur Bewertung wissenschaftlicher Publikationen. Dtsch Arztebl. 2010; 107(44):776–82.

44. Lee GY, Yamada J, Kyololo OB, Shorkey A, Stevens B. Pediatric clinical practice guidelines for acute procedural pain: a systematic review. Pediatrics. 2014; 133(3):500–15. https://doi.org/10.1542/peds.2013-2744 PMID: 24488733

45. Burda BU, Chambers AR, Johnson JC. Appraisal of guidelines developed by the World Health Organization. Public Health. 2014; 128(5):444–74. https://doi.org/10.1016/j.puhe.2014.01.002 PMID: 24856197

46. 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(4):e58625. https://doi.org/10.1371/journal.pone.0058625 PMID: 23577058

47. Lytras T, Bonovas S, Chronis C, Konstantinidis AK, Kopsachilis F, Papamichail DP, et al. Occupational asthma guidelines: a systematic quality appraisal using the AGREE II instrument. Occup Environ Med. 2014; 71(2):81–6. https://doi.org/10.1136/oemed-2013-101656 PMID: 24213564

48. Polus S, Lerberg P, Vogel J, Watananirun K, Souza JP, Mathai M, et al. Appraisal of WHO guidelines in maternal health using the AGREE II assessment tool. PLoS One. 2012; 7(8):e38891. https://doi.org/10.1371/journal.pone.0038891 PMID: 22912662

49. Ye ZK, Li C, Zhai SD. Guidelines for therapeutic drug monitoring of vancomycin: a systematic review. PLoS One. 2014; 9(6):e99044. https://doi.org/10.1371/journal.pone.0099044 PMID: 24932495

50. Brosseau L, Rahman P, Toupin-April K, Poitras S, King J, De Angelis G, et al. A systematic critical appraisal for non-pharmacological management of osteoarthritis using the Appraisal of Guidelines Research and Evaluation II instrument. PLoS One. 2014; 9(1):e82986. https://doi.org/10.1371/journal.pone.0082986 PMID: 24427268

51. Sabharwal S, Gauher S, Kyriacou S, Patel V, Holloway I, Athanasiou T. Quality assessment of guidelines on thromboprophylaxis in orthopaedic surgery. Bone Joint J. 2014; 96-B(1):19–23. https://doi.org/10.1302/0301-620X.96B1.32943 PMID: 24395305

52. Bragge P, Pattuwage L, Marshall S, Pitt V, Piccenna L, Stergiou-Kita M, et al. Quality of guidelines for cognitive rehabilitation following traumatic brain injury. J Head Trauma Rehabil. 2014; 29(4):277–89. https://doi.org/10.1097/HTR.0000000000000066 PMID: 24984092

53. Marciano NJ, Merlin TL, Bessen T, Street JM. To what extent are current guidelines for cutaneous melanoma follow up based on scientific evidence? Int J Clin Pract. 2014; 68(6):761–70. https://doi.org/10.1111/ijcp.12393 PMID: 24548269

54. Tudor KI, Kozina PN, Marusic A. Methodological rigour and transparency of clinical practice guidelines developed by neurology professional societies in Croatia. PLoS One. 2013; 8(7):e69877. https://doi.org/10.1371/journal.pone.0069877 PMID: 23894555

55. Gutarra-Vilchez RB, Barajas-Nava L, Aleman A, Sola I, Gich I, Bonfill X, et al. Systematic evaluation of the quality of clinical practice guidelines on the use of assisted reproductive techniques. Hum Fertil. 2014; 17(1):28–36.

56. Lopez-Vargas PA, Tong A, Sureshkumar P, Johnson DW, Craig JC. Prevention, detection and management of early chronic kidney disease: a systematic review of clinical practice guidelines. Nephrology. 2013; 18(9):592–604. https://doi.org/10.1111/nep.12119 PMID: 23815515

57. Santos F, Sola I, Rigau D, Arevalo-Rodriguez I, Seron P, Alonso-Coello P, et al. Quality assessment of clinical practice guidelines for the prescription of antidepressant drugs during pregnancy. Curr Clin Pharmacol. 2012; 7(1):7–14. PMID: 22299765

58. Shen J, Sun M, Zhou B, Yan J. Nonconformity in the clinical practice guidelines for subclinical Cushing's syndrome: which guidelines are trustworthy? Eur J Endocrinol. 2014; 171(4):421–31. https://doi.org/10.1530/EJE-14-0345 PMID: 24986532

59. Wang Y, Luo Q, Li Y, Wang H, Deng S, Wei S, et al. Quality assessment of clinical practice guidelines on the treatment of hepatocellular carcinoma or metastatic liver cancer. PLoS One. 2014; 9(8):e103939. https://doi.org/10.1371/journal.pone.0103939 PMID: 25105961

60. Yan J, Min J, Zhou B. Diagnosis of pheochromocytoma: a clinical practice guideline appraisal using AGREE II instrument. J Eval Clin Pract. 2013; 19(4):626–32. https://doi.org/10.1111/j.1365-2753.2012.01873.x PMID: 22809219