OPEN

What Determines the Quality of Rehabilitation Clinical Practice Guidelines?

An Overview Study

Marcel P. Dijkers, PhD, FACRM, Irene Ward, PT, DPT, NCS, Thiru Annaswamy, MD, MA, Devin Dedrick, MEd, Lilian Hoffecker, PhD, MLS, and Scott R. Millis, PhD, ABPP, CStat

Objective: The aim of the study was to determine what factors determine the quality of rehabilitation clinical practice guidelines.

Design: Six databases were searched for articles that had applied the Appraisal of Guidelines for Research & Evaluation II quality assessment tool to rehabilitation clinical practice guidelines. The 573 deduplicated abstracts were independently screened by two authors, resulting in 81 articles, the full texts of which were independently screened by two authors for Appraisal of Guidelines for Research & Evaluation II application to rehabilitation clinical practice guidelines, resulting in a final selection of 40 reviews appraising 504 clinical practice guidelines. Data were extracted from these by one author and checked by a second. Data on each clinical practice guideline included the six Appraisal of Guidelines for Research & Evaluation II domain scores, as well as the two Appraisal of Guidelines for Research & Evaluation II global evaluations.

Results: All six Appraisal of Guidelines for Research & Evaluation II domain scores were statistically significant predictors of overall clinical practice guideline quality rating; D3 (rigor of development) was the strongest and D1 (scope and purpose) the weakest (overall model $P < 0.001$, $R^2 = 0.53$). Five of the six domain scores were significant predictors of the clinical practice guideline use recommendation, with D3 the strongest predictor and D5 (applicability) the weakest (overall model $P < 0.001$, pseudo $R^2 = 0.53$).

Conclusions: Quality of rehabilitation clinical practice guidelines may be improved by addressing key domains such as rigor of development.

Key Words: Rehabilitation, Practice Guideline [Publication Type], Guidelines as Topic, Review Literature as Topic, Review [Publication Type]

(Am J Phys Med Rehabil 2021;100:790–797)

What Is Known

- Many clinical practice guidelines (CPGs) are being published, including in rehabilitation. Appraisers of CPGs frequently find them unsuitable for practice, even with modifications. What determines the quality of CPGs is unclear.

What Is New

- This overview study, using 40 reviews that appraised 504 CPGs using the Appraisal of Guidelines for Research & Evaluation II (AGREE II) tool, found that the AGREE overall quality rating is most strongly dependent on rigor of development. The recommendation for CPG use depends most on rigor, scope and purpose, and clarity of presentation. Clinical practice guideline developers should focus on improving their guidelines in these areas to have a positive reception by clinicians.

I n 2011, clinical practice guidelines (CPGs) were defined by the Institute of Medicine as "statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an appraisal of the benefits and harms of alternative care options."[1(p29)] Clinical practice guidelines make recommendations for screening, diagnosis or assessment, treatment, and management for a particular disorder or patient problem—most commonly for one of these clinical activities, but sometimes for some or all combined. In rehabilitation, CPGs are as common as elsewhere in health care—a May 19, 2020, PubMed search of "((physical medicine) OR

Quality of Rehabilitation CPGs

rehabilitation) AND (clinical practice guideline)" gave more than 7900 hits, although it is unknown how many of those represent CPGs that meet the six Institute of Medicine quality standards.[1(p5)]

Increasingly, clinicians are pressured to use CPGs in their practice, but as expressed in the title of the Institute of Medicine volume, these cannot all be trusted to be free of conflicts of interest or based on a comprehensive and through-going review of the empirical literature. As a consequence, many instruments have been developed to assist potential guideline users in critically assessing their quality. The last comprehensive review of these checklists and appraisal tools was performed by Siering et al.[2] in 2013, who identified 40 instruments in all. Since then, additional appraisal tools have been published (e.g., the checklists by Shaughnessy et al.[3] and Siebenhofer et al.[4]), but no new and comprehensive review was identified.

Siering et al.[2] extracted all questions and criteria from the 40 instruments that they found and performed a content analysis, which resulted in a list of 13 domains encompassing 33 unique items. They next coded the 40 tools to determine how many of the Siering items and of the Siering domains each covered. Appraisal of Guidelines for Research & Evaluation II (AGREE II) was the second-best instrument, covering all 13 Siering domains with one or more of its items and covering 26 of the 33 Siering items. (The quantitatively best tool was DELBI,[5] which, in contrast with AGREE II, has never been validated, and has seen little use outside of Germany.) Thus, it would seem that AGREE II, which has been used in dozens of reviews to assess hundreds of CPGs, is a good choice for appraising the quality of CPGs.

The AGREE II consists of 23 items that are combined into six "domain scores": D1, scope and purpose; D2, stakeholder involvement; D3, rigor of development; D4, clarity of presentation; D5, applicability; and D6, editorial independence; in addition, it asks for an overall quality assessment and for a recommendation for use of the CPG.[6] Overview studies that summarized across reviews that had used AGREE or AGREE II to evaluate CPGs have concluded that CPGs often are of low quality, especially where it comes to applicability, and cannot be recommended, even with modifications, in 18%–38% of cases.[7–11] A recent overview study, which was limited to rehabilitation CPGs, came to the same conclusion as the nonrehabilitation overviews did; based on 40 reviews appraising 544 CPGs published from 1994 to 2019, only 80% could be recommended, with or without modifications. The mean scores on the six AGREE II domains, on a 0–100 percentage scale, were as follows: (1) scope and purpose, 72; (2) stakeholder involvement, 53; (3) rigor of development, 56; (4) clarity of presentation, 71; (5) applicability, 34; and (6) editorial independence, 50.[12]

Given the discrepancies between the 40 reviews in the mean scores that they assigned for each of the six AGREE II domains,[12] the question arises: what do authors who appraise CPGs consider when they make a judgment on the overall quality of a CPG and decide on a recommendation for its use? Two studies have tried to answer the question. Hoffmann-Esser et al.[13] conducted an online survey of 58 German-speaking guideline appraisers and guideline users, asking them to indicate the "potential influence of the [23] AGREE II items on the two overall assessments (overall guideline quality and recommendation for use)" by rating the "strength of the influence on a Likert scale (0 = no influence to 5 = very strong influence)." (The

AGREE II Manual indicates that in making the two global assessments, scores on the 23 items and six domains should be considered, but the two should not be calculated from them. It also notes that the six domain scores "are independent and should not be aggregated into a single quality score.")[6(p9)] The items in domains 3 (rigor of development), 4 (clarity), and 6 (editorial independence) were stated by respondents to have the most impact. An open-ended question about items considered of influence tended to confirm this.[13]

A different approach was taken by Hatakeyama et al.,[14] who had three appraisers evaluate 206 Japanese CPGs using AGREE II, and used multiple regression analysis to determine the impact of the six domain scores on the overall quality assessment item. Rigor had the greatest influence ($\beta = 0.46$), followed by clarity (0.19) and applicability (0.16).[14] This clearly is not in line with what Hoffmann-Esser's respondents thought that they emphasized. Very surprising is also that her respondents gave little weight to applicability, which (at least in rehabilitation and other complex interventions) would seem to be a key CPG quality issue.[12]

The objective of the present study was to answer the question: in appraising rehabilitation CPGs, which characteristics do appraisers give the most weight in deciding on an overall quality assessment and in making a recommendation for use? The answer presumably would be useful to rehabilitation CPG authors, suggesting those areas of their guideline that need most attention to make the entire product acceptable to and be maximally useful for clinicians.

We addressed our objective using some of the data found in our previous appraisal of rehabilitation CPGs. However, we dropped reviews that did not report overall quality or recommendation for use and updated the literature search, including articles published from August 2019 to January 2020.

METHODS

This is an overview study, synthesizing information from published systematic reviews (SRs) that applied AGREE II to CPGs. It conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guideline and reports the required information accordingly (see Supplemental Checklist, Supplemental Digital Content 1, http://links.lww.com/PHM/B175).

Literature Search

The following bibliographic databases were searched for the period January 1, 2017, through January 22, 2020: *Medline* (Ovid); *Cochrane Library Databases* (Cochrane Library, Wiley); *PsycINFO* (Ovid), *Embase* (Embase.com); *CINAHL* Complete (EBSCO); and *Web of Science* (Clarivate). Search terms included the acronym and full name of the AGREE II tool; no language limits were applied. Retrieved records were organized and deduplicated using the bibliographic management software Endnote X9.

Data Selection

After deduplication, 573 abstracts remained, which were reviewed independently by two researchers (pairs made up of IW, TA, and four other rehabilitation experts), who selected articles that seemed to use AGREE II to evaluate the quality of rehabilitation CPGs. The definition of rehabilitation was modified from one developed by Levack et al.[15]: "an

intervention provided by or prescribed by rehabilitation professionals to patients to improve their functioning, maximize their independence, prevent or manage secondary complications of a chronic, disabling health condition or to manage functional implications of a chronic health condition."[(p4)] Disagreements between the two screeners were resolved by discussion or by obtaining and scrutinizing the full text if agreement could not be attained. Next, two researchers

TABLE 1. Nature and outcomes of 40 reviews using the AGREE II tool to evaluate rehabilitation CPGs

Study	Topic	No. CPGs	Mean AGREE II Domain Scores[a]						Mean Overall Rating	Recommendation, %[b]	
			D1	D2	D3	D4	D5	D6		Yes	Mod.
Andrade et al.[17] (2020)	Rehabilitation after ACL reconstruction	6	64	55	61	74	25	60	4.8	—	—
Anwer et al.[18] (2018)	Management of type 2 DM in adults	7	90	83	82	95	78	85	6.2	14	86
Appenteng et al.[19] (2018)	Management of acute pediatric TBI	17	85	58	59	82	39	53	5.2	23	77
Bhatt et al.[20] (2018)	Management of type 2 DM in children	21	69	58	47	73	49	44	—	33	19
Boaden et al.[21] (2020)	Performing videofluoroscopic swallowing studies	7	92	45	30	52	23	22	3.9	—	—
Bragge et al.[22] (2019)	Management of SCI neurogenic bladder	8	72	42	52	84	33	68	4.8	—	—
Bravo-Balado et al.[23] (2019)	Management of overactive bladder	7	60	41	54	88	23	52	4.5	—	—
Coronado-Zarco et al.[24] (2019)	Nonpharmacological osteoporosis treatment	12	62	50	53	63	33	40	4.1	0	75
Filiatreault et al.[25] (2018)	Preoperative hip fracture management	5	79	60	55	78	49	53	4.5	—	—
Gagliardi et al.[26] (2019)	Patient-centered care for women	27	89	62	60	85	43	64	—	0	67
Grammatikopoulou et al.[27] (2018)	Nutrition for adults with severe burns	8	70	41	47	74	35	55	4.3	38	38
Green et al.[28] (2019)	Treatment of acute lateral ankle ligament sprains in adults	7	66	51	32	81	8	25	4.0	—	—
Herzig et al.[29] (2018)	Acute noncancer pain management	4	73	51	63	63	31	61	4.4	75	25
Hoedl et al.[30] (2018)	Treatment of urinary incontinence in NH patients	5	67	38	58	74	28	76	4.2	—	—
Hoydonckx et al.[31] (2019)	Chronic pain intervention	4	94	50	81	84	44	53	5.2	0	75
Irajpour et al.[32] (2019)	End-of-life care	8	68	74	59	83	66	45	5.5	38	50
Jaggi et al.[33] (2018)	Neurogenic lower urinary tract management	3	86	80	82	90	69	85	6.0	33	67
Jolliffe et al.[34] (2018)	Rehabilitation after ABI	20	85	68	64	76	37	58	—	75	20
Karimi et al.[35] (2019)	Administering chemotherapy drugs	4	95	89	85	94	89	87		75	25
Kim et al.[36] (2019)	Rehabilitation after brain tumors	2	61	83	55	75	60	63	5.0	0	100
Kiriakova et al.[37] (2019)	Bone health in women with premature ovarian insufficiency	16	85	58	57	87	44	72	—	25	50
Knight et al.[38] (2019)	Rehabilitation for children with ABI	9	99	77	82	90	47	86	5.6	—	—
Lee et al.[39] (2019)	Rehabilitation after TBI	4	97	68	86	93	75	73	5.8	50	50
Lin et al.[40] (2018)	Management of musculoskeletal pain	34	72	44	47	59	26	32	3.7	—	—
Mandl et al.[41] (2019)	Poststroke rehabilitation of aphasia and dysarthria	6	44	56	39	59	28	61	3.5	0	100
O'Sullivan et al.[42] (2018)	Testing and management of various diagnostic groups	27	86	46	55	81	33	43	4.2	33	44
Parikh et al.[43] (2019)	Diagnosis and treatment of neck pain	46	68	55	47	63	31	44	4.1	—	—
Pattuwage et al.[44] (2017)	Management of spasticity in TBI	5	87	69	53	83	25	58	5.3	—	—
Perez-Panero et al.[45] (2019)	Diagnosis and management of diabetic foot	12	65	48	53	62	36	47	4.7	—	—
Reis et al.[46] (2017)	Treatment of obesity	21	89	69	71	84	49	65	4.9	—	—
Sankah et al.[47] (2019)	Exercise for hand osteoarthritis	8	90	88	77	83	43	81	4.8	63	38
Shallwani et al.[48] (2019)	Physical activity for people with cancer	20	81	64	64	77	40	67	4.6	—	—
Shetty et al.[49] (2018)	Worker's compensation disability management	1	64	67	55	75	74	69	4.5	0	100
Tamas et al.[50] (2018)	Diagnosis and treatment of dystonia	15	64	34	29	54	14	22	—	13	67
Tan et al.[51] (2019)	Treatment of venous leg ulcers	14	56	46	52	74	27	46	4.9	—	—
Uzeloto et al.[52] (2017)	PT management in respiratory disease	33	79	52	61	79	37	54	4.8	21	70
van der Ploeg et al.[53] (2019)	(Discontinuation of) statin treatment in older adults	18	72	54	55	81	50	49	4.5	—	—
Wang et al.[54] (2019)	Poststroke rehabilitation of aphasia	8	96	84	67	77	64	91	5.8	75	13

(Continued on next page)

TABLE 1. *(Continued)*

Study	Topic	No. CPGs	Mean AGREE II Domain Scores[a]						Mean Overall Rating	Recommendation, %[b]	
			D1	D2	D3	D4	D5	D6		Yes	Mod.
Zhang et al.[55] (2019)	Treatment of diabetic foot ulcers	8	89	69	69	88	53	66	5.5	75	25
Zhao et al.[56] (2019)	Nutrition for cancer patients	17	74	37	43	75	23	35	—	12	65
All mean/percentage[c]			77	56	56	75	38	53	4.6	30	53
All SD[c]			19	24	25	19	26	33	1.4	—	—

[a]Domain: D1, scope and purpose; D2, stakeholder involvement; D3, rigor of development; D4, clarity of presentation; D5, applicability; D6, editorial independence.

[b]Yes: yes, recommended; Mod: recommended with modification. "Not recommended" is omitted.

[c]The mean and SD for D1 through D6 are based on the 504 CPGs directly, NOT on the 40 means shown here. The number of cases for the overall quality rating is 384, and the number of cases for the recommendation percentages is 280.

ABI, acquired brain injury; ACL, anterior cruciate ligament; DM, diabetes mellitus; NH, nursing home; PT, physical therapy; SCI, spinal cord injury; TBI, traumatic brain injury.

independently reviewed the full texts selected by the two screeners, using these criteria:

1. AGREE II was used to evaluate existing CPGs, not to develop one.
2. The report was in English, Spanish, Portuguese, German, French, or Dutch and was published in a peer-reviewed journal.
3. All or at least most of the CPGs rated involved rehabilitation as here defined.
4. The primary target of the CPGs was a rehabilitation clinician or other healthcare provider, not the patient or a family caregiver.
5. The six AGREE II domain scores and/or the 23 item scores were reported, in tables or supplemental digital content, for each CPG.
6. For each CPG, the AGREE II "overall quality" rating and/or the recommendation for use were provided, and these had not been determined as a mathematical function of the six AGREE II domain scores

Data Extraction

A two-part Excel form was created and piloted to extract all relevant data. Part 1 focused on the review article providing appraisals of CPGs and included the following items:

1. Number of guidelines evaluated using AGREE II
2. Number of AGREE II raters for each CPG
3. Presence of overall quality ratings and their mathematical independence of the six domain scores
4. Presence of recommendations for CPG use and their independence of the six domain scores

Part 2 dealt with the CPGs themselves as reported in the review and included the following information:

1. The six AGREE II domain scores, if provided
2. The 23 AGREE II item scores, if provided
3. The AGREE II overall quality rating, on the 1–7 scale, where 1 = lowest possible quality and 7 = highest possible quality, or transformed to that scale
4. The recommendation made with respect to the CPG, using the AGREE II standard terminology or an equivalent: "recommended," "recommended with modification," and "not recommended," as provided in the review

The extraction of information was done by DD and checked by MD. In case of disagreement, discussion took place to resolve the issue. Many articles reported AGREE II "overall quality" ratings on a percent scale, not a 1–7 scale. These were back transformed, taking into account the number of raters that had been used. In instances where a consensus recommendation for a particular CPG was not reported, but only those of the two or more appraisers individually, an algorithm was used, which was similar to that developed by Hoffmann-Esser et al.[16]: "majority vote" was used to determine a "consensus" recommendation; if there was no majority, the median vote was used.

Data Processing and Synthesis

All data were uploaded to SPSS for processing and CPG description. When articles presented AGREE II item scores but not domain scores, the latter were calculated based on the former, using the AGREE II formulas.[6] Stata 16.1 statistical software was used to conduct the analysis of how the domain scores together influenced global assessments. Two mixed-effects models with a random intercept were fitted using maximum likelihood. The response variables were the overall quality score and the recommendation for use. In both models, the six domain scores were the predictors used, which were entered simultaneously. A mixed-effects linear regression model was used to predict the overall quality score. A mixed-effects ordered logit model was used to predict recommendation (ie, not recommended, recommended with modification, and recommended). The mixed-effects model was used because all but one review made more than one rating (ie, CPG AGREE II evaluation), which causes the data to be clustered. The mixed model is able to handle the clustering when estimating the model parameters.

RESULTS

The 40 articles provided AGREE II ratings for a total of 504 CPGs, with a range of 1–48 guidelines per review (mean = 12.6, SD = 10.0, median = 8). Table 1 lists these review articles, the topic area of each, and their mean scores on the six domains. Also provided is the mean overall quality score, which was provided by 33 reviews, which evaluated 384 CPGs

TABLE 2. Intercorrelations of the domain scores and their correlation with the overall rating

	D1	D2	D3	D4	D5	D6
Domain 1. Scope and purpose						
Domain 2. Stakeholder involvement	0.60					
Domain 3. Rigor of development	0.63	0.70				
Domain 4. Clarity of presentation	0.55	0.52	0.63			
Domain 5. Applicability	0.46	0.70	0.64	0.51		
Domain 6. Editorial independence	0.42	0.51	0.60	0.46	0.53	
Overall quality rating of the CPG	0.60	0.72	0.84	0.70	0.70	0.66

Based on the 383 CPGs for which an overall quality rating was available.

(on average = 11.6). The last two columns provide, by review, the percentage of CPGs that were recommended and recommended with modifications, respectively; this information was supplied by 24 reviews, which between them appraised 280 CPGs (average = 11.7). Inspection of the table shows quite some variability from one review to the next in mean domain scores, overall rating, and percentages recommended. Information to explain this variation based on the topic area covered, rater severity, or other factors is not available.

Table 2 shows that the six domain scores are all correlated with one another, at an *r* value of 0.42 or greater (all significant at the 0.0001 level). They also all are correlated with the overall CPG quality rating, at the level of 0.60 or greater. Figure 1 shows the association between the type of recommendation provided and the domain scores, as well as with the overall score. These data suggest that the influence of the six domains on the two global quality judgments is not very discrepant; however, nesting is not taken into account here.

The omnibus test for the overall quality model (Table 3) that took nesting into account was statistically significant ($P < 0.001$) with an R^2 value of 0.53. All six of the domain scores were statistically significant predictors of the overall score. D3 (rigor of development) was the strongest predictor, whereas D1 (scope and purpose) was the weakest one.

Similarly, the omnibus test for the CPG use recommendation model was statistically significant ($P < 0.001$) with a pseudo R^2 value of 0.53 (Table 3). Five of the six domain scores were significant predictors of the recommendation. D3 was also the strongest predictor in this model. D5 (applicability) was the weakest predictor ($P = 0.15$).

DISCUSSION

Ever since CPGs were first published almost half a century ago, there has been concern about their quality, including the strength of the underlying evidence; the methods used to come to recommendations; the lack of consideration of patients' values and realities; conflicts of interest of CPG developers; and inconsistent recommendations in CPGs that dealt with the same evidence to come to recommendations in the same clinical area.[1]

In this study, only 53% of the variance in overall CPG quality rating for the 384 CPGs was explained by the six domain scores, suggesting that the raters and rater teams were not very consistent and/or that they often took into account much information not covered in the 23 AGREE II items. All six domains contributed to the overall rating, with D3, rigor of development, clearly the most important factor (coefficient = 0.024), and D1, scope and purpose; D2, stakeholder involvement; and D6, editorial independence having relatively limited impact, all with a coefficient of 0.007. These findings correspond somewhat
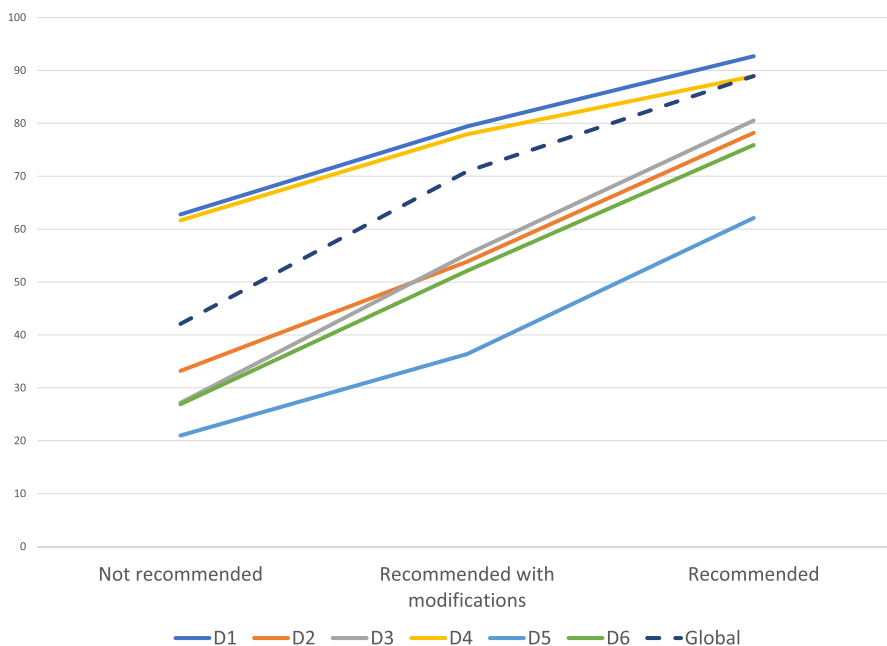
FIGURE 1. Mean score on the six AGREE II domains, and mean overall score, by type of recommendation. Note: The values of the overall quality score were multiplied by 15 to make their scale (1–7) approximately the same as that of the domain scores. Mean scores for D1 to D6 are based on, respectively: not recommended: 48 cases; recommended with modifications: 143 cases; yes, recommended: 83 cases. For the overall CPG quality rating, the number of cases is 17, 87, and 50, respectively.

TABLE 3. Results of mixed model regression of overall quality rating and CPG use recommendation on the six AGREE II domains

AGREE II Domain	Global Quality Rating				Recommendation for CPG Use			
	Coef.	SE	z	P	Coef.	SE	z	P
D1. Scope and purpose	0.007	0.002	3.53	0.000	0.056	0.019	2.98	0.003
D2. Stakeholder involvement	0.007	0.002	3.86	0.000	0.033	0.016	2.09	0.037
D3. Rigor of development	0.024	0.002	12.55	0.000	0.077	0.017	4.57	0.000
D4. Clarity of presentation	0.013	0.002	5.97	0.000	0.061	0.019	3.23	0.001
D5. Applicability	0.010	0.002	6.10	0.000	0.019	0.013	1.46	0.145
D6. Editorial independence	0.007	0.001	6.05	0.000	0.025	0.008	3.29	0.001
Constant	0.597	0.158	3.78	0.000	—	—	—	—
R^2	0.53							
Pseudo R^2					0.53			

with those of Hatakeyama et al.,[14] who reported a strong influence of D3 rigor, followed by D4 clarity of presentation and D5 applicability.

About as much variation in CPG recommendation was explained by the six domains (pseudo R^2 = 0.53). Here too, D3 was the most important predictor, but D1 and D4 (clarity of presentation) had similar coefficients. Surprisingly, applicability (D5) was not statistically significant. Hatakeyama et al.[14] did not report on determinants of CPG use recommendations, so the present findings cannot be compared with theirs. Currently, there is no explanation why the factors determining an overall rating are so dissimilar from those having an impact on CPG use recommendation.

These findings suggest that for rehabilitation CPGs, there is no single domain or factor that, if enriched, can by itself improve the likelihood that a CPG is highly rated or is recommended for use. Therefore, teams of rehabilitation clinicians, methodologists, and stakeholders working to create new CPGs or to update existing ones have to pay careful attention to all issues that are tapped by the 23 AGREE II items. Surprisingly, applicability plays no role in predicting a recommendation, which is surprising, especially given the fact that in both within rehabilitation[12] and outside of it,[7–11] applicability consistently gets the lowest scores on the six domains.

Rating the 23 items included in AGREE II can be subjective, which is why the manual recommends using for each guideline at least two appraisers, and preferably four, whose scores are "averaged" to develop the six domain scores.[6] However, a common complaint is that the manual offers no guidance for the scores intermediate between the Likert scale extremes used to rate the 23 items: 1 indicates strongly disagree and 7 indicates strongly agree, allowing for that subjectivity. The manual states: "A score between 2 and 6 is assigned when the reporting of the AGREE II item does not meet the full criteria or considerations. A score is assigned depending on the completeness and quality of reporting."[6(p8)] For the two global assessment items, no guidance is provided as to whether the scores of the multiple raters should be averaged or whether they should discuss discrepant ratings and recommendations to come to a consensus. As a consequence, many of the 40 articles provided the overall rating for each appraiser (which here were averaged) and the use recommendation by each separate assessor (which were "averaged" using the algorithm). If there is criticism that CPG development and implementation instruction

manuals leave much to be desired,[57,58] the same holds true for the most used CPGs assessment tool. The AGREE II offers guidance on how much to bank on a CPG, but its scores should not be treated by rehabilitation clinicians as the last word.

Of the six previous overview studies that compiled AGREE II appraisals from published reviews, two were descriptive, and as such had no need to use advanced statistical methods.[8,12] Three others, however, tested (implicit) hypotheses as to the association of domain scores and year of publication and should have taken this step.[7,10,11] Only Gagliardi and Brouwers[9] in testing the association between domain 5 (Applicability) scores and publication year, country of publication, and type of CPG developer used mixed effect models as appropriate in this situation. (In the case of Hatakeyama et al.,[14] there was no need for such modeling because all 206 of their CPGs were evaluated by the same team.)

It should be noted that whatever the number of items and domains in a CPG appraisal tool, none requires the assessor to independently address the nature of the evidence underlying recommendations. This has been observed by a number of authors.[59–62] Vlayen et al.[63] stated, in 2005: "in order to evaluate the quality of the clinical content and more specifically the evidence base of a CPG, verification of the completeness and the quality of the literature search and its analysis has to be added to the process of validation by an appraisal instrument."[(p239)] Given that CPGs may run into dozens if not hundreds of pages, and reviews may include 30 or more CPGs, this would be an herculean task. Instead, reviewers score "rigor of development" or a similar domain, based on the protocol the CPG authors (claimed to have) used. Thus, even a high overall CPG quality rating cum recommendation based on multiple methodology items is no guarantee that a particular guideline is making recommendations based on all the relevant evidence properly assessed and evaluated. There is an implied trust that the guideline developers were thorough in their literature search and verification of evidence. However, CPGs that are rated highly on rigor of development are likely to be better than the ones that were rater poorly on this domain, because of flawed methodology.

Developing a CPG is time-consuming and therefore expensive, even if much of the time is contributed by volunteers in academia and healthcare organizations. That almost one fifth cannot be recommended for use, even with modifications, whether in health services in general[7,8,10,11] or in rehabilitation

specifically (Table 1) means a sizeable waste of effort. (Across the 40 reviews used here, only 35% of 504 CPGs were recommended without any modification.) If these CPGs are indeed not consulted and implemented by individual clinicians and healthcare organizations, opportunities are missed to provide optimal care, eliminate useless or even harmful assessments and treatments, and reduce variations in nature and quantity of rehabilitation treatments, from organization to organization and from provider to provider. Because CPGs are such complex documents, a multitude of quality criteria can be, and have been, applied to them. Not all of these criteria may be considered important enough to actually apply, and the ones users do want to use may have uneven importance in their eyes. This analysis suggests that of the criteria used in the AGREE II tool, some are more important than others, but all play a role in telling good, bad, and indifferent CPGs apart.

Limitations

We used a functional definition of rehabilitation, which was applied to 40 review articles, which often provided minimal information on the CPGs being appraised and on the degree that rehabilitation practices or services were included in their recommendations. The reviews frequently were not clear about their use of the standardized AGREE II procedures for item scoring, overall quality rating, and making recommendations, and e-mail requests for clarification often were not answered. A strenuous effort was made, using the text in the methods section of the reviews found in the bibliographic search, to limit the secondary studies in this overview to those which offered an overall quality rating and/or a CPG use recommendation that, per AGREE II protocol, were not a mathematical derivative of the item or domain scores. We used six comprehensive bibliographic data bases to identify review articles but may have missed some that did not use the term "AGREE II" (or its full expansion) in title or abstract.

CONCLUSIONS

Based on 40 review studies that used the most often used and best validated CPG appraisal tool, AGREE II, to evaluate 504 rehabilitation CPGs, we conclude that these CPGs often have weak points and are not recommended for use 17% of the time. Although the six AGREE II domains differ somewhat in their impact on making a recommendation and on the overall quality rating, it would seem that CPG developers need to pay attention to all of them to improve the quality of rehabilitation CPGs.

ACKNOWLEDGMENTS

REFERENCES

1. Institute of Medicine: *Clinical Practice Guidelines We Can Trust*. Washington, DC, National Academies Press, 2011
2. Siering U, Eikermann M, Hausner E, et al: Appraisal tools for clinical practice guidelines: a systematic review. *PLoS One* 2013;8:e82915
3. Shaughnessy AF, Vaswani A, Andrews BK, et al: Developing a clinician friendly tool to identify useful clinical practice guidelines: G-TRUST. *Ann Fam Med* 2017;15:413–8
4. Siebenhofer A, Semlitsch T, Herborn T, et al: Validation and reliability of a guideline appraisal mini-checklist for daily practice use. *BMC Med Res Methodol* 2016;16:39
5. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. (AWMF) and Ärztliches Zentrum für Qualität in der Medizin (ÄZQ): German Instrument for Methodological Guideline Appraisal [Deutsches Instrument zur methodischen Leitlinien-Bewertung (DELBI)] Version: 2005/2006 + Domain 8 (2008). 2008. Available at: https://www.leitlinien.de/mdb/edocs/pdf/literatur/german-guideline-appraisal-instrument-delbi.pdf. Accessed May 1, 2020
6. AGREE Next Steps Consortium. The AGREE II Instrument [Electronic version]. 2017. Available at: http://www.agreetrust.org. Accessed May 9, 2020
7. Alonso-Coello P, Irfan A, Sola I, 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:e58
8. Knai C, Brusamento S, Legido-Quigley H, et al: Systematic review of the methodological quality of clinical guideline development for the management of chronic disease in Europe. *Health Policy* 2012;107(2–3):157–67
9. Gagliardi AR, Brouwers MC: Do guidelines offer implementation advice to target users? A systematic review of guideline applicability. *BMJ Open* 2015;5:e007047
10. Armstrong JJ, Goldfarb AM, Instrum RS, et al: Improvement evident but still necessary in clinical practice guideline quality: a systematic review. *J Clin Epidemiol* 2017;81:13–21
11. Rabassa M, Garcia-Ribera Ruiz S, Solà I, et al: Nutrition guidelines vary widely in methodological quality: an overview of reviews. *J Clin Epidemiol* 2018;104:62–72
12. Dijkers MP, Ward I, Annaswamy T, et al: Quality of rehabilitation clinical practice guidelines: an overview study of AGREE II appraisals. *Arch Phys Med Rehabil* 2020;101:1643–55
13. Hoffmann-Esser W, Siering U, Neugebauer EAM, et al: Guideline appraisal with AGREE II: online survey of the potential influence of AGREE II items on overall assessment of guideline quality and recommendation for use. *BMC Health Serv Res* 2018;18:143
14. Hatakeyama Y, Seto K, Amin R, et al: The structure of the quality of clinical practice guidelines with the items and overall assessment in AGREE II: a regression analysis. *BMC Health Serv Res* 2019;19:788
15. Levack WMM, Rathore FA, Pollet J, et al: One in 11 Cochrane reviews are on rehabilitation interventions, according to pragmatic inclusion criteria developed by Cochrane Rehabilitation. *Arch Phys Med Rehabil* 2019;100:1492–8
16. Hoffmann-Eßer W, Siering U, Neugebauer EA, et al: Guideline appraisal with AGREE II: systematic review of the current evidence on how users handle the 2 overall assessments. *PLoS One* 2017;12:e0174831
17. Andrade R, Pereira R, van Cingel R, et al: How should clinicians rehabilitate patients after ACL reconstruction? A systematic review of clinical practice guidelines (CPGs) with a focus on quality appraisal (AGREE II). *Br J Sports Med* 2020;54:512–9
18. Anwer MA, Al-Fahed OB, Arif SI, et al: Quality assessment of recent evidence-based clinical practice guidelines for management of type 2 diabetes mellitus in adults using the AGREE II instrument. *J Eval Clin Pract* 2018;24:166–72
19. Appenteng R, Nelp T, Abdelgadir J, et al: A systematic review and quality analysis of pediatric traumatic brain injury clinical practice guidelines. *PLoS One* 2018;13:e0201550
20. Bhatt M, Nahari A, Wang PW, et al: The quality of clinical practice guidelines for management of pediatric type 2 diabetes mellitus: a systematic review using the AGREE II instrument. *Syst Rev* 2018;7:193
21. Boaden E, Nightingale J, Bradbury C, et al: Clinical practice guidelines for videofluoroscopic swallowing studies: A systematic review. *Radiography (Lond)* 2020;26:154–62
22. Bragge P, Guy S, Boulet M, et al: A systematic review of the content and quality of clinical practice guidelines for management of the neurogenic bladder following spinal cord injury. *Spinal Cord* 2019;57:540–9
23. Bravo-Balado A, Plata M, Trujillo CG, et al: Is the development of clinical practice guidelines for non-neurogenic overactive bladder trustworthy? A critical appraisal using the Appraisal of Guidelines, Research and Evaluation (AGREE) II instrument. *BJU Int* 2019;123:921–2
24. Coronado-Zarco R, Olascoaga-Gómez de León A, García-Lara A, et al: Nonpharmacological interventions for osteoporosis treatment: systematic review of clinical practice guidelines. *Osteoporos Sarcopenia* 2019;5:69–77
25. Filiatreault S, Hodgins M, Witherspoon R: An umbrella review of clinical practice guidelines for the management of patients with hip fractures and a synthesis of recommendations for the pre-operative period. *J Adv Nurs* 2018;74:1278–88
26. Gagliardi AR, Green C, Dunn S, et al: How do and could clinical guidelines support patient-centred care for women: content analysis of guidelines. *PLoS One* 2019;14:e0224507
27. Grammatikopoulou MG, Theodoridis X, Gkiouras K, et al: AGREEing on guidelines for nutrition management of adult severe burn patients. *JPEN J Parenter Enteral Nutr* 2019;43:490–6
28. Green T, Willson G, Martin D, et al: What is the quality of clinical practice guidelines for the treatment of acute lateral ankle ligament sprains in adults? A systematic review. *BMC Musculoskelet Disord* 2019;20:394
29. Herzig SJ, Calcaterra SL, Mosher HJ, et al: Safe opioid prescribing for acute noncancer pain in hospitalized adults: a systematic review of existing guidelines. *J Hosp Med* 2018;13:256–62
30. Hoedl M, Schoberer D, Halfens RJG, et al: Adaptation of evidence-based guideline recommendations to address urinary incontinence in nursing home residents according to the ADAPTE-process. *J Clin Nurs* 2018;27(15–16):2974–83
31. Hoydonckx Y, Kumar P, Flamer D, et al: Quality of chronic pain interventional treatment guidelines from pain societies: assessment with the AGREE II instrument. *Eur J Pain* 2020;24:704–21

32. Irajpour A, Hashemi M, Taleghani F: The quality of guidelines on the end-of-life care: a systematic quality appraisal using AGREE II instrument. *Support Care Cancer* 2020;28:1555–61

33. Jaggi A, Drake M, Siddiqui E, et al: A critical appraisal of the principal guidelines for neurogenic lower urinary tract dysfunction using the AGREE II instrument. *NeurourolUrodyn* 2018;37:2945–50

34. Jolliffe L, Lannin NA, Cadilhac DA, et al: Systematic review of clinical practice guidelines to identify recommendations for rehabilitation after stroke and other acquired brain injuries. *BMJ Open* 2018;8:e018791

35. Karimi T, Bahrami M, Yadegarfar G: The critical assessment of the quality of common clinical guidelines for administering chemotherapy drugs by using AGREE II tool. *Int J Cancer Manag* 2019;12:410–6

36. Kim WJ, Novotna K, Amatya B, et al: Clinical practice guidelines for the management of brain tumours: a rehabilitation perspective. *J Rehabil Med* 2019;51:89–96

37. Kiriakova V, Cooray SD, Yeganeh L, et al: Management of bone health in women with premature ovarian insufficiency: systematic appraisal of clinical practice guidelines and algorithm development. *Maturitas* 2019;128:70–80

38. Knight S, Takagi M, Fisher E, et al: A systematic critical appraisal of evidence-based clinical practice guidelines for the rehabilitation of children with moderate or severe acquired brain injury. *Arch Phys Med Rehabil* 2019;100:711–23

39. Lee SY, Amatya B, Judson R, et al: Clinical practice guidelines for rehabilitation in traumatic brain injury: a critical appraisal. *Brain Inj* 2019;33:1263–71

40. Lin I, Wiles LK, Waller R, et al: Poor overall quality of clinical practice guidelines for musculoskeletal pain: a systematic review. *Br J Sports Med* 2018;52:337–43

41. Mandl L, Schindel D, Deutschbein J, et al: Quality of German-language guidelines for post-stroke rehabilitation of aphasia and dysarthria - results of a systematic review and of an international comparison. *Rehabilitation (Stuttg)* 2019;58:331–8

42. O'Sullivan JW, Albasri A, Koshiaris C, et al: Diagnostic test guidelines based on high-quality evidence had greater rates of adherence: a meta-epidemiological study. *J Clin Epidemiol* 2018;103:40–50

43. Parikh P, Santaguida P, Macdermid J, et al: Comparison of CPG's for the diagnosis, prognosis and management of non-specific neck pain: a systematic review. *BMC Musculoskelet Disord* 2019;20:81

44. Pattuwage L, Olver J, Martin C, et al: Management of spasticity in moderate and severe traumatic brain injury: evaluation of clinical practice guidelines. *J Head Trauma Rehabil* 2017;32:E1–12

45. Perez-Panero AJ, Ruiz-Munoz M, Cuesta-Vargas AI, et al: Prevention, assessment, diagnosis and management of diabetic foot based on clinical practice guidelines: a systematic review. *Medicine (Baltimore)* 2019;98:e16877

46. Reis ECD, Passos SRL, Santos M: Quality assessment of clinical guidelines for the treatment of obesity in adults: application of the AGREE II instrument. *Cad Saude Publica* 2018;34:e00050517

47. Sankah BEA, Stokes M, Adams J: Exercises for hand osteoarthritis: a systematic review of clinical practice guidelines and consensus recommendations. *Phys Ther Rev* 2019;24(3–4):66–81

48. Shallwani SM, King J, Thomas R, et al: Methodological quality of clinical practice guidelines with physical activity recommendations for people diagnosed with cancer: a systematic critical appraisal using the AGREE II tool. *PLoS One* 2019;14:e0214846

49. Shetty K, Raaen L, Khodyakov D, et al: Evaluation of the Work Loss Data Institute's Official Disability guidelines. *J Occup Environ Med* 2018;60:e146–51

50. Tamas G, Abrantes C, Valadas A, et al: Quality and reporting of guidelines on the diagnosis and management of dystonia. *Eur J Neurol* 2018;25:275–83

51. Tan MKH, Luo R, Onida S, et al: Venous leg ulcer clinical practice guidelines: what is AGREEd? *Eur J Vasc Endovasc Surg* 2019;57:121–9

52. Uzeloto JS, Moseley AM, Elkins MR, et al: The quality of clinical practice guidelines for chronic respiratory diseases and the reliability of the AGREE II: an observational study. *Physiotherapy* 2017;103:439–45

53. van der Ploeg MA, Floriani C, Achterberg WP, et al: Recommendations for (discontinuation of) statin treatment in older adults: review of guidelines. *J Am Geriatr Soc* 2019;30:30

54. Wang Y, Li H, Wei H, et al: Assessment of the quality and content of clinical practice guidelines for post-stroke rehabilitation of aphasia. *Medicine (Baltimore)* 2019;98:e16629

55. Zhang P, Lu Q, Li H, et al: The quality of guidelines for diabetic foot ulcers: a critical appraisal using the AGREE II instrument. *PLoS One* 2019;14:e0217555

56. Zhao XH, Yang T, Ma XD, et al: Heterogeneity of nutrition care procedures in nutrition guidelines for cancer patients. *Clin Nutr* 2020;39:1692–704

57. Gagliardi AR, Brouwers MC: Integrating guideline development and implementation: analysis of guideline development manual instructions for generating implementation advice. *Implement Sci* 2012;7:67

58. Schunemann HJ, Wiercioch W, Etxeandia I, et al: Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ* 2014;186:E123–42

59. Burgers JS: Guideline quality and guideline content: are they related? *Clin Chem* 2006;52:3–4

60. Watine J, Friedberg B, Nagy E, et al: Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. *Clin Chem* 2006;52:65–72

61. Matthys J, De Meyere M: Quality evidence important for quality guidelines. *CMAJ* 2010;182:1449–50

62. Eikermann M, Holzmann N, Siering U, et al: Tools for assessing the content of guidelines are needed to enable their effective use—a systematic comparison. *BMC Res Notes* 2014;7:853

63. Vlayen J, Aertgeerts B, Hannes K, et al: A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. *Int J Qual Health Care* 2005;17:235–42