

Expanding the Grading of Recommendations Assessment, Development, and Evaluation (Ex-GRADE) for Evidence-Based Clinical Recommendations: Validation Study

Linda Phi¹, Reem Ajaj^{1,2}, Manisha H Ramchandani¹, Xenia MC Brant^{1,3}, Oluwadayo Oluwadara¹, Olga Polinovsky¹, David Moradi¹, Andre Barkhordarian¹, Pathu Sriphanlop¹, Margaret Ong¹, Amy Giroux¹, Justin Lee¹, Muniza Siddiqui¹, Nora Ghodousi¹ and Francesco Chiappelli^{1,2,*}

¹Division of Oral Biology & Medicine, UCLA School of Dentistry, Los Angeles, CA, USA

²Division of Associated Clinical Specialties, UCLA, School of Dentistry, Los Angeles, CA, USA

³Department of Endodontics and Research and Education Center, CEO IPSEMG, UFMG, Belo Horizonte MG, Brazil

Abstract: Clinicians use general practice guidelines as a source of support for their intervention, but how much confidence should they place on these recommendations? How much confidence should patients place on these recommendations? Various instruments are available to assess the quality of evidence of research, such as the revised Wong scale (R-Wong) which examines the quality of research design, methodology and data analysis, and the revision of the assessment of multiple systematic reviews (R-AMSTAR), which examines the quality of systematic reviews.

The Grading of Recommendation Assessment, Development, and Evaluation (GRADE) Working Group developed an instrument called the GRADE system in order to grade the quality of the evidence in studies and to evaluate the strength of recommendation of the intervention that is proposed in the published article. The GRADE looks at four factors to determine the quality of the evidence: study design, study quality, consistency, and directness. After combining the four components and assessing the grade of the evidence, the strength of recommendation of the intervention is established. The GRADE, however, only makes a qualitative assessment of the evidence and does not generate quantifiable data.

In this study, we have quantified both the grading of the quality of evidence and also the strength of recommendation of the original GRADE, hence expanding the GRADE. This expansion of the GRADE (Ex-GRADE) permits the creation of a new instrument that can produce tangible data and possibly bridge the gap between evidence-based research and evidence-based clinical practice.

Keywords: GRADE, AMSTAR, Revised AMSTAR, Wong scale, Systematic Review, Vaccination and Autism, Vaccination, Whitening, Bleaching, In-office Whitening, In-office Beaching, CPAP versus Oral Appliance, CPA, Sleep Apnea, Hypoxia, Evidence-Based Decision Making, Strength of Clinical Relevance, Quality of Evidence, Strength of Recommendation, Clinical Significance, Evidence-Based Clinical Practice, Evidence-Based Dentistry, Evidence-Based Medicine.

INTRODUCTION

Progress and advancement in science through research is continually reported in scientific journals. For a scientific journal to publish an article, the article goes through multiple peer reviews and edits, where the final, published version is viewed to be of the highest merit for many. For the most part, researchers, clinicians, and individuals who read scientific articles may assume that the clinical trial, systematic review, observational study, etc. are conducted without flaw and that their data and results are correctly analyzed. For example, they may trust that the clinical trials are conducted

with blinks and placebos to ensure any possible means of preventing bias, that the Chi-Square test is used to test the homogeneity of the sampling distribution by evaluating a single variable and not multiple variables, and that the studies' t-test is used to assess the statistically significant difference between two populations of interest and not three populations. As a result, many individuals, especially readers with limited background knowledge, tend to accept the findings and recommendations of scientific papers without question.

Several instruments have been developed to assess the quality of the evidence of scientific papers such as the revised Wong scale (R-Wong) [1] and the revised scale for the "assessment of multiple systematic reviews" (R-AMSTAR) [2]. The R-Wong examines the quality of research design, methodology and data analysis [1]. The "assessment of multiple systematic reviews" (AMSTAR) revised by Kung *et al.* (2010) as the R-AMSTAR is designed to assess the quality of the methodology of systematic reviews [2]. A systematic

*Address correspondence to this author at the Evidence-Based Study Group, Evidence-Based Decisions-Practice-Based Research Network (EBD-PBRN.org) and Division of Oral Biology & Medicine, CHS 63-090, UCLA School of Dentistry, CHS 63-090, Los Angeles, CA 90095-1668; Tel: 310-794-6625; Fax: 310-794-7109; E-mail: fchiappelli@dentistry.ucla.edu

review is defined as “the product of the process of systematically reviewing the research literature pertinent to the research question” by addressing the PICO question [1-3]. The acronym PICO was developed to denote the evidence-based research question, consisting of:

- the problem patient population (P),
- the interventions (I) under
- the consideration or comparison (C),
- the clinical outcome of interest (O).

The revised Wong scale (R-Wong), which is comprised of nine questions and evaluates the methodology and quality of primary sources. Each question is scored 1, 2, or 3 (1 = inappropriate, 2 = mediocre, 3 = appropriate), with a total scale ranging from 9-27 [1].

The “Grading of Recommendations Assessment, Development, and Evaluation (GRADE)” instrument is a tool developed by the GRADE Working Group, which assesses both the quality of the evidence as well as the strength of recommendation [4]. The GRADE Working Group defines the quality of the evidence as “the extent to which one can be confident that an estimate of effect is correct” and defines the strength of a recommendation as “the extent to which one can be confident that adherence to the recommendation will do more good than harm⁴.” Four components are also evaluated when looking at a body of literature:

- study design,
- study quality,
- consistency, and directness.

Study design refers to the type of design used in the research, e.g. randomized trials and observational studies. Meanwhile, study quality refers to the methodology and execution of the study, e.g. whether a randomized trial is blinded or double-blinded. Consistency looks at the estimates of effects among the studies and the consistencies and/or inconsistencies of the results, and directness refers to the extent in which the study conditions (e.g. study population, the intervention, and outcomes) reflect the conditions of the population of interest [3].

The quality of evidence using the GRADE instrument is ultimately graded by the following scale: high, moderate, low, and very low. The overall quality of evidence, however, takes into consideration all four components of quality of evidence in the paper, and assigns these grades accordingly [4]. The strength of recommendation of a body of literature using the GRADE instrument follows this assignment: net benefits, trade-offs, uncertain trade-off, and no net benefits [4]. The GRADE is an instrument that aims to address the cost versus benefit concern that many patients and clinicians have prior to an intervention, and it is one of the various decision aids (e.g. workbooks, computer-based resources, and visuals) that help guide individuals through the decision-making process [5].

Although the GRADE draws a bridge between evidence-based health practices and clinical practices, the criteria for establishing the quality of evidence and strength of recommendation is qualitative and leaves tremendous room for

bias. The GRADE instrument is not standardized in how a clinician or a patient would evaluate a body of literature for these two outcomes, since the majority of the judgment would inevitably be based on the individual’s beliefs and opinions. As a result, we have expanded the GRADE to provide systematic and unbiased grading procedure in assessing the quality of the evidence and the strength of recommendation. The Ex-GRADE does not alter the purpose and significance of the original GRADE instrument. Rather, it provides criteria for the clinician and the patient to systematically evaluate these two components (See Appendix 1 for the grading criteria). Whether they are looking at primary or secondary sources (e.g. clinical trials and systematic reviews), patients and clinicians can gain insight into how well a research study was conducted. They are also able to gauge how substantial and reliable their data/data analysis is. By incorporating the R-Wong and the R-AMSTAR into the Ex-GRADE, this instrument is able to assess the quality of evidence and the strength of the clinical recommendation in a quantitative manner, enabling the critical analyses of the data, which in turn help to identify the qualitative significance of the research at hand [6, 7]. With this literature, we provide the validation of the “strength of recommendation” segment of Ex-GRADE with respect to evidence-based clinical decision making. By incorporating the R-Wong and the R-AMSTAR, in which both have been previously validated [1, 2] into the Ex-GRADE as the quantitative instruments in assessing the quality of evidence, no additional validation for the first component of the Ex-GRADE is necessary in validating this instrument as a whole.

METHOD

The Ex-GRADE was independently validated in three distinct and unrelated fields of health care: in-office teeth whitening, sleep apnea, and vaccination. The study regarding in-office teeth whitening aimed to assess the strength of recommendation for clinical trials in aesthetic care. The topic of sleep apnea also evaluated the strength of recommendation in the field of curative medicine, and systematic reviews were evaluated rather than clinical trials. For vaccinations, we assessed the strength of clinical recommendation of systematic reviews in preventive care.

The PICO question in each topic formed the basis for the inclusion and exclusion criteria in the literature search with various electronic databases (The National Library of Medicine (Pubmed), MEDLINE, the Cochrane Library, Google Scholar, the American Dental Association (ADA) research database, Embase, Ovid SP, Bandolier, and Web of Science). Manual search was used for additional articles not digitally posted. Gray literature and any articles that were not written in English were excluded. The remaining inclusion and exclusion criteria of the literature search were specific for the individual PICO question for each topic chosen.

Nine independent readers (three readers for the study of in-office teeth whitening, three readers for the study of sleep apnea, and three readers for the study of vaccination) were trained and standardized in order to eliminate any inconsistencies among the readers and to prevent any misinterpretations of the Ex-GRADE criteria. A preliminary trial of the

Table 1. Average Ex-GRADE Scores Across Three Independent Readers - In-Office Teeth Whitening Literature

Paper	1	2	3	4	5	6	7	8	Total
1	2.00	3.00	2.00	2.00	4.00	3.00	2.00	3.00	21.00
2	1.33	2.67	2.33	2.00	3.67	3.00	2.33	3.00	20.33
3	1.00	2.33	2.00	2.00	4.00	3.33	3.00	3.00	20.67
4	2.00	3.00	2.00	2.00	4.00	4.00	3.00	3.00	23.00
5	2.00	3.00	2.00	2.00	3.33	2.33	2.00	3.00	19.67
6	2.00	1.00	1.33	2.00	3.67	4.00	4.00	3.67	21.67
7	1.67	2.00	2.00	2.00	4.00	4.00	2.00	3.00	20.67
8	3.00	1.67	2.00	2.00	4.00	3.00	2.33	2.00	20.00
Mean	1.88	2.33	1.96	2.00	3.83	3.33	2.58	2.96	20.88
St. Dev	0.59	0.73	0.28	0.00	0.25	0.62	0.71	0.45	1.05

(Friedman non-parametric ANOVA equivalent, $p < 0.0001$)

Ex-GRADE scoring was performed to ensure that each of the nine readers read the literature critically and consistently. The trial was run with the R-Wong for the body of literature composed of primary sources, and with the R-AMSTAR for the body of literature composed of systematic reviews. Another test trial was conducted for the “Strength of Recommendation” portion of the Ex-GRADE for all papers, without discriminating between primary sources and systematic reviews. Any discrepancies in the scores were discussed until a consensus was attained for the manner in which they followed the Ex-GRADE scoring criteria. After all nine readers had been trained and their judgments were standardized, the reading and scoring of the articles were done independently among all of the readers. All readers were blind from one another’s scoring. An additional member of the research team compiled the data, averaged the scores (to obtain the means of scores), and analyzed the scores of the readers so the analyses and interpretations of the data are unbiased.

We used the Friedman test for non-parametric analysis of factorial designs using the Medical Data Analysis System (MDAS) software (EsKay Software, Pittsburg, 2004). Because the R-Wong and the R-AMSTAR are instruments that have already been validated, the analyses for scores received using the R-Wong or the R-AMSTAR were conducted according to the original methodology dictated by the creators. When evaluating primary sources that consisted of clinical trials, the scores were spanned across nine columns (corresponding to the nine questions of the R-Wong). However, if the primary sources did not contain clinical trials, the scores were inputted into the table using eight columns, as one question of the R-Wong only pertained to clinical trials [1]. When assessing the quality of evidence for systematic reviews using the R-AMSTAR, the scores were inserted across eleven columns to correspond to the test’s eleven domains [2]. For the “Strength of Recommendation” section, the means of scores among the three readers were inputted across eight columns in the table, each column corresponding to one of the eight questions of the Ex-GRADE. An alpha level (α) of 0.05 was used as the level of statistical significance. A high mean of scores and a low variance signi-

fied strength in that particular domain of clinical recommendation.

RESULTS

The R-Wong and R-AMSTAR have been previously validated [1, 2]. Therefore, the focus of this study was in validating the “strength of recommendation” portion of the Ex-GRADE, the second segment of the Ex-GRADE. It is important to keep in mind that the “strength of recommendation” assessment portion cannot be used in isolation. By looking at Fig. (1), note that either the R-Wong or the R-AMSTAR are used prior to assessing the strength of clinical recommendation, depending on the type of studies that are compiled (the R-Wong would be used for primary sources, and the R-AMSTAR would be used for systematic reviews).

In-Office Teeth Whitening

In our literature search to assess the strength of clinical recommendation regarding the use or disuse of light in in-office teeth whitening, we found 315 articles that fit our inclusion and exclusion criteria based on the PICO question. After filtering for articles that compared in-office bleaching with light and in-office bleaching without light, we had a total of 16 articles [8-23] in which we used in the validation of the Ex-GRADE.

The average scores for the in-office teeth whitening literature are shown in Table 1. Because three readers independently scored these articles, the inter-rater reliability was calculated between Reader 1 and Reader 2, between Reader 2 and Reader 3, and between Reader 1 and Reader 3. The inter-rater reliability (Pearson correlation coefficient, r) among the readers were 0.86, 0.92, and 0.91, respectively. The mean inter-rater reliability is 0.89, with a shared variance of 80% ($r^2=0.80$), indicating that the scores were indeed correlated.

As shown in Table 1, all of the Ex-GRADE scores given to the clinical trials regarding in-office bleaching fell within the 95% confidence interval of the sample (mean \pm standard deviation: 20.88 ± 1.05 ; 95% confidence interval (CI): 19.82-21.93) except for two papers. The two papers that did

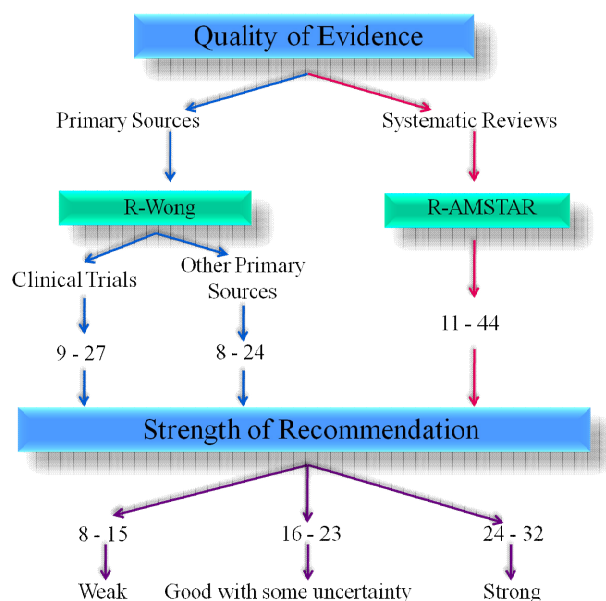


Fig. (1). Ex-GRADE.

not fall within this interval were Paper 4 (total score = 23.00) and Paper 5 (total score = 19.67). Based on the scoring criteria presented in Fig. (1), all eight papers scored within the 16- 23 point range for “good with some uncertainty.” As a translation, the scores show that for all of these papers, the intervention proposed within these bodies of literature is recommended for both the clinician, who executes the intervention (i.e., clinical relevance), and for the patient (i.e., increased patient satisfaction, increased health literacy, increased empowerment to be active participant in the treatment decision-making process).

With $p < 0.0001$, a significant difference is present in the scores among the eight domains of clinical recommendation. Table 1 shows that question 1 (the quality of evidence) and question 3 (alternative recommendation) have low mean values (1.88 ± 0.59 and 1.96 ± 0.28 , respectively), indicating that these two domains are relatively weak among these papers. With a mean score of 2.00 ± 0.00 , the strength of the domain in question 4 (availability of resources) is relatively weak, yet uniform, among this body of literature. The overall

scores in these three domains (represented by questions 1, 3 and 4) are lower compared to the other domains (represented by questions 2, 5, 6, 7, and 8), resulting in a significant difference overall among the eight domains.

Sleep Apnea

Due to the limitations of the available, published literature, no systematic reviews were found that compared patients’ compliance using continuous positive airway pressure (CPAP) versus oral appliances. As a result, we were restricted to systematic reviews that evaluated the use of CPAP (independent of oral appliances) and systematic reviews that evaluated the use of oral appliances (independent of CPAP). We had the option of using primary resources for this research topic, but systematic reviews as seen in Fig. (2) have a much higher level of evidence compared to primary sources (e.g. clinical trials) [24]. In our search for systematic reviews regarding sleep apnea, we initially found 16 systematic reviews for CPAP compliance and 4 systematic reviews for oral appliance compliance. After formulating our PICO question comparing whether CPAP or oral appliance treatment had a higher patient compliance, we excluded 10 CPAP compliance articles and 3 oral appliance compliance articles that either did not pertain to our research question or were not indeed systematic reviews, resulting in a total of 6 systematic reviews for analysis of CPAP compliance [25-30] and 1 for oral appliance compliance [31]. Table 2 and Table 3 show the average scores among the second set of three readers for the sleep apnea literature. For the scoring, two independent readers scored the 6 systematic reviews regarding CPAP use, and two independent readers scored the 1 systematic review regarding oral appliance use. The inter-rater reliability between Reader 4 and Reader 5 is 0.95 ($r^2 = 0.90$), and the inter-rater reliability between Reader 4 and Reader 6 is 0.96 (93% shared variance).

The analyses of articles were conducted in the same manner in which the analyses of the in-office whitening literature were conducted. Half of the articles for the CPAP compliance literature (paper 1, 2, and 5) fell within the 95% confidence interval of the average total means of the six papers as a whole (24.33 ± 2.96). The remaining three papers (paper 3, 4 and 6) received scores that fell outside of this confidence interval (total score of 28, 28, and 21, respectively). Using the Ex-GRADE scoring guideline as presented

Table 2. Average Scores Across Two Independent Readers – CPAP Compliance Literature

Papers	1	2	3	4	5	6	7	8	Total
1	1.00	4.00	2.50	2.00	3.00	2.00	4.00	4.00	22.50
2	3.50	2.50	3.00	2.00	3.50	3.00	3.00	3.00	23.50
3	4.00	3.00	3.00	3.00	3.00	4.00	4.00	4.00	28.00
4	4.00	3.00	3.00	2.00	4.00	4.00	4.00	4.00	28.00
5	1.00	3.00	4.00	2.00	3.00	2.00	4.00	4.00	23.00
6	1.00	3.00	3.00	2.00	3.00	2.00	4.00	3.00	21.00
Mean	2.42	3.08	3.08	2.17	3.25	2.83	3.83	3.67	24.33
St. Dev	1.56	0.49	0.49	0.41	0.42	0.98	0.41	0.52	2.96

(Friedman non-parametric ANOVA equivalent, $p = 0.1558$)

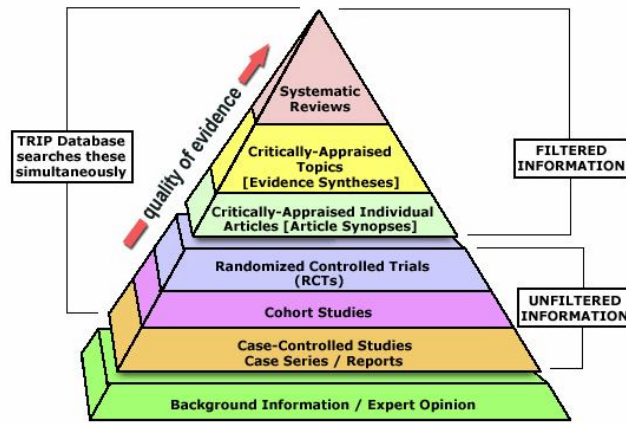


Fig. (2). Levels of quality of evidence pyramid.

in Fig. (1), papers 1, 5, and 6 were given strengths of recommendation that were “good with some uncertainty,” whereas papers 2, 3, and 4 received “strong” strengths of recommendation. Table 3 shows the average scores given to the article that focused on oral appliance compliance, having a total mean of 23.50 (mean score across questions = 2.94; standard deviation = 1.15). By receiving a total average score of 23.50 (which rounds to 24), this intervention given in this article qualifies as a “strong” recommendation for both the clinician and the patient. Because only one systematic review was found regarding patients’ compliance using oral appliances, it is difficult to formulate a general conclusion of the eight domains due to the low number of available literature.

The Friedman non-parametric p-value for both bodies of literature did not show a statistical significance across the eight domains of the Ex-GRADE (CPAP: $p = 0.1558$; oral appliance: $p = 0.5014$). In the CPAP compliance literature, question 7 (clinical significance) and 8 (patient compliance) received the highest mean scores using the Ex-GRADE crite-

ria (question 7: 3.83 ± 0.41 ; question 8: 3.67 ± 0.52), indicating that these domains were strong throughout the six papers. For the systematic review related to oral appliance, question 4 (availability of resources), 7 (clinical significance), and 8 (patient compliance) received a score of 4, indicating that these domains were well represented and addressed in this paper.

Vaccination

Our initial search for systematic reviews that investigated whether vaccinations led to the development of autism yielded 15 articles. Among the 15 articles, 11 of the articles were excluded because the articles were either reviews instead of systematic reviews, or because they did not assess the direct correlation of vaccination and autism. As a result, 4 articles [32-35] were used in our validation of the “strength of recommendation” section of the Ex-GRADE.

Table 4 contains the average scores given for the vaccination literature, with an inter-rater reliability of 0.91 between Reader 7 and Reader 8, an inter-rater reliability of 0.94 between Reader 8 and Reader 9, and an inter-rater reliability of 0.92 between Reader 7 and Reader 9. The average inter-rater reliability among the three readers is 0.92 (85% shared variance).

With an Ex-GRADE score of 21.33 (paper 1) and an Ex-GRADE score of 20.67 (paper 2), these two papers fell within the 95% confidence interval (21.08 ± 1.26). However, paper 3 received a score of 22.67 and paper 4 received a score of 19.67, lying outside of this confidence interval. Using the scoring rubric for the Ex-GRADE, all four papers qualify to have strengths of recommendation that are “good with some uncertainty.” This indicates that the interventions suggested in these articles are recommended for both the clinician and the patient. However, more research must be conducted in order to have a strong recommendation for these interventions.

A Friedman non-parametric p-value of 0.0027 signifies that there is statistical significance among the eight domains

Table 3. Average Scores Across Two Independent Readers – Oral Appliance Compliance Literature

Papers	1	2	3	4	5	6	7	8	Total	Mean	St. Dev
1	1.50	3.00	3.00	1.00	4.00	3.00	4.00	4.00	23.50	2.94	1.15
(Friedman non-parametric ANOVA equivalent, $p = 0.5014$)											

Table 4. Average Ex-GRADE Scores Across Three Independent Readers – Vaccination Literature

Papers	1	2	3	4	5	6	7	8	Total
1	1.00	3.00	3.00	1.67	3.00	3.00	2.67	4.00	21.33
2	1.00	3.00	2.67	2.00	3.00	2.00	3.00	4.00	20.67
3	1.00	3.00	3.00	2.00	3.67	3.00	3.00	4.00	22.67
4	1.00	2.00	3.00	2.00	3.33	2.00	3.00	3.33	19.67
Mean	1.00	2.75	2.92	1.92	3.25	2.50	2.92	3.83	21.08
St. Dev	0.00	0.50	0.17	0.17	0.32	0.58	0.17	0.33	1.26
(Friedman non-parametric ANOVA equivalent, $p = 0.0027$)									

of the Ex-GRADE with respect to these four systematic reviews. In a situation where the score for the Ex-GRADE is weak for various papers, those papers would be disregarded from the final analysis overall recommendation for the intervention. By disregarding low-scoring papers, we would be able to concentrate on the results from the high-quality papers, which essentially lead to a stronger recommendation. The score for Question 1 (the quality of evidence) is extremely low, with a mean of 1.00 ± 0.00 , suggesting that the quality of the evidence among all of these four papers is weak. The quality of evidence refers to the scores received using either the R-Wong [1] or the R-AMSTAR [2], so in this body of literature, a quality of the systematic reviews is weak. The mean of the scores for question 4 (availability of resources) is also low (1.92 ± 0.17) among these papers, resulting in a significant difference from the other domains which received higher scores overall (represented by questions 2, 3, 5, 6, 7, and 8).

DISCUSSION

With these collected data as provided in Tables 1-4, the Ex-GRADE does provide sufficient quantification in evaluating the quality of the evidence and the strength of the clinical recommendation. Not only does the Ex-GRADE give insight into the efficacy of the proposed intervention through the incorporation of the R-Wong [1] (which evaluates the quality of evidence of primary sources) and the R-AMSTAR [2] (which evaluates the quality of evidence of systematic reviews), it also addresses the effectiveness of the intervention through the “strength of recommendation” section of the Ex-GRADE (**Appendix 1**). The scoring intervals that signifies whether a body of literature is “weak” in terms of clinical recommendation, “good with some uncertainty” in the clinical recommendation, or “strong” clinical recommendation is equally divided among these three recommendations, where a paper receiving less than one-third of the possible points for the strength of recommendation portion (8-15 points) would be designated as “weak,” a paper that receives a score that falls within the middle third (16 -23 points) would be considered “good with some uncertainty,” and a paper that receives a score in the upper third would have a “strong” strength of recommendation. This same point distribution can also apply when an average of the scores are taken from multiple readers. In the case where a score falls between two intervals (e.g., 23.2), the score is rounded to the nearest integer.

The criteria for the strength of clinical recommendation were content-validated by several clinicians in the health care field, ensuring that these are the basis in which a recommendation is given for a particular intervention. In the clinical setting, the recommendation is not based solely on the efficacy of the intervention, but also takes into account the risk associated with the intervention, the total cost (including monetary cost), the availability of resources, and patient compliance. The clinician has a responsibility to provide their patients with the best treatments [5] and should provide alternative recommendations for the intervention in the case that the proposed intervention is unattainable or undesirable for the patient and/or clinician.

There is a terminology in the “strength of recommendation” portion of the Ex-GRADE that we want to define for clarity purposes. Question 7 asks if the results are “*clinically significant*.” Clinical significance is essentially determined by two determinants: the nature of the benefit (if the benefit is tangible or intangible), and whether this benefit is likely to be attainable. A tangible benefit can be realized in a person’s mind (e.g. a patient’s well-being or feeling). An intangible benefit is one that the individual’s mind does not consciously realize (e.g. changes in enamel mineralization level) [36]. The provided criteria that are given in Question 7 can be fulfilled by either types of benefits (tangible or intangible). The importance of this domain is to ensure that the interventions that are recommended to the patients are essentially worthwhile.

The Ex-GRADE expands the original GRADE to provide quantification of the quality of evidence and strength of recommendation of the body of literature, partaking in the final stages of evidence-based practice [37]. This quantification allows us to critically assess the validity and reliability of the evidence that is provided in available literature - an essential component of evidence-based research [38]. As a result, the specific strengths and weaknesses of the study can be critically evaluated in a single study, and the results can also be compared to the strengths and weaknesses of other bodies of literature. The total score received by the primary source or systematic review helps the patient or clinician to gauge whether the intervention or recommendation proposed in the paper is viable, or whether the individual should look into other alternatives in the domain of curative, palliative, or preventive care. In essence, the Ex-GRADE is an instrument that aims to bridge the gap between evidence-based research and evidence-based clinical practice [7], hopefully unifying these two branches and providing a means of communication between those providing insight from a research-based perspective and those physically providing the services for the intervention in the clinical setting.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGEMENTS

We would like to thank our mentor whom we believe to be one of the most influential individual and driving force in the world of Oral Biology and Evidence-Based Dentistry (EBD). The establishment of the Ex-GRADE and the foundation of EBD would not have been possible without his guidance. We would also like to thank all of those in our research team, consisting of dental clinicians, graduate students, and undergraduate pre-dental and pre-medical students who have dedicated their time and effort in expanding the domain of evidence-based research and evidence-based clinical practice throughout the years. No funding or grants have been received, and no conflict of interest is present in this study. The authors warmly thank the Evidence-Based Decisions Active Groups of Stakeholders (EBD-AGS) of the EBD-Practice-Based Research Network and the EBD Study Group for the invaluable critical contributions to this work.

APPENDIX 1

The “Strength of Recommendation” section of the Ex-GRADE (Expansion of the Grading of Recommendation Assessment, Development, and Evaluation) is graded on a point-based system, with 1 being the lowest score possible per question and 4 being the highest score possible per question. With a total of 8 questions, the minimum total score possible a primary source or systematic review will receive is 8 & the maximum total score possible is 32.

1. Are the findings and quality of evidence of the study applicable to the specific recommendation? (Score from “quality of evidence” section. *E.g. A clinical trial that receives a score of 23 using the R-Wong scale would fulfill 2 criteria: 1 criterion for scoring at least 19 & 1 criterion for scoring at least 22 → hence will receive a score of 3 using the Ex-GRADE according to the grading scale below.*)

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

For Primary Sources

R-Wong score of at least 25 for clinical trials OR at least 22 for all other primary sources
R-Wong score of at least 22 for clinical trials OR at least 20 for all other primary sources
R-Wong score of at least 19 for clinical trials OR at least 17 for all other primary sources

For Systematic Reviews:

Systematic Reviews – R-AMSTAR score of at least 40 for systematic reviews
Systematic Reviews – R-AMSTAR score of at least 36 for systematic reviews
Systematic Reviews – R-AMSTAR score of at least 31 for systematic reviews

2. Are risk and affordability considered when given the recommendation for the intervention?

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Recognition of risk for the intervention is directly stated, or acknowledgement of risk can be inferred
Recognition of possible adverse effects post-intervention is directly stated, or acknowledgement of possible adverse effects post-intervention can be inferred
Recognition of cost for the intervention is directly stated, or approximate and/or relative cost for the intervention can be inferred
Recognition of affordability is directly stated or can be inferred

3. Are alternative recommendations given, if appropriate?

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Alternative suggestions or recommendations were given with regards to risk during the intervention
--

Alternative suggestions or recommendations were given with regards to possible adverse effects following the intervention
Alternative suggestions or recommendations were given with regards to cost & affordability
Explicitly states that no alternative recommendations are appropriate with regards to risk during the intervention
Explicitly states that no alternative recommendations are appropriate with regards to possible adverse effects following the intervention
Explicitly states that no alternative recommendations are appropriate with regards to cost & affordability

4. Is availability of resources for the population of interest taken into account prior to formulating the recommendation? [Is the recommendation practical for the population of interest?]

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Insurance coverage is available for the recommended intervention at hand [<i>Some research on various insurance plans may need to be done</i>]
Other alternative funding aside from insurance is available for the recommended intervention at hand [<i>Some research for alternative funding may need to be done</i>]
Resources in terms of equipment & supplies for the recommendation are easily accessible in clinical practice [<i>This may require some prior knowledge of the equipments & supplies provided in the standard setting of the population of interest</i>]

5. Is a measurable guideline provided to monitor the intended outcome(s) of the recommendation? [Was there a method provided that can measure the effectiveness of the recommendations? How did they/will they measure the outcomes or results?]

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Method of monitoring the intended outcome of the recommendation is given
Method of monitoring the intended outcome can produce tangible data for the researcher
Method of analyzing the data produced from monitoring the intended outcome is provided

6. Are the results of the intervention statistically significant?

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Chosen methodology of the research is appropriate for the intended recommendation at hand
Methodology of the research (e.g. methodology of the clinical trial, methodology of the systematic review, etc.) is executed properly & accurately
Statistical analysis of the data shows statistical significance with $p < 0.05$

7. Are the results clinically significant?

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

For curative medicine/care, palliative medicine/care, or aesthetic/cosmetic care:

The intervention alters the pathophysiology of the disease/issue in question
The intervention can be realistically carried out & successfully executed in the clinical setting
The time it takes for noticeable results to be seen post-intervention is reasonable taking into consideration the total cost of the intervention (Cost = monetary expenses & risk, both during the intervention & post-intervention)

For preventive medicine/care:

The intervention does not alter the pathophysiology of the disease/issue in question
The intervention does not induce another pathology aside from the disease/issue in question
The intervention can be realistically carried out & successfully executed in the clinical setting

8. Is the patient likely to comply with the suggested recommendation?

- Fulfills 3 of the criteria → 4
- Fulfills 2 of the criteria → 3
- Fulfills 1 of the criteria → 2
- Fulfills 0 of the criteria → 1

CRITERIA:

Minimal level of invasiveness to the patient
Minimal level of side effects after the given intervention
Benefits of the recommendation outweigh its total cost (Cost = monetary expenses & risk, both during the intervention & post-intervention)

REFERENCES

[1] Chiappelli F, Navarro A, Moradi D, Manfrini E, Prolo P. Evidence-based research in complementary and alternative medicine III: treatment of patients with alzheimer’s disease. *Evid Based Complement Alternat Med* 2006; 3: 411-24.

[2] Kung J, Chiappelli F, Cajulis O, et al. From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (RAMSTAR) for grading of clinical relevance. *Open Dent J* 2010; 4: 84-91.

[3] Chiappelli F, Cajulis O, Newman M. Comparative effectiveness research in evidence-based dental practice. *J Evid Based Dent Pract* 2009; 9: 57-8.

[4] Atkins D, Best D, Briss P, et al. Education and debate: grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1-8.

[5] Bauer J, Spackman S, Chiappelli F, Prolo P. Model of evidence-based dental decision making. *J Evid Based Dent Pract* 2005; 5:189-97.

[6] Chiappelli F, Neagos N, Lee A, et al. Tools and methods for evidence-based research in dental practice: preparing the future. *J Evid Based Dent Pract* 2004; 4:16-23.

[7] Ajaj RA, Barkhordarian A, Phi L, Giroux A, Chiappelli F. Evidence-based dentistry: the next frontier in translational and transnational dental practice. *Dent Hypotheses* 2011; 2: 55-62.

[8] Bernardon J, Sartori N, Ballarin A, Perdigao J, Lopes GC, Baratieri L. Clinical performance of vital bleaching techniques. *Oper Dent* 2010; 35 (1): 3-10.

[9] Alomari Q, El Daraa E. A randomized clinical trial of in-office dental bleaching with or without light activation. *J Contemp Dent Pract* 2010; 11(1): E017-24.

[10] Kugel G, Ferreira S, Sharma S, Barker ML, Gerlach RW. Clinical trial assessing light enhancement of in-office tooth whitening. *J Esthet Restor Dent* 2009; 21(5): 336-47.

[11] Ontiveros JC, Paravina RD. Color change of vital teeth exposed to bleaching performed. *J Dent* 2009; 37: 840-7.

[12] Gurgan S, Cakir FY, Yazici E. Different light-activated in-office bleaching systems. *Lasers Med Sci* 2010; 25: 817-22.

[13] Polydorou O, Hellwig E, Hahn P. The efficacy of three different in-office bleaching systems. *Oper Dent* 2008; 33(5): 579-86.

[14] Marson FC, Sensi LG, Vieira LC, Araújo E. Clinical evaluation of in-office dental bleaching treatments with and without the use of light-activation sources. *Oper Dent* 2008; 33(1): 15-22.

[15] Giniger M, MacDonald J, Ziemba SL, et al. Clinical evaluation of a novel dental whitening lamp and light-catalyzed peroxide gel. *J Clin Dent* 2005; 16: 123-7.

[16] Kugel G, Papathanasiou A, Williams AJ 3rd, Anderson C, Ferreira S. Clinical evaluation of chemical and light-activated tooth whitening systems. *Compend Contin Educ Dent* 2006; 27 (1): 54-62.

[17] Luk K, Tam L, Hubert M. Effect of light energy on peroxide tooth bleaching. *J Am Dent Assoc* 2004; 135: 194-201.

[18] Papathanasiou A, Kastali S, Perry RD, Kugel G. Clinical evaluation of a 35% hydrogen peroxide in-office whitening system. *Comp Cont Dent Educ* 2002; 23 (4): 335-46.

[19] Tavares M, Stultz J, Newman M, et al. Light augments tooth whitening with peroxide. *J Am Dent Assoc* 2003; 134: 167-75.

- [20] Oteo J, Oteo C, Oteo A, Calvo MJ. Clinical efficacy of a bleaching system based on hydrogen peroxide with or without light activation. *Eur J Esthet Dent* 2010; 5(2): 216-24.
- [21] Whitman FJ, Simon JF. A clinical comparison of two bleaching systems. *J Calif Dent Assoc* 1995; 23 (12): 59-64.
- [22] Sulieman M, MacDonald E, Rees JS, Addy M. Comparison of three in-office bleaching systems. *Am J Dent* 2005; 18 (3): 194-6.
- [23] Hein DK, Ploeger BJ, Hartup JK, Waqstaff RS, Palmer TM, Hansen LD. In-office vital teeth bleaching - what do lights add. *Compend Cont Edu Dent* 2003; 24 (4A): 340-52.
- [24] Yale University School of Medicine [Accessed 11/3/2011]. Available from <http://www.ebmpyramid.org/samples/complicated.html>
- [25] Weaver TE, Sawyer AM. Adherence to continuous positive airway pressure treatment for obstructive sleep apnoea: implications for future interventions. *Indian J Med Res* 2010; 131: 245-58.
- [26] Smith I, Lasserson TJ. Pressure modification for improving usage of continuous positive airway pressure machines in adults with obstructive sleep apnoea. *Cochrane Database System Rev* 2009, Issue 4. Art. No.: CD003531.
- [27] Smith I. Interventions to improve compliance with continuous positive airway pressure for obstructive sleep apnoea. *J Clin Sleep Med* 2007; 3(7): 706-12.
- [28] Smith I, Nadig V, Lasserson TJ. Educational, supportive and behavioural interventions to improve usage of continuous positive airway pressure machines for adults with obstructive sleep apnoea. *Cochrane Database Sys Rev* 2009, Issue 2. Art. No.: CD007736.
- [29] Kakkur R, Berry R. Positive airway pressure treatment for obstructive sleep apnea. *Chest* 2007; 132: 1057-72.
- [30] Gay P, Weaver T, Loube D, *et al.* Evaluation of positive airway pressure treatment for sleep related breathing disorders in adults. *Sleep* 2006; 29(3): 381-401.
- [31] Ferguson KA, Cartwright R, Rogers R, *et al.* Oral appliances for snoring and obstructive sleep apnea: a review. *Sleep* 2006; 29(2): 244-62.
- [32] Madsen K, Vestergaard M. MMR vaccination and autism: what is the evidence for a causal association? *Drug Saf* 2004; 27(12): 831-40.
- [33] Doja A, Roberts W. Immunizations and autism: a review of the literature. *Can J Neurologic Sci* 2006; 33(4): 341-6.
- [34] Wilson K, Mills E, Ross C, *et al.* Association of autistic spectrum disorder and the measles, mumps, and rubella vaccine: a systematic review of current epidemiological evidence. *Arch Pediatr Adolesc Med* 2003; 157: 628-34.
- [35] Pursselle E. Exploring the evidence surrounding the debate on MMR and autism. *Br J Nurs* 2004; 13(14): 834-8.
- [36] Hujuel P. Levels of Clinical Significance. *J Evid Based Dent Pract* 2004; 4: 32-6.
- [37] Chiappelli F, Cajulis O. The logic model for evidence-based clinical decision making in dental practice. *J Evid Based Dent Pract* 2009; 9: 206-10.
- [38] Chiappelli F, Cajulis O. Transitioning toward evidence-based research in the health sciences for the XXI century. *Evid Based Complement Alternat Med* 2008; 5: 123-8.

Received: October 06, 2011

Revised: November 26, 2011

Accepted: November 28, 2011

© Phi *et al.*; Licensee Bentham Open.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.