Journal of Clinical Epidemiology 74 (2016) 7−18

Journal of Clinical Epidemiology

GRADE SERIES

Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format

Alonso Carrasco-Labra[a,b,c], Romina Brignardello-Petersen[c,d], Nancy Santesso[a,e,f], Ignacio Neumann[g], Reem A. Mustafa[a,h,i], Lawrence Mbuagbaw[a,j], Itziar Etxeandia Ikobaltzeta[a], Catherine De Stio[k], Lauren J. McCullagh[k], Pablo Alonso-Coello[a,l], Joerg J. Meerpohl[m], Per Olav Vandvik[n,o], Jan L. Brozek[a,e,f,p], Elie A. Akl[a,q], Patrick Bossuyt[r], Rachel Churchill[s], Claire Glenton[t,u], Sarah Rosenbaum[t,u], Peter Tugwell[v], Vivian Welch[w], Paul Garner[x], Gordon Guyatt[a,e,f,p], Holger J. Schünemann[a,e,f,p,*]

[a]*Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[b]*Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Universidad de Chile, Sergio Livingstone Pohlhammer 943, Independencia, Santiago, Chile*
[c]*Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Sergio Livingstone Pohlhammer 943, Independencia, Santiago, Chile*
[d]*Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, 4th Floor, Toronto, Ontario M5T 3M6, Canada*
[e]*Cochrane GRADEing (Applicability and Recommendations) Methods Group, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[f]*McMaster GRADE Center, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[g]*Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Región Metropolitana, Chile*
[h]*Department of Medicine/Nephrology, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108-2792, USA*
[i]*Department of Biomedical & Health Informatics, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108-2792, USA*
[j]*Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, 50 Charlton Avenue East, 3rd Floor Martha Wing, Room H321, Hamilton, Ontario L8N 4A6, Canada*
[k]*Hofstra North Shore LIJ School of Medicine, Department of Medicine, 300 Community Drive, Manhasset, NY 11030, USA*
[l]*Iberoamerican Cochrane Centre, Biomedical Research Institute Sant Pau—CIBER of Epidemiology and Public Health (CIBERESP—IIB Sant Pau), C/ Sant Antoni Maria Claret 167, Pavelló 18, planta 0, 08025 Barcelona, Spain*
[m]*German Cochrane Centre, Medical Center—University of Freiburg, Berliner Allee 29, 79110 Freiburg, Germany*
[n]*Department of Medicine, Division Gjøvik, Innlandet Hospital Trust, Kyrre Greppsgt 11, 2819 Gjøvik, Norway*
[o]*Norwegian Knowledge Centre for the Health Services, Pilestredet Park 7, 0130 Oslo, Norway*
[p]*Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[q]*Department of Internal Medicine, American University of Beirut, Riad-El-Solh Beirut 1107 2020, P.O. Box: 11-0236, Beirut, Lebanon*
[r]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, Room J2-127, PO Box 22700, 1100 DE Amsterdam, The Netherlands*
[s]*Centre for Academic Mental Health, School of Social and Community Medicine, University of Bristol, Office Barley House, Room BF11, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK*
[t]*Norwegian Branch of the Nordic Cochrane Centre, Postboks 7004 Street Olavs plass, 0130 Oslo, Norway*
[u]*Norwegian Knowledge Centre for the Health Services, P.O. Box 7004 Street Olavs plass, N-0130 Oslo, Norway*
[v]*Department of Medicine, Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada*
[w]*Bruyère Research Institute, University of Ottawa, 304b—85 Primrose Avenue, Ottawa, Ontario, Canada*
[x]*Evidence Synthesis for Global Health, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK*

Accepted 21 December 2015; Published online 11 January 2016

Funding: The Cochrane Methods Innovation Fund and GRADE Center at McMaster University funded this study. Neither of these institutions but only the investigators belonging to these institutions played a role in the planning, conducting, or publishing the study findings.

Conflict of interest: The authors of this trial declare no financial conflict of interest. However, most of them are members of the GRADE working group and the Cochrane Collaboration. H.J.S., G.G., and P.T. are convenors of Cochrane Methods Group. The views expressed in this article are those of the authors and not necessarily those of the Cochrane Collaboration or its registered entities, committees, or working groups.

Trial registration: NCT02022631 (clinicaltrials.gov).

* Corresponding author. Tel.: +1-905-525-9140x24931; fax: +1-905-522-9507.

E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

http://dx.doi.org/10.1016/j.jclinepi.2015.12.007

Abstract

Objectives: The current format of summary of findings (SoFs) tables for presenting effect estimates and associated quality of evidence improve understanding and assist users finding key information in systematic reviews. Users of SoF tables have demanded alternative formats to express findings from systematic reviews.

Study Design and Setting: We conducted a randomized controlled trial among systematic review users to compare the relative merits of a new format with the current formats of SoF tables regarding understanding, accessibility of information, satisfaction, and preference. Our primary goal was to show that the new format is not inferior to the current format.

Results: Of 390 potentially eligible subjects, 290 were randomized. Of seven items testing understanding, three showed similar results, two showed small differences favoring the new format, and two (understanding risk difference and quality of the evidence associated with a treatment effect) showed large differences favoring the new format [63% (95% confidence interval {CI}: 55, 71) and 62% (95% CI: 52, 71) more correct answers, respectively]. Respondents rated information in the alternative format as more accessible overall and preferred the new format over the current format.

Conclusions: While providing at least similar levels of understanding for some items and increased understanding for others, users prefer the new format of SoF tables.

Keywords: Summary of findings table; GRADE; Evidence summaries; Understanding; Formatting; GRADEpro; Evidence tables; Systematic reviews; Guidelines

1. Background

The "Grading of Recommendations Assessment, Development and Evaluation" (GRADE) approach [1−6] provides a structured and transparent framework to assess the quality of the evidence, also known as certainty in the evidence or confidence in the estimates of effect in systematic reviews, and the strength of recommendations in health care recommendations [7]. To facilitate the presentation of review results, for example, effect estimates and the quality of the evidence, the GRADE approach proposes the use of "summary of findings" (SoFs) tables and GRADE evidence profiles [8,9].

The overall configuration and presentation of review results in SoF tables, and evidence profiles resulted from broad-scale user-testing and stakeholder consultation as well as evidence from systematic reviews focusing on the presentation of numerical information [10−12]. Previous studies have shown that the inclusion of SoF tables in systematic reviews significantly improved readers' overall understanding [93% vs. 44% ($P = 0.003$)] along with their ability to find critical information [68% vs. 40% ($P = 0.021$)] compared to having the data only in the main text [12]. A randomized controlled trial reported that formatting modifications of GRADE evidence profiles could increase the comprehension of key findings between 5% and 47% [13].

The Cochrane Collaboration, committed to synthesize, translate, and facilitate the use of research data to inform clinical practice, has been implementing SoF tables since 2004. One limitation for implementation across review groups is the limited options currently offered to SoF table developers to display review results. For example, the current standard SoF table format does not include the option of displaying risk difference or number needed to treat. These alternative presentations of information had been requested by systematic review authors and editorial groups. The inclusion of empirically tested alternative presentations of risks and other items in SoF tables would allow authors to choose from a variety of formats; some formats may prove superior to those currently available, others may be as good as the current ones. We therefore compared the performance of several items between the currently existing and a new format of a SoF table.

2. Methods

The reporting of this study followed the latest guidance by the Consolidated Standards of Reporting Trials in its extension for reporting of noninferiority and equivalence randomized trials [14]. The protocol of the trial was registered in clinicaltrials.gov (Trial registration: NCT02022631) and published elsewhere [15].

2.1. Participants

2.1.1. Selection criteria

Participants were eligible if they considered themselves as systematic review users. We defined a user as someone who had used the Cochrane library or downloaded Cochrane or non-Cochrane systematic reviews at least twice a year to answer clinical practice questions, to inform the process of making recommendations for clinical practice guidelines, or to use reviews results for research purposes. We targeted three types of users: (1) health professionals working in primary, secondary, or tertiary care; (2) clinical practice guidelines developers; and (3) researchers. In this study, we classified as clinicians those who reported at least 50% of total time dedicated to clinical practice. To be considered clinical practice guideline developers, participants were required to have participated in the development of at least one clinical practice guideline during the last 2 years. To be considered researchers, participants were

What is new?

Key findings

- Compared to the standard summary of findings (SoFs) table, a new format of SoF tables with seven alternative items improved understanding of risk differences and helped with interpreting results and was similar to the current SoF table regarding other items in the understanding domain. In addition, the new format was more accessible and preferred by users. The results of this study also provide evidence of the potential effectiveness of the use of standardized narrative descriptions of review results.

What this adds to what was known?

- Grading of Recommendations Assessment, Development and Evaluation SoF tables have been developed to display effect estimates and the associated quality of evidence from systematic reviews in a concise and transparent manner. The current format of the tables for presenting effect estimates and quality of evidence improves understanding and assists users with finding key information from the systematic review. We found that a new alternative format increase users' flexibility on presenting and summarizing review results while maintaining or improving understanding.

What is the implication and what should change now?

- Systematic review authors can now decide on which type of format to include as SoF table to fit their audiences' needs better. In making these choices, they should bear in mind the unequivocal finding that presenting risk differences improved understanding and accessibility in this randomized trial.

required to be dedicating more than 70% of their time to conduct research (e.g., methodologists, epidemiologists, statisticians, and so forth).

2.1.2. Setting and recruitment

We recruited participants from Europe, North America, South America, and Asia. To contact participants, we used various networks: Cochrane groups and the networks of coauthors who interact with guideline developers, researchers, and systematic reviewers and attendees to workshops, conferences, and other research events. Potentially eligible participants received an invitation via e-mail along with a link to access an online questionnaire. Using this online system, we further determined participant eligibility and obtained informed consent. The Hamilton Health Sciences/Faculty of Health Sciences Research Ethics Board at McMaster University reviewed a summary of the study protocol. They classified this trial as a quality improvement study and waived the requirement for formal approval and individual consent beyond agreeing to participate.

2.2. Intervention and comparison

We compared a new format of a SoF table (Table 1) to the current SoF table (Table 2). The clinical question, patients and setting, intervention, comparator, outcomes, and the complementary information included in the explanatory footnotes were the same in both tables. We used a SoF table from a Cochrane systematic review entitled "Probiotics for the prevention of pediatric antibiotic-associated diarrhea," with only minimum modifications from the original version [16].

The differences between the current and new format of the SoF table were the methods to either show the same data in a different way or to provide supplementary data (e.g., supplementary data as risk difference). The current SoF table (Table 2) was based on a prior trial that showed the impact of the current SoF tables on understanding and accessibility of information of systematic reviews [12]. Table 3 presents the items we compared, a primary interest was to test the impact of including the risk difference which is absent from the current format. This absence was questioned by the lead investigators of this trial and led to intense discussions in GRADE working group meetings and a prior trial comparing the presence of risk differences in GRADE evidence profiles [13]. In addition, we conducted user testing and had extensive discussions in the author team that revealed other items for comparison which resulted in the items shown in Table 3.

2.2.1. Randomization

After completing background information, participants who met the inclusion criteria were stratified as clinician, guideline developer, or researcher according to self-classification. If participants classified themselves in more than one category, we asked them to indicate the profile that represents them the best. We then randomly allocated them to one of the two SoF tables in a 1:1 ratio via the "Survey Monkey" platform. The randomization scheme was automatically generated by the platform. When direct comparison between the new and current format was required, the order in which the tables were shown to participants was randomly determined.

2.2.2. Concealment of allocation

The allocation of participants to the tables was done by the "Survey Monkey" system in real time following an algorithm unknown to us, without a prespecified sequence.

Table 1. New SoF table format (Table A)

Outcomes, no of participants (studies)	Relative effects (95% CI)	Anticipated absolute effects[a] (95% CI)			Quality of the evidence (GRADE)	What happens
		Without probiotics	With probiotics	Difference		
Incidence of diarrhea: probiotic dose 5 billion CFU/d Follow-up: 10 days to 3 months Children <5 years 1,474 (7 studies) Children >5 years 624 (4 studies)	RR 0.4[b] (0.29 to 0.55) RR 0.8[b] (0.53 to 1.21)	22.3%[b] 11.2%[b]	Children <5 years 8.9% (6.5 to 12.2) Children >5 years 9% (5.9 to 13.6)	13.4% fewer children[b] (10.1 to 15.8 fewer) 2.2% fewer children[b] (5.3 fewer to 2.4 more)	⊕⊕⊕⊖ moderate[c] due to risk of bias ⊕⊕⊖⊖ low[c, d] due to risk of bias and imprecision	Probably decreases the incidence of diarrhea May decrease the incidence of diarrhea
Adverse events[e] Follow-up: 10 to 44 days 1,575 (11 studies)	—	1.8%[b]	2.3% (0.8 to 3.8)	0.5% more adverse events[f] (1 fewer to 2 more)	⊕⊕⊖⊖ low[g, h] due to risk of bias and inconsistency	There may be little or no difference in adverse events
Duration of diarrhea Follow-up: 10 days to 3 months 897 (5 studies)	—	The mean duration of diarrhea without probiotics was 4 days	—	0.6 fewer days (1.18 to 0.02 fewer days)	⊕⊕⊖⊖ low[i, j] due to imprecision and inconsistency	May decrease the duration of diarrhea
Stools per day Follow-up: 10 days to 3 months 425 (4 studies)	—	The mean stools per day without probiotics was 2.5 stools per day	—	0.3 fewer stools per day (0.6 to 0 fewer)	⊕⊕⊖⊖ low[k, l] due to imprecision and inconsistency	There may be little or no difference in stools per day

Abbreviations: CI, Confidence interval; RR, risk ratio; GRADE, Grading of Recommendations Assessment, Development and Evaluation.
Probiotics as an adjunct to antibiotics for the prevention of pediatric antibiotic-associated diarrhea in children.
Patient or population: children given antibiotics.
Settings: inpatients and outpatient.
Intervention: probiotics.
Comparison: no probiotics.

[a] The basis for the risk in the control group (e.g., the median control group risk across studies) is provided in footnotes. The risk in the intervention group (and its 95% confidence interval) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[b] Control group risk estimates come from pooled estimates of control groups. Relative effect based on available case analysis.

[c] High risk of bias due to high loss to follow-up.

[d] Imprecision due to few events and confidence intervals include appreciable benefit or harm.

[e] Side effects: rash, nausea, flatulence, vomiting, increased phlegm, chest pain, constipation, taste disturbance, and low appetite.

[f] Risks were calculated from pooled risk differences.

[g] High risk of bias. Only 11 of 16 trials reported on adverse events, suggesting a selective reporting bias.

[h] Serious inconsistency. Numerous probiotic agents and doses were evaluated among a relatively small number of trials, limiting our ability to draw conclusions on the safety of the many probiotics agents and doses administered.

[i] Serious unexplained inconsistency [large heterogeneity I^2 = 79%, *P*-value (P = 0.04), point estimates, and confidence intervals vary considerably].

[j] Serious imprecision. The upper bound of 0.02 fewer days of diarrhea is not considered patient important.

[k] Serious unexplained inconsistency [large heterogeneity I^2 = 78%, *P*-value (P = 0.05), point estimates, and confidence intervals vary considerably].

[l] Serious imprecision. The 95% confidence interval includes no effect, and lower bound of 0.60 stools per day is of questionable patient importance.

Thus, the investigators did not know in advance to which group the next participant was going to be allocated.

2.2.3. Data collection and blinding

The collection of data was done automatically by the "Survey Monkey" system. As a way to conceal the nature of the SoF tables to which participants were allocated, the tables were labeled as A or B, without any other information about their content or the study hypothesis. Participants were first exposed to one table—containing either the new or current format—and the outcomes understanding, accessibility of information, satisfaction, and preference were assessed. In the final phase of data collection, we assessed participants' preference for the new or the current format by showing the table to which they were not initially allocated. Once the data collection process was

Table 2. Current format of the SoF table (Table B)

Outcomes	Illustrative comparative risks[a] (95% CI)		Relative effect (95% CI)	No of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk	Corresponding risk				
	No probiotics	Probiotics				
Incidence of diarrhea: probiotic dose (equal to/greater than) 5 billion CFU/d Follow-up: 10 days to 3 months	Children <5 years 223 per 1,000[b]	89 per 1,000 (65 to 122)	RR 0.4[b] (0.29 to 0.55)	1,474 (7 studies)	⊕⊕⊕⊖ moderate[c]	
	Children >5 years 112 per 1,000[b]	90 per 1,000 (59 to 136)	RR 0.8[b] (0.53 to 1.21)	624 (4 studies)	⊕⊕⊖⊖ low[c, d]	
Adverse events Follow-up: 10 to 44 days	18 per 1,000[b]	23 per 1,000 (8 to 38)	Not estimable[e]	1,575 (11 studies)	⊕⊕⊖⊖ low[f, g]	Side effects: rash, nausea, gas, flatulence, vomiting, increased phlegm, chest pain, constipation, taste disturbance, and low appetite
Duration of diarrhea Follow-up: 10 days to 3 months	The mean duration of diarrhea in control groups was 4 days	0.6 fewer days (1.18 to 0.02 fewer days)		897 (5 studies)	⊕⊕⊖⊖ low[h, i]	
Stools per day Follow-up: 10 days to 3 months	The mean stools per day in control groups was 2.5 stools per day	0.3 fewer stools per day (0.6 to 0 fewer)		425 (4 studies)	⊕⊕⊖⊖ low[j, k]	

Abbreviations: CI, confidence interval; RR, risk ratio; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

Probiotics as an adjunct to antibiotics for the prevention of pediatric antibiotic-associated diarrhea in children.

Patient or population: children given antibiotics.

Settings: inpatients and outpatient.

Intervention: probiotics.

Comparison: no probiotics.

GRADE Working Group grades of evidence.

High quality: We are very confident that the true effect lies close to that of the estimate of the effect.

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.

Low quality: confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.

Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

[a] The basis for the assumed risk (e.g., the median control group risk across studies) is provided in footnotes. The corresponding risk (and its 95% confidence interval) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[b] Control group risk estimates come from pooled estimates of control groups. Relative effect based on available case analysis.

[c] High risk of bias due to high loss to follow-up.

[d] Imprecision due to few events and confidence intervals include appreciable benefit or harm.

[e] Risks were calculated from pooled risk differences.

[f] High risk of bias. Only 11 of 16 trials reported on adverse events, suggesting a selective reporting bias.

[g] Serious inconsistency. Numerous probiotic agents and doses were evaluated among a relatively small number of trials, limiting our ability to draw conclusions on the safety of the many probiotics agents and doses administered.

[h] Serious unexplained inconsistency [large heterogeneity I^2 = 79%, *P*-value (P = 0.04), point estimates, and confidence intervals vary considerably].

[i] Serious imprecision. The upper bound of 0.02 fewer days of diarrhea is not considered patient important.

[j] Serious unexplained inconsistency [large heterogeneity I^2 = 78%, *P*-value (P = 0.05), point estimates, and confidence intervals vary considerably].

[k] Serious imprecision. The 95% confidence interval includes no effect, and lower bound of 0.60 stools per day is of questionable patient importance.

completed, the database was prepared for statistical analysis in a blinded fashion.

2.3. Outcomes

We used similar outcomes (understanding, accessibility of information, satisfaction, and preference) to the ones measured in previous randomized controlled trials and other observational studies testing formats for SoF tables [11–13].

2.3.1. Primary outcome

2.3.1.1. Understanding. We defined understanding as the correct comprehension of key findings in the table. We presented participants with seven multiple-choice questions each of them with five response options, one of which

Table 3. Comparison between items included in the current and new SoF tables

	Current format (Table C)	New format (Table A)
1	Inclusion of the *N* of participants and studies column	Exclusion of the *N* of participants and studies column. Information presented in the outcomes column
2	Quality of evidence presented with symbols and labeled as high, moderate, low, or very low. Reasons for downgrading presented in the footnotes	Quality of evidence presented along with main reasons for downgrading in the same column (e.g., moderate due to imprecision)
3	"Footnotes" label	"Explanations" label
4	Baseline risk and corresponding risk expressed as natural frequencies	Baseline risk and corresponding risk expressed as percentages
5	No column presenting absolute risk reduction (risk difference) or mean difference	Inclusion of a column presenting absolute risk reduction (risk difference) or mean difference
6	Comments column included	Comments column deleted
7	No "what happens" column[a]	"What happens" column included[a]
8	Description of the GRADE working group grades of evidence definitions below the table	No description of the GRADE working group grades of evidence definitions

Abbreviations: SoF, summary of findings; GRADE, Grading of Recommendations Assessment, Development and Evaluation.
[a] The "what happens" column aims to summarize both the treatment effect and the quality of the evidence on one short narrative statement.

representing the correct answer. Each of these questions covered one alternative format under testing (Table 1). For analysis, we compared the proportion of correct answers between groups at a question level. We defined 10% as the noninferiority margin difference between groups, based on findings from previous studies that developed the Cochrane plain language summaries and the current items for SoF tables and evidence profiles [12,13,17].

2.3.2. Secondary outcomes

2.3.2.1. Accessibility of information. This outcome was composed of three self-reported domains: (1) how easy it was to find critical information in the table, (2) how easy it was to understand the information, and (3) whether the information was presented in a way that is helpful for decision making. They were measured by presenting participants' statements for which they had to indicate the degree of agreement: "It was easy to find the information about the effects," "It was easy to understand the information," and "The information is presented in a way that would help me making a decision," along with seven-point Likert scales (1 = I strongly disagree, 2 = I disagree, 3 = I somewhat disagree, 4 = Neither agree nor disagree, 5 = I somewhat agree, 6 = I agree, and 7 = I strongly agree). We also measured overall accessibility of information using a five-point Likert-type scale (1 = Very inaccessible, 2 = Inaccessible, 3 = Neither inaccessible nor accessible, 4 = Accessible, and 5 = Very accessible).

2.3.2.2. Satisfaction. We measured this dichotomous outcome (proportions per group) at an item level asking participants which formatting items satisfied them the most. For example, one question asked was as follow: "In Table A, we included a column called 'what happens.' The purpose of this column is to assist users on the interpretation of both review results and quality of the evidence." "Do you think this column should be included as an available feature in future versions of SoF tables?"

2.3.2.3. Preference. Using a seven-point Likert-type scale (1 = I strongly prefer Table A, 2 = I prefer Table A, 3 = I somewhat prefer Table A, 4 = Same preference for Table A or B, 5 = I somewhat prefer Table B, 6 = I prefer Table B, and 7 = I strongly prefer Table B), we presented participants with the following question: Between alternative (Table A) and current formats (Table B) for SoF tables, "which table do you prefer?"

2.4. Sample size calculation

Based on the primary outcome, the proportion of participants correctly answering questions about understanding in similar randomized controlled trials that tested the current SoF table format or GRADE evidence profiles ranged between 80% and 87% [12,13], and we expected, at least, the same percentage in the group of participants randomized to the new format of a SoF table. We defined a 10% noninferiority margin and an allocation of participants in a 1:1 ratio. If there was truly no difference between the current and new alternative table formats, then 280 participants were required to be 80% sure that the upper limit of a one-sided 95% confidence interval (CI) excludes a difference in favor of the current SoF table format of more than 10%. Assuming that around 10% of participants would not complete the questionnaire, we needed to recruit 308 participants.

2.5. Statistical analysis

2.5.1. Descriptive analysis

Descriptive analysis included participants' baseline characteristics and outcomes, means and mean difference (MD) standard deviations (SD) for continuous variables, and proportions for categorical variables.

2.5.2. Inferential analysis

For the primary outcome (understanding), we compared the two groups for the proportion of participants correctly answering each question separately. It was analyzed using multiple logistic regressions per question. Despite the correlation between the single questions and the key outcomes, we took a conservative approach to the analysis. To adjust for multiplicity, the *P*-value was adjusted using the Bonferroni correction for seven multiple comparisons [*P*-value to reject the null hypothesis: <0.0035, CIs were constructed with corresponding one-sided z score 2.7]. For the outcome accessibility of information, we compared the two groups for the mean answer and SDs along with 95% CIs for each of the three domains. To adjust for multiplicity here, the *P*-value was adjusted using the Bonferroni correction for four multiple comparisons (<0.0125). For the outcome "satisfaction," we present the proportion of participants satisfied with items included in Table A or Table B per group. Finally, for the outcome preference, we used linear regressions. We controlled for the order in which the tables were shown to the participants (dichotomous: 2 categories). For all models described here, we initially considered the following predictors: (1) participant strata (nominal: 3 categories), (2) years of experience (nominal: 5 categories), (3) familiarity with the GRADE approach (dichotomous: 2 categories), and (4) previous education in health research methodology or epidemiology (ordinal: 3 categories).

2.5.3. Evaluation of the models

The Harrell's method [18] was applied to define which predictors to include in the models. First, we ran the model including only one key predictor. Then, in an iterative process, each predictor was included along with the key predictor. If in any iteration a predictor changed the parameter estimate by more than 10%, we retained it in the model. The key predictor for the outcomes understanding, preference, and accessibility of information was the arm to which participants were allocated.

2.5.4. Claiming of noninferiority (CI approach)

We claimed noninferiority of the new items to the current standard items of the SoF tables regarding understanding when the upper limit of the CI was equal or lower than the noninferiority margin of 10%. We followed this approach because it is more informative as each question was related to a particular item tested in the tables. Superiority testing was applied only to the secondary outcomes although significant differences exceeding indicating improvement led us to infer superiority also for all items.

2.5.5. Dealing with dropouts and missing data

To reduce the likelihood of dropouts and missing data, we implemented the following strategies: (1) we sent only one link to participants, which included all the required questionnaires and material, to reduce multiple contacts and burden, (2) the online system randomized participants only after collecting all baseline characteristics, (3) responses to all questions were mandatory, and (4) we ensured that no more than 25 minutes were required for completion in total. If a participant was allocated to a study arm and did not complete all the questions (i.e., stopped early), we analyzed all participants for who the variables of interest were present (available case analysis).

3. Results

We sent more than 1,000 invitations and received 390 responses. Of those, 290 were randomized, 52% (151) were women, 63% (185) were between 36 and 55 years old, 42% (121) were native English speakers, 20% (58) Spanish, 12% (35) German, and 8% (23) Norwegian. In total, we included participants with more than 25 different primary languages. Although 49% (142) of the sample had formal training in epidemiology or research methods equivalent to a Master's or Doctoral degree, the remaining 51% (148) having either some formal training with no degree or no formal training at all. More than 58% (170) of the

Table 4. Baseline characteristic of participants per group

Characteristic	New format ($n = 122$)	Current format ($n = 168$)
Sex: *n* (%)		
Women	54 (44)	97 (57)
Men	68 (56)	71 (43)
Age group: *n* (%)		
<25	1 (1)	3 (2)
26–35	31 (25)	33 (20)
36–45	37 (30)	64 (38)
46–55	37 (30)	47 (28)
56–65	13 (11)	19 (11)
66–75<	3 (3)	2 (1)
Native language: *n* (%)		
English	50 (41)	71 (42)
French	5 (4)	1 (1)
German	8 (7)	27 (16)
Italian	6 (5)	3 (2)
Norwegian	10 (8)	13 (8)
Spanish	29 (24)	29 (17)
Other	14 (11)	24 (14)
Training in research methods or epidemiology: *n* (%)		
No formal training	20 (16)	24 (14)
Formal training, no degree	44 (36)	60 (36)
Formal training, MSc, PhD	58 (48)	84 (50)
Familiarity with GRADE: *n* (%)		
Not familiar at all	4 (3)	7 (4)
Very little familiar	20 (16)	23 (14)
A bit familiar	35 (29)	53 (31)
Somewhat familiar	39 (32)	43 (26)
Very familiar	24 (20)	42 (25)
Subpopulation strata: *n* (%)		
Clinicians	60 (49)	64 (38)
Guideline developers	18 (15)	24 (14)
Researchers	44 (36)	80 (48)

Abbreviation: GRADE, Grading of Recommendations Assessment, Development and Evaluation.

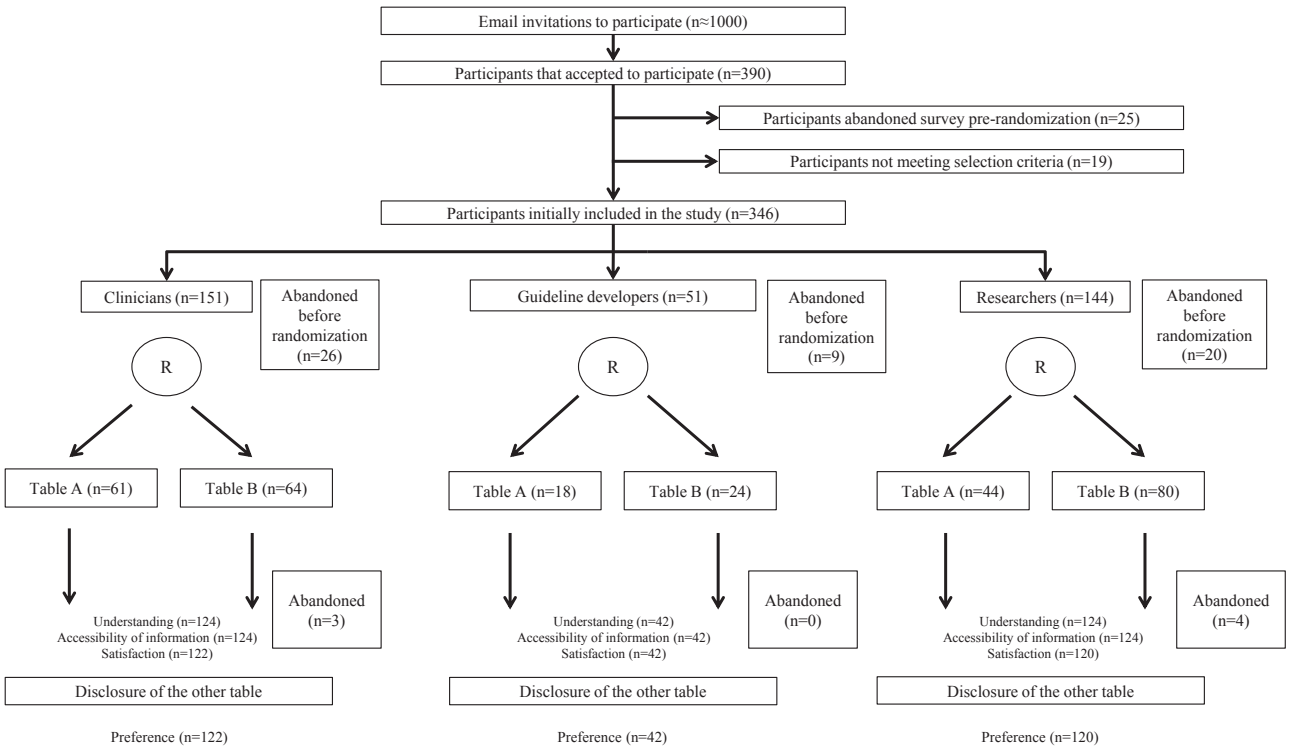

Fig. 1. Flow of participants through the study.

participants described themselves as a bit familiar or somewhat familiar with the GRADE approach (see Table 4). Participants included 125 clinicians, 42 guideline developers, and 124 researchers (see Fig. 1). The stratum to which the participants belonged did not have any impact on the study outcomes, and therefore, the results are described in an aggregated manner.

3.1. Understanding

Participants allocated to the new format consistently had a higher proportion of correct answers compared to those who were allocated to the current format (difference in proportions between groups ranging from 0% to 63%) (see Table 5). Three items showed similar results in the current and new format, two showed small differences in favor of the new format, and two showed large differences in favor of the alternative format: (1) ‘‘ability to determine a risk difference’’ [increase in proportion of correct answers 63% (95% CI: 54.6, 71.0)], and (2) ‘‘understanding of quality of evidence and treatment effect combined’’ [increase in proportion of correct answers 62% (95% CI: 52, 71)] (see Fig. 2). All new items tested were noninferior to the current ones. Using regression analyses, we determined the impact of baseline characteristics on this outcome. For question 1, only years of experience modified the estimate by more than 10% [adjusted odds ratio (OR): 1.83; 95% CI (0.91, 3.67); *P*-value = 0.088]. For question 2, years of experience, familiarity with GRADE, and level of training modified the outcome by more than 10% [adjusted OR: 0.72; 95% CI (0.20, 2.56); *P*-value = 0.6], but these modifications were not significant. For the remaining five questions, there were no covariates modifying the outcome.

3.2. Accessibility of information

Participants allocated to the new format considered, on average, that the information was more accessible across all domains assessed compared to the current formats (see Table 6). The adjusted analysis for the statement ‘‘It was easy to find the information about the effects’’ [MD 0.4; standard error (SE) 0.19; *P*-value = 0.04] and ‘‘It was easy to understand the information?’’ (MD 0.5; SE 0.20; *P*-value = 0.017) showed a nonstatistically significant difference between the two groups (*P*-value adjusted for multiple comparisons). Participants allocated to the new format considered that these items displayed review results in a way that was more helpful for decision making than the current ones (MD 0.5; SE 0.18; *P*-value = 0.011). The overall accessibility assessment per domain also favored the new format (MD 0.3; SE 0.11; *P*-value = 0.001).

3.3. Satisfaction

We asked participants which format satisfied them more and their reasons. More than 72% (203) would like to see the definition of each category for the quality of the evidence within the SoF table, 60% (171) think that

Table 5. Percentage of participants who answered correctly understanding questions

Concept	Question asked	New format (N = 122) (%)	Current format (N = 168) (%)	Risk difference (95% CI)	P-value
Ability to interpret footnotes	For the outcome adverse events, why is the quality of evidence rated as low?	89	82	7% (−2 to 15)	0.18
Ability to interpret risk	Will fewer children <5 years old have diarrhea if they take the probiotics?	96	96	0% (−5.3 to 5.4)	0.99
Ability to determine risk difference	How many fewer children <5 years will have diarrhea if they have probiotics than if they do not?	98	35	63% (54.6 to 71)	<0.001
Understanding of quality of evidence and treatment effect	Which of the following statements best represents the results informing the outcome adverse events?	88	26	62% (52 to 71)	<0.001
Understanding of quality of evidence	In children <5 years old, what result is most certain?	97	90	7% (0.1 to 12.4)	0.06
Ability to relate N of participant/studies and outcomes	How many participants and studies are informing the outcome adverse events?	95	98	−3% (−7.5 to 1.7)	1.00
Ability to quantify risk	In children >5 years old, how many fewer or more children will have diarrhea if they took probiotics as an adjunct to antibiotics compared to those who did not take probiotics?	94	88	6% (0.1 to 13.3)	0.06

Abbreviation: CI, confidence interval.

the ''number of participants/studies'' column can be eliminated and the information can be accommodated in the ''outcome'' column; 63% (178) mentioned that the ''comments'' column is not necessary, 86% (243) would like to see the reasons for downgrading the quality of evidence within the table, 88% (251) favored the inclusion of the ''what happens'' column, and 88% (250) considered that an additional column showing the risk and MDs along with their 95% CIs should be included (see Table 7).

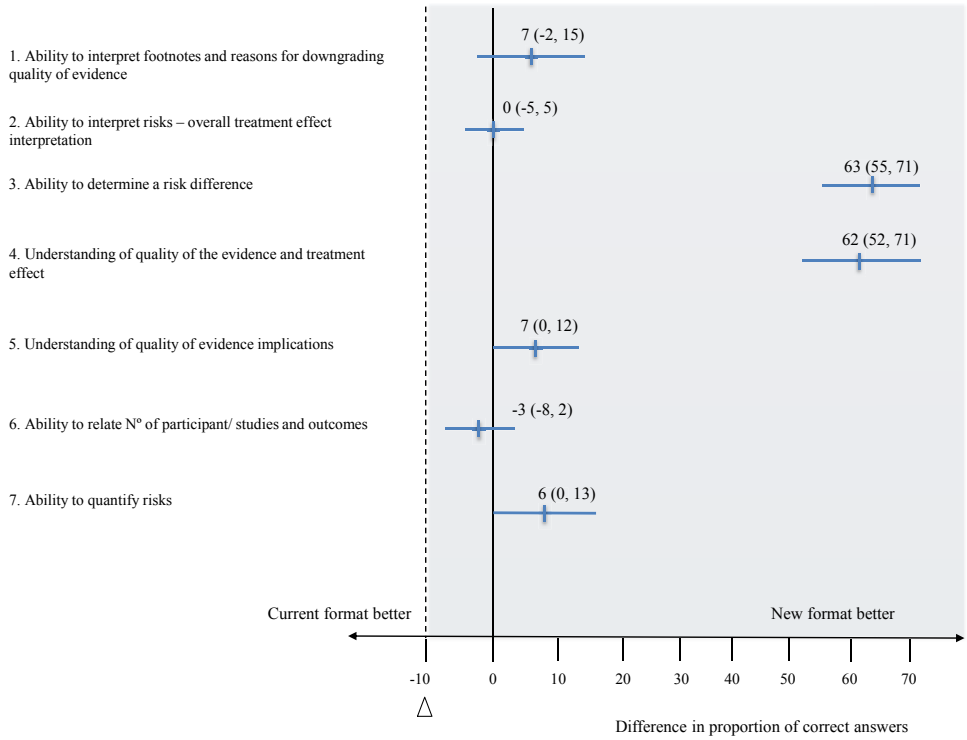

Fig. 2. Confidence intervals and noninferiority margin for the outcome understanding disaggregated at a question level. Dotted line at 10% indicates noninferiority margin; light blue tinted area to the left of the 10% margin indicates values for which the alternative formats would be considered noninferior to the current formats of SoF table. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 6. Domains[a] [mean (SD)] and overall accessibility of information

Domain	New format (n = 122)	Current format (n = 168)	Overall mean (SD) per domain	P-value	Adjusted analysis
It was easy to find the information about the effects	5.7 (1.4)	5.3 (1.5)	5.5 (1.4)	0.02	MD 0.4; SE 0.19; P = 0.04
It was easy to understand the information	5.5 (1.4)	5.1 (1.5)	5.3 (1.5)	0.02	MD 0.5; SE 0.20; P = 0.017
The information is presented in a way that would help me making a decision	5.6 (1.4)	5.1 (1.5)	5.3 (1.4)	0.004	MD 0.5; SE 0.18; P = 0.011
Overall accessibility (1–5 points) mean (SD)	4.1 (0.8)	3.7 (0.9)	3.9 (0.8)	<0.001	MD 0.3; SE 0.11; P = 0.001

Abbreviations: SD, standard deviation; MD, mean difference; SE, standard error.
[a] For each domain, the scale ranged from 1 to 7, where 1 means ''strongly disagree'' and 7 ''means strongly agree.''

3.4. Preference

Participants in both groups consistently preferred the new to the current format (mean/SD new format shown first 2.9/1.6; mean/SD current SoF table format shown first 2.8/1.7). The adjusted analysis also suggested a preference for the new over the current format. However, there were no statistically significant differences between the two groups for this outcome in either analysis. Overall, participants preferred the alternative to the current formats (MD/SD: 2.8/1.6).

4. Discussion

4.1. Main findings

This study represents an effort to provide systematic reviewers with alternative options to display review results using SoF tables. Because the current format of SoF tables improves understanding and facilitates the rapid retrieval of key findings, with an average of 90 seconds compared to a full-text review [12], we tested if seven new items in the SoF table can perform at least as effectively as the current ones. In particular, our results suggest that the following formatting changes will provide high level of understanding, accessibility, and user satisfaction with SoF tables: (1) exclusion of the N° of participants' and studies' column and the location of this data with the outcome, (2) the presentation of quality of evidence, also known as confidence in the estimates of effect or certainty in the evidence, along with the reasons for downgrading in the same column, (3) the use of the label ''explanations'' heading the footnotes, (4) the presentation of baseline risk and corresponding risk expressed as percentages, (5) the exclusion of the comments column, (6) the inclusion of a column exclusively dedicated to showing the risk difference or the MD and its 95% CI, and (7) the inclusion of a new column describing the results and the quality of the evidence using a narrative statement (''what happens'' column). The latter two items resulted in a large improvement in participants' understanding. Irrespective of the arm to which participants were allocated, respondents consistently preferred the new format to the current one.

One of the most important new item we tested in this study was the assessment of the effect on understanding and satisfaction of the ''what happens'' column. The purpose of this column is to simplify and assist systematic review users with the interpretation of both the treatment effect and the quality of the evidence in only one summary statement. It is composed of two main parts: the treatment effect and the quality of the evidence. This new item provides readers with a short but clear summary of the

Table 7. Satisfaction with new vs. current SoF table format (analysis at item level)

Question asked	Yes; n (%)	No; n (%)
Do you think it is important to have a description of the definition for each category for the quality of the evidence (GRADE working group grades of evidence)?	203 (72)	81 (28)
Do you think the ''number of participants/studies'' column can be eliminated and the information can be accommodated in the ''outcome'' column?	171 (60)	113 (40)
The ''comment'' column is missing in Table A, and instead the comments are reported in the footnotes, Do you think this ''comment'' column is necessary?	106 (37)	178 (63)
In Table A, we have included the reasons for downgrading in the ''quality of the evidence (GRADE) column. While Table B does not include this feature. Do you think Table A format is better?	243 (86)	41 (14)
In Table A, we have included a column called ''what happens'' column. The purpose of this column is to assist users on the interpretation of both review results and quality of the evidence. Do you think this column should be included as an available feature in future versions of SoF tables?	251 (88)	33 (12)
In Table A, we have included an extracolumn to display the difference between the two groups (and its 95% confidence interval). Do you think that this option of displaying the difference and its 95% confidence interval between the intervention and control group should be available in future SoF tables?	250 (88)	34 (12)

Abbreviations: SoF, summary of findings; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

evidence for a particular outcome but also could assist users when interpreting numerical expressions of risk for the same outcome.

4.2. Relation to prior work

In a previous randomized controlled trial, we showed [13] that guideline panelists preferred the presentation of risk differences to absolute risk estimates per intervention arm and allocating additional information within the tables rather than as footnotes in evidence profiles. These findings, although studied in a different population and displayed in evidence profiles are similar to the results of this study, which enhance our inferences. Previous studies reporting on the outcome accessibility to information have shown that displaying absolute risk reduction or other risk expressions derived from it (e.g., number needed to treat for benefit and harm) positively influenced users' accessibility [19].

In a large randomized trial using an online survey system, Woloshin and Schwartz [20] showed that participants (adults aged 18 years or older randomly selected from a national sample of the U.S population) reached higher levels of understanding when data about treatment effects were presented using simple percentages compared with natural frequencies, variable frequencies, percent plus natural frequencies, and percent plus variable frequencies. In our study, we found that the presentation of absolute risks using percentages showed higher levels of understanding to natural frequencies. Although the population in our trial may be more educated in the use and interpretation of risks than the one included by Woloshin and Schwartz [20], both trials show similar results and contradict the findings of our systematic review exploring this issue related to general presentation of health information [19]. The review's findings suggested that natural frequencies were better understood than percentages. However, this new evidence, in the context of summary tables for systematic reviews, suggests that percentages may be preferable to natural frequencies, particularly when presenting absolute risks. Additional work is required with different types of users.

One of the key findings of our trial is the large effect on participants' understanding of the inclusion of narrative statements that describe both the pooled estimate and the quality of the evidence. This was identified in our study as ''what happens'' column. Although participants suggested that the name for this column is not intuitive and it may need to be revised in the future, most of them preferred the inclusion of narrative statements to assist SoF table users with the interpretation of the content in the table. Narrative descriptions have been previously tested in the context of the development of the plain language summary (PLS) for Cochrane reviews [17,21]. A qualitative study by Glenton et al. showed that presentation of the magnitude of effect of interventions and the quality of the evidence, also known as confidence in the estimates of effect, in a narrative way along with the numerical data was preferred over the presentation of qualitative statements or numerical data alone [17]. A subsequent randomized trial showed that more participants understood the content of the PLS when narrative statements included information of both treatment effects and quality of the evidence and numerical data compared with the qualitative narrative statements alone (proportion of correct answers: 53% vs. 18%; $P < 0.001$) [21]. The results of these studies support the notion that narrative statements facilitate and assist users' interpretation of systematic review results.

4.3. Strengths and limitations

The present study has several strengths. First, the new format of the SoF table is the result of the collection and analysis of a wealth of stakeholder feedback, user testing, and comments from a broad audience of users with different interests and background. Second, we used outcomes that have been validated in prior trials. Third, it follows the methodological suggestions and improvements [19] for further trials conducted in the same field. Fourth, it recruited participants from more than 25 different language areas, backgrounds, and settings (i.e., clinicians, guideline developers, and researchers), which increases the generalizability of the findings.

A limitation is the remote online data collection process, which implies limited control over the environment in which the questionnaire was completed (i.e., whether it was completed by the same person that the link was sent to, whether the participant used additional material while answering the questions that measured the outcomes, and so forth). Second, some participants would likely fit into more than one of the strata defined in the study, and we used arbitrary cut offs to define their primary role. For example, clinicians might be highly involved in research, or guideline developers could also identify themselves as researchers according to our definition. Third, we chose a noninferiority design, but margins for noninferiority of our primary outcome have not clearly established. One of the reasons for choosing this design rests in the recognition that authors of reviews and users may have different preferences for presentation format and this design allowed testing acceptable choices. Fourth, although used in several trials, there is a lack of fully established and better validated outcome measures. We countered this concern by using outcome measures that were sensitive enough to change or interventions to show effects in prior trials and have good face validity. Finally, the sample corresponds to a relatively small proportion of those initially invited to participate, which may have affected generalizability.

4.4. Implications

This randomized controlled trial provides systematic review authors and users with a series of items that can be used along with the current format of the SoF table. These

items have proved to be at least as effective as the current approaches to displaying information, with higher accessibility of information and overall preference for the new SoF table. Future studies should elaborate on the wording of standardized narrative conclusion statements, comparing percentages and natural frequencies, and visual displays and test their effectiveness on systematic review users' understanding. The new format will be made available in GRADEs electronic tool GRADEpro GDT (www.gradepro.org) so that authors can choose from one of the versions and ultimately use interactive SoF tables, ideally with the items tested in this trial.

Acknowledgments

The SoF MIF authors group appreciate the valuable help provided by the Cochrane Applicability and Recommendations Methods Group, the OMERACT (Outcome Measures in Rheumatology) group, the Cochrane Effective Practice and Organization of Care Group, the Cochrane Consumers and Communication Review Group, the Cochrane Public Health Group, the Cochrane Haematological Malignancies Group, the Cochrane Colorectal Cancer Group, the Cochrane Airways Group, the Norwegian Branch of the Nordic Cochrane Center, the Cochrane Gynaecological cancer group, the Cochrane Risk of Bias Methods Group, the Cochrane Oral Health Group, the Cochrane Wounds group, the Cochrane Pain, Palliative & Supportive Care Review Group, the Cochrane Patient Reported Outcomes Methods Group, the Cochrane Musculoskeletal Group, the Cochrane Screening and Diagnostic Test Methods Group, the Cochrane Depression, Anxiety and Neurosis Review Group, and the Campbell and Cochrane Equity Methods Group.

References

[1] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.

[2] Jaeschke R, Guyatt GH, Dellinger P, Schunemann H, Levy MM, Kunz R, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive. BMJ 2008;337:a744.

[3] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.

[4] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336:995−8.

[5] Guyatt GH, Oxman AD, Kunz R, Jaeschke R, Helfand M, Liberati A, et al. Incorporating considerations of resources use into grading recommendations. BMJ 2008;336:1170−3.

[6] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ 2008;336:1049−51.

[7] Schunemann, HJ, Best D, Vist G, Oxman AD, GRADE Working Group. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;169:677−80.

[8] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66:158−72.

[9] Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66:173−83.

[10] Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. J Clin Epidemiol 2007;60:1298−305.

[11] Rosenbaum SE, Glenton C, Nylund HK, Oxman AD. User testing and stakeholder feedback contributed to the development of understandable and useful summary of findings tables for Cochrane reviews. J Clin Epidemiol 2010;63:607−19.

[12] Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. J Clin Epidemiol 2010;63:620−6.

[13] Vandvik PO, Santesso N, Akl EA, You J, Mulla S, Spencer FA, et al. Formatting modifications in GRADE evidence profiles improved guideline panelists comprehension and accessibility to information. A randomized trial. J Clin Epidemiol 2012;65:748−55.

[14] Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:2594−604.

[15] Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Protocol: comparison between standard and new alternative formats of the summary-of-findings tables in Cochrane review users: a non-inferiority randomized controlled trial. Trials 2015;16:164.

[16] Johnston BC, Goldenberg JZ, Vandvik PO, Sun X, Guyatt GH. Probiotics for the prevention of pediatric antibiotic-associated diarrhea. Cochrane Database Syst Rev 2011;(11):CD004827.

[17] Glenton C, Santesso N, Rosenbaum S, Nilsen ES, Rader T, Ciapponi A, et al. Presenting the results of Cochrane systematic reviews to a consumer audience: a qualitative study. Med Decis Making 2010;30:566−77.

[18] Harrell FE. Regression Modeling Strategies: with applications to linear, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

[19] Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev 2011;(3):CD006776.

[20] Woloshin S, Schwartz LM. Communicating data about the benefits and harms of treatment: a randomized trial. Ann Intern Med 2011;155:87−96.

[21] Santesso N, Rader T, Nilsen ES, Glenton C, Rosenbaum S, Ciapponi A, et al. A summary to communicate evidence from systematic reviews to the public improved understanding and accessibility of information: a randomized controlled trial. J Clin Epidemiol 2015;68:182−90.