Journal of Clinical Epidemiology 74 (2016) 19–27

Journal of Clinical Epidemiology

Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed

Miranda Langendam[a], Alonso Carrasco-Labra[b,c,d], Nancy Santesso[b], Reem A. Mustafa[b,e,f], Romina Brignardello-Petersen[d,g], Matthew Ventresca[b], Pauline Heus[h], Toby Lasserson[i], Rasmus Moustgaard[j], Jan Brozek[b,k], Holger J. Schünemann[b,k,l,m,*]

[a]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, PO Box 22660, J1B-211, 1100 DD Amsterdam, The Netherlands*

[b]*Department of Clinical Epidemiology & Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*

[c]*Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Universidad de Chile, Sergio Livingstone Pohlhammer 943, Independencia, Santiago, Chile*

[d]*Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Sergio Livingstone Pohlhammer 943, Independencia, Santiago, Chile*

[e]*Department of Medicine, University of Missouri, 2411 Holmes St., Kansas City, MO 64108-2792, USA*

[f]*Department of Biomedical and Health Informatics, University of Missouri, 2411 Holmes St., Kansas City, MO 64108-2792, USA*

[g]*Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College St, 4th Floor, Toronto, Ontario M5T 3M6, Canada*

[h]*Dutch Cochrane Centre, Julius Center—UMC Utrecht, Huispostnummer Str. 6.131, Postbus 85500, 3508 GA Utrecht*

[i]*Cochrane Editorial Unit, St Albans House, 57-59 Haymarket, London SW1Y 4QX, United Kingdom*

[j]*Nordic Cochrane Centre, Blegdamsvej 9, 7811, 2100 Copenhagen, Denmark*

[k]*Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*

[l]*Cochrane Applicability and Recommendations Methods Group, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*

[m]*McMaster GRADE Center, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*

Accepted 22 December 2015; Published online 11 January 2016

Abstract

Objectives: The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group has developed GRADE evidence profiles (EP) and summary of findings (SoF) tables to present evidence summaries in systematic reviews, clinical guidelines, and health technology assessments. Explanatory notes are used to explain choices and judgments in these summaries, for example, on rating of the quality of evidence.

Study Design and Setting: A systematic survey of the explanations in SoF tables in 132 randomly selected Cochrane Intervention reviews and in EPs of 10 guidelines. We analyzed the content of 1,291 explanations using a predefined list of criteria.

Results: Most explanations were used to describe or communicate results and to explain downgrading of the quality of evidence, in particular for risk of bias and imprecision. Addressing the source of baseline risk (observational data or control group risk) was often missing. For judgments about downgrading the quality of evidence, the percentage of informative explanations ranged between 41% (imprecision) and 79% (indirectness).

Conclusion: We found that by and large explanations were informative but detected several areas for improvement (e.g., source of baseline risk and judgments on imprecision). Guidance about explanatory footnotes and comments will be provided in the last article in this series.

Keywords: GRADE; Quality of evidence; Systematic reviews; Health technology assessment; Risk of bias; Summary of findings tables; GRADEpro

1. Introduction

For clinicians, guideline developers, or policy makers, a summary of evidence is an essential element to inform health care decisions [1,2]. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach, used by more than 90 organizations worldwide,

Conflict of interest: The authors of this study declare no financial conflict of interest. However, most of them are members of the GRADE working group and the Cochrane Collaboration.

Funding: This study was supported by the Cochrane Collaboration's Methods Innovation Fund and the McMaster GRADE Center. Neither of these institutions played a role in the planning, conduct, or publishing the study findings except through the authors' affiliation.

* Corresponding author. Tel.: +1 905 525 9140; fax: +1 905 522 9507.
E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

http://dx.doi.org/10.1016/j.jclinepi.2015.12.008
0895-4356/

What is new?

Key findings

- By and large explanations, formerly called footnotes, in summary of findings (SoFs) tables and Grading of Recommendations Assessment, Development, and Evaluation (GRADE) evidence profiles are informative, but there are several areas for improvement including addressing source of baseline risk, incomplete and incorrect judgments on imprecision, inconsistency and publication bias, and redundant and overly detailed information.

What this adds to what was known?

- This empirical study adds knowledge on how authors explain the reasons for downgrading and upgrading according to GRADE and where guidance is needed to create concise, clear, accurate, and relevant explanatory footnotes and comments.

What is the implication and what should change now?

- Guidance is needed on how to formulate explanations about judgments when using GRADE and developing SoFs tables, in particular for the domains imprecision and inconsistency and the comment column.

offers a transparent and structured process for developing and presenting summaries of evidence.

The GRADE working group presents such evidence summaries as GRADE summary of findings (SoFs) tables and evidence profiles (EPs) [3]. SoF tables are typically used to summarize the findings of systematic reviews. SoF tables provide a SoFs and an overall rating of the quality of evidence for each of the included outcomes in a quick and accessible format. SoF tables are an integral part of Cochrane systematic reviews with more than 1,300 SoF tables published in Cochrane reviews up to December 2014. For new Cochrane reviews, Cochrane standards describe that an assessment of the quality of evidence based on GRADE domains is mandatory and including a SoF table highly desirable (http://editorial-unit.cochrane.org/mecir). GRADE EPs, although very similar to SoF tables, contain more detailed information about the SoFs for each outcome, the assessment of each of the GRADE criteria, and the overall rating of the quality of that evidence. EPs are typically used to present evidence to guideline panels when making recommendations or decisions and can also be used by authors of systematic reviews when preparing SoF tables.

Where SoF tables can refer to the full text of the review, EPs sometimes are published or used as a stand-alone table, which means that more details are necessary. Creation of (interactive) SoF tables or EPs is supported through utilization of the GRADEpro Guideline Development Tool software (www.gradepro.org), which provides structure and help functions for users.

Key elements of both tables are explanatory notes, known as ‘‘footnotes’’ and ‘‘comments.’’ As a result of user testing on SoF tables, we refer to what was originally called footnotes as ‘‘explanations’’ or ‘‘explanatory notes.’’ Authors use explanations and the comments column to clarify information, to explain choices and judgments, or to add information that facilitates interpretation of the results and rating of the quality of evidence. They can be used to justify downgrading or to clarify the decision not to downgrade despite some concern about the evidence. In fact, such explanations can be as helpful as explanations that support a decision to downgrade. For guideline panels and decision makers, that is, users of the tables, explanations, and comments are critical to understanding and interpreting the evidence. Despite their importance, explanations and comments have received little attention. For example, in the GRADE guidance JCE series, we have provided detailed guidance on how to rate the quality of evidence, but not how to explain the reasons for downgrading and upgrading in concise, clear, accurate, and relevant explanations [4,5]. A recent audit of two cohorts of Cochrane Reviews published in 2013 and 2014 also identified current challenges in understanding downgrading decisions in many SoF tables, including unclear specification of the domain and the number of levels that had been downgraded (http://editorial-unit.cochrane.org/mecir). Our work with authors of Cochrane and other reviews and guideline developers suggested that many explanations lack basic characteristics that would make them informative. We observed, for example, large variation in the level of detail of the explanations and decisions on downgrading and upgrading that were not always in line with GRADE guidance [6]. We also realized that explanations and comments not only create transparency about judgments, but also give detailed insight into how systematic reviewers communicate the evidence, how they cope with GRADE, and also how they can draw on information in the text of the systematic review and present in explanations.

The purpose of this investigation was to document how authors make use of explanations and comments in GRADE evidence tables and explain judgments about the quality of evidence. This article is the second in a series of three articles that focus on improving GRADE evidence tables and were partially supported by the Cochrane Collaboration Methods Innovation Fund. The first article reported the results of a randomized controlled trial (RCT) comparing a new format of SoF tables to the standard format [7]. This article describes the systematic survey of explanations and comments in SoF tables in Cochrane reviews and in EPs in guidelines from various organizations, including the World Health Organization. The next

and final article will provide examples of both good and poor explanations and guidance on how to write useful explanations for GRADE evidence tables.

2. Method

2.1. Summary of study design

We conducted a systematic survey of the explanations (as stated above, previously called footnotes and comments) in SoF tables in randomly selected new and updated Cochrane intervention reviews and EPs in a selection of guidelines and analyzed the content of the explanations and comments using a predefined list of criteria related to the assessment and presentation of evidence. Although GRADE has used the terms factors, criteria or domains to describe issues related to downgrading and upgrading, GRADE now uses ''domains'' to describe upgrading and downgrading considerations with items referring to the specific issues to consider within the domains to make decisions about increasing or lowing the certainty in the evidence. Thus, will use domains in this article.

2.2. Sources of explanations and comments in SoF tables and GRADE EPs

Three sources of SoF tables and EPs were included in this study.

1. Cochrane reviews with at least one complete SoF table published up to Cochrane Library Issue 3, 2012 (502 reviews). If in a SoF table, the outcomes were not listed by row, or one or more standard SoF columns were missing, or if explanations on downgrading were not stated for evidence that was not high, the SoF table was considered incomplete. We extracted explanations and comments from all SoF tables in a random sample of 107 Cochrane reviews. We updated this set with a random sample of 25 Cochrane reviews published between April and September 2014 (148 reviews). The total number of included reviews and guidelines was determined by reaching saturation, that is, sampling more data did not lead to more information relating to the research question.

   Because upgrading occurs infrequently, we reviewed all SoF tables in our database up to 2012 to identify relevant examples. Search terms were ''rating up,'' ''upgrading,'' ''large effect,'' ''magnitude effect,'' ''dose-response,'' and ''confounding.''
2. GRADE EPs in a selection of 10 guidelines from a variety of organizations: the American College of Chest Physicians [8,9], American Thoracic Society [10,11], World Allergy Organization [12], and World Health Organization [13–17]. These guidelines were chosen because the organizations who produced them adopted GRADE relatively early and guideline development was supported by members of the GRADE working group.
3. GRADE EPs produced for one of the reviews assessed in the GRADE reliability study [18]. This data source enabled us to investigate variability in writing explanations as 25 different study participants evaluated the same evidence from a review [19].

2.3. Data extraction

We developed and piloted data extraction forms for the SoF tables and EPs in Microsoft Excel 2010. The forms included the following variables: Cochrane Review or guideline ID (if applicable); number of SoF tables/EPs per review/guideline, number of explanations, or comments; content of each explanation or comment; column in the SoF table or EP to which the explanation referred; and whether the explanation or comment referred to single study evidence, nonquantitative evidence, or upgrading.

For the GRADE reliability study, we extracted the explanations from the EPs and grouped them by outcome and GRADE domain to create an overview of all explanations addressing the same GRADE domain for the same body of evidence.

2.4. Analysis

2.4.1. Aggregation by theme

We coded the extracted explanations from all SoF tables in the 132 randomly selected Cochrane reviews and the 17 EPs from the 10 selected guidelines and grouped them by predefined theme. One investigator extracted the data, and a second verified. The themes were based on the different elements of the SoF table, the GRADE domains to assess the quality of evidence (e.g., risk of bias, imprecision, indirectness, inconsistency, among others), and communicating results. The pilot of the data extraction form was used to test the completeness of the themes and the coding instructions. We calculated frequencies of the themes.

The themes were as follows:

1. Downgrading and upgrading domains: risk of bias, indirectness, inconsistency, imprecision, publication bias, large magnitude of effect, plausible bias, and dose-response gradient.
2. Justification for no downgrading (specifying minor concerns about the evidence).
3. Judgment across domains (footnote covers more than one domain).
4. Baseline risk (information regarding source of baseline risk or the control group risk, e.g., from observational studies).
5. Outcome (information regarding outcome definition, follow-up, and so forth).
6. Effect measure (information regarding the chosen effect measure, for example, relative risk, hazard ratio).

7. Study design (information regarding type of study, for example, crossover trials or cohort studies).
8. Describing or communicating results (information regarding interpretation, for example, NNT).
9. Describing subgroup or sensitivity analysis.
10. Other than 1−9.

For the GRADE working group's reliability study, selected examples of explanations for common situations, like many unclear risk of bias items, to illustrate how authors address these issues.

2.4.2. Informative explanations and comments

To explore usefulness, each explanation and comment was rated as being informative or not informative by two raters (M.L., P.H., R.B-.P., M.V., and N.S.). The judgment was made by one investigator and carefully checked by a second. Explanations and comments that were not informative were those that were unclear or did not explain a judgment; we specified the criteria for each domain (see Appendix A at www.jclinepi.com and Section 3). We also identified explanations that had too much detail, that is, those with unnecessary details. We performed a reliability study on a random sample of 50 explanations. Two raters (M.L. and H.J.S.) rated footnotes as informative or not informative. The agreement was 74% and the chance corrected agreement (weighted kappa) 0.47 indicating moderate agreement. After very brief discussion to resolve disagreement, agreement increased to 94% and only 3 of 50 evaluations would have required further discussion or review by a third person.

To investigate if the explanations followed current published guidance on downgrading and upgrading, the text of the explanation was checked for misconceptions and errors. We did not go back to the text, analyses, and tables of the review.

2.4.3. Roles of explanations and comments

On review of the results, we used group meetings and group consensus to agree on key roles of explanations based on the intended messages and our prior knowledge about explanations.

3. Results

3.1. SoF tables from Cochrane reviews

Most (86%) of the 502 Cochrane reviews from 2012 included complete SoF tables, and in 2014, this percentage increased (92% of 148 reviews). Table 1 lists the reasons for incompleteness. We randomly selected 132 reviews (241 SoF tables) and extracted 1,027 explanations (771 footnotes and 256 comments). The median number of explanations per SoF table was 4 (range 1−21). Of the 241 SoF tables, 77 included comments. Median number was 2 comments (range 1−11).

All presented results are not specified in explanations and comments (unless stated otherwise). To increase readability, we will use the term ''explanations'' to indicate this. We identified three main roles for the explanations based on the analysis of their content: (1) to provide additional information or sources of the information; (2) to provide judgments about the quality of the evidence using the GRADE domains and items; and (3) to communicate and interpret the findings.

Table 2 presents the distribution of the explanations by theme and by role, see Appendix B at www.jclinepi.com for the distribution in explanations and comments separately. Most explanations were used to describe or communicate results (29% of all extracted explanations) and to explain downgrading (59%), in particular for risk of bias and imprecision.

Overall, across all roles and GRADE domains and depending on the criteria, the percentage of explanations that was informative ranged from 15% (justification for no downgrading when indicated) to 98% (baseline risk). For role 2, providing judgments for the quality of evidence, the percentage of explanations that were informative ranged from 41% (imprecision) to 79% (indirectness).

3.1.1. Role 1: additional information or sources of information: baseline risk, choice of outcome, effect measure, and study design

Of all extracted explanations, 12% were used to give additional information or describe sources of information.

Authors are asked to document the source for the assumed baseline risk, for example, the control group risk of trials in the systematic review or observational studies including representative populations [20]. Only 25% of the reviews (41 of 1,027 explanations; 32 of 130 reviews) provided this information. Informative explanations described the source, for example, the risk in the placebo groups, mean change from baseline in control group or prognostic model, if applicable for low and high risk groups.

Explanations for choice of outcome, effect measure, and study design were used infrequently (Table 2). Explanations that were not informative included information that was obvious from the table, for example, zero studies measured the outcome.

3.1.2. Role 2: judgments for the quality of evidence using the GRADE domains

Out of the sample of 1,027 explanations, 59% addressed downgrading or upgrading a body of evidence.

3.1.2.1. Risk of bias. Review authors must make an overall judgment about risk of bias across studies that contribute data to an outcome [21]. About one-third of the explanations on rating the quality of the evidence referred to risk of bias. An informative explanation says whether the risk of bias criterion was likely to influence the results by indicating if the criterion (e.g., lack of allocation concealment) was present or not in the studies or influenced the results

Table 1. Reviews with complete and incomplete summary of findings (SoFs) tables

Summary of findings	Reviews (*n*, %) 2012[a]	2014[b]
Complete SoF tables	431 (85.9)	136 (91.9)
Quality of evidence assessed but justification for downgrading missing	22 (4.3)	6 (4.1)
Other columns missing	13 (2.6)	3 (2.0)
Not a SoF table	36 (7.2)	3 (2.0)
Total	502	148

[a] Issue 3, 2012.
[b] Cochrane reviews published between April and September 2014.

(e.g., a large study having that problem and the results being different).

According to these criteria, 53% of the explanations addressing risk of bias (and based on evidence from more than one study) was informative (e.g., ''we downgraded because lack of blinding of patients and providers in 4 of 5 studies; it was unclear whether allocation was concealed in 2 studies; and only one study clearly used intention to treat analysis'').

In most explanations, authors provided information about either the number of studies with high risk of bias and/or the specific risk of bias item for which there was concern. In contrast, some authors did not provide any details and simply stated that there was ''high risk of bias in the trial(s) included.'' However, none of the explanations on risk of bias could be labeled as misconception or error.

Overly detailed explanations, providing information that is better placed in the text of the review, provided references to the individual studies or explained in detail that there were no limitations. Some explanations included industry funding as a risk of bias. Industry funding is not currently considered a direct source of bias but may be a surrogate for other reasons to downgrade according to GRADE (e.g., reporting bias, directness, or publication bias).

Table 2. Explanations and comments: aggregation by theme (based on 132 reviews)[a]

Theme	Frequency total = 1,027	% Informative per theme
Role 1: To provide additional information or sources of the information		
Baseline risk	41	98
Effect measure	8	50
Outcome	21	81
Study design	22	77
Role 2: To provide judgments for the quality of evidence using the GRADE domains		
Downgrading/upgrading	611	
Risk of bias	205	
Based on single study evidence	71	92
Based on more than 1 study	134	53
Imprecision	148	41
Inconsistency	83	48
Indirectness	53	79
Publication bias	19	74
Magnitude of effect[b]	6	100
Opposing plausible bias[b]	0	
Dose-response[b]	0	
More than one GRADE domain	78	
GRADE domain unclear	19	
Justification for no downgrading	68[c]	16
Role 3: To communicate and interpret findings		
Communicating/interpreting results, subgroup/sensitivity analysis	294	82
Other	30	

Abbreviation: GRADE, Grading of Recommendations Assessment, Development, and Evaluation.
[a] An explanation can refer to more than one outcome.
[b] Based on random sample of 132 reviews, analysis of upgrading was based on total of 502 reviews.
[c] Different GRADE domains.

3.1.2.2. Imprecision. Detailed GRADE guidance [22] suggests that authors of systematic reviews should consider downgrading for imprecision if the 95% confidence interval (CI) crosses important thresholds, the sample size or number of events is small or the optimal information size (OIS) is not met.

Downgrading (or justification for no downgrading) for imprecision was explained by referring to the width of the CI (65% of the imprecision explanations), sparse data (24%), or the OIS (10%). Regarding the width of the CI, most authors referred to clinical relevance thresholds implicitly (e.g., ''95% CIs include both negligible effect and appreciable benefit'') and rarely explicitly [e.g., ''serious imprecision. The 95% CI of the pooled estimate includes appreciable benefit (<0.75) and nonappreciable benefit (≥ 0.75 and ≤ 1.00) with (drug name)'']. Noninformative explanations were too superficial and did not define clinically important thresholds, the OIS, or provide more information about the number of events, for example, ''wide confidence interval,'' ''small study,'' and ''very few events.''

We identified several misconceptions (12 reviews): judgment based on statistical significance only (instead of taking into account whether the CI crosses the null effect and appreciable benefit or harm), uninformed downgrading for imprecision with large sample size and limited number of events (downgrading for imprecision not needed when sample size is very large, e.g. more than 2000), downgrading only if CI includes a relative effect of 0.75 and 1.25 (instead of 0.75 or 1.25 and null effect, i.e., one of the thresholds), and downgrading based on the CIs of the individual studies (should be the pooled effect).

3.1.2.3. Inconsistency. GRADE suggests considering rating down if large inconsistency (heterogeneity) in study results remains after exploration of a priori hypotheses. Inconsistency can be explored by evaluating four items [23].

Only one explanation took these four items into account (''serious inconsistency: chi-square for

heterogeneity = 0.04, I^2 = 63%. Risk ratios ranged from 0.58 to 1.03. The CIs for the trials showing most extreme effects overlapped to only a small extent. Too few trials to explore this heterogeneity'') and two explanations considered three of the four items.

Half of the explanations were based on statistical items only, mostly by reporting I^2 [e.g., ''high unexplained heterogeneity (75% > I^2 >50%)'' or ''homogeneous data (I^2 = 0%)'']. Other explanations were based on similarity of point estimates only [e.g., ''direction of one study favored (intervention), whereas rest favored placebo''] or point estimates and statistical items [e.g., ''heterogeneity between studies was considered substantial, but all individual estimate effects from primary studies had the same direction which favored the early (intervention) group'']. The extent of overlap of CIs was mentioned only twice (e.g., ''unexplained heterogeneity of I^2 = 86%; two studies suggested benefit; however, the CIs do not overlap'').

Explanations were considered informative if the level of heterogeneity was reported. Noninformative explanations were mainly too open (e.g., ''unexplained heterogeneity of results,'' ''significant heterogeneity between the trials,'' or ''clinical and methodological heterogeneity'').

We found three misunderstandings. One misunderstanding was that inconsistency could not be evaluated because the number of studies was too low or because there was only one study (results of two studies can be inconsistent; a body of evidence based on a single study cannot be inconsistent). The second misunderstanding was that inconsistency is related to statistically significance of the results of the individual studies (e.g., ''no serious inconsistency: neither trial found a statistically significant difference in time to hospital discharge''). Statistical significance of the effect estimate is not an item in the assessment of inconsistency; however, the authors meant to point at wide and therefore overlapping CIs. The third misunderstanding was judgment of inconsistency for explained instead of unexplained heterogeneity (e.g., ''clinical and methodological heterogeneity,'' ''inconsistent results (I^2 = 61%) which can be explained partly by medication subgroups''). For many explanations, it was not clear whether they were used to describe the level of heterogeneity or the reason for downgrading (or no downgrading).

3.1.2.4. Indirectness. Quality of evidence may decrease when substantial differences exist between the population, the intervention, the comparators or the outcomes measured in relevant research studies and those under consideration in a guideline or systematic review. Other reasons for downgrading for indirectness arise in case head-to-head comparisons are unavailable or when there is uncertainty about estimate on background risk [24].

Explanations on indirectness were considered informative if the Population, Intervention, Comparisons and Outcomes element(s) that caused the indirectness was clearly stated. Almost all (79%) the explanations addressing indirectness were informative, with a varying degree of detail. Uncertainty about the baseline risk was never used to indicate indirectness.

3.1.2.5. Publication bias. The absence of publication bias cannot be demonstrated, and it is almost equally difficult to know where to place the threshold to rate down for its likely presence [25].

Reflecting these challenges, only 19 of the 1,027 explanations (15 reviews) referred to publication bias, 10 explanations were used to explain that publication bias was suspected. Most explanations were brief. Given the challenges with rating publication bias, we considered explanations in which an evidence-specific reason was provided informative (e.g., ''less than 10 studies, so creating a funnel plot was not applicable,'' ''asymmetric funnel plot,'' ''some evidence for publication bias''). General statements were considered not informative (''it is difficult to exclude the risk of publication bias'').

3.1.2.6. Reasons for upgrading. Three domains can be used for rating up the certainty of the effects: large magnitude of effect, dose-response gradient, and the effect of opposing plausible confounding or bias. These domains primarily apply to observational studies. Although it is theoretically possible to rate up results from RCTs, there no compelling examples of such an instance [26].

In four SoF tables of 502 reviews from 2012, upgrading was explicitly stated. In one SoF table, the evidence was based on one uncontrolled interrupted time series study. In the three other reviews, the quality of evidence of RCTs was first downgraded and then upgraded because of a large magnitude of effect. In two SoF tables, however, the upper bound of the 95% CI was higher than 0.50 and the risk of bias could have caused the large effects. The third review was updated in May 2012, and upgrading was replaced by not downgrading for another domain.

In nine SoF tables, explanations were formulated referring to magnitude of effect without specifically mentioning upgrading. It was often not clear whether the authors were just describing their results or if they upgraded the quality of the evidence. In three SoF tables, we found explanations indicating dose-response gradients. We did not find any explanation referring to opposing plausible confounding or bias.

3.1.2.7. Justification for no downgrading. Explanations can be used to clarify why authors do not downgrade despite some concern about the evidence. Of the explanations that could be clearly labeled as justification of no downgrading, most were used to report that there was no concern, which we consider not informative (e.g., ''the two studies in this analysis had low risk of bias'').

3.1.2.8. Single RCT evidence. Evidence from single studies was common. GRADE suggests especially careful

scrutiny of all relevant issues (risk of bias, precision, directness, and publication bias) when only a single RCT addresses a particular question [21]. Possibly redundant explanations were referring to inconsistency (e.g., "no inconsistency: only one trial") or "single trial only" as sole explanation.

3.1.2.9. Narrative analyses (no pooling). Of the random sample of 132 reviews, 11 reviews (8%) had narrative analyses (see first sentence). Out of a total of 567 reviews from 2012 to 2014, 28 reviews had SoF tables with missing explanations (Table 1). Of these 28 reviews, 6 reviews (21%) included narrative analyses. Explanations referring to outcomes for which results could not be pooled were rare (46 explanations in 11 reviews). Effects were either not reported, described in the comments column (e.g., "2 RCTs: one showed small benefits in provider knowledge (family planning) with supervision, whereas one study (prescribing knowledge) was inconclusive"), or presented in an adapted layout of the SoF table.

3.1.3. Role 3: to communicate and interpret findings

About one-quarter of the explanations were used to communicate and interpret findings or to describe subgroup or sensitivity analyses, mostly as comments (Table 2). Comments were used to enhance interpretation of the results, for example, providing alternative effect measures (e.g., number needed to treat, standardized mean difference), subgroup or sensitivity analyses (e.g., "the design of some of the studies put the results at some risk of bias. A sensitivity analysis which removed studies at a high risk of selection bias gave a result that was much closer to unity"), describing the effect [e.g., "(drug name) increases the chance of cessation of vomiting"], additional information on the outcome (e.g., "diarrhea was reported as a side effect in all the studies"), and explanations of a not estimable effect (e.g., "no events," "not measured in included studies"). Statements about the statistical significance level (e.g., "not statistically significant"), information that could be derived from other columns in the SoF table (e.g., "estimated effect based on only two studies"), or details on the analysis (e.g., I^2 65%, 122 total events) were considered not informative.

Examples of explanations from the GRADE reliability study, when many items of the risk of bias assessment are unclear, and different approaches of imprecision and inconsistency are shown in Appendix C at www.jclinepi.com.

3.2. EPs in guidelines

We extracted 264 explanatory notes (formerly "footnotes") from 17 EPs in 10 guidelines. The mean number of explanations per EP was 15.5 (standard deviation, 5.9; range 8–31). Appendix D at www.jclinepi.com presents an overview of the themes and the assessment of how informative the explanations were. The proportion of explanations that were informative was in line with the results from the SoF tables. Nine explanations (three guidelines) suggested misconception: "without pooled estimate, it was not possible to assess the consistency or precision of the results" and upgrading of downgraded RCTs, imprecision based on individual studies. Table 3 compares the themes between the guidelines and SoF tables.

4. Discussion

We conducted a systematic survey of the use of explanations in GRADE evidence tables, a widely used approach to summarize findings from systematic reviews or health technology assessments that inform decision makers such as guideline panels. We found that explanations were often informative but detected several areas for improvement. These areas include addressing source of baseline risk, incomplete and incorrect judgments on imprecision, inconsistency and publication bias, and redundant and overly detailed information. Providing explanations for narrative analyses is an important area to monitor and explore.

4.1. Strengths

This survey was based on a large sample of GRADE evidence tables. The sample of SoF tables is representative for Cochrane reviews; the sample of EPs is likely to represent an optimistic picture, as the sampling frame was guideline organizations that are relatively familiar with using GRADE. Furthermore, this research was performed by a

Table 3. Comparison of distribution of explanations by theme

	EPs guidelines		SoF tables Cochrane reviews (2012 + 2014)	
Theme	***n***	**%**	***n***	**%**
Baseline risk	3	1.1	41	4.0
Effect measure, outcome, study design	20	7.5	51	5.0
Downgrading/upgrading	160	60.6	611	59.4
Communicating results/subgroup/ sensitivity analysis	66	23.9	294	28.6
Other	15	5.7	30	2.9
Total	264		1,027	

Abbreviations: EP, evidence profile; SoF, summary of finding.

large group of experienced authors, and the analysis of the extracted explanations was systematic and detailed.

4.2. Weaknesses

The unit of analysis was the individual explanation. This means there could be dependency between explanations within the same SoF table because they were developed by one group of review authors. This has no consequences for the evaluation of the presented examples, as they were chosen to cover the whole range of what authors intend to communicate. However, because of the possibility of dependency, the presented frequencies have to be interpreted with some caution. They underestimate or overestimate the problem.

The identified misunderstandings with interpreting GRADE were based on the content of the explanation. We did not compare the content of the explanation with the results in the original review, as this was beyond the scope of this study. Finally, some of the authors of this article developed several of the selected clinical guidelines. These authors, however, were not involved in analyzing the individual explanations.

4.3. Summary of the results

The magnitude of the problem was similar for SoF tables and EPs, but there was a slight increase in the proportion of complete SoF tables in Cochrane systematic reviews over time. The results suggest that guidance is primarily needed for the GRADE domains imprecision and inconsistency and appropriate use of the comment column. For the GRADE domains risk of bias and indirectness, we found higher percentages of informative explanations. We speculate that this could be due to the existing tools like the Cochrane risk of bias tool and the indirectness tool in GRADEpro, which assist authors in making the judgment [27]. Taken together with findings from the CEU audit of Cochrane Reviews (http://editorial-unit.cochrane.org/mecir), our survey provides some supportive evidence that the specific issues relating to imprecision likely reflect wider challenges of interpreting results of relative and absolute measures of effect and overreliance on statistical significance more generally.

Although GRADE suggests the initial separate consideration of the eight GRADE domains, with a yes/no decision regarding rating up or down in each case, the final rating of overall evidence quality occurs while considering the judgments about the domains in context. For example, if there is some concern about risk of bias and some concern about precision, one might decide to downgrade one level for some concern with both domains. Or when very serious inconsistency is present, it may lead to an imprecise summary effect estimate. In these instances, following careful consideration of the relationship between both concerns, downgrading by one level rather than two may well constitute a more reasonable approach.

The balancing of domains should be reflected in the explanation of downgrading together with clarification as to the number of levels the evidence has been downgraded by. Our data indicate that explanations were mostly used to address the domain risk of bias. It seems likely that situations as described in the examples above will occur more frequently; review authors will need more guidance on how to arrive at a final rating of the quality of evidence and how to explain that. The results suggest that misconceptions and errors are primarily related to imprecision and inconsistency, domains that will benefit from more guidance while they are considered.

To provide such guidance, the next article in this series will build on these results and describe best practice for explanations in GRADE evidence tables, a critical pathway to making evidence assessment more informative.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.12.008.

References

[1] 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.

[2] 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.

[3] Langendam MW, Akl EA, Dahm P, Glasziou P, Guyatt G, Schunemann HJ. Assessing and presenting summaries of evidence in Cochrane Reviews. Syst Rev 2013;2:81.

[4] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[5] Schunemann H, Oxman A, Higgins J, Vist G, Glasziou P, Guyatt G. Chapter 11: presenting results and 'summary of findings' tables. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011). The Cochrane Collaboration; 2011. Available at www.cochrane-handbook.org.

[6] Langendam MW, Mustafa R, Santesso N, Carrasco-Labra A, Moustgaard R, Ventresca M, et al. Harmonization of explanations for common judgments about the quality of evidence in summary of findings tables. Cochrane Database Syst Rev 2013;1.

[7] Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa R, Mbuagbaw L, et al. Improving GRADE evidence tables: A randomized trial shows improved understanding of content in Summary-of-Findings Tables with a new format. J Clin Epidemiol 2016;74:7−18.

[8] Holbrook A, Schulman S, Witt DM, Vandvik PO, Fish J, Kovacs MJ, et al. Evidence-based management of anticoagulant therapy: antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141. e152S-84S.

[9] Taichman DB, Ornelas J, Chung L, Klinger JR, Lewis S, Mandel J, et al. Pharmacologic therapy for pulmonary arterial hypertension in

adults: CHEST guideline and expert panel report. Chest 2014;146: 449−75.
[10] Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW, Casale TB, et al. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. J Allergy Clin Immunol 2010;126: 466−76.
[11] Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 2011;183:788−824.
[12] Fiocchi A, Schunemann HJ, Brozek J, Restani P, Beyer K, Troncone R, et al. Diagnosis and Rationale for Action against Cow's Milk allergy (DRACMA): a summary report. J Allergy Clin Immunol 2010;126. 1119−1128 e12.
[13] WHO. The Use of Bedaquiline in the Treatment of Multidrug-Resistant Tuberculosis: Interim Policy Guidance. Geneva: World Health Organization Copyright (c) World Health Organization 2013; 2013.
[14] WHO. WHO Recommendations for Augmentation of Labour. Geneva: World Health Organization Coypright (c) World Health Organization 2014; 2014.
[15] WHO. WHO Recommendation on Community Mobilization through Facilitated Participatory Learning and Action Cycles with Women's Groups for Maternal and Newborn Health. Geneva: World Health Organization Copyright (c) World Health Organization 2014; 2014.
[16] WHO. Guidelines for the Identification and Management of Substance Use and Substance Use Disorders in Pregnancy. Geneva: World Health Organization Copyright (c) World Health Organization 2014; 2014.
[17] Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis 2007;7:21−31.
[18] Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol 2013;66:736−42. quiz 42 e1−5.
[19] Rosner S, Hackl-Herrwerth A, Leucht S, Lehert P, Vecchi S, Soyka M. Acamprosate for alcohol dependence. Cochrane Database Syst Rev 2010;CD004332.
[20] 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.
[21] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence−study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[22] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence −imprecision. J Clin Epidemiol 2011;64:1283−93.
[23] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence −inconsistency. J Clin Epidemiol 2011;64:1294−302.
[24] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence−indirectness. J Clin Epidemiol 2011;64:1303−10.
[25] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence−publication bias. J Clin Epidemiol 2011;64:1277−82.
[26] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311−6.
[27] Schünemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:49−62.