ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 173−183

Journal of Clinical Epidemiology

GRADE guidelines: 13. Preparing Summary of Findings tables and evidence profiles—continuous outcomes

Gordon H. Guyatt[a,b,*], Kristian Thorlund[a], Andrew D. Oxman[c], Stephen D. Walter[a], Donald Patrick[d], Toshi A. Furukawa[e], Bradley C. Johnston[a], Paul Karanicolas[f], Elie A. Akl[g], Gunn Vist[c], Regina Kunz[h], Jan Brozek[a], Lawrence L. Kupper[i], Sandra L. Martin[j], Joerg J. Meerpohl[k,l], Pablo Alonso-Coello[m], Robin Christensen[n], Holger J. Schunemann[a,b]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, Oslo 0130, Norway*
[d]*Department of Health Services, Seattle Quality of Life Group/Center for Disability Policy and Research, University of Washington, Box 359455, Seattle, WA 98195-9455, USA*
[e]*Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan*
[f]*Department of Surgery, University of Toronto, Toronto, Ontario, Canada*
[g]*Department of Medicine, State University of New York at Buffalo, Buffalo, NY, USA*
[h]*Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, Basel 4031, Switzerland*
[i]*Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA*
[j]*Department of Maternal and Child Health, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA*
[k]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg 79110, Germany*
[l]*Pediatric Hematology and Oncology, Center for Pediatrics and Adolescent Medicine, University Medical Center Freiburg, Freiburg 79106, Germany*
[m]*Iberoamerican Cochrane Center-Servicio de Epidemiología Clínica y Salud Pública and CIBER de Epidemiología y Salud Pública (CIBERESP), Hospital de Sant Pau, Universidad Autónoma de Barcelona, Barcelona 08041, Spain*
[n]*Musculoskeletal Statistics Unit, The Parker Institute, Copenhagen University Hospital, Frederiksberg, Denmark*

Accepted 12 August 2012; Published online 30 October 2012

Abstract

Presenting continuous outcomes in Summary of Findings tables presents particular challenges to interpretation. When each study uses the same outcome measure, and the units of that measure are intuitively interpretable (e.g., duration of hospitalization, duration of symptoms), presenting differences in means is usually desirable. When the natural units of the outcome measure are not easily interpretable, choosing a threshold to create a binary outcome and presenting relative and absolute effects become a more attractive alternative.

When studies use different measures of the same construct, calculating summary measures requires converting to the same units of measurement for each study. The longest standing and most widely used approach is to divide the difference in means in each study by its standard deviation and present pooled results in standard deviation units (standardized mean difference). Disadvantages of this approach include vulnerability to varying degrees of heterogeneity in the underlying populations and difficulties in interpretation. Alternatives include presenting results in the units of the most popular or interpretable measure, converting to dichotomous measures and presenting relative and absolute effects, presenting the ratio of the means of intervention and control groups, and presenting the results in minimally important difference units. We outline the merits and limitations of each alternative and provide guidance for meta-analysts and guideline developers.

Keywords: GRADE; Effect size; Standardized mean difference; Minimal important difference; Meta-analysis; Continuous outcomes

The GRADE system has been developed by the GRADE Working Group. The named authors drafted and revised this article. A complete list of contributors to this series can be found on the *Journal of Clinical Epidemiology* Web site.

* Corresponding author. CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada. Tel.: 905-527-4322; fax: 905-523-8781.

E-mail address: guyatt@mcmaster.ca (G.H. Guyatt).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2012.08.001

1. Introduction

The first 12 articles in this series introduced the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to systematic reviews and guideline development [1], discussed the framing of the question [2], presented GRADE's concept of quality of evidence and how to apply it [3–9] presented GRADEs approach to resource use considerations [10], described how to make overall ratings of confidence [11], and discussed Summary of Findings (SoF) tables presenting the results of binary outcomes [12]. In this thirteenth article, we address issues specific to SoF tables that report results of continuous outcomes.

Our recommendations will differ according to whether

1. investigators have all used the same measure that is familiar to the target audiences
2. investigators have all used the same or very similar measures that are less familiar to the target audiences
3. investigators have used different measures

2. Options when investigators have all used the same measure that is familiar to the target audiences

In the simplest situation, authors of primary studies have all used the same measure of the continuous outcome of interest, and the target audiences will easily interpret that outcome. This is likely to be true, for instance, of durations of events, such as hospitalization or symptoms for conditions such as sore throat, otitis media, or influenza. For such outcomes, the SoF table should include a weighted difference of means.

Table 1 presents examples of such outcomes from systematic reviews in SoF format. Each of these is easily understood and has a straightforward interpretation. For instance, supportive employment led to an increase in days of competitive employment of approximately 46 days with a confidence interval (CI) that many would consider narrow those interested in an increase of 46 days would probably be interested in an increase of 35. This modest effect may be important to the target population.

In the second example in Table 1, approximately 10 fewer hours of diarrhea—that is, using the range of control group means, a reduction of from 59 to 49 or 170 to 160—may be important to some parents and children and not to others. Ideally, empirical research would have established peoples' attitude toward relatively small decreases in the duration of diarrhea. One way of interpreting the findings of such research would be in terms of a minimally important difference (MID) [13–15], the smallest reduction in diarrhea that people would consider important.

Our attitude toward zinc treatment might change depending on whether the effect is less than 1 hour or greater

Table 1. Examples of mean differences in easily understood units[a]

Patients, interventions, comparators	Participants (studies), follow-up	Quality of the evidence (GRADE)	Comparator	Intervention vs. comparator mean difference (95% CI)
Schizophrenia[a] Supportive employment vs. other vocational approaches	843 participants (five studies) 12–24 mo (mean = 19 mo) of follow-up	⊕⊕⊕◯ **Moderate** because of risk of bias (suspicion of selective reporting bias)	**32.3 d** of competitive employment	**45.9 d (95% CI: 34.7, 57.1) longer** in competitive employment
Children with acute diarrhea[b] Zinc vs. placebo	4,242 participants (13 studies)	⊕⊕⊕◯ **Moderate** because of inconsistency	The mean diarrhea duration (hr) ranged across control groups from **59 to 170 hr**	**9.60 (95% CI: 18.25, 0.96) fewer** hours of diarrhea
Common cold[c] NSAIDs vs. no NSAIDs	214 participants (two studies)	⊕⊕◯◯ **Low** because of imprecision and heterogeneity	**7.33 d**	**0.23 (95% CI: 1.75 fewer to 1.29 more) fewer** days of cold symptoms
Surgery[d] Supplemental perioperative oxygen vs. routine oxygen administration	2,963 participants (four studies)	⊕⊕⊕◯ **Moderate** because of imprecision	The mean hospital stay (d) across control groups ranged from **6.4 to 11.9 d**	**0.86 d (95% CI: −0.29, 2.00) longer** hospital stay

Abbreviations: CI, confidence interval; NSAID, nonsteroidal anti-inflammatory drugs.

[a] Kinoshita Y MD PhD, Furukawa T MD PhD, Omori IM MD PhD, Watanabe N MD PhD, Marshall M MD, Bond GR MD, Huxley P MD, Kingdon D MD. Supported employment for adults with severe mental illness. Cochrane Database of Systematic Reviews (submitted).

[b] Kim SY, Chang YJ, Cho HM, Hwang YW, Moon YS. Non-steroidal anti-inflammatory drugs for the common cold. Cochrane Database Syst Rev 2009, Issue 3. Art. No.: CD006362.

[c] Lazzerini M, Ronfani L. Oral zinc for treating diarrhoea in children. Cochrane Database Syst Rev 2008, Issue 3. Art. No.: CD005436. DOI: 10.1002/14651858.CD005436.pub2.

[d] Garcia-Alamina J BScN, Devereaux PJ MD PhD, Sessler D MD, Leslie K MBBS, MD, M Epi, Perera R MSc PhD, Alonso-Coello P MD PhD. Supplemented perioperative oxygen to reduce the incidence of surgical site infection: systematic review and meta-analysis of randomized controlled trials (submitted).

Key points

- Summary of Findings tables provide succinct presentations of evidence quality and magnitude of effects.
- Summarizing the findings of continuous outcomes presents special challenges to interpretation that become daunting when individual trials use different measures for the same construct.
- The most commonly used approach to providing pooled estimates for different measures, presenting results in standard deviation units, has limitations related to both statistical properties and interpretability.
- Potentially preferable alternatives include presenting results in the natural units of the most popular measure, transforming into a binary outcome and presenting relative and absolute effects, presenting the ratio of the means of intervention and control groups, and presenting results in preestablished minimally important difference units.

than 18 hours (the 95% CI limits) and the relation of those limits to the MID. If so, rating down confidence in effect estimates for imprecision as well as inconsistency—for which the authors decreased their confidence—may be appropriate. A similar logic applies to the other examples in Table 1, although it suffers from a limitation we will raise subsequently: a mean difference less than the MID may still be consistent with a substantial proportion of participants experiencing an important benefit.

3. Options when investigators have all used the same or very similar measures that are less familiar to the target audiences

Transparency becomes more challenging when clinicians and patients are unfamiliar with the units of the outcome measure. For instance, Table 2 presents data derived from a systematic review addressing the impact of compression stockings for people taking long flights [16]. Outcomes include the presence of edema. Because each study used the same measurement tool for assessing edema, it is possible to make the pooled difference between the groups (the ''weighted mean difference'') of 4.7 units more interpretable by noting that edema is measured on a scale of 0 (no edema) to 10 (maximum edema). Intuitively, on such a scale, 4.7 units represent a large difference.

Sometimes, the meaning of changes in score becomes even more obscure. This is often true, for instance, of measures of health-related quality of life (HRQL). In these situations, the MID—in this case, the smallest change in

Table 2. Summary of Findings: compression stockings compared with no compression stockings for people taking long flights

Outcomes	Absolute risks (95% CI)		Relative effect, OR (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
	Estimated risk	Corresponding risk				
	Without stockings	With stockings (95% CI)				
Symptomatic deep vein thrombosis: inferred from surrogate, symptomless deep vein thrombosis	**Low-risk population** **5 per 10,000** **High-risk population** **18 per 10,000**	**0.5 per 10,000** (0–1.25) **1.8 per 10,000** (1–8)	**0.10** (0.04, 0.25)	2,637 (nine studies)	⊕⊕⊕○ **Moderate** because of indirectness [4]	
Edema: post-flight values measured on a scale from 0 (no edema) to 10 (maximum edema)	The mean edema score ranged across control groups from **6.4 to 8.9**	The mean edema score in the intervention groups was on average **−4.72 lower** (95% CI: −4.91, −4.52)		1,246 (six studies)	⊕⊕○○ **Low** [4] because of risk of bias (unblinded, unvalidated measure)	
Adverse effects	See comment	See comment	Not estimable	1,182 (four studies)	See comment	The tolerability of the stockings was described as very good with no complaints of side effects in four studies [5]

Abbreviations: CI, confidence interval; OR, odds ratio.
Patients or population: anyone taking a long flight (lasting more than 6 hours); Settings: international air travel; Intervention: compression stockings [1]; Comparison: without stockings.

HRQL score that patients would consider important—may be very helpful [14]. For instance, in a measure of chronic lung disease in which possible scores in HRQL range from 1 to 7, 0.5 represents the MID [15]. When they have access to such information, authors should include it in the SoF table or evidence profile.

It may be preferable, if data from individual studies permit, to set a threshold and present results as a dichotomy. For instance, studies of the impact of thrombolytic therapy after stroke used the Rankin instrument that classifies patients into one of six categories of disability from no symptoms to severe handicap. Authors of a systematic review evaluating the impact of thrombolytic therapy on inpatients with stroke dichotomized the Rankin instrument, creating a ''bad outcome'' category of those dead or moderately or severely disabled by Rankin criteria [17].

The reviewers found that thrombolytic therapy significantly reduced the proportion of patients who were dead or dependent at the end of follow-up (odds ratio [OR], 0.84; 95% CI: 0.75, 0.95). Making this fully interpretable requires the corresponding absolute reduction. Given the control group risk of approximately 60%, the OR of 0.84 corresponds to a reduction in the risk of death or dependency of more than 4% (number needed to treat [NNT], 25). In such instances, it is likely to be helpful enlisting clinician users of the instrument to help in choosing the threshold for the dichotomy.

In another example of this approach, a review addressing the impact of flavonoids on symptoms in patients with hemorrhoids [18], trials did not use the same symptom measures. All but one, however, recorded the proportions of patients free from symptoms, with symptom improvement (both classified, by authors of the review as improved), still symptomatic, or worse (both classified as not improved). In the primary analysis, the authors pooled on the a priori expectation of a similar magnitude and direction of treatment effect across studies. They reported a relative risk of 0.42—that is, a 58% relative risk reduction—in the adverse outcome of failure to achieve symptomatic improvement (95% CI: 0.28, 0.61).

Although applicable to any ordinal scales, whatever their length, one possible limitation of this approach is its apparent susceptibility to bias: if more than one threshold is reasonable, it is possible for reviewers to choose a threshold that provides the most (or the least) optimistic estimate of treatment effect. However, at least in some circumstances [19]—and perhaps in most [20]—choice of threshold makes little difference in apparent magnitude of relative effect. Even so, the choice of threshold may influence the level of statistical significance. Reviewers should therefore choose and justify their threshold, provide results for all reasonable thresholds, or both. Preferably, such a threshold approach should be introduced at the protocol stage to reduce the risk of biased selection of a threshold.

In deciding whether to dichotomize the outcome, review authors should also consider possible loss of statistical power [21]. This will be a particular concern if a result that reaches the conventional threshold for statistical significance as a continuous variable loses significance when converted to a dichotomy. You (and more importantly, patients and their families) may be primarily interested in whether patients have crossed a meaningful threshold. For example, you may believe that there is an important dividing line between depressed and nondepressed and whether patients have crossed that line is your question of interest. In these circumstances, a loss of power or statistical significance from converting a continuous outcome to a dichotomous outcome accurately reflects uncertainty about the proportion of people who would cross that dividing line.

If, on the other hand, your use of dichotomization is simply to enhance interpretability, one option for dealing with this situation is to report the statistical significance of the result when analyzed as a continuous variable and present the estimate of relative or absolute effect from the dichotomy only as an aid to the interpretation of the magnitude of the effect.

4. Options when investigators have used different measures

Reviewers face further challenges when studies measure the same concept but use different measurement instruments. For instance, one set of trials may have measured depression using the Beck Depression Inventory-II [22], and another set may have used the Hamilton Rating Scale for Depression [23]. Under these circumstances, providing pooled estimates of effect and making results interpretable mandates use of one of five available approaches. Table 3 summarizes the merits of each approach and our associated recommendations. We refer readers interested in an in-depth examination, including details of the derivation and statistical properties, to a separate article [24].

Tables 4 and 5 illustrate the application of the approaches to two examples: dexamethasone for pain in patients undergoing laparoscopic cholecystectomy [25] (Table 4) and respiratory rehabilitation for chronic obstructive pulmonary disease [26] (Table 5).

4.1. Standard deviation units: standardized mean difference

One way of generating a pooled estimate when trials have measured the same construct with different instruments is to divide the difference between the intervention and control means (i.e., the difference in means) in each trial by the estimated between-person standard deviation (SD) (row (A) in Tables 4 and 5) [27]. This measure is often referred to as the standardized mean difference (SMD) or Cohen effect size.

Presenting results in SD units (as an SMD) is the longest standing and most widely used approach and is recommended in the Cochrane Handbook [27]. Calculating and

Table 3. Five approaches to presenting results of continuous variables when primary studies have used different instruments to measure the same construct

Approach	Advantages	Disadvantages	Recommendation
SD units (standardized mean difference; effect size)	Widely used	Interpretation challenging Can be misleading depending on whether population is very homogenous or heterogeneous	Do not use as the only approach
Present as natural units	May be viewed as closer to primary data	Few instruments sufficiently used in clinical practice to make units easily interpretable	Approaches to conversion to natural units include those based on SD units and rescaling approaches. We suggest the latter. In rare situations when instrument very familiar to frontline clinicians, seriously consider this presentation
Relative and absolute effects	Very familiar to clinical audiences and thus facilitate understanding Can apply GRADE guidance for large and very large effects	Involve assumptions that may be questionable (particularly methods based on SD units)	If the MID is known, use this strategy in preference to relying on SD units Always seriously consider this option
Ratio of means	May be easily interpretable to clinical audiences Involves fewer questionable assumptions than some other approaches Can apply GRADE guidance for large and very large effects	Cannot be applied when measure is change and therefore negative values possible Interpretation requires knowledge and interpretation of control group mean	Consider as complementing other approaches, particularly the presentation of relative and absolute effects
MID units	May be easily interpretable to audiences Not vulnerable to population heterogeneity	Only applicable when MID is known To the extent that MID is uncertain, this approach will be less attractive	Consider as complementing other approaches, particularly the presentation of relative and absolute effects

Abbreviations: SD, standard deviation; MID, minimally important difference.

presenting results in SD units has, however, major limitations. First, clinicians and their patients are unlikely to be able to relate to this way of presenting results [28]. Second, if the variability or heterogeneity in the severity of patients' condition (and thus the variability in scores on the chosen outcome) varies between trials, the SDs will also vary. As a result, trials that enroll heterogeneous groups of patients will yield smaller SMDs than those enrolling less heterogeneous patients, even if the actual (not standardized) mean difference estimates—and thus the absolute estimate of the magnitude of treatment effect—is similar across all trials. Finally, as one example we will present demonstrates, if very homogenous populations are enrolled, SD units can give a misleading inflated impression of the magnitude of treatment effect.

In both Tables 4 and 5, the presentations in SD units suggest a large treatment effect. The structure of the SoF table, however, is not well suited to this presentation. If authors use the SMD, it is not sensible to present absolute values in the intervention and comparison groups because studies have used different measurement instruments with different units. One approach to this dilemma, presented in Tables 4 and 5, is to present the SMD in place of the two columns usually devoted to absolute rates. An alternative is to present the median value from the studies that used the most familiar measure of the concept in the control group column and the SMD in the intervention group column. To aid interpretability of a metric unfamiliar to clinicians or patients, a comment provides a rule-of-thumb guide to the significance of various effect sizes [29] (row (A) in Tables 4 and 5).

4.2. Conversion into units of the most commonly used instrument

A second approach (row (B) in Tables 4 and 5) converts the effect size back into the natural units of the outcome measure most familiar to the target audiences. There are two statistical approaches to making the conversion. One (illustrated in Table 4) calculates the absolute difference in means by multiplying the SMD by an estimate of the SD associated with the most familiar instrument.

To make this calculation, one needs to choose an SD to use. From each study, one can calculate a weighted average of the control and intervention SDs (either change or posttest); we suggest using the median of these SDs. There are also options for estimating the CI around the mean in natural units that we describe in our more statistically detailed article [24].

In this case, the chosen measure is a 100-unit visual analog scale, and the magnitude of effect is 8.1. This result would be of limited use without knowledge of the MID, and thus the comment includes the estimated MID (10 units [30]), suggesting a modest effect (row (B) in Table 4).

The other statistical approach (presented in Table 5) makes a simple conversion—before pooling and without calculating the SMD—of other instruments to the units of the most familiar instrument [24]. In this case, we chose the

Table 4. Application of approaches to dexamethasone for pain after laparoscopic cholecystectomy example

Outcomes	Estimated risk or estimated score/value with placebo	Absolute reduction in risk or reduction in score/value with dexamethasone	Relative effect (95% CI)	Number of participants (studies)	Confidence in effect estimate[a]	Comments
(A) Postoperative pain, SD units: investigators measured pain using different instruments. Lower scores mean less pain	The pain score in the dexamethasone groups was on average **0.79 SDs (1.41–0.17) lower** than in the placebo groups		—	539 (5)	⊕⊕◯◯[b,c] Low	As a rule of thumb, 0.2 SD represents a small difference, 0.5 a moderate, and 0.8 a large
(B) Postoperative pain, natural units: measured on a scale from 0 (no pain) to 100 (worst pain imaginable)	The mean postoperative pain scores with placebo ranged from 43 to 54	The mean pain score in the intervention groups was on average **8.1 (1.8–14.5) lower**	—	539 (5)	⊕⊕◯◯ Low[b,c]	Scores estimated based on an SMD of 0.79 (95% CI: −1.41, −0.17) The minimally important difference on the 0–100 pain scale is approximately 10
(C) Substantial postoperative pain: investigators measured pain using different instruments	20 per 100[d]	More patients in dexamethasone group achieved important improvement in pain score **0.15 (95% CI: 0.19, 0.04)**	RR = 0.25 (95% CI: 0.05, 0.75)	539 (5)	⊕⊕◯◯[b,c] Low	Scores estimated based on an SMD of 0.79 (95% CI: −1.41, −0.17) Method assumes that distributions in intervention and control groups are normally distributed and variances are similar
(D) Postoperative pain: investigators measured pain using different instruments. Lower scores mean less pain	28.1[e]	3.7 lower pain score (6.1 lower 0.6 lower)	Ratio of means = 0.87 (0.78–0.98)	539 (5)	⊕⊕◯◯[b,c] Low	Weighted average of the mean pain score in dexamethasone group divided by mean pain score in placebo
(E) Postoperative pain: investigators measured pain using different instruments	The pain score in the dexamethasone groups was on average **0.40 (95% CI: 0.74, 0.07) minimally important difference units** less than in the control group		—	539 (5)	⊕⊕◯◯[b,c] Low	An effect less than half the minimally important difference suggests a small or very small effect

Abbreviations: CI, confidence interval; SD, standard deviation; SMD, standardized mean difference.

[a] Quality rated from 1 (very low quality) to 4 (high quality).

[b] Evidence limited by heterogeneity between studies.

[c] Evidence limited by imprecise data (small sample size or event rate).

[d] The 20% comes from the proportion in the control group requiring rescue analgesia.

[e] Crude (arithmetic) means of the postoperative pain mean responses across all five trials when transformed to a 100-point scale.

Table 5. Application of approaches to chronic respiratory rehabilitation for health-related quality-of-life impairment in patients with chronic airflow limitation

Outcomes	Estimated baseline score/ proportion improving in control patients	Absolute increase in proportion improving in patients receiving respiratory rehabilitation	Relative effect (95% CI)	Number of participants (studies)	Confidence in effect estimate[a]	Comments
(A) HRQL: investigators measured HRQL using different instruments. Higher scores mean better HRQL	The HRQL score in the respiratory rehabilitation group improved on average **0.72 (95% CI: 0.48, 0.96) SDs** more in the respiratory rehabilitation patients than in the control patients		—	818 (16)	⊕⊕⊕⊕ High	As a rule of thumb, 0.2 SD represents a small difference, 0.5 a moderate, and 0.8 a large
(B) HRQL measured on a scale of 1—7	Control group baseline, 4.5[a] Average improvement in control, 0.04	HRQL improved on average **0.71 (95% CI: 0.48, 0.94)** more in the respiratory rehabilitation patients than in the control patients	—	818 (16)	⊕⊕⊕⊕ High	Calculated by transforming all scores to the CRQ in which the minimally important difference is 0.5
(C) Proportion of patients with important improvement in HRQL	**0.30[b]**	Differences in proportion achieving important improvement **0.31 (95% CI: 0.22, 0.40)** in favor of rehabilitation	**OR = 3.36 (95% CI: 2.31, 4.86)**	818 (16)	⊕⊕⊕⊕ High	Calculation uses established minimally important difference of 0.5 units on the CRQ and 4 units on the St. George's Respiratory Questionnaire
(D) The currently recommended approach to ratio of means relies on posttest only and is therefore not applicable to change scores, which are the focus of results from these trials						
(E) HRQL measured in minimally important difference units	HRQL improved on average 1.75 (95% CI: 1.37, 2.13) **minimally important difference units** more in the respiratory rehabilitation than in the control group		—	818 (16)	⊕⊕⊕⊕ High	An effect of close to two times the minimally important difference suggests a moderate to large effect

Abbreviations: CI, confidence interval; HRQL, health-related quality of life; SD, standard deviation; CRQ, Chronic Respiratory Questionnaire; OR, odds ratio.

[a] Approximate average of baseline control group scores in the studies that reported the baseline score.

[b] This represents the median of the proportion of patients in the control group who achieved an improvement greater than the minimally important difference. That is, in the study at the center of the distribution of change, 30% of the control group achieved an improvement of more than 0.5 (CRQ) or 4 (St. George's).

Chronic Respiratory Questionnaire (CRQ), with units of 1–7, and rescaled the mean and SD of the other instruments to CRQ units. Given the MID of the CRQ (0.5 [15], presented in comments), the mean difference in change of 0.71 suggests a substantial effect of rehabilitation.

This second approach, presenting in units of the most familiar instrument, may be the most desirable when the target audiences have extensive experience with that instrument, particularly if the MID is well established [13]. Nevertheless, the natural unit presentation may, in relation to the MID, still be misleading. For instance, had the difference between rehabilitation and control been 0.4 on the 7-point scale in which 0.5 represents the MID, clinicians are at risk of interpreting the result as indicating that no one benefits, and the treatment is not worth administering. This is almost certainly an inaccurate interpretation as conversion into an absolute difference and NNT would demonstrate [31]. For instance, in one study, a mean difference of 0.43 units in the CRQ translated into a proportion benefiting of 34% and thus an NNT of approximately 3 [31].

4.3. Conversion to relative and absolute effects

A third approach (row (C) in Tables 4 and 5) converts the continuous measure into a dichotomy and thus allows calculation of relative and absolute effects on a binary scale. One method to generate a dichotomy from continuous data relies on the SMD and assumes that results of both treatment and control groups are normally distributed and have equal variances [20,32]. Meta-analysts usually make these assumptions when they calculate SMDs. We have used this approach in row (C) in Table 4, and it suggests a very large relative effect and a substantial absolute effect, particularly when the baseline risk is high.

This approach has the advantage that it can be applied easily by consulting Table 6, which provides the relation between the SMD and the risk difference. In Table 6, the top panel presents the conversion when the outcome is undesirable (e.g., pain) and the bottom panel when the outcome is desirable (e.g., response to treatment).

The approach, however, suffers from three important limitations. First, the dichotomous outcome that the intervention is decreasing is often not self-evident from the continuous outcome from which it is derived. We obtain a difference in the proportion of patients in intervention and control groups above some threshold, but the choice of that threshold may be arbitrary. In this example (row (C) in Table 4), we have characterized the threshold as ''substantial postoperative pain.''

Second, the approach requires investigators to specify the proportion of control patients with adverse outcomes—in this case, the proportion above the pain threshold. Choosing this proportion may also be difficult. For instance, if one knows that control group pain scores varied from 43 to 54, with SDs around 15, how is one to decide the proportion of patients who failed to experience an important improvement with placebo? In this case, we have used the proportion requiring rescue analgesia. The latter problem is ameliorated to some extent because only at the extremes of control proportions do the proportions benefiting change substantially.

Third, the approach, by relying on the SMD, is vulnerable to whether study populations had very similar scores on the outcome of interest or whether scores were widely variable.

Other statistical approaches also rely on the SMD to generate dichotomous presentations for continuous outcomes [33,34]. They share similar limitations, with the exception that they do not require specification of control group risk, and one approach becomes unstable when the underlying risk is less than 20% or greater than 80% [34].

Another strategy for creating dichotomies and generating estimates of relative and absolute effect relies on knowledge of the MID. In applying the approach, we assume normal distributions of data and then calculate the proportions of participants in the intervention and control groups in each study that demonstrated an improvement greater than the MID [24]. The results are then pooled across studies. Applying this approach in Table 5, findings

Table 6. Risk difference derived from SMD[a]

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
For situations in which the event is undesirable, reduction (or increase if intervention harmful) in adverse events with the intervention									
Control group response rate (SMD)									
−0.2	−0.03	−0.05	−0.07	−0.08	−0.08	−0.08	−0.07	−0.06	−0.040
−0.5	−0.06	−0.11	−0.15	−0.17	−0.19	−0.20	−0.20	−0.17	−0.12
−0.8	−0.08	−0.15	−0.21	−0.25	−0.29	−0.31	−0.31	−0.28	−0.22
−1.0	−0.09	−0.17	−0.24	−0.23	−0.34	−0.37	−0.38	−0.36	−0.29
For situations in which the event is desirable, increase (or decrease if intervention harmful) in positive responses to the intervention									
Control group response rate (SMD)									
0.2	0.04	0.61	0.07	0.08	0.08	0.08	0.07	0.05	0.03
0.5	0.12	0.17	0.19	0.20	0.19	0.17	0.15	0.11	0.06
0.8	0.22	0.28	0.31	0.31	0.29	0.25	0.21	0.15	0.08
1.0	0.29	0.36	0.38	0.38	0.34	0.30	0.24	0.17	0.09

Abbreviation: SMD, standardized mean difference.

[a] Approach from Furukawa [32].

suggest substantial relative and absolute benefit in HRQL as a result of respiratory rehabilitation.

If one only has posttest data (rather than magnitude of change), one can apply this approach if evidence exists regarding meaningful thresholds. For instance, if one knows that people with scores less than 8 on the Hamilton Rating Scale for Depression (HAM-D) are considered to be not depressed, one could examine the proportion of individuals below that threshold.

If such meaningful thresholds do not exist, one can still use posttest data if one assumes that the minimally important change within an individual corresponds, on average, to the MID between individuals. Making this assumption, one can calculate the difference in the proportion who benefit in intervention and control. To do this, one takes the mean value in the control group plus one MID unit and calculates the proportion of patients in each group above that threshold.

4.4. Ratio of means

A fourth, thus far infrequently used, approach (row (D) in Table 4) may appeal to clinicians: calculate a ratio of means (RoM) between the intervention and control groups [35]. Advantages of RoM include the ability to pool studies with outcomes expressed in different units, avoiding the vulnerability of heterogeneous populations that limits approaches that rely on SD units, and ease of clinical interpretation. However, the published RoM method is designed for posttest scores only and is therefore omitted from Table 5, which presents changes from baseline.

It is possible to calculate a ratio of change scores if both intervention and control groups change in the same direction in each relevant study, and this ratio may sometimes be informative. Limitations include (1) the unlikelihood of intervention and control group changes in the same direction in all studies and (2) the possibility of misleading results if the control group change is very small—in which case, even a modest change in the intervention group will yield a large and therefore misleading RoM changes.

In the dexamethasone for postoperative pain example (Table 4), the RoM approach suggests a relative reduction in pain of only 13%, meaning that those receiving steroids have a pain severity of 87% as severe as those in the control group, an effect that strikes us as modest.

4.5. MID units

A final strategy pools across studies in the same way as the SMD, but instead of dividing the mean difference of each study by its SD, it divides by the MID associated with that outcome [36]. The final output, instead of being in SD units, is in MID units. This approach avoids the problem of varying SDs across studies that may distort estimates of effect in approaches that rely on the SMD. It may, in addition, be more easily interpretable, although it risks the possibility that a difference less than the MID may be interpreted as trivial when a substantial proportion of patients have achieved an important benefit. In addition, to the extent that the MID estimate is not based on secure evidence, the approach becomes more questionable.

As stated in the comments column in Table 4, the result for dexamethasone for pain is an effect less than half of one MID, suggesting a small or very small effect. Table 5, in contrast, shows an effect of close to two MID units, suggesting a substantial benefit in HRQL as a result of respiratory rehabilitation.

5. Reflections on the interpretation of the five methods

The prior discussion makes evident that there is no ideal method for making results of continuous variables interpretable, particularly when studies have used different measurement tools for the same construct (e.g., pain, physical function, emotional function). Given the sometimes questionable assumptions that each approach makes, it would be reassuring if the methods led to essentially the same inferences. This is true for the respiratory rehabilitation example: all approaches suggest a moderate to large absolute effect of respiratory rehabilitation on HRQL.

This is not the case, however, for dexamethasone and pain. Here, the SMD (A) and the relative and absolute effects (C) suggest large benefit, whereas the other three approaches suggest small or even trivial pain reduction. The explanation for this is the homogeneity of patients enrolled with respect to their pain, leading to a very small SD. That small SD then suggests a large effect when expressed in SD units.

This limitation of the SD approach is highlighted by calculation of relative and absolute effects using the MID approach to dichotomizing the data described above in the respiratory rehabilitation example. Applying that approach to the pain after cholecystectomy data results in very different estimates of relative effect (relative risk [RR], 0.64; 95% CI: 0.34, 1.17) and absolute effects (risk difference [RD], 0.03; 95% CI: 0.01, 0.07), contrasting with the large point estimates and relatively narrow CIs around both relative and absolute effects in row (C) in Table 4. This highlights the vulnerability of methods that rely on SD units. This vulnerability is reflected in our recommendations below.

6. Recommendations for enhancing interpretability in meta-analyses in which primary studies use different instruments to measure the same underlying construct

We have described five approaches to enhancing the interpretability of continuous variables in meta-analyses in which primary studies have used different instruments. Review authors will have to tailor their approach to the individual situation but may find the following guides helpful:

1. Using more than one presentation is likely to be both informative and, if the clinical message is similar,

reassuring. It can also reduce the risk of biased selection of which presentation to use when the messages are different. If the messages are different, and it is not clear which to believe, review authors could consider rating down their confidence for inconsistency. Tables 4 and 5 present a model for presenting more than one approach within a single SoF table.

2. When one instrument is in use in regular clinical practice and is familiar to most consumers of a systematic review or guideline, a presentation in natural units of that instrument should be one of the options chosen.
3. Comments should be geared to helping with interpretation (e.g., rules of thumb for interpreting SMD and stating the MID if established)
4. If possible, choose methods that do not rely on SD units. If SD units are chosen, provide some guide for interpretation. In approach (B), the rescaling option would be preferable to multiplying the effect in SD units by the SD of the most popular instrument. In approach (C), generating relative and absolute effects using the MID is, if it is available, preferable to using any of the approaches that rely on units.
5. In most instances, one should seriously consider expressing the magnitude of effect as both an OR or relative risk as well as a risk difference. The advantages include familiarity for clinicians and ability to apply GRADE guidance for large and very large effects (for relative effect) and usefulness for clinical decision making (for absolute effects) (Table 3). Because presentation of relative effects alone may be misleading, in particular when relative effects are large but absolute effects small, the summary should ensure communication of the magnitude of absolute effect.
6. Reviewers should aim at transparency, citing the source of MIDs and SDs used, and the underlying assumptions.

7. Conclusion

Summarizing continuous variables in ways that are both valid and interpretable is challenging. To achieve these goals, systematic review authors and guideline developers should carefully consider the approaches we have suggested.

References

[1] 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.
[2] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[3] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J. GRADE guidelines 10 - Considering resource use and rating the quality of economic evidence. J Clin Epidemiol 2013;66:140−50 [in this issue].
[11] Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P. GRADE guidelines 11 - Making an overall rating of evidence for a single outcome and for all outcomes. J Clin Epidemiol 2013;66:151−7 [in this issue].
[12] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines 12 - Preparing summary of findings tables (SOF) - binary outcomes. J Clin Epidemiol 2013;66:158−72 [in this issue].
[13] Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77(4):371−83.
[14] Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? Qual Life Res 2007;16:1097−105.
[15] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407−15.
[16] Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrom M. Compression stockings for preventing deep vein thrombosis in airline passengers. Cochrane Database Syst Rev 2009;3.
[17] Wardlaw J, del Zoppo G, Yamaguchi T, Berge E. Thrombolysis for acute ischaemic stroke. Cochrane Database Syst Rev 2003;3: CD000213. Art. No. 10.1002/14651858.CD000213
[18] Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. Br J Surg 2006;93(8):909−20.
[19] Furukawa TA, Akechi T, Wagenpfeil S, Leucht S. Relative indices of treatment effect may be constant across different definitions of response in schizophrenia trials. Schizophr Res;126(1−3):212−219.
[20] Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol 1991;44:241−8.
[21] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006;332:1080.
[22] Beck AT, Steer RA, Brown G. BDI-II: Beck Depression Inventory Manual. 2nd ed. San Antonio, TX: The Psychological Corporation; 1996.
[23] Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967;6:278−96.
[24] Thorlund K, Walter S, Johnston B, Furukawa TA, Guyatt G. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. Res Synth Methods 2011;2(3):188−203.

[25] Karanicolas PJ, Smith SE, Kanbur B, Davies E, Guyatt GH. The impact of prophylactic dexamethasone on nausea and vomiting after laparoscopic cholecystectomy: a systematic review and meta-analysis. Ann Surg 2008;248(5):751−62.
[26] Lacasse Y, Goldstein R, Lasserson T, Martin S. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2006;18.
[27] Deeks J, Higgins J, Altman D. Analyzing data and undertaking meta-analyses. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0. Chichester, UK: Wiley; 2011.
[28] Fern E, Monroe K. Effect-size estimates: issues and problems in their interpretation. J Consum Res 1996;23:89−105.
[29] Cohen J. Statistical power analysis in the behavioral sciences. Hillsdale, NJ: Erlbaum; 1988.
[30] Dworkin R, Turk D, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008;9:105−21.
[31] Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ 1998;316: 690−3.
[32] Furukawa T. From effect size into number needed to treat. Lancet 1999;353:1680.
[33] Cox D, Snell E, editors. Analysis of binary data. London, UK: Chapman and Hall; 1989.
[34] Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995;117(1):167−78.
[35] Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Med Res Methodol 2008;8:32.
[36] Johnston B, Thorlund K, Schunemann H, Xie F, Murad M, Montori V, et al. Improving the interpretation of health-related quality of life evidence in meta-analysis: the application of minimal important difference units. Health Qual Life Outcomes 2010;11(8): 116.