BMJ Open

Presentation approaches for enhancing interpretability of patient-reported outcomes (PROs) in meta-analysis: a protocol for a systematic survey of Cochrane reviews

Tahira Devji,[1] Bradley C Johnston,[1,2,3,4,5] Donald L Patrick,[6] Mohit Bhandari,[1,7] Lehana Thabane,[1] Gordon H Guyatt[1,8]

To cite: Devji T, Johnston BC, Patrick DL, *et al*. Presentation approaches for enhancing interpretability of patient-reported outcomes (PROs) in meta-analysis: a protocol for a systematic survey of Cochrane reviews. *BMJ Open* 2017;**7**:e017138. doi:10.1136/bmjopen-2017-017138

► Prepublication history and additional material for this paper are available online. To view, please visit the journal (http://dx.doi.org/10.1136/bmjopen-2017-017138).

Received 6 April 2017
Revised 23 June 2017
Accepted 27 July 2017

For numbered affiliations see end of article.

Correspondence to
Tahira Devji;
devjits@mcmaster.ca

ABSTRACT

Introduction Meta-analyses of clinical trials often provide sufficient information for decision-makers to evaluate whether chance can explain apparent differences between interventions. Interpretation of the magnitude and importance of treatment effects beyond statistical significance can, however, be challenging, particularly for patient-reported outcomes (PROs) measured using questionnaires with which clinicians have limited familiarity. The objectives of our study are to systematically evaluate Cochrane systematic review authors' approaches to calculation, reporting and interpretation of pooled estimates of patient-reported outcome measures (PROMs) in meta-analyses.

Methods and analysis We will conduct a methodological survey of a random sample of Cochrane systematic reviews published from 1 January 2015 to 1 April 2017 that report at least one statistically significant pooled result for at least one PRO in the abstract. Author pairs will independently review all titles, abstracts and full texts identified by the literature search, and they will extract data using a standardised data extraction form. We will extract the following: year of publication, number of included trials, number of included participants, clinical area, type of intervention(s) and control(s), type of meta-analysis and use of the Grading of Recommendations, Assessment, Development and Evaluation approach to rate the quality of evidence, as well as information regarding the characteristics of PROMs, calculation and presentation of PROM effect estimates and interpretation of PROM effect estimates. We will document and summarise the methods used for the analysis, reporting and interpretation of each summary effect measure. We will summarise categorical variables with frequencies and percentages and continuous outcomes as means and/or medians and associated measures of dispersion.

Ethics and dissemination Ethics approval for this study is not required. We will disseminate the results of this review in peer-reviewed publications and conference presentations.

Strengths and limitations of this study

- Robust methodology using a systematic process to identify a comprehensive sample of Cochrane systematic reviews.
- Duplicate screening and data extraction.
- Detailed criteria for making judgements regarding authors' interpretation of the magnitude of pooled patient-reported outcome measure effect estimates that will ensure reproducible and accurate inferences.
- We will not extract data on patient-reported outcomes that are not statistically significant.

INTRODUCTION

Clinicians increasingly rely on summary effect estimates from systematic reviews and meta-analyses to guide their clinical decisions and to provide information for shared decision-making. When contemplating a recommendation, guideline developers also require best current evidence from systematic reviews to inform their decisions.

Systematic reviews and meta-analyses of clinical trials evaluating the effects of medical treatments and health interventions often include patient-reported outcome measures (PROMs), ideally with established measurement properties (eg, validity, responsiveness). Meta-analyses of clinical trials typically provide sufficient information for decision-makers to evaluate whether chance can explain apparent differences between interventions—this is true for PROMs as well as other measures. Interpretation of the magnitude of treatment effects, if authors present both relative and absolute effects, is relatively straightforward for binary outcomes and for continuous outcomes in which natural units are familiar to the target audience (eg, length of hospitalisation). For other continuous outcomes, and particularly for PROMs, interpretation can be much more difficult.

Challenges for PROM interpretation include clinicians' and patients' unfamiliarity with the measurement instruments. When the outcome in all studies is measured with the same instrument, the most straightforward analytic approach is to present the mean difference (MD), the absolute difference between the intervention and control mean responses in natural units. This may be straightforward, but still problematic in terms of the target audience intuitively understanding the magnitude of effect. For instance, without further information, clinicians may find it difficult to grasp the importance of a three-point difference in the St George's Respiratory Questionnaire or a one-point difference on a visual analogue scale for anxiety. Do these differences represent effects that are trivial, small but important, moderate or large in magnitude?

The situation becomes even more challenging when individual trials included in meta-analyses use different instruments to measure the same constructs. For example, one set of trials may have measured pain in patients with knee osteoarthritis using instruments with multiple domains (eg, pain, physical function) such as the Western Ontario and McMaster University Osteoarthritis Index,[1] and others may have used the Knee Injury and Osteoarthritis Outcome Score[2] or an instrument with a single domain, a visual analogue scale for pain intensity on movement.

When trials measure the same construct but the measurement instruments differ, systematic reviewers may generate a pooled estimate by dividing the difference between the intervention and control means (ie, the difference in means) in each trial by the estimated between-person SD for that trial.[3 4] The resulting summary effect measure, often referred to as the standardised mean difference (SMD) or effect size, is the longest standing and most widely used approach and is recommended in the Cochrane handbook for systematic reviews of interventions.[5]

This approach, however, has limitations. First, the SMD is expressed in SD units, which is not an intuitive summary effect measure for patients or clinicians.[6 7] Second, the SMD is vulnerable to differential variability in populations. That is, if the variability or heterogeneity in the severity of patients' condition (and thus the variability in scores on the chosen outcome) varies between trials, their SDs will also vary. As a result, clinical trials that enrol a heterogeneous group of patients will yield smaller SMDs than those that enrol less heterogeneous patients, even if the actual magnitude of treatment effects is similar.[8 9]

Many research groups have proposed alternative statistical approaches for presenting continuous outcomes from meta-analyses that may be more easily interpreted by clinicians than are these standard measures.[10–18] The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group has published an article providing an overview of methods for presenting pooled continuous outcomes in Summary of Findings (SoF) tables and evidence profiles.[8 9] The authors summarised the merits and limitations of each alternative and offered guidance for meta-analysts and guideline developers. Available evidence suggests that for clinician audiences, some ways of expressing effects (such as risk differences) are more easily understood, and more appealing, than others (such as the SMD).[7] Given the potential implications for decision-making in healthcare, it is important to explore how data from PROMs are summarised in published systematic reviews.

The objectives of this study are therefore to systematically evaluate how Cochrane reviews that provide summaries of PROMs suggesting underlying non-zero treatment effects calculate, present and interpret the results.

METHODS

Design overview

We will conduct a methodological survey of Cochrane systematic reviews. We will use standard methodology for conducting systematic reviews,[5] as described in previous protocols from our group.[19–23]

Definition of PROM instruments

For the purpose of this project, PROMs are self-report instruments on a continuous scale that address patient-important outcomes such as health-related quality of life (HRQoL), functional ability, symptom severity, satisfaction, psychological distress and well-being.[24]

Summary statistics for continuous outcomes

Table 1 summarises the categories of options available to systematic review authors in generating pooled estimates for continuous outcomes, in particular PROMs and their relative merits. In the following, we enumerate the approaches, with additional comments.

Mean difference

The MD, the absolute difference between the intervention and control mean responses, is typically used in circumstances in which investigators in the primary studies have used the same PROM. Interpretability of the MD can be facilitated by consideration of (1) the relation between the MD and the total range of possible instrument scores, (2) the minimal important difference (MID), the smallest change in instrument score that patients, on average, consider important and (3) referring to the wide experience of clinicians using the instrument in their clinical practice (if, as is not typical, such experience exists).

SMD

We have previously noted how to compute an SMD and the strengths and limitations of this approach, also summarised in table 1.

To aid interpretability of a metric unfamiliar to clinicians or patients, Cohen provided a rule of thumb to guide the significance of various effect sizes. An SMD in the range of 0.2 represents a small effect, in the range of 0.5 represents a moderate effect and in the range of 0.8 a large effect.[3] Some studies have suggested that an

Table 1 Approaches to presenting results of continuous variables in meta-analysis

Approach	Advantages	Disadvantages
When primary studies have used the same instrument		
Mean difference	Data are presented on the scale of the instrument Easier to interpret if instrument is well known	Few instruments sufficiently used in clinical practice to make units easily interpretable
When primary studies have used different instruments to measure the same construct		
SD units (standardised mean difference; effect size)	Widely used	Interpretation challenging Can be misleading depending on whether population is very homogeneous or heterogeneous
Present as natural units	May be viewed as closer to primary data	Few instruments sufficiently used in clinical practice to make units easily interpretable
Relative and absolute effects (eg, relative risk, OR, risk difference)	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)
Ratio of means	May be easily interpretable to clinical audiences Involves fewer questionable assumptions than some other approaches	Cannot be applied when measure is change and therefore negative values possible Interpretation requires knowledge and interpretation of control group mean
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

Table reproduced from GRADE guidelines: 13[8]
GRADE, Grading of Recommendations, Assessment, Development and Evaluation; MID, minimally important difference.

effect size of 0.5, or half a SD, roughly corresponds to the MID.[25 26] Other investigators suggest this rule may be excessively simple, and further studies are needed to determine its usefulness.[16]

Conversion into units of the most commonly used instrument

When primary studies have used different PROMs to measure the same construct, one can generate a MD by converting back to natural units of the most popular PROM or the PROM with the best measurement properties. One way of implementing this approach is a direct conversion from the SMD: one chooses the SD of the selected instrument and multiplies that value by the SMD. Far preferable, because it avoids vulnerability to varying SDs, is to convert scores to units of the most popular instrument for each individual study and then pool across studies.[18]

Conversion to relative and absolute effects

To enhance interpretability, systematic review authors may convert a continuous measure into a dichotomy and calculate relative or absolute effects on a binary scale. A set of methods to generate a dichotomy from continuous data rely on the SMD; these typically assume that results of both treatment and control groups are normally distributed and have equal variances.[14 27] These approaches typically require an estimate of response rates in the control group or in the treatment group and allow transformation of an SMD into either relative effects, typically an OR, or absolute effects, typically a risk difference (RD), the difference between the observed risk in the experimental and the control groups. There are, however, other statistical approaches that also rely on the SMD to generate dichotomous presentations for continuous outcomes but do not require specification of the control group response rate.[28 29]

The number needed to treat (NNT), the inverse of the RD, represents the number of persons who need to be treated to prevent one additional outcome and provides an alternative way of expressing the RD. An excel spreadsheet is available to calculate NNT for any effect size and any response rate (see NNT Calculator two at http://ebmh.med.kyoto-u.ac.jp/toolbox.html).

Another strategy for creating dichotomies and generating estimates of relative and absolute effect relies on knowledge of the MID. Using this approach, one assumes normal distributions of data and then, for each trial, calculates the probability of experiencing a treatment effect larger than the MID in intervention and control groups and then pools the resulting proportions across studies.[18] This approach has the advantage of avoiding the SMD's vulnerability to differing heterogeneity of patients across studies and thus differing SDs.

Ratio of means

The ratio of means (RoM) method produces a relative measure of comparative effect by dividing the mean response in the intervention group by the mean response in the control group.[12] One limitation of the RoM method is that it is designed for post-test scores, as mean values of the intervention and control groups must both be in the same direction (both intervention and control group change being simultaneously positive or negative).[9]

MID units

This strategy is similar to the SMD approach in that it pools across studies by standardising the MD, but instead of dividing the MD of each study by its SD, it divides by the MID associated with the PROM used in that study. A natural, but misleading, interpretation of MID units would be that if the result is below 1, the treatment effect is unimportant. This interpretation assumes, erroneously, that the effect will be identical in every patient. Given that this is not the case, even if the average effect is smaller than the MID, there is still possibility that a sizeable proportion of patients experience an effect greater than or equal to the MID.[17] This argues for complementing presentation in MID units with calculation and presentation of absolute effects, the RD and, potentially, the corresponding NNT.

Eligibility criteria

We will include systematic reviews published in the Cochrane database of systematic reviews meeting the following criteria:

1. Described as a 'meta-analysis' of randomised controlled trials (RCTs);
2. Published from 1 January 2015 to 1 April 2017;
3. Includes a comparison of an intervention with another intervention in human participants;
4. Reports in the abstract a statistically significant measure of effect (p value<0.05 or CI excluding a null effect) for at least one continuous outcome, specifically a PROM, from a pooled analysis.

We will exclude network meta-analyses.

Literature search

We will search the Cochrane Database of Systematic Reviews in the Cochrane Library for potentially eligible systematic reviews. We will limit the search to reports published from 1 January 2015 to 1 April 2017. The search strategy is presented in online appendix 1 .

Review process

Teams of two trained reviewers will perform title and abstract and full-text screening and data abstraction independently and in duplicate, including the selection of the PROM (using prespecified criteria—see below). Each team will resolve disagreements through discussion and, when unsuccessful, through consultation with one of two arbitrators (TD, GHG). To ensure consistency across reviewers we will, prior to commencing the review, conduct calibration exercises until reviewers achieve near-perfect agreement. We will use Microsoft Excel for eligibility screening and data extraction. These forms will be standardised and pilot tested, and we will provide reviewers with detailed written instructions to assist with study screening and extraction.

Choosing eligible studies

Using the 'RAND' function in excel, we will randomly sample aliquots of 100 citations from our search results. Teams of two reviewers will screen titles and abstracts for potential eligibility and citations identified as potentially eligible by either reviewer will proceed to full-text review. In the title and abstract screening, reviewers will judge if the study is a systematic review of randomised trials evaluating treatment effects in human participants, and if the authors report at least one apparent PROM that is statistically significant. Reviewers will independently review full texts of citations flagged as potentially eligible in duplicate to determine final eligibility. We will continue the random sampling process until the number of eligible studies meets our required sample size.

Data abstraction

Reviewer teams will abstract data from eligible reviews using a pilot-tested standardised data abstraction form (see online appendix 2) with corresponding detailed instructions. As with study screening, we will perform calibration exercises prior to commencing data extraction. Reviewers will obtain data for all patient-reported outcomes (PROs) reported as statistically significant in the report. If more than one pairwise comparison with a statistically significant result for the same PROM is reported, reviewers will select the comparison that reports the largest number of patients included in the analysis. When multiple meta-analyses for outcomes of interest are performed within a single review, we will use data only from the primary analysis, and not subgroup or sensitivity analyses, with the exception of sensitivity analyses for alternative presentation approaches.

Study characteristics

For all included systematic reviews, we will extract the following information:

1. Year of publication;
2. Number of included trials;
3. Number of included participants in the intervention and control arms;
4. Clinical area;
5. Type of intervention and control;
6. Type of meta-analysis (standard meta-analysis vs individual participant data meta-analysis);
7. Use of the GRADE approach to rate confidence in effect estimates[30];
8. Outcomes of morbidity and mortality, reported as either primary or secondary outcomes and authors' conclusions about the magnitude of effect.

Characteristics of PROMs

We will document the construct that represents the continuous outcome of interest (eg, pain, function, HRQoL, etc) and the name of the instrument(s) measuring the construct(s). We will record whether authors described or provided a citation reporting the measurement properties, including evidence of MID estimates, of the PROM(s). We will document whether investigators describe their approach to PROM selection when more than one PROM capturing the same construct(s) is reported within a single trial. For instance, the authors may describe a hierarchical approach that may be based on evidence of PROM validity and responsiveness. For each PRO of interest (ie, each construct), we will note whether systematic reviewers pooled the same or different PROMs.

Calculation and presentation of PROM estimate of effects

We will record the type of summary effect measure (MD, SMD, RoM, MID units, dichotomous absolute (RD, NNT) and relative effects (RR, OR)) for the selected PROM(s). If more than one is reported, we will summarise all effect measures reported, and whether the authors have specified one as the primary analysis. We will document the corresponding point estimates and 95% CIs for each summary effect measure. We will explore how authors calculated pooled estimates for different measures (eg, obtaining absolute effects using an MID threshold to dichotomise patients). We will document whether the authors pooled differences in post-test scores or changes from baseline between the intervention and control or a combination of both. For the reviews presenting pooled estimates in MID units, or creating dichotomies and generating absolute or relative effects from an SMD or MD, we will document the number of estimates of effect presented for a range of MID estimates or plausible control group response rates and, if available, the source of these response rates or MID estimates. If needed, we will contact authors for additional information.

Interpretation of PROM effect estimates

Regarding interpretation, we will document the extent to which authors discuss their interpretation of the main effect of interest. Possibilities will include no comment on magnitude of effect or characterisation of effects as trivial, small but important, moderate or large. Another possibility is commenting on the size of positive effects in relation to the harms or burdens associated with the intervention (magnitude of effect outweighing, or not outweighing, burdens and harms). When authors have made inferences regarding magnitude of effect, we will document the basis of these inferences. These might include, for example, reference to an MID (or some other meaningful threshold); Cohen's interpretation of effect sizes; the instrument's total score range; clinicians' or patients' intuition.

Developing criteria for making decisions regarding authors' interpretation will be challenging. We have experience with this sort of judgement. For instance, we developed detailed criteria for classifying inferences regarding subgroup effects as strong claim, claim of likely effect and claim of a suggestion of possible effect.[31] We anticipate an iterative process based on examination of the wording authors use to communicate their inferences regarding magnitude of effect, ultimately leading to guidance that allows reproducible judgments. Box 1 presents preliminary criteria to inform judgments about authors' interpretation of pooled estimates of PROMs.

Box 1 Criteria for judging authors' interpretation of pooled estimates of patient-reported outcome measures (PROMs)

Criteria

- Did the review authors dichotomise continuous PROM data and present as a pooled relative or absolute estimate informed by a minimally important difference (MID) or some other meaningful threshold?
- Did the review authors present summary effect estimates in MID units?
- Did the review authors present summary effect estimates as a ratio of means?
- Did the review authors characterise the magnitude of effect in reference to an MID, some other meaningful threshold, Cohen's interpretation of effect sizes, PROM's total score range, clinician's or patient's intuition?
- Did the authors use only descriptive words (eg, trivial, small but important, moderate, large) to characterise the magnitude of effects?
- Did the authors comment on the magnitude of effect outweighing or not outweighing burdens or harms associated with the intervention?
- Did the investigators indicate the need for empirically determined thresholds (eg, MID, responder criteria) to quantify the importance of apparent effects?

For reviews that report more than one presentation approach, either as primary or sensitivity analyses, we will document whether the authors discuss if results are or are not congruent, and if they do what they conclude. We will document authors' discussion of limitations or uncertainties regarding their characterisations of magnitude of effect.

Information regarding interpretation may come from either the text of the articles, including results and discussion or described as a comment or footnote in the GRADE SoF table. We will document authors' conclusions regarding the treatments under investigation; that is, whether review authors make a recommendation either explicitly or implicitly for or against a particular treatment or no recommendation at all and the results that formed the basis of the conclusions.

Sample size

The primary aim is to estimate the proportion of reviews that provide interpretation of magnitude of effect. We will estimate our sample size to achieve a 95% CI estimate with a margin of error of +/-0.05 around the estimated proportion of reviews that provide interpretation of magnitude of effect. Assuming a conservative prior estimate of the proportion to be 0.5, we would need 200 reviews to achieve the desired confidence interval (0.43, 0.57).

Analysis

Agreement

We will assess agreement between reviewers for study inclusion at the full-text screening stage and reviewers' judgements about authors' interpretation of PROM effect estimates. We will calculate both crude agreement and chance-correlated agreement (kappa statistic). If fewer than 15% or more than 85% of citations are included in this study, we will measure agreement using chance-independent agreement (phi statistic). We will interpret the agreement statistics using the guidelines proposed by Landis and Koch[32] : kappa values of 0 to 0.2 represent slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement and greater than 0.80 almost perfect agreement. We will use these same thresholds for interpreting phi.

Description of the data

We will provide a summary of the number of included trials, number of participants, clinical area, type of intervention and control, type of meta-analysis and use of GRADE. We will conduct a descriptive analysis of all variables. We will summarise categorical variables with frequencies and percentages and continuous outcomes as means and/or medians and associated measures of dispersion (SD, IQR, range).

Methods for calculation and presentation of PROM estimate of effects

We will calculate the proportion of systematic reviews presenting a pooled estimate for each summary effect measure (SMD, RR, RD, etc). For reviews creating dichotomies and generating absolute or relative effects from an SMD or MD, we will calculate the proportion of systematic reviews that provided a source for control group response rates or MID estimates. We will also calculate the proportion of reviews that use a single value versus a range of plausible values for control group response rates or MID. For reviews pooling different PROMs and generating an MD by converting back to natural units, we will summarise investigators approach to PROM selection.

Interpretation of PROM effect estimates

We will calculate the proportion of systematic reviews that discuss the interpretation of PROM results beyond reporting statistical significance. We will summarise authors' inferences regarding magnitude of effect and the basis of these inferences. We will evaluate the relation between the results and authors' inferences by comparing the consistency, or lack there of, in authors' interpretation of magnitude of summary effect estimates from each presentation format across systematic reviews. We will also assess how authors' inferences regarding the magnitude of the treatment effect estimate compares to a guide we have previously developed for categorising effect sizes that correspond to small and large treatment effects (see online appendix 3).[7] We will also summarise whether review authors made a recommendation either explicitly or implicitly for or against a particular treatment or no recommendation at all and the relation of such conclusions to the inferences made from the magnitude of PROM and non-PROM (eg, morbidity and mortality) effect estimates.

For reviews that report more than one presentation approach to enhance interpretability, we will calculate the proportion that discuss congruency or lack there of in the results and document details from authors' discussion about congruency. We will calculate the proportion of reviews that provide a discussion of limitations regarding their characterisations of magnitude of effect.

DISCUSSION

Main objectives of our study

Our review will systematically evaluate how Cochrane systematic reviews with summaries of PROs suggesting underlying non-zero treatment effects calculate, present and interpret the results. By publishing our detailed study protocol we are reflecting our commitment to making the objectives and design of methodological studies more transparent.

Strengths and limitations

Our study has several strengths. Our systematic survey will be the first to evaluate current practice of Cochrane reviewers in summarising evidence from PROs and their practice in interpreting results regarding the magnitude of effects. In this empirical study, we will use robust methodology including explicit eligibility criteria, a systematic process to identify a comprehensive sample of eligible Cochrane systematic reviews, the use of standardised forms for study screening and data abstraction and detailed criteria for making judgements regarding authors interpretation that will ensure reproducible and accurate inferences. We will pilot these forms and develop detailed instructions for both study screening and data extraction and achieve near-perfect agreement between reviewers during calibration exercises before commencing study selection and data abstraction. We will evaluate each of the reviews and extract data in duplicate and independently, confirming the reproducibility of judgements.

Our study has potential limitations. First, it will involve several reviewers' judgements at each step of the process. However, detailed instructions, piloting and calibration exercises will minimise disagreement. Second, some of the reviewers may be less experienced in MID, PRO and meta-analysis methods than others. To overcome this limitation, we will partner less experienced reviewers with those who are more experienced. Third, we focus only on RCTs and have excluded Non-randomized Sudies (NRS).

Lastly, two of the available presentation methods, relative and absolute dichotomised effects using a MID and MID units are only applicable when an MID is known. In addition, to the extent that the MID estimate is not based on empirically sound evidence, these approaches become less trustworthy. At the time of writing this protocol, our group is conducting several projects to advance MID methods, including the development of an instrument for evaluating the credibility of anchor-based MID estimates, as well as a systematic review to identify published anchor-based MIDs for all known PRO instruments.[33] Prior to completion of our proposed study, should preliminary data from our compendium of anchor-based MIDs allow us to determine the availability of MIDs for PROs evaluated in eligible systematic reviews, we will be able to provide a more accurate depiction of systematic review authors' choice of presentation approaches. In the absence of such information, when authors do not refer to an MID, it will remain unclear whether authors simply prefer other presentation approaches that do not rely on the MID or if an anchor-based MID is not established for the PROMs of interest and thus, precluded authors from conducting MID dependent analyses.

Implications

PROs provide patients' insights on the impact of a disease or treatment, and investigators increasingly rely on these outcomes as key endpoints in clinical trials. Given the pervasiveness and influence of PROs in both clinical and research practice, enhanced interpretation of PRO results from systematic reviews and meta-analyses is required to inform optimal shared decision-making. The findings of this study will inform the systematic review community regarding the current practice of summarising and presenting effect estimates for continuous variables, specifically PROs, in Cochrane systematic reviews. Our findings with regard to possible underuse of available methods, possible deficiencies in interpretation and consistency or inconsistency of apparent magnitude of effect will influence recommendations on reporting, conduct and interpretation of PROs. Our results are likely to be of interest for systematic review authors and guideline developers, funding agencies, health decision-makers and journal editors.

Ethics and dissemination

Ethics approval for this study is not required. We will disseminate the results of this review in peer-reviewed publications and conference presentations.

Author affiliations
[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[2]Systematic Overviews through Advancing Research Technology (SORT), Child Health Evaluative Sciences, The Research Institute, The Hospital For Sick Children, Peter Gilgan Centre for Research and Learning, Toronto, Canada
[3]Institute of Health Policy Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[4]Department of Anesthesia and Pain Medicine, The Hospital For Sick Children, Toronto, Canada
[5]Department of Community Health and Epidemiology, Dalhousie University, Centre for Clinical Research, Halifax, Canada
[6]Department of Health Services, University of Washington, Seattle, USA
[7]Division of Orthopaedic Surgery, McMaster University, Hamilton, Canada
[8]Department of Medicine, McMaster University, Hamilton, Canada

Contributors TSD, DLP and GHG conceptualised the study. All authors contributed to design of the study and read and approved the manuscript. TSD developed the first draft of the manuscript and incorporated comments from authors for successive drafts.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sector. TSD, BCJ and GHG are Canadian Institutes of Health Research, Knowledge Synthesis grant (DC0190SR) recipients for two projects related to MID methods.

Competing interests None declared.

Patient consent Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Bellamy N, Buchanan WW, Goldsmith CH, *et al*. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;15:1833–40.
2. Roos EM, Roos HP, Lohmander LS, *et al*. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88–96.
3. Cohen J. *Statistical power analysis for the behavioural sciences. Hillside*. NJ: Lawrence Earlbaum Associates, 1988.
4. Hedges LV, Olkin I. *Statistical methods for meta-analysis*: Academic press, 2014.
5. Higgins JP, Green S. *Cochrane handbook for systematic reviews of interventions*: Wiley Online Library,, 2008.
6. Fern EF, Monroe KB. Effect-Size Estimates: Issues and Problems in Interpretation. *J Consum Res* 1996;23:89–105.
7. Johnston BC, Alonso-Coello P, Friedrich JO, *et al*. Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. *CMAJ* 2016;188:25–32.
8. Guyatt GH, Thorlund K, Oxman AD, *et al*. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol* 2013;66:173–83.
9. Johnston BC, Patrick DL, Thorlund K, *et al*. Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. *Health Qual Life Outcomes* 2013;11:211.
10. Anzures-Cabrera J, Sarpatwari A, Higgins JP. Expressing findings from meta-analyses of continuous outcomes in terms of risks. *Stat Med* 2011;30:2967–85.
11. da Costa BR, Rutjes AW, Johnston BC, *et al*. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. *Int J Epidemiol* 2012;41:1445–59.
12. 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.
13. Friedrich JO, Adhikari NK, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol* 2011;64:556–64.
14. Furukawa TA. From effect size into number needed to treat. *The Lancet* 1999;353:1680.
15. Hasselblad V, McCrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;15:81–96.
16. Johnston BC, Thorlund K, da Costa BR, *et al*. New methods can extend the use of minimal important difference units in meta-analyses of continuous outcome measures. *J Clin Epidemiol* 2012;65:817–26.
17. Johnston BC, Thorlund K, Schünemann HJ, *et al*. Improving the interpretation of quality of life evidence in meta-analyses: the

application of minimal important difference units. *Health Qual Life Outcomes* 2010;8:116.

18. Thorlund K, Walter SD, Johnston BC, *et al*. Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. *Res Synth Methods* 2011;2:188–203.
19. Akl EA, Briel M, You JJ, *et al*. LOST to follow-up Information in Trials (LOST-IT): a protocol on the potential impact. *Trials* 2009;10:40.
20. Alonso-Coello P, Carrasco-Labra A, Brignardello-Petersen R, *et al*. A methodological survey of the analysis, reporting and interpretation of Absolute Risk ReductiOn in systematic revieWs (ARROW): a study protocol. *Syst Rev* 2013;2:113.
21. Briel M, Lane M, Montori VM, *et al*. Stopping randomized trials early for benefit: a protocol of the Study Of Trial Policy Of Interim Truncation-2 (STOPIT-2). *Trials* 2009;10:49.
22. Kasenda B, von Elm EB, You J, *et al*. Learning from failure--rationale and design for a study about discontinuation of randomized trials (DISCO study). *BMC Med Res Methodol* 2012;12:131.
23. Sun X, Briel M, Busse JW, *et al*. Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. *Trials* 2009;10:101.
24. Locklear T. *Reaching consensus on patient-centered definitions: a report from the Patient-Reported Outcomes PCORnet Task Force*, 2015.
25. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92.
26. Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *COPD* 2005;2:57–62.
27. Suissa S. Binary methods for continuous outcomes: a parametric alternative. *J Clin Epidemiol* 1991;44:241–8.
28. Cox DR, Snell EJ. *Analysis of binary data*: CRC Press, 1989.
29. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167–78.
30. Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–94.
31. Sun X, Briel M, Busse JW, *et al*. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012;344:e1553.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
33. Johnston BC, Ebrahim S, Carrasco-Labra A, *et al*. Minimally important difference estimates and methods: a protocol. *BMJ Open* 2015;5:e007953.