Journal of Clinical Epidemiology 68 (2015) 489–497

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Grades for quality of evidence were associated with distinct likelihoods that treatment effects will remain stable

Gerald Gartlehner[a,b,*], Isolde Sommer[a], Tammeka Swinson Evans[b], Kylie Thaler[a], Kathleen N. Lohr[b]

[a]*Department for Clinical Epidemiology an Evidence-based Medicine, Austrian Cochrane Branch, Danube University Krems, Dr. Karl Dorrekstrasse 30, 3500 Krems, Austria*

[b]*RTI-UNC Evidence-based Practice Center, Research Triangle Institute International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709-2194, USA*

Accepted 24 September 2014; Published online 4 December 2014

Abstract

Objectives: We sought to determine whether producers or users of systematic reviews using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach or a close variation give the same meanings to terms intended to convey uncertainty about treatment effects when interpreting grades for the quality or strength of evidence.

Study Design and Setting: Following exploratory interviews with stakeholders and user testing, we conducted an international Web-based survey among producers and users of systematic reviews. For each quality grade (high, moderate, low, very low/insufficient), we asked participants to assign a minimum likelihood that treatment effects will not change substantially as new studies emerge. Using multivariate analysis of covariance, we tested whether the estimated likelihoods differed between producers and users.

Results: In all, 244 participants completed the survey. The associated minimum likelihoods that treatment effects will not change substantially for high, moderate, and low grades of quality of evidence (QOE) were 86% [95% confidence interval (CI): 85%, 87%], 61% (95% CI: 59%, 63%), and 34% (95% CI: 32%, 36%), respectively (very low was preset at 0%). Likelihoods for each grade were similar between producers and users of systematic reviews ($P > 0.05$ for all comparisons).

Conclusion: GRADE is, in general, a suitable method to convey uncertainties for systematic review producers to users. The wide ranges of likelihoods associated with GRADE terms suggest that current definitions of levels of QOE that rely exclusively on qualitative certainty expressions should be augmented by numerical predictions once such data are available.

Keywords: GRADE; Systematic reviews; Quality of evidence; Survey; Stability of treatment effects; Certainty of treatment effects

1. Introduction

Systematic reviews strive to provide scientifically rigorous, independent, and accurate summaries of the scientific evidence with respect to a specific question of interest. Because the quality of scientific evidence varies, grades that convey the strength of the evidence with respect to a specific outcome of interest have become important tools to communicate certainties and uncertainties to readers and other stakeholders [1,2].

Over the past decade, many organizations producing or using systematic reviews have adopted the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach as a rating system that explicitly and transparently reflects the confidence that researchers have in an available body of evidence [3]. Some organizations—such as the Evidence-based Practice Center (EPC) program of the US Agency for Healthcare Research and Quality (AHRQ)—have made small adaptations to the GRADE system to meet their specific needs [4,5]. In this

Funding: The survey and the analyses received no external funding. The exploratory interviews with stakeholders were funded by the Agency for Healthcare Research and Quality under contract number HHSA290201200008i to RTI International. The authors of this report are responsible for its content. Statements in this manuscript should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services.

Conflict of interest: G.G. and K.T. are members of the GRADE Working Group. G.G. and K.N.L. are coauthors of the AHRQ guidance for grading the strength of evidence. The other authors have no disclosures to report.

* Corresponding author. Tel.: +43 2732 893 2911; fax:+43 2732 893 4910.

E-mail address: gerald.gartlehner@donau-uni.ac.at (G. Gartlehner).

http://dx.doi.org/10.1016/j.jclinepi.2014.09.018
0895-4356/

What is new?

Key findings

- Survey responders expect that effect estimates graded as high quality of evidence (QOE) have a likelihood of remaining stable that ranges from 86% to 100%. For effect estimates graded as moderate QOE, the likelihood of remaining stable as new studies emerge ranges from 61% to 85%, for low QOE from 34% to 60%, and for very low QOE 0% to 33%. Associated likelihoods for each grade (high, moderate, low, very low/insufficient) were similar between producers and users of systematic reviews using Grading of Recommendations Assessment, Development and Evaluation (GRADE). However, individual answers within the groups of producers and users ranged widely for each grade.

What this adds to what was known?

- GRADE is a suitable method for producers of systematic reviews to convey uncertainties to users because both groups give similar meanings to quality of evidence grades.

What is the implication and what should change now?

- Current definitions of QOE in GRADE rely on qualitative expressions of certainty (e.g., ''likely''), and these terms are open to individual interpretations. Future research needs to assess the stability of treatment effects from bodies of evidence with different grades of QOE to quantify the proportion of treatment effects that can be expected to remain stable as new studies emerge. Adding such numerical forecasts of certainty to definitions of grades could reduce variation in how readers interpret those grades.

manuscript, we refer to both approaches collectively as the GRADE approach.

GRADE's conceptual framework uses information about risk of bias, imprecision, inconsistency, indirectness, and reporting bias to communicate the confidence that systematic reviewers have in estimates of treatment effects; GRADE calls this ''quality of evidence'' (QOE) [6]. In the context of systematic reviews, GRADE defines QOE as the extent of the confidence that the estimates of an effect are correct [6]. Likewise, guidance for EPCs defines QOE (which they refer to as ''strength of evidence'') as the degree of confidence that estimates are close to the true effect and the likelihood that findings will remain stable over time (i.e., the likelihood that future studies will not have an important impact on the estimate of an effect) [4]. Table 1 lists the definitions of the four levels of QOE of GRADE and the AHRQ guidance for GRADE.

Semantically, key elements of these definitions include the concepts of truth, confidence (in effect estimates), modifiers of levels of confidence (e.g., very, moderately, or limited), qualitative forecasts of certainty (e.g., likely, very likely), and deficiencies. Of these constructs, deficiencies are the easiest to understand. They refer to problems related to the risk of bias in studies included in systematic reviews and issues related to the consistency, precision, and other attributes of whole bodies of evidence. Experts developing systematic reviews typically use specific criteria for documenting deficiencies even if assigning ratings of risk of bias involves a considerable element of judgment.

By contrast, the other constructs reflect relatively unquantifiable ideas, which turn on the concept of true effects and certainty (or uncertainty) about true effects. A true effect can be viewed as the effect size that we would observe if a study had an infinitely large sample size (and thus no sampling error) [7]. Given the definitions of different grades of QOE, a rating of high QOE, for example, means that the observed effect is close to the true effect and thus will remain relatively unchanged even as new studies emerge; that is, it will remain stable. By contrast, a rating of low QOE means that the likelihood is high that future studies will have a substantial impact on the direction or magnitude of the estimate of effect of a given outcome.

To date, the interpretation of QOE grades by producers and users of systematic reviews has not been tested. The concern is that both systematic reviewers and users of such reports give different meanings to terms intended to convey certainty (or uncertainty) when interpreting the ratings of QOE. This factor may lead to different interpretations of the same information and, subsequently, to different decisions by users of that information. For example, producers of systematic reviews for clinical practice guidelines could grade critical outcomes of a new promising intervention as low QOE because of concerns of directness and an imprecise effect estimate. However, they could still view the treatment effect as clinically important and recommend the new intervention. On a policy level, decision makers might decide against reimbursing this intervention because they interpret ''low quality evidence'' as not solid enough to capture ''the true effect'' and to be used for a change in reimbursement policies.

Realistically, a true treatment effect can rarely be determined as a reference standard. For that reason, here, we equate true effect with stability of effect as new studies emerge, a concept that can be measured. The objective of our study was to determine to what extent the likelihoods that effects will remain stable associated with each QOE grade are similar for producers and users of systematic reviews.

2. Methods

To achieve our objective, we conducted an international Web-based survey in English among producers and users of