



# High quality (certainty) evidence changes less often than low-quality evidence, but the magnitude of effect size does not systematically differ between studies with low versus high-quality evidence

Benjamin Djulbegovic MD, PhD<sup>1</sup> | Muhammad Muneeb Ahmed MS<sup>2</sup> |  
Iztok Hozo PhD<sup>3</sup> | Despina Koletsi DDS<sup>4</sup> | Lars Hemkens MD, MPH<sup>5,6,7</sup> |  
Amy Price PhD<sup>8</sup> | Rachel Riera MD<sup>9</sup> | Paulo Nadanovsky PhD<sup>10</sup> |  
Ana Paula Pires dos Santos PhD<sup>11</sup> | Daniela Melo PhD<sup>12</sup> | Ranjan Pathak MD<sup>13</sup> |  
Rafael Leite Pacheco MD<sup>14</sup> | Luis Eduardo Fontes MD<sup>14,15</sup> |  
Anderson Miranda MSc<sup>15</sup> | David Nunan PhD<sup>15,16</sup>

<sup>1</sup>Department of Computational & Quantitative Medicine, Beckman Research Institute, City of Hope, Duarte, California, USA

<sup>2</sup>Michael G. DeGroot School of Medicine, McMaster University, Hamilton, Ontario, Canada

<sup>3</sup>Department of Mathematics, Indiana University Northwest, Gary, Indiana, USA

<sup>4</sup>Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

<sup>5</sup>Department of Clinical Research, University of Basel, Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Basel, Switzerland

<sup>6</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA

<sup>7</sup>Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany

<sup>8</sup>Anesthesia Informatics and Media Lab, Stanford University, Stanford, California, USA

<sup>9</sup>Universidade Federal de São Paulo, Escola Paulista de Medicina, Brazil (Unifesp), São Paulo, Brazil

<sup>10</sup>Department of Epidemiology and Quantitative Methods in Health, National School of Public Health, Fundação Oswaldo Cruz (FIOCRUZ) - Department of Epidemiology, Institute of Social Medicine, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

<sup>11</sup>Department of Pharmaceutical Sciences, Universidade Federal de São Paulo (Unifesp), Rio de Janeiro, Brazil

<sup>12</sup>Department of Community and Preventive Dentistry, Faculty of Dentistry, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

<sup>13</sup>Department of Medical Oncology and Therapeutics Research, City of Hope, Duarte, California, USA

<sup>14</sup>Centro Universitário São Camilo, Researcher at the Center of Health Technology Assessment, Hospital Sirio-Libanês, São Paulo, Brazil

<sup>15</sup>Department of Intensive Care, and Emergency Medicine at Faculdade de Medicina de Petrópolis, in Petrópolis, Rio de Janeiro, Brazil

<sup>16</sup>Kellogg College, University of Oxford, Oxford, UK

## Correspondence

Benjamin Djulbegovic, MD, PhD, Department of Computational & Quantitative Medicine, Beckman Research Institute, City of Hope, 1500 E Duarte Road, Duarte, CA 91010-3000, USA.

Email: [bdjulbegovic@coh.org](mailto:bdjulbegovic@coh.org)

## Abstract

**Rationale, Aims, and Objectives:** It is generally believed that evidence from low quality of evidence generate inaccurate estimates about treatment effects more often than evidence from high (certainty) quality evidence (CoE). As a result, we would expect that (a) estimates of effects of health interventions initially based on

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Journal of Evaluation in Clinical Practice* published by John Wiley & Sons Ltd.

**Funding information**

Agency for Healthcare Research and Quality,  
Grant/Award Number: R01HS024917

high CoE change less frequently than the effects estimated by lower CoE (b) the estimates of magnitude of effect size differ between high and low CoE. Empirical assessment of these foundational principles of evidence-based medicine has been lacking.

**Methods:** We reviewed the Cochrane Database of Systematic Reviews from January 2016 through May 2021 for pairs of original and updated reviews for change in CoE assessments based on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) method. We assessed the difference in effect sizes between the original versus updated reviews as a function of change in CoE, which we report as a ratio of odds ratio (ROR). We compared ROR generated in the studies in which CoE changed from very low/low (VL/L) to moderate/high (M/H) versus M/H to VL/L. Heterogeneity and inconsistency were assessed using the tau and  $I^2$  statistic. We also assessed the change in precision of effect estimates (by calculating the ratio of standard errors) (seR), and the absolute deviation in estimates of treatment effects (aROR).

**Results:** Four hundred and nineteen pairs of reviews were included of which 414 ( $207 \times 2$ ) informed the CoE appraisal and 384 ( $192 \times 2$ ) the assessment of effect size. We found that CoE originally appraised as VL/L had 2.1 [95% confidence interval (CI): 1.19–4.12;  $p = 0.0091$ ] times higher odds to be changed in the future studies than M/H CoE. However, the effect size was not different ( $p = 1$ ) when CoE changed from VL/L  $\rightarrow$  M/H [ROR = 1.02 (95% CI: 0.74–1.39)] compared with M/H  $\rightarrow$  VL/L (ROR = 1.02 [95% CI: 0.44–2.37]). Similar overlap in aROR between the VL/L  $\rightarrow$  M/H versus M/H  $\rightarrow$  VL/L subgroups was observed [median (IQR): 1.12 (1.07–1.57) vs. 1.21 (1.12–2.43)]. We observed large inconsistency across ROR estimates ( $I^2 = 99\%$ ). There was larger imprecision in treatment effects when CoE changed from VL/L  $\rightarrow$  M/H (seR = 1.46) than when it changed from M/H  $\rightarrow$  VL/L (seR = 0.72).

**Conclusions:** We found that low-quality evidence changes more often than high CoE. However, the effect size did not systematically differ between the studies with low versus high CoE. The finding that the effect size did not differ between low and high CoE indicate urgent need to refine current EBM critical appraisal methods.

**KEYWORDS**

critical appraisal-bias, evidence-based medicine, meta-epidemiology, observational studies, random error, randomized trials, systematic review

**1 | INTRODUCTION**

A foundational epistemological principle underpinning evidence-based medicine (EBM) is based on the assumption that the estimates of the effects of health interventions are closer to the 'truth' if they are based on higher than on lower quality (certainty) of evidence (CoE).<sup>1</sup> If the estimated treatment effects are close to the 'true' effects, this would also imply that they would less likely to change as evidence accumulates after new studies are completed. Conversely, because its relation to the 'truth' is less certain, this also implies that the estimated effects when evidence is of low quality would more

likely change in future research. Research to date indicates that guideline panels are willing to issue stronger recommendations when they deem evidence to be of high quality, thus indirectly affirming this central EBM assumption.<sup>2–5</sup>

However, whether this indirect assessment of quality of evidence based on guidelines panels' decision-making is accurate is not known. It is possible that current methods of critical appraisal of CoE do not discriminate well between 'true' accurate from inaccurate estimates of treatment effects. That is, the effects of health interventions based on low quality of evidence may turn out to reflect 'true effects' by testing in subsequent studies. On the other hand,



what was originally deemed as high-quality evidence may be undermined by future studies more often than initially expected. Thus, it is not known if low-quality evidence is more often revised than high-quality evidence. Empirical evidence supporting this foundational principle of EBM is lacking.

The main purpose of this report is to assess if (a) low certainty evidence is more often revised than high certainty evidence in subsequent studies and if (b) the magnitude of effect size differs between high and low CoE.

## 2 | METHODS

We assessed the change in CoE between the original and updated Cochrane systematic reviews, which reported rating of CoE as per the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for critical appraisal of medical evidence.<sup>6</sup> We used GRADE as this has been widely recognized as the most advanced system for operationalization of fundamental principles of EBM and critical evaluation of medical evidence.<sup>1,7,8</sup> GRADE was developed in the first decade of 21st Century after critical appraisal of 106 systems for rating of the quality of medical research evidence showed that none of them was capable of distinguishing low from high-quality evidence.<sup>1,9,10</sup>

We focused on the assessment of systematic reviews, rather on individual trials, because the second important EBM principle is that assessment of the true effects of health interventions is best accomplished by evaluating total evidence on the topic rather than based on a study selected to favour a particular claim.<sup>1</sup> GRADE is also considered a suitable method to assess certainty of evidence at the level of systematic review/meta-analysis.<sup>8</sup> Thus, the unit of our analysis was a systematic review/meta-analysis (SR/MA).

Cochrane Reviews are regularly updated providing a unique opportunity to assess when and whether the assessment of CoE changes between the original and updated reviews as a result of new evidence generated between two reviews. Since 2013 Cochrane Reviews have mandated the use of GRADE Summary of Findings (SoF)<sup>11</sup> to summarize CoE and magnitude effects of interventions that the reviews assessed. We evaluated all Cochrane reviews published in the last 5 years in the Cochrane Database of Systematic Reviews [<https://www.cochranelibrary.com/cdsr/about-cdsr>].

We used SoFs from the original and updated reviews to extract data for the primary outcome related to CoE and to assess the magnitude and direction of effect. (In case of multiple primary outcomes, the data were extracted from the first one listed in SoF table that contained data in both original and updated review). Eligible SR/MAs were divided into five groups; data were extracted from each group by pairs of independent reviewers. Kappa interrater agreement was calculated for each pair regarding CoE. As explained, we recorded CoE according to GRADE criteria (very low, low, moderate and high).<sup>1,12</sup>

We also extracted summary meta-analytic estimates for the primary outcome from each pair of reviews, that is, point estimates,

dispersion (e.g., 95% confidence interval), metric used (e.g., relative risk, odds ratio, hazard ratio, standardized mean differences, etc.), number of trials per meta-analysis, number of participants, type of comparator (active vs. placebo/no treatment), type of treatment (pharmaceutical vs. non-pharmaceutical), whether the authorship of the original and updated reviews changed (to capture potential differences in judgment of CoE by the review team), and type of studies (randomized controlled trials vs. observational studies) that were meta-analyzed.

We converted all effect estimates into odds ratio (OR). We also converted all effect sizes in the same direction, with OR < 1 indicating reduction of undesirable outcomes (i.e., more beneficial treatment). Because GRADE separates recommendations as strong versus weak based on the CoE,<sup>13</sup> typically endorsing strong versus weak (conditional) recommendations based on moderate/high versus low/very low, respectively,<sup>4,14</sup> our key analysis focused on the differences in effect sizes between these subgroups. We conducted McNemar's test for paired (before vs. after) data to reject the null hypothesis of equal probability that CoE remained the same, that is, in very low/low CoE versus moderate/high CoE groups. To test for linear trend in change of CoE over all categories—from very low to high—we employed a symmetry test with marginal homogeneity tests (which reduces to McNemar's test for two non-independent categories of observations).

To assess for differences in the magnitude of effect size between original and updated evidence as a function of change in the assessment of CoE we calculated the ratio of odds ratio (ROR) across meta-analytic estimates.<sup>15</sup> ROR compares intervention effects in meta-analysis of trials with very low/low versus those with moderate/high CoE (or vice versa).<sup>15</sup> Thus, if the comparison referred to OR with very low/low versus those with moderate/high CoE pertains to ROR < 1, this would mean that treatment effects were more beneficial in meta-analysis of trials with very low/low CoE, while ROR > 1 would indicate the opposite.<sup>15,16</sup> A test of interactions was performed to assess the hypothesis of no difference between the subgroups (i.e. treatments effects in very low/low vs. moderate/high CoE).<sup>17</sup> Because of assumed correlations in comparison of treatment effects, we calculated standard errors for ROR by correlating the effect sizes observed in the original versus updated reviews.<sup>17</sup> We obtained the values for correlation coefficients from the data. We performed sensitivity analyses by: (a) assuming one correlation coefficient between effect sizes in the original versus updated reviews and (b) calculating correlation coefficients for each subgroup according to direction of treatment effects (i.e., we calculated separate correlation coefficients for the subgroup showing positive, negative and no change in direction of effects between the original versus updated review—three correlation coefficients in total). We also repeated all analyses assuming no correlations between the effect sizes. Since we observed no differences in the results regardless of the postulated assumptions, we report the default analysis based on calculation with three different correlation coefficients.

Our hypothesis was that ROR between the subgroups would differ; in addition, we would expect that the effect size would be

larger if CoE change from moderate/high to very low/low than other way around.

The analyses were based on using random effect Sidik-Jonkman model. We assessed heterogeneity, that is, dispersion of effect size across the meta-analytic estimates by calculating  $\tau$  (tau) statistic.<sup>16</sup> We used  $I^2$  statistic to assess inconsistency;  $I^2$  represents the estimated proportion of the observed variance in true effect sizes across individual meta-analyses rather than sampling error<sup>16</sup>; it depends both on heterogeneity and total variation in the estimates between the analyses.<sup>16,18</sup> We complemented assessment of heterogeneity with calculation of the absolute deviation of treatment effects (aROR) as a function of change in CoE.<sup>19</sup> By definition, aROR is positive and reflects the x-fold deviation of treatment effect from OR = 1 on the OR scale. Thus, if ROR = 0.8 or ROR = 1.25, the absolute deviation is equal to aROR = 1.25. aROR across all SR/MAs was expressed as (unweighted) median and interquartile range (IQR).<sup>19</sup> We also evaluated how the precision of the estimates changed by calculating the ratio of standard errors for each subgroup summarized as (unweighted) median and IQR.<sup>19</sup> Values >1 indicate larger standard errors (less precision) associated with given category (e.g., very low/low vs. moderate/high) of CoE.<sup>19</sup>

A number of subgroup analyses—all defined a priori and published in the protocol to provide further methodological details<sup>20</sup>—were performed. These include assessment of differences between patient-oriented (e.g., mortality, quality of life, etc.) versus disease-oriented outcomes (e.g., disease response, laboratory outcomes, etc.), effect of a change in authorship between the original and updated reviews, effect of comparator intervention (active treatment vs. placebo/no treatment control) and type of treatment category (pharmaceutical vs. non-pharmaceutical). Finally, in some cases, the SRs included observational studies along with randomized controlled trials (RCTs) and implausibly large ORs generated in conversion processes from standardized mean differences. We further analyzed these results by performing sensitivity analyses excluding SRs with observational studies and large ORs from the analysis.

This paper is reported per PRISMA guidelines.<sup>21</sup> All analyses were conducted with the Stata, ver17 statistical package.<sup>22</sup>

### 3 | RESULTS

The original search, performed on 20 October 2020, identified 3323 potentially eligible reviews of which 419 SR were included in the final analysis (Figure 1). Of these, 414 (207 × 2) and 384 (192 × 2) pairs of the reviews were eligible for the analysis of CoE and effect size, respectively. Total number of trials included in 414 reviews was 4217 (1814 before and 2403 after); mean number of trials per meta-analysis was 10 (minimum: 1, maximum: 133). Total number of participants was 3,057,956; mean number of participants per meta-analysis was 10,506 (minimum: 16; maximum: 1,202,382). Interrater kappa agreement between the reviewers varied from 0.79 to 0.97.

Figure 2 shows comparison of CoE in the original and updated Cochrane reviews across all categories of CoE (Figure 2A) and from very low/low to moderate/high (Figure 2B) according to GRADE criteria. Consistent with EBM principles, evidence judged to be of very low/low CoE had 2.1 (1.19–4.12;  $p = 0.0065$ ) times higher odds to be upgraded in the future studies than moderate/high CoE (Figure 2B). Similarly, across all categories of CoE, the test for trend was highly significant, indicating an increased probability of change in CoE from very low to high CoE ( $p = 0.0021$  for linear trend). We observed no instance in which high or moderate quality evidence was re-assessed as very low-quality evidence in the updated SR, while very low CoE was upgraded to moderate or high CoE in 9/39 of updated SR (Figure 2A).

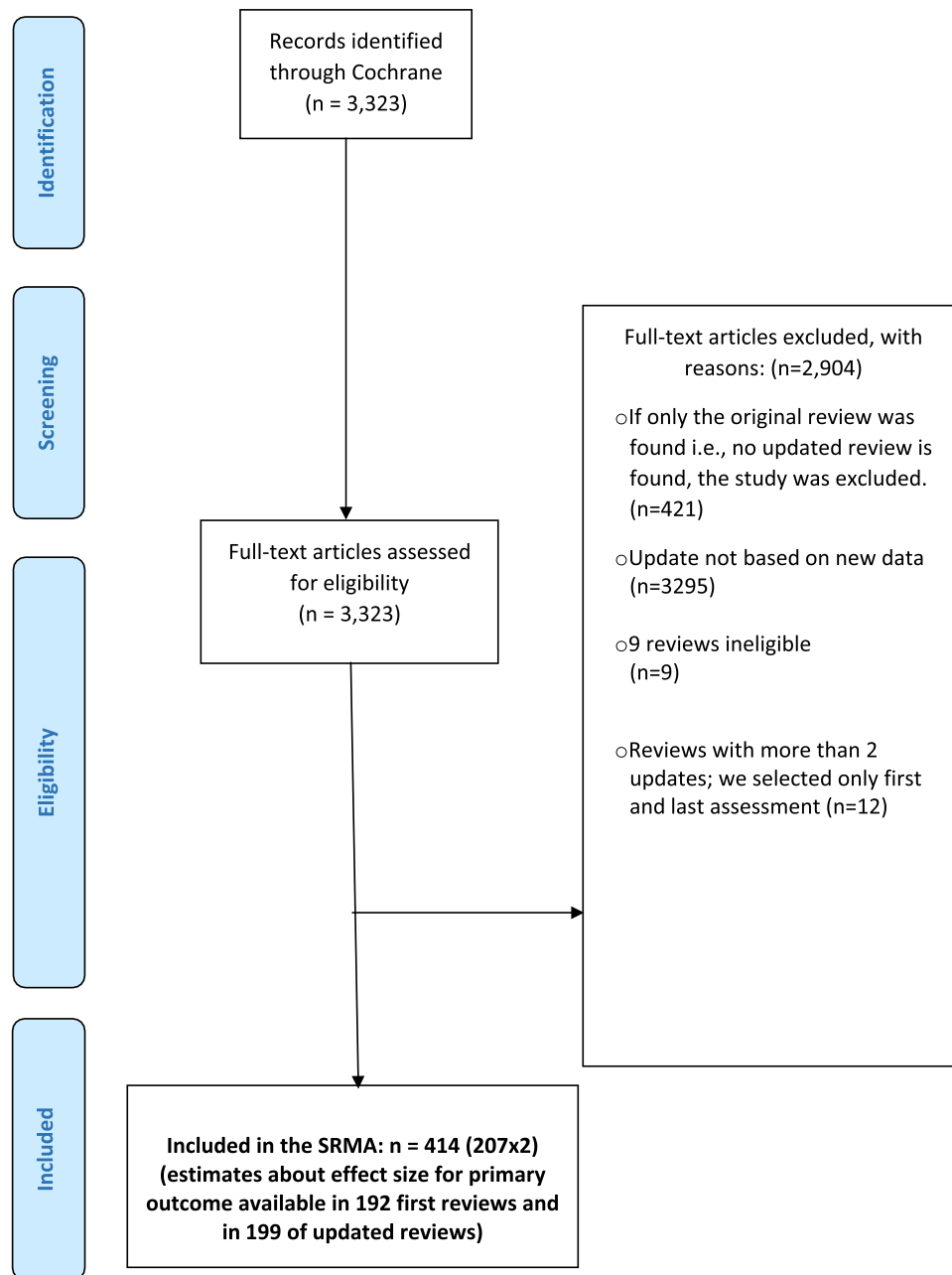
However, we detected no effect of change in CoE on the magnitude of treatment effects [ROR = 1.02 (95% CI: 0.74–1.39) for change of CoE from very low/low to moderate/high versus 1.02 (95% CI: 0.44–2.37) for moderate/high to very low/low CoE]. Test between the subgroups was not significant ( $p = 1$ ). (Figure 3) Although, as explained earlier, from guidelines recommendations perspectives, GRADE typically groups CoE as moderate/high versus low/very low, we also tried to compare the effect sizes at the two extremes of CoE: very low versus high. Because we observed no study with high CoE that changed into very low CoE (Figure 2A), ROR was impossible to calculate for this comparison.

Nevertheless, there was larger dispersion in ROR in meta-analyses where CoE changed from moderate/high to very low/low than in the opposite direction. This was probably driven by low power for the analysis instead of the hypothesis that effect size would be larger if CoE changed from moderate/high to very low/low than other way around. [We had half as many of meta-analyses available for the assessment of ROR based on change of CoE from moderate/high to very low/low ( $n = 16$ ) as those in which CoE changed from very low/low to moderate high ( $n = 33$ ).]

aROR was similar between the subgroups [median (IQR): 1.12 (1.07–1.57) vs. 1.21 (1.12–2.43)] (Figure 4A, Table 1). As in case of ROR, we observed larger dispersion in aROR in meta-analyses where CoE changed from moderate/high to very low/low than in the opposite direction (Figures 4A,B).

The meta-analyses with no change in CoE had similar ROR [ROR = 1.01 (95% CI: 0.85 to 1.21)] (Figure 3B) and aROR [median (IQR): 1.13 (1.04–1.66)] (Table 1, App Figure S4 and App Figure SA) to those MAs in which CoE changed (Figure 4 and App Figure SA). Inconsistency was large across all meta-analytic estimates ( $I^2 = 99\%$ ). There was larger imprecision in treatment effects when CoE changed from VL/L → M/H (seR = 1.46) than when it changed from M/H → VL/L (seR = 0.72).

Qualitative analysis indicated that direction of the effect changed in 6 SR/MAs only: two in the reviews in which CoE changed from very low/low to moderate/high (of which one was statistically significant) and in 4 SR/MAs with no change in the assessment of CoE (of which one was statistically significant) (Figure 5, App Figures S12 and S13).



**FIGURE 1** PRISMA diagram (study flow diagram for evidence source and selection)

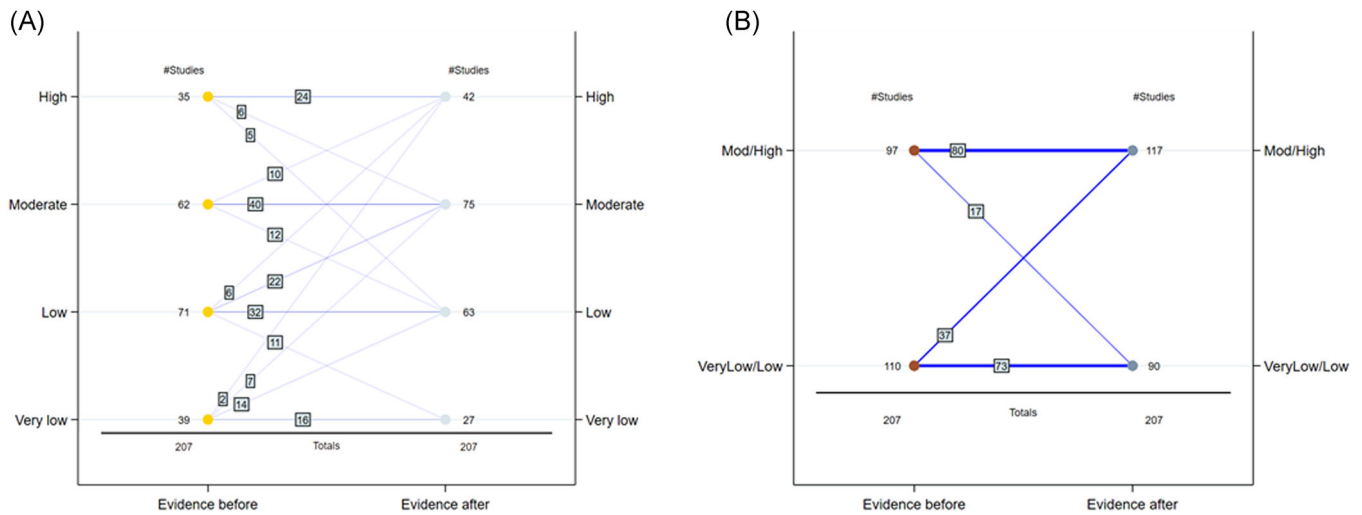
Sensitivity analyses for all pre-defined subgroups showed no change in the results. In fact, when non-randomized studies or outliers were excluded from the analyses, no statistically significant changes were seen in any of the analyses (Appendix).

## 4 | DISCUSSION

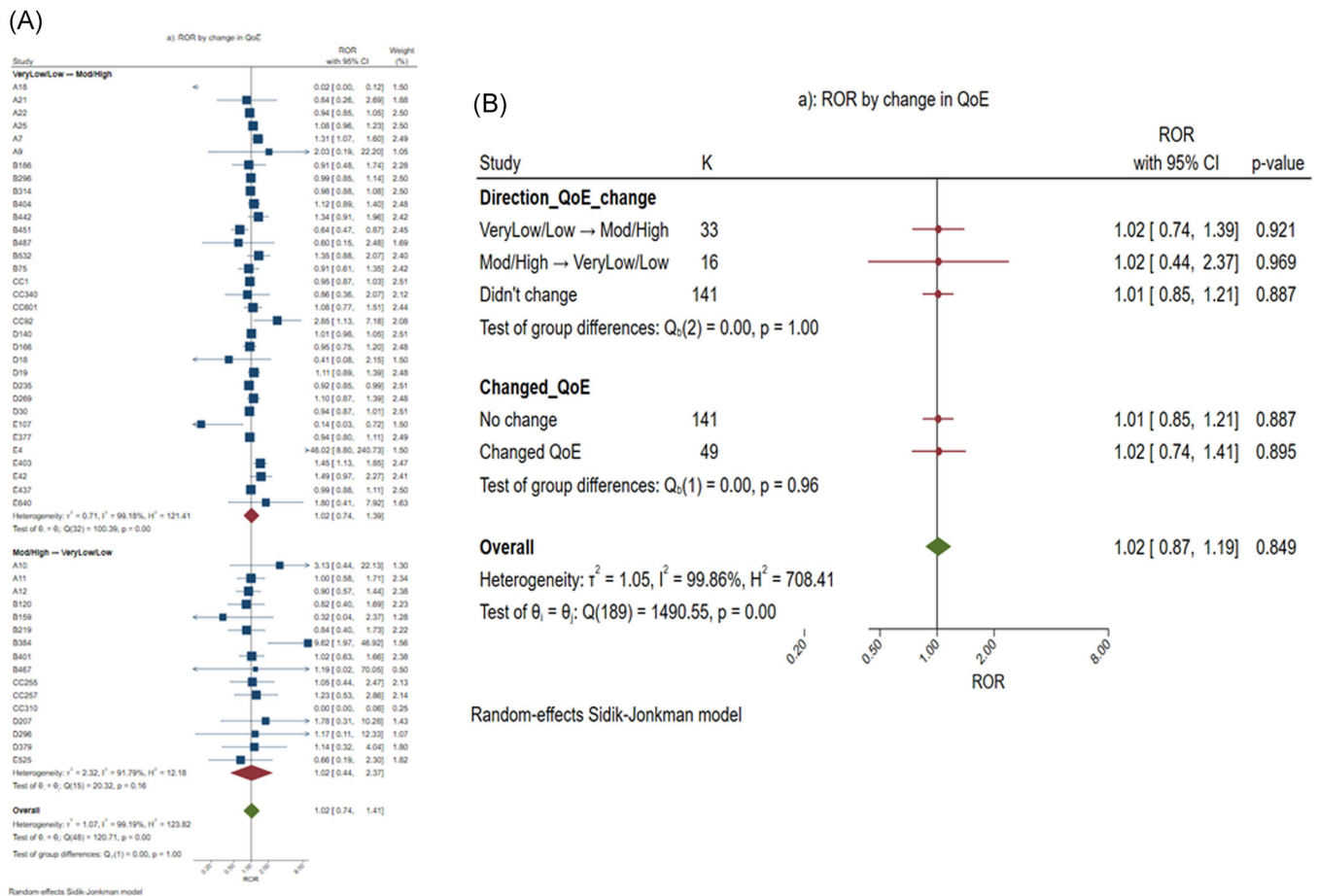
Almost 30 years ago, EBM<sup>23</sup> was introduced to wide medical audience, subsequently being assessed to represent one of the most important medical milestones of the last 160 years, in the same category as innovations such as antibiotics and anesthesia.<sup>24</sup> At the

heart of EBM is notion that ‘not all evidence is created equal’—some evidence is more credible than others; the higher quality of evidence, the more accurate and trustworthy are our estimates about true effects of health interventions.<sup>1</sup> Surprisingly, however, the relationship between CoE and estimates of treatment effects has not been empirically evaluated.

Here, we provide the first empirical support for the foundational EBM principle that low-quality evidence changes more often than high CoE (Figure 2). However, we found no difference in effect sizes between studies appraised as very low versus high [or, very low/low versus moderate/high CoE (Figure 3)]. This implies that effects that are assessed as less trustworthy/potentially unreliable (as when CoE

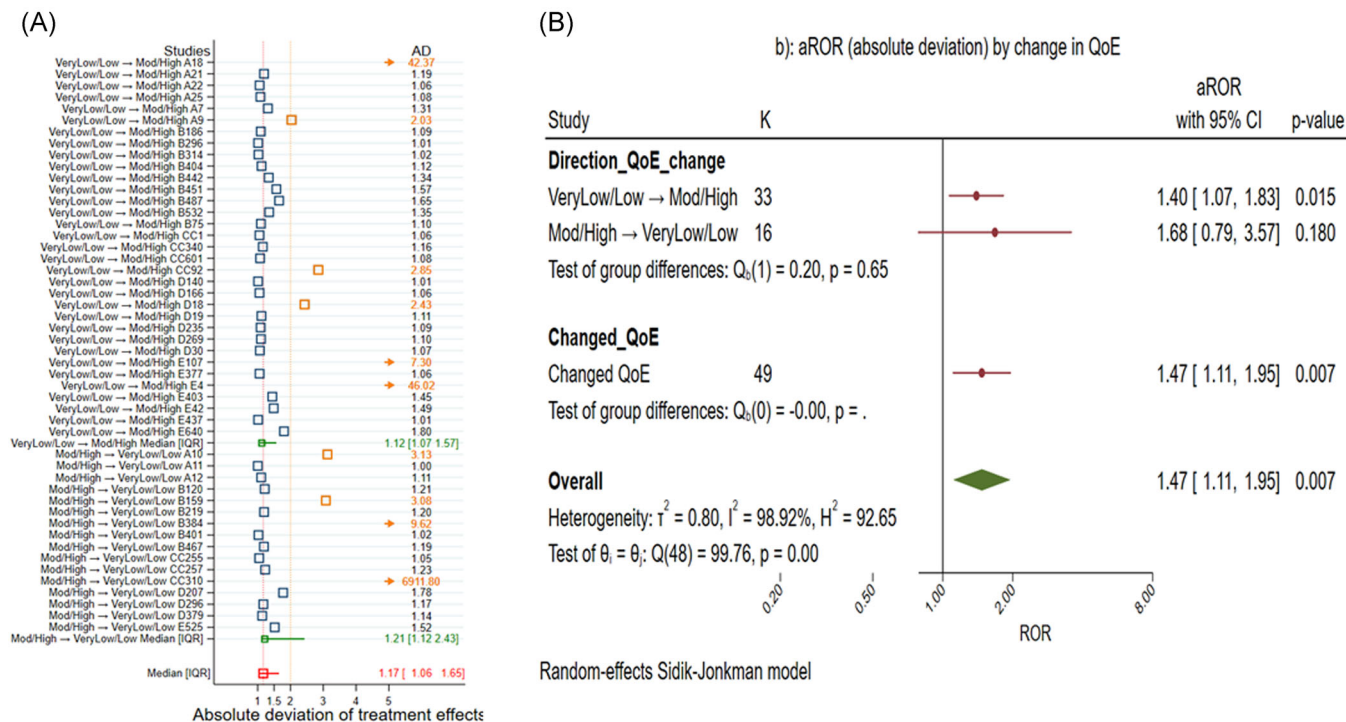


**FIGURE 2** Change in certainty of evidence (CoE) in original and updated Cochrane systematic review. (A) across all categories of CoE as characterized by GRADE; (B) grouped as very low/low versus moderate/high-quality evidence



**FIGURE 3** Comparison of effects of health interventions in meta-analyses in which certainty of evidence (CoE) changed from very low/low to moderate/high versus effects in meta-analyses where CoE changed from moderate/high to very low/low (A); (B) summary of studies shown in (A) with addition of comparison of meta-analyses where CoE did not change. ROR-ratio of odds ratio;  $\tau^2$  ( $\tau^2$ ) statistic and  $H^2$ , measures of heterogeneity;  $I^2$  statistic, measure of inconsistency





**FIGURE 4** (A) Absolute deviation (AD) of treatment effects (aROR) in meta-analyses in which certainty of evidence (CoE) changed from very low/low to moderate/high versus effects in meta-analyses where CoE changed from moderate/high to very low/low; (B) summary of aROR by change in CoE (For graph displaying aROR for all studies, including those that did not have change in CoE, see Supporting Information Appendix, App Figure S4 and App S4a)

**TABLE 1** Summary of aROR (absolute deviation of treatment effects away from OR = 1)

All data	After dropping outliers <sup>a</sup>
All studies, median [IQR]: 1.14 [1.05 1.65]	All studies, median [IQR]: 1.12 [1.03 1.40]
VeryLow/Low → Mod/High, median [IQR]: 1.12 [1.07 1.57]	VeryLow/Low → Mod/High, median [IQR]: 1.11 [1.06 1.47]
Mod/High → VeryLow/Low, median [IQR]: 1.21 [1.12 2.43]	Mod/High → VeryLow/Low, median [IQR]: 1.19 [1.11 1.52]
CoE didn't change, median [IQR]: 1.13 [1.04 1.66]	CoE didn't change, median [IQR]: 1.12 [1.03 1.39]

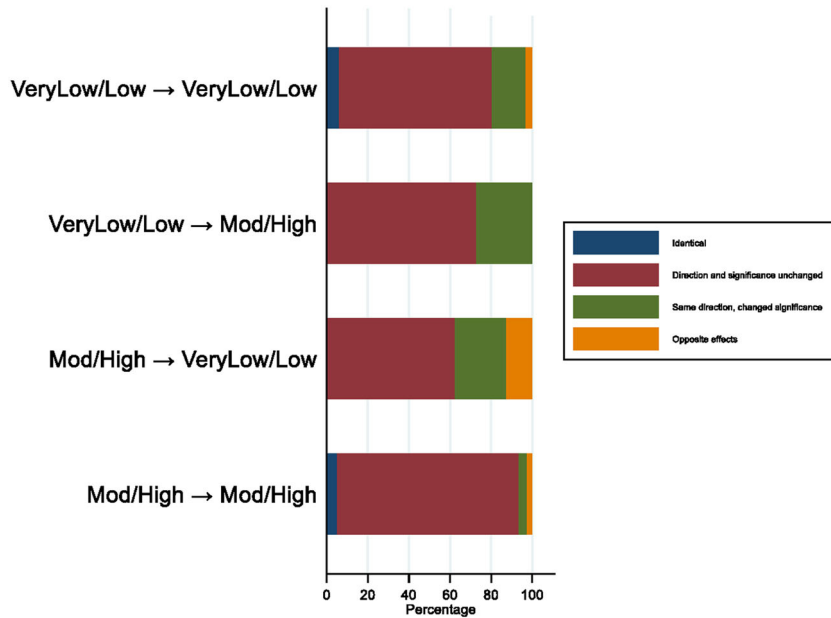
<sup>a</sup>After dropping studies that were converted to OR from studies that originally used standardized mean difference [SMD] (n = 20) and mean difference [MD] (n = 19) metrics to summarize treatment effects.

is low) cannot be distinguished from those assessments, which are presumably more trustworthy/accurate (as when CoE is high). If the magnitude of treatment effects cannot be meaningfully distinguished from evidence appraised as high versus low quality, then the core principle of EBM seems to be challenged.

Our 'negative' results should not be construed as a challenge to sound, normative EBM epistemological principles, which hold that optimal practice of medicine requires explicit and conscientious attention to the nature of medical evidence.<sup>1,25,26</sup> Rather, in assessing the relationship between CoE and 'true' effects of health interventions, more salient question is to ask if the current appraisal methods capture CoE as intended by the EBM principles. Critical appraisal of CoE is integral aspect of conduct of systematic reviews, guidelines development and is widely integrated in the curricula in most medical and allied professional schools across the world. Over the years,

many critical appraisal methods have been developed<sup>1</sup> to eventually culminate in development of GRADE methodology, which has been endorsed by more than 110 professional organizations.<sup>7</sup> However, as we demonstrate here, despite GRADE's capacity to distinguish CoE across its categories, it could not—and we suspect none of other appraisal methods that GRADE has replaced—reliably discerned the influence of CoE on the estimates of treatment effects. The results agree with those of Gartlehner et al who, based on cumulative meta-analysis of 37 Cochrane reviews, found<sup>27</sup> limited value of GRADE in predicting stability of strength of evidence as new studies emerged. Other authors also questioned validity of GRADE as the system that is sufficiently empirically justified to ensure that our judgments are proportional to underlying (quality) of evidence.<sup>28,29</sup>

The finding that the magnitude of effect size is not reflected in a change of CoE is surprising as elucidating bias effects that resulted in



**FIGURE 5** Change in effect size, qualitative analysis (see also App Figures S12 and S13)

misleading advices to patients has been one of the key reasons for the rise of EBM. For example, a large body of observational evidence indicated that hormone replacement therapy (HRT) can reduce heart attack by 40%–50%, which resulted in advice to millions of women to take HRT to prevent heart attack.<sup>30</sup> However, when high quality of evidence was generated, the opposite was observed: more women died from heart attack if they took HRT than from placebo.<sup>30</sup> Similarly, thousands of women with breast cancer were advised to undergo highly toxic stem cell transplant based on unreliable observational evidence indicating improvement in disease-free survival by about 50% compared with historical control<sup>31</sup>—the findings that were overturned once high-quality randomized trials were done.<sup>32,33</sup>

In addition, previous meta-epidemiological studies showed that various study limitations that affect CoE significantly influence estimates of treatment effects<sup>34</sup> (although not always consistently<sup>16</sup>). For example, as measured by ROR, inadequate or unclear (vs. adequate) random-sequence generation, inadequate or unclear (vs. adequate) allocation concealment, or lack of or unclear double-blinding (vs. double-blinding) led to statistically significant exaggeration of treatment effects by 11%, 7% and 13%, respectively.<sup>34</sup> These study limitations are taken into account in rating of CoE using GRADE method,<sup>6</sup> so one would expect that effect size would differ between low versus high CoE in the GRADE assessment. However, on further examination, we observe that GRADE combines the study limitations such as adequacy of allocation concealment, blinding, etc. (risk of bias) with the assessment of inconsistency, imprecision, indirectness and publication bias to assign the final rating of CoE (from very low to high quality) in additive fashion.<sup>12,35</sup> It appears that using additive means to report the properties of negative and positive changes in treatment effect could unhelpfully neutralize this effect and cause imprecision in the overall estimate. Thus, one can have the same estimates of treatment effects but completely different GRADE ratings. This is, however, problematic because central assumption of

GRADE is that estimates underpinned by high CoE are unlikely to change, whereas the very low/low CoE estimates are more likely to change.

A potential limitation of our study is that we have not collected data on the individual factors that drove assessment of CoE (i.e., study limitations/risk of bias vs. inconsistency, imprecision, or indirectness, for example). However, the present empirical report targets, for first time, the end-stage level assessment of CoE, according to GRADE specifications, which is how CoE is used in practice to aid interpretation of evidence and affect development of clinical guidelines.

We also detected imprecision in the estimates of effects sizes and relatively wide ROR confidence intervals, particularly in the subgroup of meta-analyses describing treatment effects when CoE changed from moderate/high to low/very low. It may be argued that the current methods of CoE appraisal are simply not sensitive enough and that with much larger sample size of SR/MAs, we would be able to differentiate between effect sizes across categories of CoE. This point was made by Howick and colleagues<sup>36</sup> who showed no change in the CoE between original and updated reviews in a set of the 48 trials they examined, albeit they made no attempt to identify changes in effect sizes. We also found that in 71 cases the updated reviews were based on inclusion of only 1 extra trial, which might not be enough to overturn or appreciably revise the effect estimate. However, sensitivity analyses comparing the changes in effect size as a function of the number of trials added in the updated meta-analyses showed no difference in the results, regardless of the choice of cut-off for the inclusion of these additional trials in the analysis (e.g., 1 vs.  $\geq 3$ , or any other way). Importantly, critical appraisal (and GRADE) applies to both evidence obtained in single and multiple trials and is required in the Cochrane Reviews regardless of the quantity of existing evidence. Obtaining the larger sample sizes is also unrealistic given that we reviewed almost all SRs in the Cochrane database since





the GRADE assessment of CoE was mandated (up to May 2021). Finally, few Cochrane Reviews we analyzed included observational studies. It is possible that GRADE may not differentiate the quality of randomized evidence well but that it may perform better if the comparison is made between randomized versus observational studies. The Cochrane Reviews, however, are typically based on randomized trials. Therefore, categorization of CoE based on currently mandated critical appraisal system using GRADE in the Cochrane Reviews does not meaningfully separate effect sizes across the existing gradation of CoE (although, capacity of GRADE to distinguish the magnitude of effect size between randomized and observational studies outside of the purview of Cochrane Reviews remains a worthwhile goal for further empirical research).

Given that studies can be well done, and correctly estimated treatment effects, but be poorly reported,<sup>37,38</sup> it is also possible that we could not detect influence of CoE on the estimates of treatment effects because current critical appraisal methods depend on the quality of reporting of the trials that are selected for meta-analysis. However, if we believe that quality of reporting does not matter, then the entire critical appraisal efforts can be considered misplaced to begin with.

## 5 | CONCLUSIONS

To the extent that the central to the epistemology of EBM is that what is justifiable or reasonable to believe depends on CoE,<sup>1</sup> our findings indicate urgent need to refine current EBM critical appraisal methods. If EBM is going to flourish, it is crucial to develop methods with capacity to categorize CoE to reliably differentiate between magnitude effects that are potentially biased from those that are accurate and trustworthy. The major opportunity, therein, lies in addressing the main limitations of this study—carefully and painstakingly discerning various aspects of CoE (from the components related to study limitations/risk of bias to inconsistency, imprecision, or indirectness) to better characterize CoE and its relationship to the magnitude of effects of health interventions.

## ACKNOWLEDGEMENTS

This project was supported in part by grant number R01HS024917 from the Agency for Healthcare Research and Quality (Dr. Djulbegovic). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

## CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

## AUTHOR CONTRIBUTIONS


The authors are notable as an interdisciplinary team of EBM practitioners and instructors who are respected as clinicians, mathematicians, epidemiologists, statisticians, methodologists, and researchers across academic institutions, hospitals and clinics in the UK, Can, USA, Brazil and Switzerland. Their research experience ranges from recently acquired doctorates to over 40 years in research and clinical practice. All authors contributed to the methods, commented on the analysis and contributed to writing and revising the manuscript. Our sources and selection criteria are contained within the document, the data is publicly available from the Cochrane Database and our statistical methods are outlined in the methods, figures and tables. PRISMA was used to report our findings. BD serves as the guarantor of the article. *A conceptual idea:* Benjamin Djulbegovic; *Design:* Benjamin Djulbegovic and David Nunan; *Protocol development:* Benjamin Djulbegovic, Muhammad Muneeb Ahmed, David Nunan, Lars Hemkens, Despina Koletsi, Amy Price, Rachel Riera, Paulo Nadanovsky, Ana Paula Pires dos Santos, Daniela Melo, Rafael Leite Pacheco, Luis Eduardo Fontes; *Data acquisition:* Muhammad Muneeb Ahmed, Despina Koletsi, Amy Price, Rachel Riera, Paulo Nadanovsky, Ana Paula Pires dos Santos, Daniela Melo, Rafael Leite Pacheco, Luis Eduardo Fontes, Ranjan Pathak. *Statistical analysis:* Iztok Hozo, Benjamin Djulbegovic, Lars Hemkens; *Drafting manuscript:* Benjamin Djulbegovic; *Critical revision of the manuscript for important intellectual content:* Benjamin Djulbegovic, Lars Hemkens, David Nunan, Amy Price, Despina Koletsi, Rachel Riera, Paulo Nadanovsky, Ana Paula Pires dos Santos, Daniela Melo, Rafael Leite Pacheco, Luis Eduardo Fontes, Ranjan Pathak. *Administrative, technical, or material support:* Benjamin Djulbegovic, Muhammad Muneeb Ahmed. *Supervision:* Benjamin Djulbegovic.

## DATA AVAILABILITY STATEMENT

Data are available from the authors upon request.

## ORCID

Benjamin Djulbegovic  <http://orcid.org/0000-0003-0671-1447>

Muhammad Muneeb Ahmed  <http://orcid.org/0000-0003-4208-6247>

## REFERENCES

1. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet*. 2017;390(10092):415-423.
2. Djulbegovic B, Trikalinos TA, Roback J, Chen R, Guyatt G. Impact of quality of evidence on the strength of recommendations: an empirical study. *BMC Health Serv Res*. 2009;9(1):120.
3. Djulbegovic B, Kumar A, Kaufman RM, Tobian A, Guyatt GH. Quality of evidence is a key determinant for making a strong guidelines recommendation. *J Clin Epidemiol*. 2015;68(7):727-732.
4. Djulbegovic B, Reljic T, Elqayam S, et al. Structured decision-making drives guidelines panels' recommendations "for" but not "against" health interventions. *J Clin Epidemiol*. 2019;110:23-33.
5. Djulbegovic B, Hozo I, Li S-A, Razavi M, Cuker A, Guyatt G. Certainty of evidence and intervention's benefits and harms are key determinants of guidelines' recommendations. *J Clin Epidemiol*. 2021; 136:1-9.

6. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407-415.
7. GRADE Working Group. GRADE; 2021. Accessed June 26, 2021. <https://www.gradeworkinggroup.org/>
8. Gartlehner G, Sommer I, Evans TS, Thaler K, Lohr KN. Grades for quality of evidence were associated with distinct likelihoods that treatment effects will remain stable. *J Clin Epidemiol*. 2015;68(5):489-497.
9. West S, King V, Carey T, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No 02-E016. 2002:64-88.
10. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res*. 2004;4(1):38.
11. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, et al. Comparison between the standard and a new alternative format of the Summary-of-Findings tables in Cochrane review users: study protocol for a randomized controlled trial. *Trials*. 2015;16:164.
12. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *J Clin Epidemiol*. 2011;64(12):1277-1282.
13. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.
14. Djulbegovic B, Hozo I, Li SA, Razavi M, Cuker A, Guyatt G. Certainty of evidence and intervention's benefits & harms are key determinants of guidelines' recommendations. *J Clin Epidemiol*. 2021;136:1-9.
15. Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med*. 2002;21(11):1513-1524.
16. Moustgaard H, Jones HE, Savović J, et al. Ten questions to consider when interpreting results of a meta-epidemiological study—the MetaBLIND study as a case. *Res Synth Methods*. 2020;11(2):260-274.
17. Higgins JPT, Green S. Cochrane Collaboration. *Cochrane handbook for systematic reviews of interventions*. Wiley-Blackwell; 2011.
18. Higgins J, Thompson S. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558.
19. Ewald H, Klerings I, Wagner G, et al. Abbreviated and comprehensive literature searches led to identical or very similar effect estimates: a meta-epidemiological study. *J Clin Epidemiol*. 2020;128:1-12.
20. Will high quality (certainty) evidence change less often than low-quality evidence after new data is collected?; 2020. Accessed July 14, 2021. <https://osf.io/84qgc/>
21. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*. 2009;62:e1-e34.
22. STATA, ver. 17 [computer program]. College Station, TX; 2021.
23. Evidence-based medicine working group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992; 268:2420-2425.
24. Dickersin K, Straus SE, Bero LA. Evidence based medicine: increasing, not dictating, choice. *BMJ*. 2007;334(suppl 1):s10.
25. Djulbegovic B, Guyatt GH, Ashcroft RE. Epistemologic inquiries in evidence-based medicine. *Cancer Control*. 2009;16(2):158-168.
26. Sackett D, Rosenberg W, Muir Gray J, Haynes R, Richardson W. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312: 71-72.
27. Gartlehner G, Dobrescu A, Evans TS, et al. AHRQ Methods for Effective Health Care. In: Assessing the Predictive Validity of Strength of Evidence Grades: A Meta-Epidemiological Study. Rockville (MD): Agency for Healthcare Research and Quality (US); 2015.
28. Mercuri M, Baigrie BS. What confidence should we have in GRADE? *J Eval Clin Pract*. 2018;24(5):1240-1246.
29. Mercuri M, Baigrie B, Upshur REG. Going from evidence to recommendations: can GRADE get us there? *J Eval Clin Pract*. 2018; 24(5):1232-1239.
30. Investigators WGFtWsHI. Risks and benefits of estrogen plus progestin in healthy postmenopausal women principal results from the women's health initiative randomized controlled trial. *JAMA*. 2002; 288(3):321-333.
31. Peters WP, Ross M, Vredenburgh JJ, et al. High-dose chemotherapy and autologous bone marrow support as consolidation after standard-dose adjuvant therapy for high-risk primary breast cancer. *J Clin Oncol*. 1993;11:1132-1143.
32. Tallman MS, Gray R, Robert NJ, et al. Conventional adjuvant chemotherapy with or without high-dose chemotherapy and autologous stem-cell transplantation in high-risk breast cancer. *N Engl J Med*. 2003;349(1):17-26.
33. Rettig RA, Jacobson PD, Farquhar CM, Aubry WM. *False hope: Bone marrow transplantation for breast cancer*. Oxford University Press; 2007.
34. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157(6): 429-438.
35. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol*. 2011;64(12): 1294-1302.
36. Howick J, Koletsis D, Pandis N, et al. The quality of evidence for medical interventions does not improve or worsen: a metaepidemiological study of Cochrane reviews. *J Clin Epidemiol*. 2020;126: 154-159.
37. Soares HP, Daniels S, Kumar A, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *BMJ*. 2003;328:22-25.
38. Mhaskar R, Djulbegovic B, Magazin A, Soares HP, Kumar A. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. *J Clin Epidemiol*. 2012;65(6):602-609.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Djulbegovic B, Ahmed MM, Hozo I, et al. High quality (certainty) evidence changes less often than low-quality evidence, but the magnitude of effect size does not systematically differ between studies with low versus high-quality evidence. *J Eval Clin Pract*. 2022;28:353-362. doi:10.1111/jep.13657