Journal of Clinical Epidemiology 107 (2019) 51−59

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Stratification by quality induced selection bias in a meta-analysis of clinical trials

Jennifer Stone[a,1], Usha Gurunathan[b,c,1], Kathryn Glass[d], Zachary Munn[e], Peter Tugwell[f], Suhail A.R. Doi[g,*]

[a]*Department of Health Services Research and Policy, Research School of Population Health, Australian National University, Canberra, ACT, Australia*
[b]*Department of Anaesthesia, The Prince Charles Hospital, Brisbane, Queensland, Australia*
[c]*School of Population Health, University of Queensland, Brisbane, Queensland, Australia*
[d]*National Centre for Epidemiology and Population Health, Research School of Population Health, Australian National University, Canberra, ACT, Australia*
[e]*The Joanna Briggs Institute, The University of Adelaide, Adelaide, South Australia, Australia*
[f]*Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada*
[g]*Department of Population Medicine, College of Medicine, Qatar University, Doha, Qatar*

Accepted 14 November 2018; Published online 17 November 2018

Abstract

Objectives: The inconsistency demonstrated across strata when using different scales has been attributed to quality scores, and stratification continues to be done using risk of bias domain judgments. This study examines if restricting primary meta-analyses to studies at low risk of bias or presenting meta-analyses stratified according to risk of bias is indeed the right approach to explore potential methodological bias.

Study Design and Setting: Reanalysis of the impact of quality subgroupings in an existing meta-analysis based on 25 different scales.

Results: We demonstrate that quality stratification itself is the problem because it induces a spurious association between effect size and precision within stratum. Studies with larger effects or lesser precision tend to be of lower quality—a form of collider-stratification bias (stratum being the common effect of the reasons for these two outcomes) that leads to inconsistent results across scales. We also show that the extent of this association determines the variability in effect size and statistical significance across strata when conditioning on quality.

Conclusions: We conclude that stratification by quality leads to a form of selection bias (collider-stratification bias) and should be avoided. We demonstrate consistent results with an alternative method that includes all studies.

Keywords: Meta-analysis; Risk of bias; Quality score; Stratification; Heparin

1. Introduction

Clinical practitioners and policymakers making decisions on therapy need some form of summary judgment on the trustworthiness of studies included in meta-analyses. This is commonly done by presenting meta-analyses stratified according to risk of bias or restricting the primary meta-analysis to studies at low risk of bias after risk of bias assessment [1]. In this respect, stratification using a scale was reported to have made a difference in the meta-analysis by Nurmohamed et al. [2] comparing low molecular weight heparin (LMWH) and unfractionated heparin (UFH) or standard heparin for the prevention of postoperative deep vein thrombosis (DVT). Nurmohamed had reported a 25−30% reduction in DVT risk with LMWH compared with standard heparin. However, when they stratified by score on a quality scale with eight safeguards listed, they reported no significant difference between the two heparins for the high-quality (with seven to eight safeguards) general surgery studies. They concluded that there was no convincing evidence that LMWH was clinically superior to standard heparin in the thromboprophylaxis of general surgery patients.

This observation led to the study by Jüni et al. [3] who investigated whether the choice of quality assessment scale could have affected the conclusions of this meta-analysis. They used author defined cutoffs for high and low quality

Conflict of interest: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.

[1] These authors contributed equally.

* Corresponding author. Tel.: +974 66001271; fax: +974 4403 7854.
E-mail address: sardoi@gmx.net (S.A.R. Doi).

https://doi.org/10.1016/j.jclinepi.2018.11.015
0895-4356/

What is new?

Key findings

- Studies with either biased effects or lesser precision tend to belong to lower quality strata (one of possible subdivisions based on quality). This is because the quality stratum is a common effect of the reasons for both these outcomes.
- Categorizing trials into high or low quality based on quality assessment to ascertain the impact of quality leads to selection (collider) bias, which may lead to inconsistent results.

What this adds to what was known?

- Alternative methods for incorporating quality assessment into meta-analyses should be sought, such as weighting by quality using the quality effects model or direct corrections for possible bias based on either elicitation by experts or empirically based prior distributions.

What is the implication and what should change now?

- The common strategy of assessing the impact of quality in meta-analysis by excluding lower or including higher quality studies should be abandoned.

(if undefined, they cut off at the median score). In their reanalysis, Jüni et al. [3] observed inconsistency in the effect estimated between high- and low-quality clinical trials across the scales used in the reanalysis of data from the Nurmohamed meta-analysis. They thus concluded that there was no association between summary quality scores and treatment effects and suggested that identification and assessment of methodological aspects of clinical trials and their influence on effect sizes are more important than using summary scores to categorize clinical trials into high or low quality [3]. Although this caused stratification by quality score to be abandoned, stratification by quality judgments has remained in use for the same purpose.

In keeping with this recommendation, groups such as Cochrane and GRADE [4] recommend such stratification based on a summary assessment of low, moderate, or high risk of bias for each clinical trial. This involves assessing safeguards in key domains; when many are present, the study is deemed low risk, and when many are missing, the study is deemed at high risk of bias [1]. Clinical practitioners and decision makers see such an approach as a measure of the trustworthiness of meta-analyses leading to widespread and continued use of summary judgments for stratification by quality [5−7].

Quality scales allow enumeration of safeguards into a score (usually converted into study ranks with a common highest rank when multiple scales are in use so that scales are comparable regardless of number of safeguards). However, although domain judgments enumerate safeguards in a similar way to study ranking through scores, there have been no direct comparisons despite the recommendation not to use quality scales for the same purpose [1]. Part of the confusion around this was researchers viewing quality scores as an absolute measure of quality and not rescaling these scores as ranks.

Both these approaches (be it a judgment or an enumerated score) may be missing a key element of implementing best practice methods by introducing collider-stratification bias. If we take a closer look at quality stratification, it implies conditioning on quality regardless of whether this is being done through strata defined by quality judgments or a quality scale. In either case, if quality acts as a collider, conditioning on it can lead to the inconsistency of effects previously mentioned through a spurious association between study size and results, a form of selection bias [8]. It should be pointed out that such an association between precision and study effect may also occur without selection bias because of heterogeneity and even by chance [9,10]. Nevertheless, because one of the reasons for this association is indeed one of the forms of selection bias, the presence of such an association in meta-analysis tends to raise the alert regarding potential bias. The association manifests as study asymmetry [9,10], which can be checked for using tools such as funnel [11] and Doi plots [10] developed for this purpose.

In this article, we examine the role of collider-stratification bias in the etiology of the previously documented inconsistency of effects using quantitative measures of plot asymmetry. To do so, this study will reanalyze the 17 general surgery clinical trials from the meta-analysis by Nurmohamed et al. [2] using the same scales used by Juni et al. [3] We aim to ascertain if the variable results from quality stratification previously reported can be explained by collider-stratification bias.

2. Methods

2.1. Data description

We used the studies reported by Nurmohamed et al. [2] comparing the effectiveness of LMWH and UFH as prophylaxis against postoperative thromboembolism. There were 17 general surgery clinical trials [12−28] also used by Jüni et al. [3] for their assessment. We based our analysis on the same data set to allow comparability with Jüni et al. [3]. All original articles were retrieved for our analysis, including an article in German [13] (which was translated) and five abstracts [12,16,19,24,29], of which three [19,21,30] have subsequently been published. DVT was chosen as the major endpoint for our analysis. The

randomized controlled clinical trials included in the original meta-analysis by Nurmohamed et al. [2] were published between January 1984 and April 1991. The general surgery clinical trials included abdominothoracic and gynecological procedures. All general surgery clinical trials used standard doses of heparins [2] and I^{125} fibrinogen leg scanning as a mandatory screening procedure. The major endpoints reported were DVT, pulmonary embolism, and major bleeding. All clinical trials involved adult patients with baseline comparability in demographic factors and risk factors for DVT between the LMWH and UFH groups.

One study by Samama et al. [25] involved three doses of LMWH namely 20, 40, and 60 mg, each with separate control groups and hence was considered as three clinical trials in our analysis, resulting in 19 datasets in total. Kakkar et al. [22] conducted a double-blind clinical trial and an open clinical trial with LMWH, and we chose the double-blind clinical trial results for our analysis. Three studies [14,21,26] involved the use of dihydroergotamine mesylate along with heparins in both the groups. In other studies [28,30] where more than one dose of LMWH or UFH were used, the more commonly prescribed dose or the dose comparable to other studies was chosen for analysis. The results from fibrinogen scanning were used if there was more than one diagnostic test to detect the DVT outcome. All the pertinent data including study details, study design, type of surgery, description of patient population, intervention, diagnosis, and main outcome were extracted from the articles and tabulated (Supplementary Tables S1a and S1b).

2.2. Quality assessment

The 25 quality assessment scales [2,31−54] mentioned by Moher et al. [55] and subsequently used by Jüni et al. [3] in their article were examined for the development of a single composite scale. All the original articles cited by Moher et al. [55] were retrieved for this purpose. The full version of a scale mentioned in a then-unpublished article [56] was retrieved by electronic search from a later publication in 2003 [51].

A composite scale was also developed using all unique internal validity safeguard items from the 25 scales included in this analysis. In doing so, the composite scale can be used to judge the *agreement* (intraclass correlation) between each of the 25 scales and the composite and to show that this agreement does not influence the estimates of effect. Across all scales, we only included items that focus on the internal validity of the clinical trial rather than the quality of reporting of an article (Supplementary Table S2). The items mentioned by Moher et al. [55] as influencing internal validity such as randomization, blinding, patient follow-up, and statistical analysis among others were included in the scale. The final version of the 25 scales was created following discussion and consensus among two authors excluding (from the scales) items that were either repeated or irrelevant to assess the methodological strength of a clinical trial. Across all scales, this resulted in 25 unique questions with each representing a single safeguard that was then counted (1 point). The exception was the randomization safeguard for which being well described and mention of allocation concealment were also counted as two additional safeguards (see Table S2). This made the maximum count of safeguards (quality score) obtainable from the largest (composite) scale equal to 27 points, but this varied from scale to scale. However, we converted scores to the study rank Q_i (defined as a score for a study/maximum score for the study on the scale) to allow comparability for stratification across scales with different numbers of safeguards. This then removes the possibility of a score on one scale implying high quality, whereas the same score on another implying low quality.

All the scales addressed five categories of internal validity: (1) design-specific sources of bias (excluding confounding), (2) selection bias, (3) confounding, (4) information bias, and (5) analytic bias (excluding the methods used for confounding). Scoring from each scale was decided with agreement from two authors. Any discrepancies or differences were resolved by discussion. For example, question 3 in the composite scale was essentially similar to a question from the scales of Kleijnen et al., [36] Goodman et al. [45], and Levine et al. [54]. The original scale from Goodman et al. [45] was limited to questions 3, 4, 6, 9, 10, 20, and 23 in the composite scale. This resulted in an abbreviated version of Goodman et al.'s [45] scale with seven items in it (the rest not deemed to be bias safeguards).

2.3. Analysis

The two-way mixed effects (consistency) intraclass correlation coefficient (ICC) was calculated for the quality score of each scale with the composite scale. Two quality strata were created for every scale: low and high quality, defined by sorting data sets by quality rank and assigning the lowest ranking nine datasets to the low-quality group and the remaining 10 to the high-quality group. This allowed for the same number of low-quality clinical trials across scales, and ties were ignored when the lowest ranking nine datasets were selected. Given that we were using a common ranking scheme across the same 17 studies for each scale, this implies a similar cutoff for each scale when we cut off at a fixed number of lower ranked studies.

The relative risk (RR) for DVT with LMWH compared with UFH was the effect size for the meta-analyses. Meta-analyses were performed using fixed effect (FE) models as studies were homogenous. The FE analysis was performed in subgroups defined by quality, with a separate meta-analysis performed on each subgroup across every scale. Meta-analysis using the quality effects model was also run to allow for comparison with an alternative to quality stratification using all studies. Association

between precision and effect size was quantified using the Luis Furuya-Kanamori (LFK) index and was visually shown using the Doi plot [10] and funnel plot. Egger's regression *P* value [11] was also computed for comparison (Supplementary Material) but was not used as the low-quality subgroup had less than 10 studies, and power of the test is low when compared with the LFK index [10]. A conditional logistic regression model was run using the 52 stratum results to assess univariable and multivariable odds ratios (ORs) of a statistically significant result (the presence/absence of statistically significant results in the stratum) according to their symmetry and quality. ''Symmetry'' was coded as LFK index categories of no asymmetry and minor asymmetry vs. major asymmetry and quality was coded high vs. low. The analysis was conditional on the scale used for stratification (group was the scale). Quality is the assessment of the study while scale is the tool used to assess the study quality, and thus, conditioning on scale ensures that study assessment is independent of the scale used. Conditional logistic regression was run in Stata version 13 (StataCorp, College Station, TX, USA), and meta-analysis was run using MetaXL version 5.3 software (www.epigear.com).

3. Results

3.1. Quality assessment scales

The 25 scales were widely different in terms of what they addressed. Some scales dealt with the quality of clinical trials in general [32,35,37,40−43,45−49,54]. Other scales were designed for a specific meta-analysis on topics such as pain [31,33,53], homeopathy [34,36,53], laser therapy [39], respiratory muscle training [50], smoking [44], liver disorders [38], and antibiotic prophylaxis [52]. Nurmohamed et al.'s [2] scale was the only one among the 25 scales that was designed for the meta-analysis of heparins.

When the ICC of all the 25 scales was calculated in relation to the composite scale, ICCs of the scales were found to range from 0.21 to 0.87. The scales with ICC $\leq$0.6 were classified as poor agreement scales, and the rest of the scales were good in terms of agreement with the composite scale. This resulted in eight poor scales and 17 good scales in terms of their ICC. The asymmetry was not influenced by the ICC of the scale used to create the strata (Table 1).

3.2. Effect estimates

Under the FE model, the pooled effect estimate for all clinical trials was RR 0.75 (95% CI 0.62−0.91). The Cochran's Q statistic $P = 0.91$ indicated the studies were homogenous. The overall LFK index was 0.89 (no asymmetry), and Egger's *P* concurred ($P = 0.448$). Given symmetry, there was no association between precision and effect size overall.

With meta-analysis stratified by high and low quality based on the individual scales, the pooled effect estimates (Table 1) across 25 scales for the high-quality clinical trials (FE model) ranged from 0.71 to 0.83 (median = 0.78) and for the low-quality clinical trials ranged from 0.64 to 0.86 (median = 0.67). Most scales demonstrated a larger pooled effect within the lower quality strata (except nine scales [34,36,38,42,45,46,48,52,54]).

3.3. Asymmetry and collider-stratification bias

Fig. 1 depicts the directed acyclic graph that shows how a relationship between study precision and effects can be induced by quality stratification. In keeping with this expectation, studies belonging to the lower quality stratum would be expected to have a larger pooled effect, given the expectation of biased larger studies through collider-stratification bias (as mentioned previously). In addition, we found that smaller studies were also in this stratum (which had an average sample size about half that of studies in the higher quality stratum [216 vs. 526]).

Of the 52 strata (includes the composite scale), 26 each were of high or low quality. There was no major asymmetry in the high-quality strata, and the majority (20/26) had significant results (Table 1). The creation of an association between effect magnitude and precision (LFK index) was seen across half (14/26) of the low-quality strata, which demonstrated gross asymmetry (LFK $\geq$ 2), and almost all (13/14) had nonsignificant results. The remaining 12 strata were symmetrical, and half (6/12) had statistically nonsignificant results (Table 1).

To ensure that the effect on the results was indeed driven by this association (asymmetry) rather than quality or the particular scale, we looked at the discrepancy across scales resulting from asymmetry (LFK index) or quality after conditioning on scale (Table 2). The conditional logistic regression (group variable was scale) result demonstrated that symmetrical (including no asymmetry and minor asymmetry combined) strata had a 13-fold increase in odds of significant results compared with strata with major asymmetry. In contrast, the high-quality strata had only a threefold increase in the odds of significant results compared with low-quality strata and lost importance in the multivariable model (Table 2). Clearly, asymmetry is the main driver of inconsistent results, and a stratum of high-quality or low-quality studies would become less statistically significant if they lose symmetry (Table 2). This explains the discordance seen in Table 1 between strata of studies by quality. The same was demonstrable when pooled effect size groups (categorized at the median) were looked at across strata. Both high-quality (OR: 1.89; $P = 0.123$) and symmetrical (OR: 2.50; $P = 0.121$) strata had an increased odds of smaller effect sizes.

Finally, analysis without stratification using the quality effects model demonstrates that results are consistent and match those of the symmetrical strata (Table 1).

Table 1. Pooled estimates of effect across studies by 25 quality assessment scales (high- and low-quality strata)

Scale	ICC	FE high (95% CI)	LFK high	FE low (95% CI)	LFK low	QE (95% CI)
Imperiale and McCullough [38]	0.21	0.74 (0.60–0.92)	−0.6	0.80 (0.54–1.20)	1.25	0.76 (0.62–0.92)
Onghena and van Houdenhove [33]	0.37	0.77 (0.62–0.96)	1.31	0.70 (0.48–1.04)	1.52	0.79 (0.64–0.97)
Jonas [51]	0.40	0.76 (0.59–0.98)	−1.24	0.74 (0.56–0.99)	1.74	0.78 (0.63–0.97)
Brown [46]	0.42	0.75 (0.60–0.94)	−1.67	0.77 (0.54–1.08)	2.43	0.77 (0.62–0.96)
Linde et al. [34]	0.48	0.71 (0.56–0.89)	−0.07	0.86 (0.62–1.19)	1.29	0.76 (0.61–0.94)
Goodman et al. [45]	0.53	0.74 (0.60–0.92)	−0.63	0.80 (0.53–1.20)	1.28	0.77 (0.63–0.93)
Andrew [48]	0.54	0.72 (0.57–0.90)	0.52	0.84 (0.60–1.18)	1.17	0.77 (0.63–0.94)
Evans and Pollock [52]	0.58	0.72 (0.57–0.90)	0.52	0.84 (0.60–1.18)	1.17	0.77 (0.63–0.95)
Chalmers et al. [37]	0.60	0.83 (0.65–1.05)	0.69	0.64 (0.46–0.88)	1.58	0.80 (0.64–0.99)
Gøtzsche [49]	0.61	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.79 (0.64–0.98)
Levine [54]	0.61	0.75 (0.60–0.94)	−1.67	0.77 (0.54–1.08)	2.43	0.77 (0.63–0.94)
Nurmohamed et al. [2]	0.61	0.80 (0.63–1.01)	−1.22	0.67 (0.48–0.94)	2.82	0.78 (0.64–0.96)
Smith et al. [50]	0.66	0.83 (0.65–1.05)	0.69	0.64 (0.46–0.88)	1.58	0.80 (0.65–0.99)
ter Riet et al. [31]	0.66	0.83 (0.65–1.05)	0.69	0.64 (0.46–0.88)	1.58	0.78 (0.64–0.95)
Kleijnen et al. [36]	0.68	0.75 (0.60–0.94)	−1.67	0.77 (0.54–1.08)	2.43	0.77 (0.63–0.95)
Beckerman et al. [39]	0.69	0.83 (0.65–1.05)	0.69	0.64 (0.46–0.88)	1.58	0.79 (0.64–0.97)
Colditz et al. [41]	0.70	0.83 (0.65–1.05)	0.69	0.64 (0.46–0.88)	1.58	0.79 (0.64–0.98)
Spitzer et al. [44]	0.71	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.79 (0.64–0.97)
Jadad et al. [35]	0.78	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.78 (0.64–0.95)
Chalmers et al. [47]	0.82	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.78 (0.64–0.97)
Cho and Bero [43]	0.82	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.79 (0.64–0.97)
Detsky et al [42]	0.83	0.75 (0.60–0.94)	−1.67	0.77 (0.54–1.08)	2.43	0.77 (0.62–0.94)
Koes et al [53]	0.83	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.79 (0.64–0.97)
Poynard [40]	0.83	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.78 (0.63–0.96)
Reisch et al [32]	0.87	0.78 (0.63–0.96)	−0.55	0.67 (0.44–1.03)	3.24	0.79 (0.64–0.97)

Abbreviations: CI: confidence interval; ICC, Intraclass correlation coefficient; FE, fixed effect; FE high, pooled estimate with the FE model for high-quality studies; FE low, pooled estimate with the FE model for low-quality studies; LFK $\geq \pm 1$ represents minor asymmetry and $\geq \pm 2$ implies major asymmetry.

The results from the quality effects (QE) model align with the symmetrical strata. In this case, it happens to be mainly in the high-quality strata; however, this may not necessarily be the case in other meta-analyses.

4. Discussion

In this study, we confirm what Jüni et al. [3] had reported, which is that the stratification of the clinical trials into high and low quality based on a score-based criterion (converted to relative ranks) results in discrepancies in the pooled estimates between the high- and low-quality clinical trials. However, we also demonstrate that such discrepancies are largely because of conditioning on quality, which allows precision and effect size to associate within strata by quality. It is well known that both smaller clinical trials [9] and biased clinical trials tend to be of poorer quality although smaller clinical trials are not necessarily biased, and biased clinical trials are not necessarily small. Normally there is therefore no association between study size (precision) and effect size in a meta-analysis unless there is a systematic bias toward selection of only lower or higher quality studies thus creating asymmetry between larger and smaller clinical trials because of this association. Although this was not the case in the overall meta-analysis (Fig. 2), this was induced by quality stratification, and Figures 3 and 4 demonstrate this because small studies tend to bin into different strata depending on if they are biased or not. These results suggest therefore that (1) it is incorrect to draw any general conclusions relating to the impact of quality on a pooled effect size from studies stratified by quality and (2) the interpretation by Jüni et al. [3] and others [57–59] of their empirical observations were in error.

The quality rank assessed by the scales was based on the number of safeguards reported in or determinable from the published article. The purpose of the quality score therefore is not to judge possible magnitudes of bias but to enumerate each safeguard present based on our assessment and then create a single summary quality score based on counts. These scores therefore do not aim to capture a "*multidimensional quality space*," but rather to capture the relative ranking or judgment of these studies in terms of number of safeguard items, implicitly assuming that this ranking

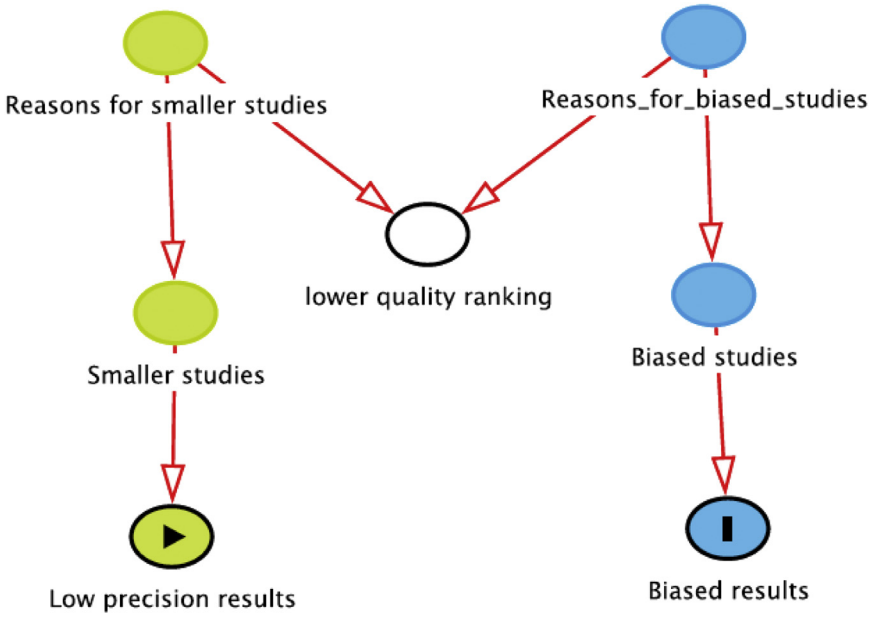

Fig. 1. Directed acyclic graph showing how collider bias occurs when stratifying by quality (unfilled circle). Quality is what we stratify on so it links biased and small studies and therefore creates asymmetry (open circle depicts adjustment/stratification on that variable and therefore cannot be removed).

correlates with the probability that they may be credible. This rank can be used to stratify studies in much the same way as domain judgments if we enumerate the low-risk judgments across the several domains in the risk of bias scheme as suggested by Higgins et al. [1]. Regardless of how this is done, the higher ranked studies will not necessarily be devoid of bias, and the lower ranked studies will not necessarily be biased. What the ranks really tell us is the relative probability (compared to the highest ranked study) that they may be credible.

Although this study clearly demonstrates that the worry about the application of such scores based on empirical data was misplaced, there still remains the objection based on theoretical considerations first suggested by Greenland [60]. The latter was based on the premise that the quality score indicates effect size bias in a quantitative fashion such that quality components may cancel each other out [61], making summary scores poorly informative [60,61]. However, the impact of such quality components on the direction or magnitude of change in an effect size is unknown, and we believe that such a use of these scores remains unsupported. We therefore do not consider such a direct connection to the effect size to be the appropriate use of quality scores or even of quality assessment in general. Quality scores should essentially enumerate study safeguards and should then only be used to rank studies, relative to the best study, by the probability that they may be credible. We advocate that such ranks be standardized to start from 1 as we have done in this study (by dividing each score by the maximum score) as this would allow ranks to be comparable across scales. Perhaps, the way to improve on such enumerated scores in the future is to conduct meta-epidemiological studies to generate data that allow such scores to be weighted by some measure of impact, prior to conversion to ranks.

Table 2. Univariable and multivariable odds ratios[a] of a statistically significant stratum result according to their symmetry and quality

	Univariable			Multivariable		
Stratum variable	**Odds ratio**	**95% CI**	***P***	**Odds ratio**	**95% CI**	***P***
Symmetry						
Gross asymmetry	1			1		
Minor or no asymmetry	13.00	1.70–99.37	0.013	10.8	1.03–114.15	0.047
Quality						
Low quality	1			1		
High quality	3.17	1.26–7.93	0.014	1.2	0.37–3.93	0.763

Abbreviation: CI, confidence interval.

[a] Conditional logistic regression with scale as the group variable. Univariable: symmetry and quality in separate models. Multivariable: symmetry and quality in the same model.

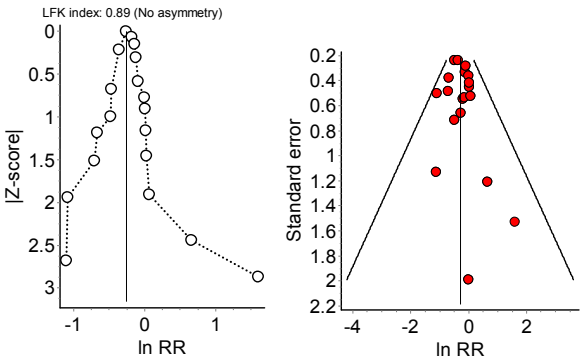

Fig. 2. Doi and funnel plots showing symmetry overall in a meta-analysis of all trials. Note that this plot only demonstrates (a)symmetry but does not define its cause/source (see other analyses and discussion regarding this). Asymmetry on the Doi plot requires consideration of asymmetry of both area under each limb on the plot and the number of studies in each limb. RR, relative risk.

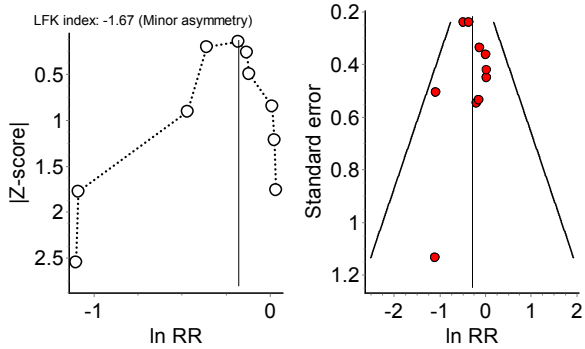

Fig. 4. Doi and funnel plots showing no asymmetry in the high-quality stratum of trials using the scale by Levine et al. Note that this plot only demonstrates (a)symmetry but does not define its cause/source (see other analyses and discussion regarding this). Asymmetry on the Doi plot requires consideration of asymmetry of both area under each limb on the plot and the number of studies in each limb. RR, relative risk.

Possible solutions that replace stratification by quality for assessing the trustworthiness of meta-analyses are weighting by quality using the quality effects model [62,63] (regression-based methods using quality components or ranks [64,65] should be avoided as this is similar to stratification) and direct corrections for possible bias based on either elicitation by experts [66,67] or empirically based prior distributions [68,69]. The easiest to implement is the quality effects model, and this was run on this dataset with results shown in Table 1. Across all scales, the results were consistent regardless of the scales ICC and in keeping with the results from the symmetrical strata.

In conclusion, categorizing clinical trials into higher or lower quality based on a cutoff score or risk of bias assessment (without a score) in an attempt to assess bias in meta-analysis is a flawed approach. The latter results in erroneous conclusions because the magnitude and significance of the pooled estimate will vary depending on the selection bias induced by such stratification and has little to do with the impact of quality per se. This common approach taken in meta-analysis should be avoided, and alternative methods that include all studies to draw conclusions should be sought [63,70].

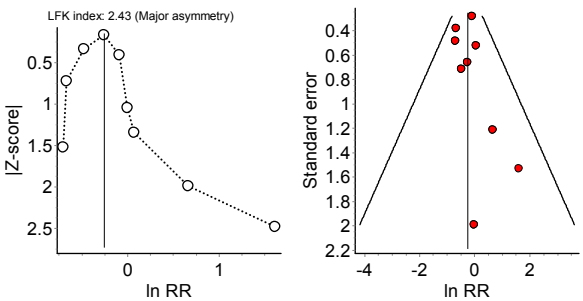

Fig. 3. Doi and funnel plots showing major asymmetry in the low-quality stratum of trials using the scale by Levine et al. Note that this plot only demonstrates (a)symmetry but does not define its cause/source (see other analyses and discussion regarding this). Asymmetry on the Doi plot requires consideration of asymmetry of both area under each limb on the plot and the number of studies in each limb. RR, relative risk.

Acknowledgments

The authors are extremely grateful to anonymous reviewers for helpful comments on a previous version. The responsibility for the content of the article and views expressed are ours.

Authors' contributors: J.S., U.G., and S.D. conceived the study concept and design and the analytic strategy. J.S. and U.G. undertook the initial drafting of the article and statistical analysis. J.S., U.G., K.G., Z.M., P.T., and S.D. interpreted data and undertook critical revision of the article for important intellectual content. K.G., Z.M., P.T., and S.D. provided administrative, technical, or material support. S.D. is the guarantor, had full access to all data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

J.C.S. was supported by the Australian National University, Australia Higher Research Degree Scholarship. This study was not supported by any funding source.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.11.015.

References

[1] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[2] Nurmohamed MT, Rosendaal FR, Büller HR, Dekker E, Hommes DW, Vandenbroucke JP, et al. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. Lancet 1992;340:152–6.

[3] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:6.
[4] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[5] Myung S-K, Ju W, Cho B, Oh SW, Park SM, Koo BK, et al. Efficacy of vitamin and antioxidant supplements in prevention of cardiovascular disease: systematic review and meta-analysis of randomised controlled trials. BMJ 2013;346:f10.
[6] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557−60.
[7] de Souza RJ, Mente A, Maroleanu A, Cozma AI, Ha V, Kishibe T, et al. Intake of saturated and trans unsaturated fatty acids and risk of all cause mortality, cardiovascular disease, and type 2 diabetes: systematic review and meta-analysis of observational studies. BMJ 2015;351:h3978.
[8] Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol 2010;39:417−20.
[9] Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ 2010;341:c3515.
[10] Furuya-Kanamori L, Barendregt JJ, Doi SAR. A new improved graphical and quantitative method for detecting bias in meta-analysis. Int J Evid Based Healthc 2018;16:195−203.
[11] Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002.
[12] Blum A, Desruennes E, Elias A, Lagrange G, Loriferne J. DVT Prophylaxis in surgery for digestive-tract cancer comparing the LMW heparinoid ORG- 10172 (Lomoparan) with calcium heparin (abstr.). Thromb Haemost 1989;62:126.
[13] Adolf J, Knee H, Roder JD, van de Flierdt E, Siewert JR. [Thromboembolism prophylaxis with low molecular weight heparin in abdominal surgery]. Dtsch Med Wochenschr 1989;114:48−53.
[14] Baumgartner A, Jacot N, Moser G, Krähenbühl B. Prevention of postoperative deep vein thrombosis by one daily injection of low molecular weight heparin and dihydroergotamine. Vasa 1989;18:152−6.
[15] Caen JP. A randomized double-blind study between a low molecular weight heparin Kabi 2165 and standard heparin in the prevention of deep vein thrombosis in general surgery. A French multicenter trial. Thromb Haemost 1988;59:216−20.
[16] Dahan M, Levasseur P, Bogaty J, Boneu B, Samama M. Prevention of post-operative deep vein thrombosis (DVT) in malignant patients by Fraxiparine (a low molecular weight heparin). A cooperative trial(abstr.). Throb Haemost 1989;62:519.
[17] Encke A, Breddin K. Comparison of a low molecular weight heparin and unfractionated heparin for the prevention of deep vein thrombosis in patients undergoing abdominal surgery. Br J Surg 1988;75:1058−63.
[18] Fricker JP, Vergnes Y, Schach R, Heitz A, Eber M, Grunebaum L, et al. Low dose heparin versus low molecular weight heparin (Kabi 2165, Fragmin) in the prophylaxis of thromboembolic complications of abdominal oncological surgery. Eur J Clin Invest 1988;18:561−7.
[19] Gallus A, Cade J, Ockelford P, Hepburn S, Maas M, Magnani H, et al. Orgaran (Org 10172) or heparin for preventing venous thrombosis after elective surgery for malignant disease? A double-blind, randomised, multicentre comparison. ANZ-Organon Investigators' Group. Thromb Haemost 1993;70:562−7.
[20] Hartl P, Brücke P, Dienstl E, Vinazzer H. Prophylaxis of thromboembolism in general surgery: comparison between standard heparin and Fragmin. Thromb Res 1990;57:577−84.
[21] Kakkar VV, Stringer MD, Hedges AR, Parker CJ, Welzel D, Ward VP, et al. Fixed combinations of low-molecular weight or unfractionated heparin plus dihydroergotamine in the prevention of postoperative deep vein thrombosis. Am J Surg 1989;157:413−8.
[22] Kakkar VV, Murray WJ. Efficacy and safety of low-molecular-weight heparin (CY216) in preventing postoperative venous thrombo-embolism: a co-operative study. Br J Surg 1985;72:786−91.
[23] Koller M, Schoch U, Buchmann P, Largiadèr F, von Felten A, Frick PG. Low molecular weight heparin (KABI 2165) as thromboprophylaxis in elective visceral surgery. A randomized, double-blind study versus unfractionated heparin. Thromb Haemost 1986;56:243−6.
[24] Leizorovicz A. Comparison of two doses of low molecular weight heparin in the prevention of post-operative vein thrombosis (DVT) (abstr.). Thromb Haemost 1989;62:1−647.
[25] Samama M, Bernard P, Bonnardot J, Combe-Tamzali S, Lanson Y, Tissot E. Low molecular weight heparin compared with unfractionated heparin in prevention of postoperative thrombosis. Br J Surg 1988;75:128−31.
[26] Sasahara AA, Koppenhagen K, Häring R, Welzel D, Wolf H. Low molecular weight heparin plus dihydroergotamine for prophylaxis of postoperative deep vein thrombosis. Br J Surg 1986;73:697−700.
[27] Verardi S, Cortese F, Baroni B, Boffo V, Casciani C, Palazmni E. Deep vein thrombosis prevention in surgical patients: effectiveness and safety of a new low-molecular-weight heparin. Curr Ther Res 1989;46:366−72.
[28] Verardi S, Casciani CU, Nicora E, Forzano F, Origone A, Valle I, et al. A multicentre study on LMW-heparin effectiveness in preventing postsurgical thrombosis. Int Angiol 1988;7:19−24.
[29] Welzel D, Stringer M, Hedges A, Parker CJ, Welzel D, Ward VP, et al. Fixed combinations of low-molecular-weight or unfractionated heparin plus dihydroergotamine in the prevention of postoperative deep-vein thrombosis. Thromb Haemost 1989;62:5−19.
[30] Liezorovicz A, Picolet H, Peyrieux JC, Boissel JP. Prevention of perioperative deep vein thrombosis in general surgery: a multicentre double blind study comparing two doses of Logiparin and standard heparin. H.B.P.M. Research Group. Br J Surg 1991;78:412−6.
[31] ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. J Clin Epidemiol 1990;43:1191−9.
[32] Reisch J, Tyson J, Mize S. Aid to the evaluation of therapeutic studies. Pediatrics 1989;84:815−27.
[33] Onghena P, Van Houdenhove B. Antidepressant-induced analgesia in chronic non-malignant pain: a meta-analysis of 39 placebo-controlled studies. Pain 1992;49:205−19.
[34] Linde K, Clausius N, Ramirez G, Melchart D, Eitel F, Hedges LV, et al. Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. Lancet 1997;350:834−43.
[35] Jadad A, Moore R, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1−12.
[36] Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homoeopathy. BMJ 1991;302:316−23.
[37] Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, et al. A cohort study of summary reports of controlled trials. JAMA 1990;263:1401−5.
[38] Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. Ann Intern Med 1990;113:299−307.
[39] Beckerman H, de Bie RA, Bouter LM, De Cuyper HJ, Oostendorp RA. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. Phys Ther 1992;72:483−91.
[40] Poynard T. Evaluation de la qualite methodologique des assais therapeutiques randomises. Presse Med 1988;17:315−8.

[41] Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med 1989;8:441−54.
[42] Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992;45:255−65.
[43] Cho M, Bero L. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 1994;272:101−4.
[44] Spitzer WO, Lawrence V, Dales R, Hill G, Archer MC, Clark P, et al. Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking. Clin Invest Med 1990;13:17−42.
[45] Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. Ann Intern Med 1994;121:11−21.
[46] Brown SA. Measurement of quality of primary studies for meta-analysis. Nurs Res 1991;40:352−5.
[47] Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981;2:31−49.
[48] Andrew E. Method for assessment of the reporting standard of clinical trials with roentgen contrast media. Acta Radiol Diagn (Stockh) 1984;25:55−8.
[49] Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Control Clin Trials 1989;10:31−56.
[50] Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. Am Rev Respir Dis 1992;145:533−9.
[51] Jonas WB, Chez RA. The role and importance of definitions and standards in healing research. Altern Ther Health Med 2003;9:A5−9.
[52] Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. Br J Surg 1985;72:256−60.
[53] Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. BMJ 1991;303:1298−303.
[54] Levine J. Trial assessment procedure scale (TAPS). In: Spilker B, editor. Guide to Clinical Trials. New York: Raven Press; 1991.
[55] Moher D, Jadad A, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62−73.
[56] Jonas WB. The likelihood of validity evaluation method 1993: Unpublished manuscript.
[57] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249−56.
[58] Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973−82.
[59] Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. J Clin Epidemiol 1999;52:631−6.
[60] Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140:290−6.
[61] Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001;2(4):463−71.
[62] Doi SAR, Thalib L. A quality-effects model for meta-analysis. Epidemiology 2008;19:94−100.
[63] Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials II: the quality effects model. Contemp Clin Trials 2015;45:123−9.
[64] Sterne JA, Juni 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;21:1513−24.
[65] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609−13.
[66] Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc 2009;172:21−47.
[67] Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, et al. A proposed method of bias adjustment for meta-analyses of published observational studies. Int J Epidemiol 2011;40:765−77.
[68] Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. J R Stat Soc Ser A Stat Soc 2009;172:119−36.
[69] Rhodes KM, Turner RM, Savovic J, Jones HE, Mawdsley D, Higgins JPT. Between-trial heterogeneity in meta-analyses may be partially explained by reported design characteristics. J Clin Epidemiol 2018;95:45−54.
[70] EFSA Scientific Colloquium 23 − Joint European Food safety Authority and Evidence-Based Toxicology Collaboration Colloquium evidence integration in risk assessment: the science of combining apples and oranges 25−26 October 2017 Lisbon, Portugal. EFSA Support Publ 2018;15:1396E.