European Journal of Epidemiology (2018) 33:1025–1031
https://doi.org/10.1007/s10654-018-0443-3

Case study in major quotation errors: a critical commentary on the Newcastle–Ottawa scale

Andreas Stang[1,2] · Stephan Jonas[3] · Charles Poole[4]

Received: 25 July 2018 / Accepted: 8 September 2018 / Published online: 26 September 2018

Abstract
The Newcastle-Ottawa scale (NOS) is one of many scales used to judge the quality of observational studies in systematic reviews. It was criticized for its arbitrary definitions of quality items in a commentary in 2010 in this journal. That commentary was cited 1,250 times through December 2016. We examined the citation history of this commentary in a random sample of 100 full papers citing it, according to the Web of Science. Of these, 96 were systematic reviews, none of which quoted the commentary directly. All but 2 of the 96 indirect quotations (98%) portrayed the commentary as supporting use of the NOS in systematic reviews when, in fact, the opposite was the case. It appears that the vast majority of systematic review authors who cited this commentary did not read it. Journal reviewers and editors did not recognize and correct these major quotation errors. Authors should read each source they cite to make sure their direct and indirect quotations are accurate. Reviewers and editors should do a better job of checking citations and quotations for accuracy. It might help somewhat for commentaries to include abstracts, so that the basic content can be conveyed by PubMed and other bibliographic resources.

The Newcastle–Ottawa scale (NOS) is one of many scales used to judge the quality of observational studies [1]. It allows the qualitative assessment of cohort and case–control studies. The Cochrane Handbook observed that the NOS contains just eight items, neglects crucial elements of selection bias, requires customization to the needs of specific reviews, and is difficult to apply [2]. Deeks et al. [3] noted that it has no items on inclusion or exclusion criteria, baseline comparability of compared groups, or the internal validity domain of data analysis. Nonetheless, it has become popular enough to have its own Wikipedia page (https://en.wikipedia.org/wiki/Newcastle%E2%80%93Ottawa_scale, accessed Feb 9, 2017).

In September 2010, one of the authors (AS) published a critical commentary about the NOS, called "the commentary" for the remainder of this article [4]. The commentary clearly concluded that the NOS is unfit for use in systematic reviews. It ended with this statement:

> "…Wells et al. provide a quality score that has unknown validity at best, or that includes quality items that are even invalid. The current version appears to be inacceptable for the quality ranking of both case-control studies and cohort studies in meta-analyses. The use of this score in evidence-based reviews and meta-analyses may produce highly arbitrary results."

Of note, the commentary and therefore the PubMed entry for this commentary did not contain an abstract. The commentary has been cited 1250 times as of December 22, 2016 according to the Web of Science (Thomssen-Reuters). According to Google scholar (www.scholar.google.de, accessed Feb 9, 2017), this article has been cited 6550 times. It is one of the most frequently cited papers of the European Journal of Epidemiology according to the editor-

✉ Andreas Stang
andreas.stang@uk-essen.de

[1] Director of the Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Hufelandstr. 55, 45147 Essen, Germany

[2] Department of Epidemiology, Boston University School of Public Health, 715 Albany St, Boston, MA 02118, USA

[3] Department of Medical Informatics, RWTH Aachen University, Pauwelsstr. 30, 52057 Aachen, Germany

[4] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, CB #7435, 135 Dauer Drive, Chapel Hill, NC 27599-7435, USA

in-chief (Hofman A, personal communication, September 22, 2016).

Reference errors have been classified into two basic types: errors of citation and errors of quotation. Quotation errors occur when a "referenced statement does not reflect the content of its source." [5]. de Lacey et al. (1985) [6] classified quotation errors as "seriously misleading" for "incorporating an error seriously misrepresenting or bearing no resemblance to the original source." Similarly, Eichorn and Yankauer (1987) defined a "major error of quotation (indirect rather than direct)" as one in which "the cited reference either failed to substantiate, was unrelated to, or even contradicted the author's assertion. Minor errors were those which did not seriously affect the authors' assertion, such as oversimplification or drawing conclusions which the authors of the cited reference were unwilling to do." Subsequent studies of quotation errors have employed the classification of Eichorn and Yankauer [7] with little variation [5, 8, 9].

Arguably, the most serious of the major quotation errors occurs when a cited source flatly contradicts an assertion attributed to it. A clear example would consist of any citation of the aforementioned commentary [4] as though it supported use of the NOS in systematic reviews. The aim of this note is

(1) To provide insights into the citation history of the commentary,
(2) To analyze in which ways authors quoted it and used the results of their quality assessments, and finally,
(3) To give a summary resume about the unusual quotation history of this comment and an outlook about ways to make future meta-analyses more informative.

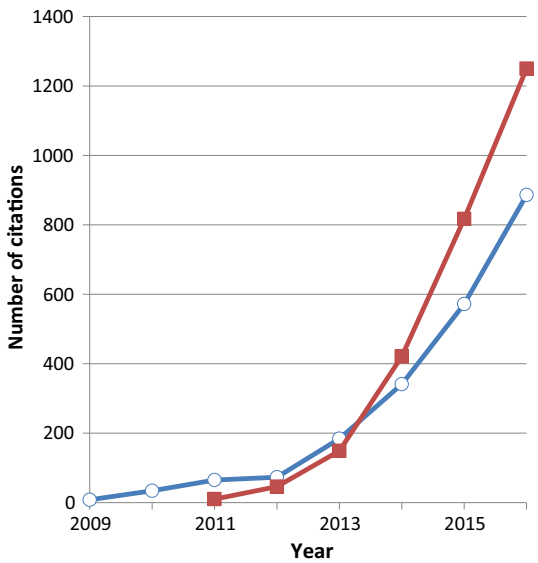

Fig. 1 Cumulative number of abstracts in PubMed that contain "New Castle Ottawa Scale" (blue graph and circles) and cumulative number of citations of the commentary in the Web of Science (red graph and squares), 2009 through 2016. (Color figure online)

Web of Science citation history

We used the Web of Science (http://www.webofknowledge.com, accessed December 22, 2016) to derive citation statistics by year, country, first author, journal, and web of science category. The cumulative number of citations of the commentary in the Web of Science follows a similar trend as the cumulative number of abstracts in PubMed that contain the phrase "Newcastle–Ottawa Scale" over the years 2009 through 2016 (Fig. 1). The vast majority of citations of the commentary come from China (n = 866, 69%), followed by the US (n = 201, 16%) and England (n = 65, 5%). Authors from all other countries cited the commentary fewer than 50 times. As the total number of systematic reviews differs considerably by countries, we also estimated the citation proportion of the commentary per 1000 systematic review article of each country. This proportion was highest in Asian countries including Thailand (29.1 per 1000), China (17.7 per 1000 articles), and South Korea (10.5 per 1000 articles) (Table 1).

The most frequent Web of Science categories that included a citation of the commentary were "multidisciplinary sciences" (n = 172, 14%), "oncology" (n = 170, 14%), "surgery" (n = 120, 10%), "gastroenterology and hepatology" (n = 105, 8%) among others. An analysis by journals revealed that PLOS One (n = 123, 10%), International Journal of Clinical and Experimental Medicine (n = 47, 4%), Scientific Reports (n = 41, 3%), and Tumor Biology (n = 37, 3%) were the most frequent journals that cited the commentary.

Review of the full texts of 100 randomly selected papers citing the NOS commentary

We exported the 1250 references that cited the commentary from the Web of Science (as of December 22, 2016) and drew a random sample of 100 references. We retrieved all 100 full papers and entered several items into a database including PubMed identification number (PMID), study type, country of first author, description of the NOS in the methods section, the number of observational studies and RCTs included in the systematic review, the minimum and maximum NOS score in each systematic review, whether authors used their quality assessment in further analyses, whether authors categorized the NOS scores, and whether

Table 1 Number and proportion of citations of the commentary by country

Country	Commentary citations	Systematic reviews	Proportion (per 1000)
China	866	48,817	17.7
USA	201	183,347	1.1
United Kingdom	65	11,358	5.7
Thailand	47	1616	29.1
Italy	39	38,700	1.0
Canada	37	32,313	1.1
Netherlands	30	21,437	1.4
Australia	28	29,114	1.0
Spain	20	17,075	1.2
South Korea	18	1707	10.5
Japan	17	22,160	0.8
Switzerland	17	12,059	1.4
Brazil	14	10,355	1.4
Germany	13	35,113	0.4
France	12	29,625	0.4

Legend: statistics only for countries of affiliation with 10 or more citations of the commentary; the number of systematic reviews in PubMed between July 2010 and December 2016 was searched by: "(Meta-analysis OR metaanalysis OR systematic review) AND (2010/7:2016/12 [dp]) AND (country [AD])" on September 19th, 2017

authors restricted the inclusion of observational studies for their systematic review based on NOS.

Two articles did not cite the commentary (error in the Web of Science) [PMID 24138388; 26860011]. In the remaining 98 articles, none directly quoted the commentary. One was a protocol for a planned systematic review [PMID 22071020] and another a commentary [PMID 27444273] that chided a group of reviewers for not using the NOS in their systematic review, which in our view was the right decision. The remaining 96 articles were systematic reviews. These systematic reviews included the NOS quality assessment of overall 1395 observational studies. The median number of included studies per systematic review was 11 (10th percentile 5, 90th percentile 27).

Overall 94 out of 96 systematic reviews indirectly quoted the commentary incorrectly. All of these articles gave the impression that the commentary supported the use of the NOS in systematic reviews. Hence, the vast majority of articles citing the commentary committed major quotation errors. Eighteen (19%) of the 94 articles gave an additional citation in support of using the NOS. Only in 9, it was the Wells report. In 5, it were previous reviews in which NOS had been used. Among the remaining four, one article [PMID 25618311] additionally cited the PRISMA statement [10], which made no mention of the NOS. Another article [PMID 26938805] additionally quoted Jüni et al. (1999) who allegedly found the NOS items "reliable and easy to interpret," even though this paper could not have referred to the NOS, as it was published while the NOS was still under development, and even though the authors drew general conclusions that "the use of summary scores to identify trials of high quality is problematic" and "[r]elevant methodological aspects should be assessed individually and their influence on effect sizes explored," views we endorse wholeheartedly. A further article [PMID 24975405] additionally cited the famous article by DerSimonian and Laird [11] as though it supported use of the NOS, even though that article was published at least a decade before the appearance of the NOS, made no mention of summary quality scores, and endorsed meta-regression of individual study characteristics. The fourth article [PMID 27149861] cited a HTA report on the evaluation of non-randomized intervention studies [3] .

Two out of 96 systematic reviews contained a critical comment about the validity of the NOS. One stated that the NOS had "received positive endorsement" from Deeks et al. [12] but that it had "received criticism regarding its validity and applicability in meta-analysis cohort trial quality assessment" from the commentary. These authors also noted that "detailed psychometric properties have not been published in peer-reviewed journals to date" [PMID 27149861]. The other article, after indirectly quoting the commentary in the Methods section in favor of using the NOS in systematic reviews, indirectly quoted it again in the Discussion section as pointing out "limitations of NOS," which the authors blamed for discrepant quality rankings of some studies in their review and a previous one, which had also used the NOS [PMID 24365211]. We judged this

review as a review that contained a critical comment about the validity of the NOS.

In their original material, Wells et al. [1] explicitly stated that a threshold score that distinguishes between 'good' and 'poor' quality studies has to be identified. Until now, Wells et al. did not suggest this threshold. However, 51 out of 96 systematic reviews (53%) mentioned a categorization of the NOS accompanied by qualitative labels related to the study quality: 40 dichotomized the NOS with varying cutoffs (> 4, > 5, > 6, > 7, > 8) for 'high quality' studies and 11 introduced more than two categories (3 categories: n = 10, 4 categories: n = 1). Furthermore, 7 out of 96 meta-analyses (7%) included studies only if the NOS was larger than an a priori defined threshold (> 2: n = 1, > 4: n = 1, > 5: n = 3, > 6: n = 2). Overall 21 out of 96 meta-analyses (22%) ran a sensitivity analysis among high quality studies, a stratified analysis by NOS quality, or a meta-regression analysis to study the influence of the NOS score on the meta-analytic results.

One article [PMID 22770982] claimed that the NOS ranges between 0 and 10, however, it actually ranges between 0 and 9 according to Wells et al. A few articles called the NOS "modified NOS" without clarifying what they meant by "modified" [PMID 22071020, 22770982, 24487609, no PMID NA_2]. Two articles modified the NOS (cohort study) so that it can also be used for cross-sectional studies [PMID 26055921; no PMID NA_1]. One article did not report how many RCTs and observational studies were finally included in the systematic review so that the distribution of NOS scores (only for observational studies) is uninterpretable [PMID 23905841].

Identification of influential articles and authors

We created a citation network within all articles citing the commentary to identify influences between citing articles. For this analysis, the citations were extracted on October 3, 2017. A total of 1856 articles citing the commentary was identified. A total of 486 citations from one article directly citing the commentary to another one directly citing the commentary were found. Using these 486 citations, a directed citation network was created by introducing articles as nodes and edges from citing to cited article. Within this network, we searched for the most cited articles. Since articles either received 1–4 citations (98%) or 6 and more citations (2%), we considered the 9 articles with 6 or more citations as "highly cited article". The citation range within the 9 highly cited articles was from 6 [PMID 25354465] to 26 [PMID 23683848]. The latter article was the only one that included a correct and direct quotation. These 9 articles were responsible for 123 (23%) of the 467 citations between first-generation articles. Notably, 6 out of these 8 articles were by the same first author. Upon closer investigation, we found this author on the author list of 59 (3%) of the 1856 articles citing the commentary. Articles with this author received 92 (20%) of the 335 total citations among the articles citing the commentary.

In addition, we visualized the citation network following just the 9 highly cited articles. We created a hierarchy of all articles with a "citation path" to the commentary through one of these 9 articles. A citation path is an iterative connection through the citation network from citing article to cited article. The paths span a tree-like structure (Fig. 2). The distance of a layer in the tree to the trunk can be seen as a proxy of time, as each node in the tree was published prior to its children. In the first distance (articles citing the commentary and none of the other 1856 articles), the highly cited articles made up 1 (0.07%) and 5 (0.3%) of all articles citing the commentary in this tree layer for the correctly and incorrectly citing articles respectively. However, these articles had a high influence on other articles, as their descendants made up 22 (8%) and 29 (11%) of articles in the second, and 5 (8%) and 25 (41%) of all articles in the third tree layer for correctly and incorrectly citing articles respectively. While only the 9 most influential articles were evaluated regarding their correctness of quotation and we cannot make any assumptions on the correctness of their descendants, the relatively high numbers of their descendants indicate that the first six articles were highly influential on later articles and might contribute to the propagation of quotation errors that were shown in earlier parts of the analysis.

Summary and outlook about ways to make future meta-analyses more informative

Our quotation analysis shows that authors of systematic reviews who cite the commentary obviously do not take much care of references that they quote. Obviously, they use the commentary just to give a published reference for the NOS score. The temptation to cite the commentary without having read it may have been triggered by three aspects: (1) Wells et al. never published a peer-reviewed paper about the NOS. Instead, they only provided a webpage that contains all material related to the NOS. In contrast, the commentary is an easy-to-find article in PubMed that enables easily referencing the NOS by PubMed, (2) the Cochrane Collaboration considers the Downs and Black instrument and the Newcastle–Ottawa Scale "the two most useful tools" but does not suggest that either one is very useful in absolute terms [2], (3) the commentary was published without an abstract and therefore the PubMed entry gave no indication of the negative

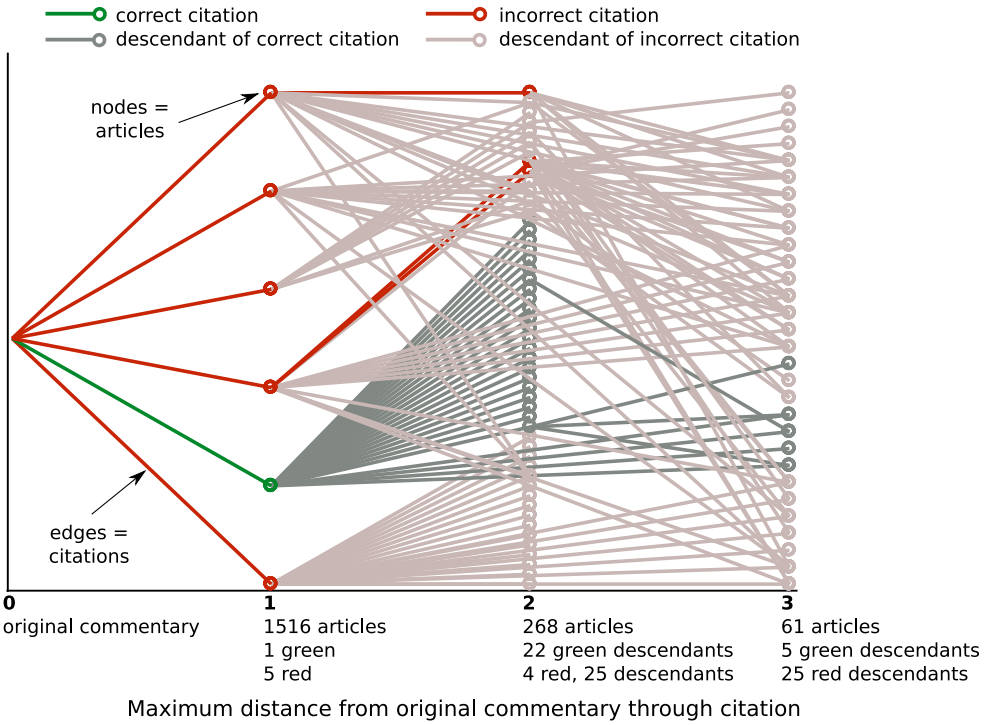

Fig. 2 Citation network restricted to the 9 highly cited articles and their descendants. All articles cited the original commentary, however, these connections were omitted for distance larger than one for readability

views of the NOS expressed in the commentary besides the title of the commentary.

Outlook

One major limitation of quality scales such as the NOS is the conceptual nebulousness of "quality." In an oft-cited review of quality in philosophy, economics, marketing and operations management, Garvin (1984) [13] identified five main perspectives. One of them seems less than useful in considering scales such as the NOS. That is the *transcendent* view of quality as a characteristic that can be recognized but not defined, as Plato wrote of beauty [14] and United States Supreme Court Justice Potter Stewart wrote of pornography [15]. Garvin's other four quality concepts seem more relevant for present purposes. The first is a *product-based* approach, in which quality consists of one or more precisely measurable attributes of a product. The second is a *manufacturing-based* approach, in which the desired attributes are imparted to the product by strict adherence to detailed specifications for design and production [16]. The third is a *value-based* approach, in which the desirability of each quality component is weighed against the cost of attaining it [17]. Perhaps the most important and, unfortunately, most neglected is a *user-based* definition of quality as the capacity to satisfy consumers' wants [18].

These considerations combine to support a perspective on the NOS and similar scales as sets of design and production specifications that stipulate more or less accurately measured features of research design, conduct and analysis. The selection and weighting of the quality items reflect the values of the experts who devise the scales. Elements of a user-based approach have been largely, if not entirely, neglected. In the present context, such an approach would require preference surveys of those who rely on health research in applied settings. To our knowledge, no such surveys have been conducted in the development or evaluation of the NOS or any similar scale. To take an important example, the assignment of relatively high weights in these scales to design features that bias results toward the null [19] might reflect a degree of valuative discord between the developers and users of quality scales.

Over time, the authors of the *Cochrane Handbook for Systematic Reviews of Interventions* have been moving slowly but steadily away from quality scores. At present, "This *Handbook* draws a distinction between assessment of methodological quality and assessment of risk of bias, and recommends a focus on the latter" (Sect. 8.2.2) [20]. The *Handbook* further notes that some quality score items, such as conducting a sample size analysis or obtaining ethical approval, "are unlikely to have direct implications for risk of bias." (Sect. 8.2.2) [20]. Even for those quality items that would be expected to affect validity, the aggregation into a single score can obscure important differences [21, 22]. As has been noted, two studies could receive

identical scores even though one is substantially biased toward the null and the other is biased sharply away from the null [21, 22].

Donabedian, finding similar difficulties in attempts to define quality of medical care, concluded that they "convey vividly the impression that the criteria of quality are nothing more than value judgments that are applied to several aspects, properties, ingredients or dimensions of a process called medical care. As such, the definition of quality may be almost anything anyone wishes it to be, although it is, ordinarily, a reflection of values and goals current in the medical care system and in the larger society of which it is a part." [23].

The upshot is to see wisdom in the conclusion of Jüni et al. [24] that "the use of summary scores to identify trials of high quality is problematic. Relevant methodological aspects should be assessed individually and their influence on effect sizes explored." At the meta-analytic level, separate assessments of associations between the results of studies and their validity features can be obtained by stratified systematic review and meta-regression analysis. The *Cochrane Collaboration*'s RevMan software does not yet include a meta-regression module. Nonetheless, the *Cochrane Handbook* (Sect. 9.6.5) [12] refers approvingly to that approach, while emphasizing the frequent limitation of small numbers of studies and highly clustered design features. When these limitations are not strong, multiple meta-regression can be used to construct what Rubin [25] calls a "response surface" from which variable effect-measure estimates can be obtained. The predictors in such models can include etiologic modifiers of effect measures (which Rubin calls "X factors") as well as validity features of studies' design, conduct and analysis (which he calls "Z factors").

At the individual study level, and therefore in individual patient data (IPD) meta-analysis, sensitivity analysis and bias modeling can be used to enhance validity [26]. As with stratified meta-analysis and meta-regression, these analyses can be conducted on individual or multiple biases. Ebert and Drake [27] provide a good example of a systematic review with simple sensitivity analyses of unmeasured confounders, selection bias and exposure measurement error in individual studies.

In conclusion, the vast majority of indirect quotations of the commentary have been misleading. It appears that authors of systematic reviews who quote the commentary most likely did not read it. Obviously, reviewers and editors of the journals in which articles with the misquotations have been published did not recognize these misleading quotations. To reduce misquotation of commentaries, and perhaps editorials as well, they should be published with abstracts so that the PubMed entries can give authors something more than a title from which to discern the general tenor of the cited publication's contents.

Acknowledgements This work was supported by the German Federal Ministry of Education and Science (BMBF) [Grant no. 01ER1704]. The funding source had no role in the study design, in the collection, analysis and interpretation of data, in the writing of the report, and in the decision to submit the paper for publication.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle–Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses. http://www.ohrica/programs/clinical_epidemiology/oxfordasp. 2009.
2. Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: Including non-randomized studies. In: Higgins JBT, Green S, editors. Cochrane handbook of systematic reviews of interventions, version 510 (updated March 2011). www.handbook.cochrane.org: The cochrane collaboration. 2011.
3. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7:iii–173.
4. Stang A. Critical evaluation of the Newcastle–Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25:603–5.
5. Lee SY, Lee JS. A survey of quotation accuracy in two Korean dermatological journals. Ann Dermatol. 1995;7:236–9.
6. de Lacey G, Record C, Wade J. How accurate are quotations and references in medical journals? Br Med J (Clin Res Ed). 1985;291:884–6.
7. Eichorn P, Yankauer A. Do authors check their references? A survey of accuracy of references in three public health journals. Am J Public Health. 1987;77:1011–2.
8. Evans JT, Nadjari HI, Burchell SA. Quotational and reference accuracy in surgical journals. A continuing peer review problem. JAMA. 1990;263:1353–4.
9. Tfelt-Hansen P. The qualitative problem of major quotation errors, as illustrated by 10 different examples in the headache literature. Headache. 2015;55:419–26.
10. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. 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(10):e1–34. https://doi.org/10.1016/j.jclinepi.2009.06.006.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
12. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, editors. Cochrane handbook of systematic reviews of interventions, version 510 (updated March 2011). www.handbook.cochrane.org: The cochrane collaboration. 2011.
13. Garvin DA. What does product quality really mean? Sloan Manag Rev. 1984;26:25–43.
14. Dickie G. Aesthetics: an introduction. New York: The Bobbs-Merrill Company, Inc.; 1971.

15. On Gerwitz P. I know it when I see it. Yale Law J. 1996;105:1023–47.
16. Crosby PB. Quality is free. New York: McGraw-Hill; 1979.
17. Broh RA. Managing quality for higher profits. New York: McGraw-Hill; 1982.
18. Kuehn AA, Day RL. Strategy of product quality. Harv Bus Rev. 1962;40:100–10.
19. Rodgers A, MacMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: implications for the interpretation and design of thromboprophylaxis trials. Thromb Haemost. 1995;73:167–71.
20. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing risk of bias in included studies. Cochrane handbook for systematic reviews of interventions, version 510 (updated March 2011). www.handbook.cochrane.de: The cochrane collaboration. 2011.
21. Greenland S. Quality scores are useless and potentially misleading—Reply to Re—a critical-look at some popular analytic methods. Am J Epidemiol. 1994;140:300–1.
22. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001;2:463–71.
23. Donabedian A. Evaluating the quality of medical care. Milbank Q. 1966;44:166–203.
24. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282:1054–60.
25. Rubin DR. Meta-analysis: literature synthesis or effect-size surface estimation? J Educ Stat. 1992;17:363–74.
26. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Dordrecht: Springer; 2009.
27. Ebert CS Jr, Drake AF. The impact of sleep-disordered breathing on cognition and behavior in children: a review and meta-synthesis of the literature. Otolaryngol Head Neck Surg. 2004;131:814–26.