BMJ Open

Examining the quality of evidence to support the effectiveness of interventions: an analysis of systematic reviews

Robert L Kane,[1] Mary Butler,[1] Weiwen Ng[2]

To cite: Kane RL, Butler M, Ng W. Examining the quality of evidence to support the effectiveness of interventions: an analysis of systematic reviews. *BMJ Open* 2016;**6**:e011051. doi:10.1136/bmjopen-2016-011051

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-011051).

Received 6 January 2016
Revised 19 April 2016
Accepted 20 April 2016

[1]Minnesota Evidence-based Practice Center, University of Minnesota School of Public Health, Minneapolis, Minnesota, USA
[2]Health Services Research, Policy, and Administration, University of Minnesota School of Public Health, Minneapolis, Minnesota, USA

Correspondence to
Professor Robert L Kane;
kanex001@umn.edu

ABSTRACT

Objective: This analysis examines the quality of evidence (QOE) for 1472 outcomes linked to interventions where the QOE was rated in 42 systematic reviews of randomised clinical trials and/or observational studies across different topics.

Setting: Not applicable.

Participants: 76 systematic reviews.

Primary and secondary outcome measures: Strength of evidence ratings by initial reviewers.

Results: Among 76 systematic reviews, QOE ratings were available for only 42, netting 1472 comparisons. Of these, 57% included observational studies; 4% were rated as high and 12% as moderate; the rest were low or insufficient. The ratings varied by topic: 74% of the surgical study pairs were rated as low or insufficient, compared with 82% of pharmaceuticals and 86% of device studies, 88% of organisational, 91% of lifestyle studies, and 94% of psychosocial interventions.

Conclusions: We are some distance from being able to claim evidence-based practice. The press for individual-level data will make this challenge even harder.

Strengths and limitations of this study

- This is an empirical review of the prevalence of strength of evidence (quality of evidence, QOE).
- It raises issues that are too often ignored.
- It reflects current standards on QOE.
- The sample of reviews was not random.
- The level of agreement across studies on what constitutes QOE could not be determined.

INTRODUCTION

In medical care, evidence-based practice (EBP) is championed, but is it feasible? Comparative effectiveness research is heralded as a way to "learn what forms of health care work best so that we can abandon those that are ineffective and adopt those diagnostic tests, treatments, and approaches to prevention that do the most to improve health."[1] As medicine aspires towards EBP, comparative effectiveness research provides that evidence. Guideline writers are exhorted to base their work on evidence. Rules have been written about how to collect and present evidence (see EQUATOR network http://www.equator-network.org).[2] [3] The fundamental unspoken question is, 'How good is the evidence?' Is an adequate evidence base available to support most recommendations for care? A recent *Lancet* editorial asserts that we are a long way from having evidence to support EBP because much of the information published is not correct.[4]

Proponents of the conceptual framework for EBP have recognised that not all practices can be based on strong evidence. They characterise EBP as practice that uses the best available evidence to inform decisions and considers patients' values and preferences.[5] Those charged with converting evidence reports into guidelines must still rely substantially on judgement. Guideline writers are admonished to distinguish when evidence is strong and when it is not.[6]

However, this formulation leaves open the question of how much evidence is needed to claim an evidence base for practice. Clearly, there will never be empirical support for all decisions in the complex world of practice, but the situation today is well short of that goal. Today we hear complaints that relatively little evidence exists on improving health outcomes. Efforts to create guide books for patients based on systematic reviews were hampered by the limited evidence on final outcomes.[7]

Indeed, the quality of research and research reporting leave much room for improvement. For example, Ioannidis has noted that a substantial number of highly cited studies were later contradicted by research that found weaker effects, or effects in the opposite direction, than the original studies.[8] Some would even assert that because of the uncertainties inherent in the research, blindly applied evidence-based medicine can be hazarardous.[9] Nearly two-thirds (62%) of publications cited by the National Institute for Health and Care Excellence (NICE) to support primary care recommendations were judged of uncertain relevance to patients in primary care.[10]

This study examines a set of systematic reviews to assess the quality of evidence (QOE) that is available to form the basis for EBP. For this purpose, we use the GRADE (Grading of Recommendations Assessment, Development, and Evaluation Working Group) framework to define QOE as the confidence one can place that the effect described is real.[11]

We chose to look within systematic reviews because they gather together relevant studies on a given topic and subject them to critical appraisal, including assessing the overall strength of the evidence they present. In the evidence hierarchy, the upper ranks are populated by systematic reviews and meta-analyses. Therefore, we expected that reviews employing meta-analysis might include better evidence.

METHODS

We used a convenience sample of systematic reviews deliberately stratified by source to capture a range of reviewer organisations and approaches. We reviewed the 10 most recent systematic reviews of interventions published in each of four major journals, chosen because they frequently publish such reviews: *Annals of Internal Medicine, The BMJ, JAMA* and *Pediatrics*. We supplemented this sample with 10 recent reviews issued in reports from the Cochrane Collaborative and 16 from Evidence-based Practice Center (EPC) funded by the Agency for Healthcare Research and Policy (AHRQ) that did not duplicate our selected journal systematic reviews. We selected only reviews that searched for randomised trials or observational studies in multiple databases and laid out clear criteria for inclusion or exclusion of articles. We excluded reviews funded by the US Preventive Services Task Force because we wanted to emphasise treatment interventions. We determined if each review included any observational studies.

Whenever a QOE assessment was performed, we examined the QOE assigned to each intervention/outcome pair to an intervention category. We relied on the original reviewers' judgements about QOE; we did not attempt to make our own judgements about QOE. We dropped reviews where such judgements were not made. Ratings of QOE are based on the body of identified studies that examined the effects a treatment (or intervention) has on a given outcome and provides a rating to communicate the confidence a person may have in the stability of an identified effect. Two closely related methods for ascertaining QOE are GRADE and the AHRQ EPC criteria. GRADE couches QOE in terms of the quality of the evidence and incorporates applicability,[11] while the AHRQ EPC programme separates out applicability for later assessment and uses the term strength of evidence.[12] Both use a common set of criteria including risk of bias (a reflection of the overall quality of included studies), consistency (whether the findings from different studies showed the same pattern of effect), directness (how directly the evidence links the intervention to the desired, ultimate health outcome) and precision (the degree to which the various studies showed comparable variance around the main effect).[13] QOE is typically classified as high, moderate or low, indicating the confidence around the direction and magnitude of the effect. QOE may also be rated as very low or insufficient, which denotes the presence of few studies of quality and the lack of knowledge about the intervention's effect. We collapsed very low and insufficient into an insufficient category. When reviewers attempted to make a comparison, but found no evidence, we assigned a QOE of 'none'.

We classified the intervention categories post hoc into pharmaceuticals, surgery, medical devices, organisational, psychosocial, lifestyle and dental. With the exception of lifestyle interventions, these categories reflect the nature of the intervention. Lifestyle reflects more the target of the intervention; namely, changes in personal health behaviours like eating, exercise and weight loss. Medical devices included those implanted through surgery as well as those used externally, like radiotherapies and durable medical equipment. Organisational interventions were changes to the context in which care was delivered. We assigned each intervention/outcome pair to only one category.

We classified outcomes into intermediate or surrogate outcomes (primarily physiological measures) and final outcomes presumed to affect patients' lives and wellbeing. Examples of intermediate/surrogate outcomes include glycated haemoglobin level, blood pressure and preventive care processes in general. We classified all non-emergency department outpatient visits and all disease complications as intermediate outcomes. Examples of final outcomes included mortality, inpatient admission, emergency department visits, quality of life and symptoms. We excluded adverse events related to the treatment as outcomes.

When reviewers described the QOE for an intervention/outcome pair as 'low to moderate' (two cases) or 'moderate to high' (one case), we classified the QOE as low or moderate, respectively.

We tabulated the responses by category and used a Pearson's χ^2 statistic to examine differences in patterns. We hypothesised, post hoc, that systematic reviews that used meta-analysis might find stronger QOE, and that reviews which included observational studies might find lower QOE.

RESULTS

Of the 76 reviews, 34 (45%) did not use a systematic rating scheme. From the remaining 42 reviews, we abstracted 1472 outcomes linked to a specific intervention. Of those rated, 50% were final outcomes and 49% were intermediate outcomes; 1% could not be classified because they were composite outcomes. Reflecting the sample, of the studies that rated QOE, 39 used the methods endorsed by the AHRQ EPCs and13 used GRADE. A small number used a Bayesian method, which did not explicitly assess the QOE according to either AHRQ or GRADE criteria.

Figure 1 summarises the distribution of the QOE ratings among intervention/outcome pairs for strength of evidence (SOE); 84% were low or insufficient. Only 4% were rated high, and 12% were rated moderate.

Table 1 contrasts the QOE ratings by whether the analysis included observational studies. Just over 70% of intervention/outcome pairs came from reviews that included observational studies. Those without such studies fared somewhat better at the higher ratings but were also more often rated insufficient. The distributions of ratings when observational studies were included versus not were significantly different (χ^2 266.6, p<0.000).

Table 2 summarises the distribution of intervention/outcome pairs across the intervention categories for each level of SOE. Each row shows the QOE distribution for each intervention category. The first column shows the relative frequency of that intervention. Drug studies accounted for more than half of all pairs. Psychosocial interventions accounted for 13%. Surgical studies, device studies, lifestyle interventions and organisational studies each accounted for just under 10%. Evaluations of dental procedures were rare. Surgical studies had the highest rate of high QOE; nonetheless, 74% of the pairs were rated as low or insufficient. Pharmaceuticals and devices were next with 82% and 86%, respectively. Organisational and lifestyle had 88% and 91%, respectively. Psychosocial interventions fared the worst, with 94% of the evaluated intervention/outcome pairs rated as low or insufficient. The distributions of QOE ratings across intervention categories were statistically significant (χ^2 266.6, p<0.0000).

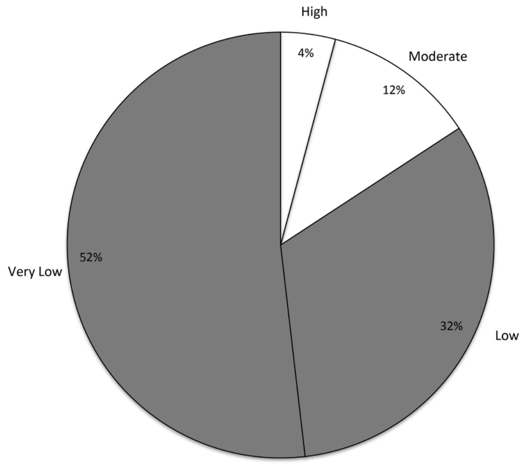


Figure 1 Quality of evidence ratings.

Table 1 QOE ratings for intervention/outcome pairs by inclusion of observational studies

	Observational studies included		
SOE rating	No (N=433)	Yes (N=1039)	Grand total (N=1472
High	5.8% (25)	3.5% (36)	4.1% (61)
Moderate	15% (65)	10.2% (106)	11.6% (171)
Low	22.6% (98)	36.5% (379)	32.4% (477)
Insufficient*	56.6% (245)	49.9% (518)	51.8% (763)
Total	433	1039	1472

*Insufficient is equivalent to GRADE's very low.
GRADE, Grading of Recommendations Assessment, Development, and Evaluation Working Group; QOE, quality of evidence; SOE, strength of evidence.

Figure 2 summarises the effect of meta-analysis on QOE. Contrary to what might be expected, the meta-analysed interventions were less likely to have high or moderate QOE than interventions that were not meta-analysed. Here again the differences were highly significant (χ^2 531.4, p<0.0000).

DISCUSSION

Many people make a distinction between EBP and the quality of the evidence. While utilising available evidence to make informed clinical decisions is laudable, some evidence is better than none. But the unanswered question remains how much is sufficient to assert a claim of EBP.

In his pioneering book on the subject, David Sackett defined the goal of evidence-based medicine as providing "…clinicians with the best scientifically derived information on which to make clinical decisions."[14] By that definition the subsequent efforts to create evidence have moved us forward, but, as this paper illustrates, we still have some distance to go before we can say we are basing our treatment decisions on evidence. Guideline developers struggle to extract the most they can from available research, but much of what they do continues to be based on thoughtful extrapolation.

This assessment of the QOE of interventions suggests that great caution should be used when talking about EBP. Only 42 of the 76 reviews rated QOE; among those, most assessments of effectiveness were based on weak evidence. Few bodies of evidence rose to the level of substantially reducing uncertainty in results. We also need better reviews. The limited number of systematic reviews that assessed strength (or quality) of evidence is cause for concern. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)[15] criteria mandate this step. (http://www.equator-network.org/reporting-

Table 2 Quality of evidence rating by type of intervention

Intervention category	Distribution by category (%, N)	QOE rating (%, N)			
		High	Moderate	Low	Insufficient
Surgical	6 (95)	8.4 (8)	16.8 (16)	24.2 (23)	50.6 (48)
Pharmaceutical	59 (870)	5.4 (47)	12.9 (112)	35.7 (311)	46.0 (400)
Device	9 (127)	1.6 (2)	12.6 (16)	31.5 (40)	54.3 (69)
Organisational	7 (106)	1.9 (2)	10.4 (11)	26.4 (28)	61.3 (65)
Lifestyle	6 (81)	0.0 (0)	8.6 (7)	54.3 (44)	37.1 (30)
Psychosocial	13 (192)	1.0 (2)	4.7 (9)	16.1 (31)	78.2 (150)
Total	100 (1472)	4 (61)	12 (171)	33 (477)	51 (762)

guidelines/prisma/).[3] All systematic reviews should use GRADE or a similar set of guidelines to rate QOE.

We acknowledge that QOE ratings are not strongly reliable[16] and that other reviewers might come to different conclusions about QOE when faced with the same results.[17] The differences by meta-analysis use also suggest a problem in rating QOE. The results seem paradoxical. Analyses that included meta-analysis had more instances of insufficient and low evidence. While poolability per se is not a sign of quality, one might have expected that studies that were poolable would have more consistent data and hence provide a stronger source of evidence. Further, topics with decisional equipoise may be more likely to be reviewed, but equipoise may be in part a result of conflicting studies. However, variations in judgement are not likely the sole reason for findings of this magnitude.

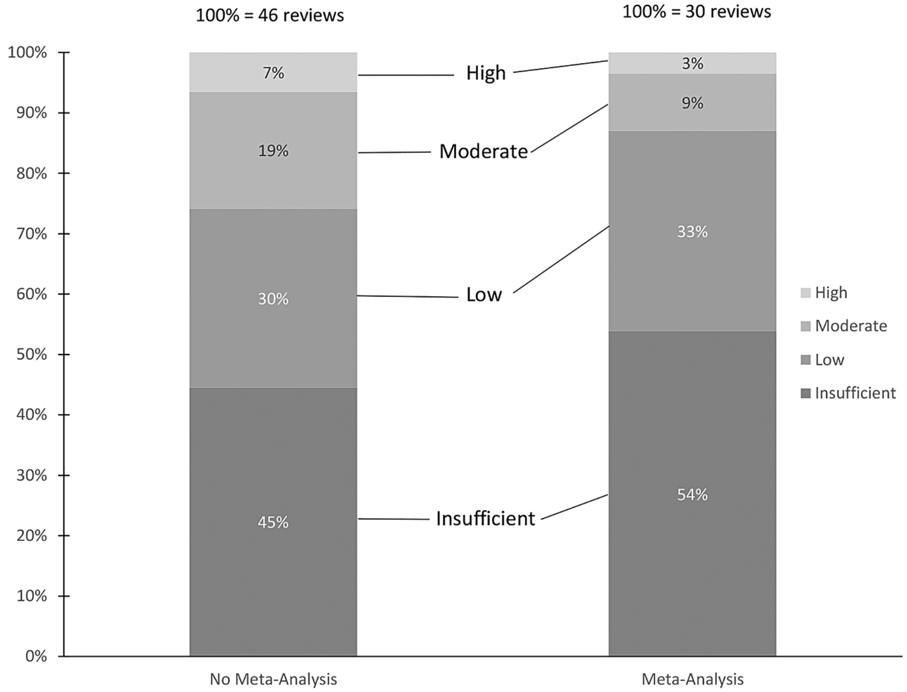
Evidence-based medicine has moved the practice of healthcare to a higher standard, but it is time to assess its status. Developing clinical guidelines still relies heavily on judgements beyond the available evidence. A recent study on the reproducibility of psychological experiments suggests that we need not simply more research, but better research.[18]

Improving the methodological quality of new randomised controlled trials (RCTs) should help improve the evidence base considerably, but new research should also consider the methodological shortcomings of the existing efforts and specifically remedy them. More attention needs to be placed on reproducibility using consistent methods and measures.

Currently, even when there is evidence, it is usually limited to providing the mean effect of mean treatment on the mean patient. RCTs may be the first line of analysis, but few RCTs are large enough to permit subgroup analysis. Making RCT data publicly available for meta-analyses that might yield patient-level results is a promising effort. However, it is still a ways off, and it will be limited in scope.

We will need to rely on analyses of large clinical data sets and deal with all the inherent concerns about confounding.[19] [20] The challenges of creating high-quality

Figure 2 Differences in strength of evidence (SOE) ratings by meta-analysis.



evidence will become more complex as we move to integrate 'big data' that provide more individualised information beyond the mean effect of the mean treatment on the mean patient. Some suggestions have been made about how to integrate RCTs and big data,[21] but the vast majority of big data analyses will rely on observational methods.

Since few trials directly compare the effectiveness of alternative treatments head-to-head, we frequently rely on inferences from indirect comparisons that use Bayesian methods. Therefore, researchers must strive to assess conclusions drawn from Bayesian methodology consistent with the existing framework that researchers and decision-makers have come to rely on for SOE.

Medical care must ultimately decide what constitutes evidence. Given the debates around how to judge the study quality of non-RCTs[22] [23] and how to test and assess complex interventions,[24] the challenges to creating a strong evidence base are substantial. Medical care will likely always contain some elements of art and messiness. Because patients and clinicians need to make life-affecting decisions, they will have to act with less certainty; but they still need a reasonable amount. As clinicians and patients increasingly demand patient-level data, the field will be forced to adopt other techniques. We need more reliable methods that allow for (1) experimentation at the local level, (2) collection of at least minimally necessary data and outcomes to allow appropriate assessment, and (3) aggregation of results to feed back into the learning process.

Part of the question depends on whether the goal is elucidating causation or simply predicting outcomes. The latter can be achieved under less stringent conditions. To support decision-making, many patients and their clinicians may need only to estimate the likelihood of a successful outcome of a given treatment without fully knowing the underlying mechanism.

In the meantime it behooves us to be more modest in our claims to practice evidence-based medicine.

Contributors RLK and MB designed the study. WN abstracted the data and developed preliminary material. All authors wrote the paper.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. McClellan M, Benner J, Garber A, *et al*. Implementing Comparative Effectiveness Research: Priorities, Methods, and Impact. The Brookings Institute, 2009.
2. Eden J, Levit L, Berg A, Morton S, eds. *Finding what works in health care: standards for systematic reviews*. Washington DC: National Academies Press, 2011.
3. Hales S, Lesher-Trevino A, Ford N, *et al*. Reporting Guidelines for Implementation and Operational Research. *Bull World Health Organ* 2016;94:58–64.
4. Horton R. Offline: what is medicine's 5 sigma? *Lancet* 2015;385:1380.
5. Guyatt G, Rennie D, Meade M, *et al*. *Users' guides to the medical literature. A manual for evidence-based clinical practice*. 3rd edn. American Medical Association, 2015.
6. Djulbegovic B, Guyatt GH. Evidence-based practice is not synonymous with delivery of uniform health care. *JAMA* 2014;312:1293–4.
7. Fahey T, NicLiam B. Assembling the evidence for patient centred care. *BMJ* 2014;349:g4855.
8. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218–28.
9. Livingston EH, McNutt RA. The hazards of evidence-based medicine: assessing variations in care. *JAMA* 2011;306:762–3.
10. Steel N, Abdelhamid A, Stokes T, *et al*. A review of clinical practice guidelines found that they were often based on evidence of uncertain relevance to primary care patients. *J Clin Epidemiol* 2014;67:1251–7.
11. Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–94.
12. Berkman ND, Lohr KN, Ansari MT, *et al*. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. *J Clin Epidemiol* 2015;68:1312–24.
13. Berkman ND, Lohr K, Ansari M, *et al*. *Grading the strength of a body of evidence when assessing health care interventions for the effective health care program of the Agency for Healthcare Research and Quality. An update*. Rockville, MD: Agency for Healthcare Research and Quality, 2013. Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13(14)-EHC130-EF. http://www.effectivehealthcare.ahrq.gov/reports/final.cfm
14. Sackett DL, Richardson WS, Rosenberg WMC, *et al*. *Evidence-based medicine; how to practice and teach EBM*. New York: Churchill Livingstone Press, 1997.
15. Preferred Reporting Items for Systematic Reviews and Meta-Ananlyses (PRISMA) [Web site]. Ottawa University of Oxford, 2009 [updated copyright 2015; cited 2016 2/18/2016]. Website for PRISMA containing PRISMA checklists, flow diagram, statement, E & E statement]. http://www.prisma-statement.org/Default.aspx
16. Berkman ND, Lohr KN, Morgan LC, *et al*. Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews: Methods Research Report. (Prepared by RTI International-University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 12-EHC-67-EF. Rockville, MD: Agency for Healthcare Research and Quality: 2012 May.
17. Alexander PE, Gionfriddo MR, Li SA, *et al*. A number of factors explain why WHO guideline developers make strong recommendations inconsistent with GRADE guidance. *J Clin Epidemiol* 2016;70:111–22.
18. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 2015;349:aac4716.
19. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA* 1998;279:1089–93.
20. Deeks JJ, Dinnes J, D'Amico R, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7:iii–x, 1–173.
21. Angus DC. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA* 2015;314:767–8.
22. Lash TL, Fox MP, MacLehose RF, *et al*. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969–85.
23. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615–25.
24. Guise JM, Chang C, Viswanathan M, *et al*. Agency for Healthcare Research and Quality Evidence-based Practice Center methods for systematically reviewing complex multicomponent health care interventions. *J Clin Epidemiol* 2014;67:1181–91.