ANN MED

Annals of Medicine, 2015; 47: 1–5

ISSN 0785-3890 print/ISSN 1365-2060 online
DOI: 10.3109/07853890.2014.969766

informa
healthcare

EDITORIAL

Methodological considerations of the GRADE method

Antti Malmivaara

Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki, Finland

The GRADE method (Grading of Recommendations, Assessment, Development, and Evaluation) provides a tool for rating the quality of evidence for systematic reviews and clinical guidelines. This article aims to analyse conceptually how well grounded the GRADE method is, and to suggest improvements. The eight criteria for rating the quality of evidence as proposed by GRADE are here analysed in terms of each criterion's potential to provide valid information for grading evidence. Secondly, the GRADE method of allocating weights and summarizing the values of the criteria is considered. It is concluded that three GRADE criteria have an appropriate conceptual basis to be used as indicators of confidence in research evidence in systematic reviews: internal validity of a study, consistency of the findings, and publication bias. In network meta-analyses, the indirectness of evidence may also be considered. It is here proposed that the grade for the internal validity of a study could in some instances justifiably decrease the overall grade by three grades (e.g. from high to very low) instead of the up to two grade decrease, as suggested by the GRADE method.

Key words: Evidence-based medicine, GRADE, methodology, quality of evidence, risk of bias, systematic reviews, validity

Key message

- The quality of evidence during systematic reviews should be based on the degree of internal validity of each study and the consistency of findings across clinically homogeneous studies and, when feasible, also on publication bias.

Background

The GRADE method (Grading of Recommendations, Assessment, Development and Evaluation) aims to provide a tool for rating the quality of evidence (particularly for effectiveness) and grading the strength of recommendations (1,2). The tool is intended for use by those summarizing evidence for systematic reviews, as well as in clinical practice guidelines and health technology assessments. The method for rating the quality of evidence has significant implications for patients, health care professionals, policy-makers, and researchers. The GRADE method has been endorsed by many well-known organizations around the world (3).

Five of the eight criteria proposed in the GRADE method have the potential to decrease one's confidence in the correctness of the effect estimates: risk of bias, inconsistency of results across studies, indirectness of evidence, imprecision, and publication bias (4–11). Three further criteria are proposed that have the potential to increase this confidence: a large magnitude of effect with no plausible confounders, a dose-response gradient, and a conclusion that all plausible residual confounding would further support inferences regarding treatment effect. GRADE proposes these three criteria should be considered particularly in observational studies (10). The GRADE method proposes four levels for expressing the quality of evidence: high, moderate, low, and very low.

The aim of this article is, firstly, to describe the conceptual meaning of each of the eight GRADE criteria, and to consider their ability to increase or decrease confidence in estimates of outcome of a systematic review. A second aim is to consider the conceptual homogeneity of the GRADE criteria, the rationale for weighting the GRADE criteria, and the rationale for summarizing the values decided for each criterion in order to reach the overall rating of confidence in the effect estimate.

The eight GRADE criteria for rating evidence

The eight GRADE criteria, their potential to increase or decrease the grade of evidence, and the author's interpretation and conclusions are presented in Table I.

Criteria that may decrease confidence in the results

The first GRADE criterion—risk of bias—is conceptually a matter of the internal validity of a scientific study. The degree of risk of bias can be determined by reading carefully the methods section of each original study and assessing how well the planning and execution of the study was carried out. This has been a universal

Correspondence: Antti Malmivaara, MD, PhD, Chief Physician, Centre for Health and Social Economics, National Institute for Health and Welfare, Mannerheimintie 166, 00270 Helsinki, Finland. E-mail: antti.malmivaara@thl.fi

(Received 4 June 2014; accepted 15 September 2014)

Table I. The GRADE criteria and the author's interpretations and conclusions.

GRADE criteria	Author's interpretation of the GRADE criteria	Author's conclusion of the GRADE criteria	GRADE criteria for assigning level of evidence	Author's conclusion of the GRADE criteria for assigning level of evidence
Criteria which may decrease confidence in results				
Limitations to study quality (risk of bias)	Reflects the (lack of) internal validity of the study. The foremost quality criterion in science	Agree with the criterion	Decrease with 1 or 2 levels if serious limitations (−1) or very serious limitations (−2) to study quality	Decreasing with 1 or 2 levels is appropriate. In some cases decreasing with even 3 levels (e.g. evidence from high to very low) is justified
Inconsistency	Reflects the (lack of) consistency of the results of a study	Agree with the criterion	Decrease with 1 or 2 levels if inconsistency is serious (−1) or very serious (−2)	Decreasing with 1 or 2 levels is appropriate
Indirectness of evidence	All studies synthesized in a systematic review should have similar patient populations, interventions, control interventions, and outcomes. It is not appropriate to do a meta-analysis combining direct and indirect evidence posing different hypotheses, except in network meta-analyses	Mostly disagree with the criterion	Decrease with 1 or 2 levels if serious (−1) or very serious indirectness (−2)	Decreasing level of evidence is not appropriate when based on summarizing results from incommensurable studies. Network meta-analyses may allow decisions for decreasing level of evidence based on indirectness
Imprecision	Reflects random error in outcome estimates. The wideness of confidence intervals is one result of a study or meta-analysis and should not be used as a quality criterion	Disagree with the criterion	Decrease with 1 or 2 levels if serious imprecision (−1) or very serious imprecision (−2)	Decreasing level of evidence based on degree of random error in the outcome estimates is not appropriate; the limitation shown by wide confidence intervals is a result of a systematic review
Probability of publication bias	Selective reporting of outcomes is a matter of internal validity of the study and belongs to the 'limitations to study quality' criterion. When individual studies are not at all published, the results of a systematic review are potentially biased	Agree with the criterion	Decrease with 1 or 2 levels if publication bias likely (−1) or very likely (−2)	Decreasing level of evidence is appropriate
Criteria which may increase confidence in results				
Large magnitude of effect	This is a result of a study or meta-analysis and should not be used as a quality criterion. Large magnitude of effect may imply a high risk of biased results rather than increased confidence in results	Disagree with the criterion	Increase with 1 or 2 levels if large (+1) or very large (+2) evidence of association	Increasing level of evidence based on large magnitude of effect is not appropriate, because of a risk for biased conclusions
Dose-response gradient	Dose-response gradient often exists in studies assessing etiology of disease, but effectiveness of an intervention usually does not show a linear dose-response pattern	Mostly disagree with the criterion	Increase with 1 level if evidence of a dose response gradient (+1)	Increasing level of evidence based on a dose-response gradient is rarely appropriate when assessing effectiveness of an intervention
Residual confounding would further support inferences regarding treatment effect	If some plausible confounders have not been documented, there is no credible way to determine how adjusting these parameters would alter the effectiveness estimates.	Disagree with the criterion	Increase with 1 level if all plausible confounders would reduce a demonstrated effect (+1) or would suggest a spurious effect if no effect was observed (+1).	Increasing level of evidence is not appropriate because the confounders cannot be documented. Consequently there is a risk for biased conclusions

way in science to assess confidence in the trustworthiness of a particular study. There is also empirical evidence in medicine on how methodological imperfections in a randomized trial can influence effect estimates, which usually become exaggerated (12).

The second GRADE criterion—the inconsistency of results across studies included in a systematic review—means that the results deviate from each other, and this naturally leads to decreased confidence in the effectiveness estimates. If the original studies in a systematic review are clinically homogeneous (answering the same research question) and are all of high methodological quality but there is major inconsistency in the results, then statistical testing will probably show that there is heterogeneity in the results. In these cases, it seems evident that given the inconsistency in the results, confidence in the effectiveness estimates must be lower than if the results from each study were similar. Consistency across populations, interventions, comparison interventions, and outcomes may occur in its purest form in cases where a particular study has been undertaken to test the reproducibility of the findings. The reproducibility of results along with a high internal validity has been the universal criterion to determine the confidence of research findings in science.

The third GRADE criterion—indirectness—refers to any deviation in the research question or its operationalization between studies that are included in a systematic review; in other words, differences may be found in the PICOs (population characteristics, interventions, control interventions, or in outcome measures). Conceptually, only direct outcomes produce evidence that can be considered of adequate credibility for answering the research question. Surrogate outcomes may not be associated with the primary outcome. As an example, an intervention to treat osteoporosis may increase bone mineral density (surrogate outcome) but may not be associated with decreased occurrence of hip fracture (primary outcome). Thus, rather than assess the degree of indirectness it seems more plausible to analyse the direct and indirect outcomes separately and to produce two separate evidence propositions: taking the previous example further, one for the intervention's effectiveness on osteoporosis in increasing bone mineral density, and another for its effectiveness in preventing hip fracture. In this scenario the patients should know that there is, for example, very low confidence in treatment effectiveness for prevention of hip fracture, but high confidence for increasing the bone mineral density. The physician and the patient should then discuss the implications of this evidence in order to reach an appropriate treatment decision. To sum up, the study question according to PICO has to be followed consistently, i.e. all studies in the systematic review must have similar PICOs. However, in a network meta-analysis studies having both direct (e.g. occurrence of a hip fracture) and indirect (e.g. bone mineral density) outcomes can in some cases be used to obtain a summary estimate of the patient-relevant direct outcome measure (occurrence of a hip fracture). If evidence is produced by network meta-analyses, confidence in the effectiveness estimates may be lowered. Network meta-analysis is a promising tool for systematic reviews, but still requires conceptual and operational development (13).

The fourth GRADE criterion—imprecision—reflects conceptually the random variation in outcome estimates due to chance and is distinct from internal validity, which reflects the potential risk of obtaining biased estimates. If the original studies in a systematic review are clinically homogeneous (they have similar research questions and they are similar in regard to the patient, the intervention, the control intervention, the outcome, and the successful execution of the trial) and all have a low risk of bias, it is appropriate to undertake a meta-analysis and obtain a summary estimate of effectiveness. The 95% confidence interval is often interpreted as indicating a range within which we can be 95% certain that the true effect lies (14), and it allows also an interpretation of whether the clinically important results lay within or outside these confidence intervals. This information on the wideness of the confidence intervals can be used as a basis for clinical inferences, e.g. it allows a conclusion that confidence intervals have exceeded the minimal clinically important difference. It is not necessary to include imprecision as a separate criterion for confidence in the effectiveness estimate in a systematic review, but the confidence in results should concomitantly cover both the point estimate and the respective confidence interval. Whether confidence intervals are wide or narrow, the same probability exists of the point estimate lying within the limits of these confidence intervals. The uncertainness of the point estimate and its confidence interval is related to the risk of bias within the original studies, and to the imprecision of the results, respectively. For example, all else being equal, when the point estimate and confidence interval in one case is 0.5 (0.4–0.6) and in another case 0.5 (0.1–3.0) the degree of confidence in each result (taking into consideration both the point estimate and respective confidence interval) is the same—in the former case the point estimate is 0.5 and precision is very good (0.4–0.6), and in the latter case the point estimate is 0.5 but precision is very poor (0.1–3.0).

The fifth GRADE criterion that can decrease confidence in the outcomes of a systematic review is publication bias. Selective reporting of outcomes is a matter of the internal validity of an individual study, and should be included in the criterion for 'limitations to the study quality'. When individual studies are not published at all, biased results in systematic reviews may emerge. The existence of publication bias is one of the potential sources of risk of bias in systematic reviews. Conceptually publication bias decreases the internal validity of a systematic review. Whether or not this bias exists remains often speculative, even though indications of publication bias can be traced using e.g. funnel plot graphics. Mandatory registration of trials has increased the possibilities of identifying publication bias.

Criteria that may increase confidence in results

GRADE suggests considering rating up the quality of evidence in case of methodologically rigorous observational studies (10).

The first GRADE criterion that has the potential to increase one's confidence in the results is a large magnitude of effect. Conceptually this criterion refers to the properties of the study object, i.e. how large an effect the cause (intervention) can bring about in the outcome of a particular study population and setting. Secondly it shows the degree to which this effect has been seen in a particular study. The first conceptualization poses a problem: How can the inherent property of a cause–effect relationship be used as a criterion for confidence in the effect estimates? There is also a further concern to be raised based on empirical studies: There is evidence that deficiencies in the internal validity of a study may lead to exaggerated treatment effects (12). This has been illustrated in observational studies that indicate quite convincingly that postmenopausal oestrogen therapy protects from cardiovascular disease (15), while later randomized controlled trials indicate the opposite (16,17). Another example of exaggerated effectiveness is that of selective prescribing leading to overestimation of the benefits of lipid-lowering drugs (18). Numerous observational studies have reported strong relationships of the effect of prevention on health outcomes that have later been contradicted by randomized controlled trials (19). A multitude of methodological research has identified sources of bias in observational studies that are related to patient behaviours or underlying patient characteristics, the healthy-user effect, the healthy-adherer effect, confounding by

functional status or cognitive impairment, and confounding by selective prescribing (19). Better adjustment for confounders in observational studies may not even be feasible without external validation studies (20).

A specific category of large magnitude of effect is created by circumstances where deterioration of the patient's condition is inevitable but treatment provides instantaneous or steadily increasing and clinically predictably improvement—e.g. epinephrine for anaphylactic shock or dialysis for increasing life expectancy in terminal renal failure. In these cases the effectiveness of the treatment is obvious, and there is usually no controversy among professionals. In estimating the degree of effectiveness in these particular cases one may use the method proposed by Glasziou et al. (21). However, there are cases in which there is apparently a very large magnitude of effect, but a deeper insight indicates that e.g. due to variation in the severity of the disease the magnitude of the effect is actually not clear. For example hip replacement for severe osteoarthritis increases function and decreases pain. In this case, due to lack of randomized trials assessing effectiveness of hip replacement in comparison with natural course or conservative treatment, the magnitude of effect in severe osteoarthritis is uncertain, and in milder cases of osteoarthritis the effectiveness is not at all clear. Effectiveness of hip replacement, at least for the less severe cases of osteoarthritis, should be studied in randomized trials, as findings from observational studies will remain uncertain due to potential for confounding by indication, by other baseline differences, and due to other potential biases related to observational studies. To sum up, while a large magnitude of effect is according to GRADE supposed to increase one's confidence in the effectiveness estimate, there is plenty of evidence indicating that it should rather make one wary of potential bias.

The second GRADE criterion that can be considered to increase one's confidence in the evidence of effectiveness is a dose-response gradient between the intervention and outcome. Conceptually this is a matter of the relationship between cause and effect and thus a property of the studied causal relationship, the establishment of which is one of the objectives of the systematic review. For this reason, it is conceptually problematic to use this criterion as an external measure of one's confidence in the results. Empirically, a dose-response gradient is found mainly in studies assessing the etiology of diseases, e.g. there is a dose-response gradient between tobacco smoking and the risk of lung cancer. On the contrary, when dealing with the effectiveness of interventions, a dose-response gradient is often lacking, e.g. drug treatments may not exhibit a linear relationship between dose and effect (22). Furthermore, the dose-response gradient can be confounded, e.g. by severity of the disease which may be a more powerful effect modifier than the dose-response pattern (23). Furthermore, when there is evidence of a dose-response relationship between treatment and outcome, no distinction needs to be made between evidence coming from a randomized trial or from an observational study when deciding whether to increase one's confidence in the evidence of effectiveness.

The third GRADE criterion that has the potential to increase one's confidence in the evidence is the case where one can reasonably assume that all plausible residual confounding would further support inferences regarding treatment effect. This criterion is one that would be considered in the discussion section of a paper, since the study itself is not able to provide an empirical documentation to assess whether this criterion has been actualized. Conceptually it is a matter of the internal validity of the study in situations where it has not been possible to document all relevant confounders, which therefore cannot be adjusted for in the statistical analysis. The concern here is how one can know whether the hidden confounders have been actualized or not in a particular study—and furthermore, in cases where the confounders would have been documented and adjusted for, what would have been the impact on the effect estimate? This difficulty is illustrated above in the case of observational studies versus randomized controlled trials.

The GRADE method for summarizing the quality of evidence

The GRADE method gives a weighting to the eight criteria. In the case of risk of bias, an inconsistency in results across studies, serious deficiencies in the indirectness of evidence, imprecision, and publication bias would lead to a one-level decrease in the grade given to the quality of evidence, while very serious deficiencies would lead to a two-level decrease (in the latter case, for example, high-level evidence will be downgraded to low-level evidence) (4). In cases where there is a large magnitude of effect, the level of evidence may be increased by one level, and in cases where there is a very large magnitude of effect, it would be increased by two levels. If there is a dose-response gradient, the level of evidence can be increased by one level. In cases where hypothetical control for all plausible residual confounding would be expected to support inferences regarding a treatment effect, the level of evidence can be increased by one level.

Finally the points are summarized together in order to reach the final rating of the quality of evidence. Randomized controlled trials start from a high level and observational studies start from a low level.

My concern in the calculation of the final confidence in the evidence is, as described above, firstly related to the potential of all the eight items to reflect accurately the trustworthiness of the causal relationships. In my opinion, only three of the criteria—risk of bias, inconsistency of the findings, and publication bias—are valid for an assessment of the grade of evidence in systematic reviews. Decreasing the grades of evidence based on these three criteria is thus justified—although there is a need to consider the weighting of these criteria. If the risk of bias in all randomized trials is very high indeed, decreasing the grade of evidence by only two grades (e.g. evidence quality changes from high level to low level) may not be enough, but rather the most appropriate decision could be to decrease the grade of evidence by three grades (e.g. evidence quality move from high level to very low level). An obvious example of this is a situation where all the 12 relevant quality items for a randomized trial (24) are considered, but no trial meets any of these or all trials include fatal flaws (e.g. more than 50% loss to follow-up).

Another concern is related to the conceptual heterogeneity of the eight items in calculating the decisive level for confidence in the point estimate. When conceptually different entities are summarized together, defining the precise nature of that confidence in the evidence becomes impossible. In addition, presentations of these different concepts obtain similar weights (-1 or -2, zero, $+1$, $+2$), which are then summarized to reach the final estimate for confidence in the results. I think that the conceptual and empirical basis for this weighting, the suggested threshold values for obtaining each weight, as well as the way of simply summarizing together values representing different concepts have not been justified in a satisfactory manner in the articles describing the GRADE method. In a paper on the overall ratings of confidence in effect estimates, the potential for arriving at a non-plausible grading of evidence in summing up the points of each of the

validity criteria has been addressed (25). In these cases it is suggested that the gestalt of the confidence in estimates of effect is considered before arriving at the final decision. I think that the need for this recent additional guidance illustrates the problems related to the conceptual heterogeneity of the eight items and to the method for weighting these items and finally calculating the level of evidence.

I consider the GRADE method laudable for having brought also observational studies into the assessment of the effectiveness of interventions. I think, however, that the methodological rigour (lack of risk of bias) of each particular study should be the primary criterion by which to judge the reliability of produced evidence. Observational studies have an inherent advantage over randomized trials in the often superior adherence to the interventions compared with the experimental designs. Moreover, methodological developments recently introduced have used instrumental variables to compensate for the lack of randomization (26). I think that qualities specific to observational studies should be considered when determining the quality of evidence that these studies provide (27). Although the risk of bias in observational studies is usually higher than in randomized trials, observational studies may, for some study questions, provide more reliable results. Thus appraising observational studies consistently and in a uniform manner as methodologically weak and starting at the grade of low quality may not reflect each particular research context appropriately.

Conclusions

It is suggested that assessing the quality of evidence in systematic reviews should be based on the degree of internal validity of each study and the consistency of findings across clinically homogeneous studies and, when feasible, also on publication bias. In cases of very high risk of bias, the grade of evidence should be decreased by three grades (e.g. from high level to very low level) instead of decreasing by only two grades, as suggested by the GRADE method.

Acknowledgements

The author thanks Asko Lukinmaa, MD, PhD, for comments on the manuscript and Mark Phillips, BA, for checking the English language.

Funding: No outside funding.

Declaration of interest: The author declares no support from any organization for the submitted work; no financial relationships with any organization that might have an interest in the submitted work in the previous three years; and no other relationships or activities that could appear to have influenced the submitted work, except being a member of the Editorial Board of the Cochrane Collaboration Back Review Group.

References

1. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;3287454:1490.
2. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;644:380–82.
3. Kavanagh BP. The GRADE system for rating clinical guidelines. PLoS Med. 2009;69:e1000094.
4. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;644:401–6.
5. 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;644:407–15.
6. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol. 2011;6412:1294–302.
7. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol. 2011;6412:1303–10.
8. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol. 2011;6412:1283–93.
9. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J Clin Epidemiol. 2011;6412:1277–82.
10. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;6412:1311–16.
11. Goldet G, Howick J. Understanding GRADE: an introduction. J Evid Based Med. 2013;6:50–4.
12. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;3367644:601–5.
13. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. BMJ. 2013;346:f2914.
14. Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org..
15. Stampfer MJ, Colditz GA, Willett WC, Manson JE, Rosner B, Speizer FE, et al. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the nurses' health study. N Engl J Med. 1991;32511:756–62.
16. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA. 1998;2807:605–13.
17. Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med. 2003;3496:523–34.
18. Glynn RJ, Schneeweiss S, Wang PS, Levin R, Avorn J. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. J Clin Epid. 2006;59:819–28.
19. Shrank WH, Patrick AR, Brookhart A. Healthy User and Related Biases in Observational Studies of Preventive Interventions: A Primer for Physicians. J Gen Intern Med. 2011;26:546–50.
20. Sturmer T, Glynn RJ, Rothman KJ, Schneeweiss S. Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information. Med Care. 2007;45:S158–65.
21. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;3347589:349–51.
22. Davis JM, Chen N. Dose response and dose equivalence of antipsychotics. J Clin Psychopharmacol. 2004;24:192–208.
23. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration. Plos Med. 2008;5:260–7.
24. Furlan AD, Pennick V, Bombardier C, van Tulder M; Editorial Board, Cochrane Back Review Group. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine (Phila Pa 1976). 2009;34:1929–41.
25. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;662:151–7.
26. Xian Y, Holloway RG, Chan PS, Noyes K, Shah MN, Ting HH, et al. Association between stroke center hospitalization for acute ischemic stroke and mortality. JAMA. 2011;3054:373–80.
27. Vandenbroucke J. When are observational studies as credible as randomised trials? Lancet. 2004;3639422:1728–31.