EDITOR'S CHOICE

Random variation and rankability of hospitals using outcome indicators

Anne-Margreet van Dishoeck, Hester F Lingsma, Johan P Mackenbach, Ewout W Steyerberg

Department of Public Health, Center for Medical Decision Making, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

Correspondence to
Anne-Margreet van Dishoeck, Room Ae-138, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands; a.m.vandishoeck@erasmusmc.nl

Accepted 19 April 2011
Published Online First
3 June 2011

ABSTRACT

Objective: There is a growing focus on quality and safety in healthcare. Outcome indicators are increasingly used to compare hospital performance and to rank hospitals, but the reliability of ranking (rankability) is under debate. This study aims to quantify the rankability of several outcome indicators of hospital performance currently used by the Dutch government.

Methods: From 52 indicators used by the Netherlands Inspectorate, the authors selected nine outcome indicators presenting a fraction and absolute numbers. Of these indicators, four were combined into two, resulting in seven indicators for analysis. The official data of 97 Dutch hospitals for the year 2007 were used. Uncertainty in the observed outcomes within the hospitals (within hospital variance, σ^2) was estimated using fixed effect logistic regression models. Heterogeneity (between hospital variance, τ^2) was measured with random effect logistic regression models. Subsequently, the rankability was calculated by relating heterogeneity to uncertainty within and between hospitals ($\tau^2/(\tau^2 + \text{median } \sigma^2)$).

Results: Sample sizes varied but were typically around 200 per hospital (range of median 90–277) with a median of 2–21 cases, causing a substantial uncertainty in outcomes per hospital. Although fourfold to eightfold differences between hospitals were noted, the uncertainty within hospitals caused a poor (<50%) rankability in three indicators and moderate rankability (50–75%) in the other four indicators.

Conclusion: The currently used Dutch outcome indicators are not suitable for ranking hospitals. When judging hospital quality the influence of random variation must be accounted for to avoid overinterpretation of the numbers in the quest for more transparency in healthcare. Adequate sample size is a prerequisite in attempting reliable ranking.

INTRODUCTION

There is a growing focus on quality and safety in healthcare. Increasingly, indicators are used to assess hospital performance. In different countries nationwide systems have been set up to monitor the performance of healthcare institutions using a framework of structure, process and outcome indicators.[1,2]

Public disclosure of the results of hospital performance leads to several inconsistent comparisons and rankings and there is concern among professionals about the value and reliability of such rankings.[3–10] Although rankings seem to be simple, they ignore the chance variability in differences between hospitals and the magnitude of differences.[11] In this research the authors focus on the suitability of indicators, specifically outcome indicators, to provide reliable hospital comparisons.

Two core components determine the reliability of hospital comparisons: within-hospital uncertainty (how reliable the estimates are for each hospital) and between-hospital heterogeneity (how large the differences are between hospitals).

The amount of uncertainty in the analysis of hospital performance is higher than intuition might suggest.[12] For low-incidence outcome and for smaller subgroups in the population, uncertainty can be large.[13] The smallest hospitals would likely experience five to seven times more uncertainty about their true performance.[14]

The second component is heterogeneity between hospitals.[15] Heterogeneity relates to the true differences beyond chance between hospitals and can be estimated with random effect models.

Both components determine the reliability of ranking with an indicator, the 'rankability'. The term rankability, which was first used by van Houwelingen *et al* (web-published research[16]), measures what part of the variation between the crude hospital effects is due to unexplained differences as opposed to uncertainty. The authors loosely interpret rankability as the signal to (statistical) noise ratio.

Because there are no minimal sample size requirements for the indicators used by the Dutch government, the numbers may be small, making ranking attempts less reliable. This study aims to quantify the rankability of several outcome indicators of hospital performance in the Netherlands.

METHODS

Data

Data were obtained from the Netherlands Inspectorate's indicator set. The inspectorate uses this set to assess possible flaws in the quality of care in Dutch hospitals. This obligatory set includes 21 areas with 52 performance indicators, of which 14 are outcome indicators presenting both fraction and absolute numbers. Five indicators were excluded because of clear evidence of registration bias, such as extrapolation of a limited sample in time or patient groups, leaving nine outcome indicators (table 1). Data from 2007 were used, which are publicly available (http://www.ziekenhuizentransparant.nl/). For acute myocardial infarction (AMI), the majority of hospitals reported in-hospital mortality instead of 30-day mortality. Several hospitals reported both. Using these data, the 30-day mortality was multiplied by 0.74 to give data for the five hospitals that only reported 30-day mortality.

Uncertainty

Numerator and denominator data were used for each hospital to create a patient level dataset. A coefficient for unfavourable outcome was estimated for each hospital and was compared to the overall average using a fixed effect logistic regression model with an offset variable and hospital as a categorical variable. The SE of the estimated coefficient (σ^2) indicates the uncertainty of the estimate, or the within-hospital variance. The median σ^2 over all hospitals was taken to represent the within-hospital variance. The median was used because of the skewed distribution of the σ^2.

Heterogeneity

A random effect logistic regression model was used to estimate unexplained heterogeneity, indicated by τ^2 (the between-hospital variance). Unlike the fixed effect model, the random effect model accounts for the fact that the observed outcomes for smaller hospitals can take on extreme values because of random variation. The variance indicates the differences between hospitals beyond chance.[17]

For the interpretation of τ^2, a 95% range of ORs was calculated for the hospitals compared with the average as follows: $\exp(-1.96*\tau^2)$; $\exp(1.96*\tau^2)$.[18]

Rankability

To estimate rankability, the following formula was used:

$$\rho = \tau^2/(\tau^2 + \text{median}\sigma^2)$$

Rankability relates the heterogeneity τ^2 from the random effect logistic regression model (differences

Table 1 Outcome indicators and their description

Indicator	Numerator	Denominator
Nosocomial pressure ulcer prevalence among hospitalised patients	Number of patients with a pressure ulcer gr. 2–4	All hospitalised patients who were examined for the presence of a pressure ulcer
Pressure ulcer incidence after total hip replacement	Number of patients with a pressure ulcer gr. 2–4	All total hip replacement patients
Bile duct leakage within 30 days after cholecystectomy	Number of patients with bile duct leakage within 30 days of cholecystectomy	All patients with a cholecystectomy
Unintended reoperation after colorectal surgery	Number of unintended reoperation after colorectal surgery	All colorectal operations excluding appendix
In hospital mortality after AMI for patients younger than 65 years	Number of patients younger than 65 years who died during hospitalisation because of AMI	All patients younger than 65 years hospitalised because of AMI
In hospital mortality after AMI for patients 65 years and older	Number of patients 65 years and older who died during hospitalisation because of AMI	All patients 65 years and older hospitalised because of AMI
Readmission after heart failure for patients younger than 75 year	Number of readmissions after heart failure within 12 weeks after hospital discharge in patients younger than 75 years	All patients younger than 75 years admitted for heart failure
Readmission after heart failure for patients 75 years and older	Number of readmissions after heart failure within 12 weeks of hospital discharge in patients 75 years and older	All patients 75 years and older admitted for heart failure
Remaining cancer tissue after breast-conserving lumpectomy	Number of patients in whom cancer tissue is left after an initial local excision of a malignant breast tumour	All patients who had local excision of a malignant breast tumour

AMI, acute myocardial infarction; gr., grade.

between the hospitals) to the SE σ^2 of the individual hospitals from the fixed effect logistic regression model. Rankability can be interpreted as the part of heterogeneity between hospitals that is due to unexplained differences, and the rest is due to natural variation or chance. Therefore, rankability describes the reliability of ranking.

Case-mix adjustment

The data on performance indicators did not include patient characteristics, except for two outcomes: AMI mortality and heart failure readmission. The original indicators were stratified by age. The indicators AMI <65 years plus ≥65 years; and heart failure <75 years plus ≥75 years were combined in two datasets and a limited age adjustment was applied by putting age group in the fixed part of the random effect model.

The statistical analysis was performed with R statistical software (version 2.7.1, R Foundation for Statistical Computing, Vienna, Austria), using the lme4 library to fit random effect logistic regression models.

RESULTS

Nine outcome indicators (table 1) were studied, of which four indicators were combined into two.

Within-hospital uncertainty

The number of cases and the total number of patients per hospital varied widely for the different indicators (table 2). For instance, pressure ulcer prevalence varied from 0 to 39 cases, while the number of patients ranged from 59 to 548. For cholecystectomy, the number of cases with bile duct leakage was very small (median 2). A considerable number of hospitals reported 0 cases (29 out of 97), resulting in a median incidence of leakage of the bile duct of 0.5%. The within-hospital uncertainty was largest for cholecystectomy (σ 1.01), and pressure ulcer incidence (σ 0.85) because of the small number of cases (table 3).

Between-hospital heterogeneity

Heterogeneity between hospitals varied from none (τ^2 0) for cholecystectomy to τ^2 0.29 for colorectal surgery. The corresponding 95% range of the ORs was 0.35 and 2.86 for colorectal surgery, meaning that hospitals at the higher end of the distribution had a 2.86 higher chance of reoperation than in the average hospital. Similarly, at the lower end of the distribution, patients had a 0.35 lower chance of reoperation. This was equivalent to an eightfold difference between the hospitals for this indicator.

Rankability

Because of the large between-hospital differences, rankability was the highest (71%) for colorectal surgery and the lowest (<50%) for the indicators pressure ulcer prevalence, pressure ulcer incidence, and cholecystectomy. For pressure ulcer the rankability was relatively low despite a σ^2 of 0.19 related to the small between-hospital differences (τ^2). Rankability was moderate (50–75%) for the indicators colorectal surgery, AMI, heart failure readmission, and breast-saving lumpectomy.

Adjustment for case mix revealed that a part of the heterogeneity in the AMI indicator was by age. For heart failure readmission, age was borderline significant.

Table 2 Descriptive statistics

Indicator	Number of hospitals	Median cases (range)	Median N (range)	Median outcome % (range)
Nosocomial pressure ulcer prevalence	93	10 (0–39)	233 (59–548)	3.7 (0–11.1)
Nosocomial pressure ulcer incidence after total hip replacement	90	2 (0–23)	197 (26–1131)	1.1 (0–8.9)
Leakage of the bile duct within 30 days of cholecystectomy	95	2 (0–7)	255 (109–625)	0.5 (0–3.63)
Unintended reoperation after colorectal surgery	94	15 (0–47)	209 (57–557)	6.9 (0–18.4)
In-hospital mortality after AMI, age <65 years	88	1 (0–17)	85.5 (4–720)	1.1 (0–6.8)
In-hospital mortality after AMI, age ≥65 years	88	10 (0–46)	117.5 (28–541)	8.6 (0–20.8)
Readmission after heart failure, age <75 years	93	6 (0–30)	77 (13–389)	7.9 (0–22.6)
Readmission after heart failure, age ≥75 years	93	10 (0–50)	133 (13–376)	8.0 (0–23.1)
Remaining cancer tissue after breast-saving lumpectomy	94	7 (1–46)	76 (14–300)	10.5 (1.2–35.7)

AMI, acute myocardial infarction.

Table 3 Rankability

Indicator	σ^2	τ^2	95% range OR −	95% range OR +	Rankability
Nosocomial pressure ulcer prevalence	0.19	0.11	0.52	1.91	37%
Nosocomial pressure ulcer incidence after total hip replacement	0.85	0.16	0.46	2.17	38%
Leakage of the bile duct within 30 days of cholecystectomy	1.01	0.00	1	1	0%
Unintended reoperation after colorectal surgery	0.12	0.29	0.35	2.86	71%
In-hospital mortality after AMI, age groups combined*	0.19	0.27	0.36	2.76	58%
Readmission after heart failure, age groups combined*	0.14	0.15	0.47	2.11	51%
Remaining cancer tissue after breast-saving lumpectomy	0.25	0.28	0.35	2.82	53%

*Results for the combined age groups are adjusted for age.
AMI, acute myocardial infarction.

Rankability for the combined indicator AMI was 58% and for heart failure 51%.

DISCUSSION

Several outcome indicators were tested to assess their reliability for ranking hospitals using the concept of rankability. Rankability indicates what extent the variation between crude hospital effects is due to true differences (as opposed to measurement error). Combining fixed effect logistic regression models and random effect logistic regression models, the uncertainty within individual hospitals and the unexplained heterogeneity between hospitals could be estimated. Considerable variability was found to be due to chance alone within hospitals. However, the unexplained differences between hospitals were small for some indicators. Both lead to low rankability.

It should be noted that ranking is a specific form of hospital comparison. Although the amount of uncertainty is an important factor in all hospital comparisons, ranking also ignores the magnitude of the differences. For example, when the random effect estimates of 10 hospitals show that they all have very similar outcomes, ranking them from 1 to 10 ignores the similarity. Therefore, reporting rankability is even more relevant for rankings.

The indicators in this research showed substantial uncertainty that influenced rankability. For cholecystectomy, there were no differences other than those by chance alone between the hospitals. Using this indicator for ranking hospitals is therefore ineffective. This adds to the criticism by de Reuver and Gouma about this indicator.[19] Substantial heterogeneity led to larger rankability in the colorectal surgery indicator (71%). Nevertheless, for this indicator it remains unclear how much of these differences are caused by case mix. It is plausible that a different indication for surgery, such as traumatic injury or colorectal cancer, may play a role in reoperation rate. Case-mix correction should be performed before using this indicator to rank hospitals on their performance. The lack of heterogeneity influences the rankability of the pressure ulcer prevalence. For AMI and heart failure, a simple stratification was performed for two age groups. Combining both age groups resulted in a larger number of cases and total numbers. While rankability of the group of patients younger than 65 was low due to the limited number of cases, the pooled data stratified for age had a moderate rankability (51%).

In order for rankability to be large, the between variance needs to dominate the within variance. Therefore measuring performance should be precise and with adequate sample size if we want to distinguish between hospitals. Rankability combines both the within variance and the between variance. If the between-variance (heterogeneity) is large, we can accept more within-variance to still be able to distinguish between hospitals.

The measurement of rankability provides a way of assessing reliability of ranking. We might compare rankability with the signal-to-noise ratio that is used for electrical signals and is defined as the power ratio between a signal (meaningful information) and the background noise (unwanted signal). So, an indicator provides a signal on quality of care, which is corrupted by random variation. The problem with ranking on crude hospital performance occurs when a rare event is chosen for the indicator, like mortality. Some hospitals have small sample sizes that make the statistics for the performance unstable and the rank order unlikely to replicate. One might also argue that ranking should be avoided. Furthermore, if for 'pay for performance' or 'quality bonus' initiatives are attempted, the signal to noise ratio should be large not to falsely accuse hospitals or individuals.

Lingsma *et al* used rankability to assess the ranking of a small number of in-vitro fertilisation (IVF) clinics.[20] They found considerable heterogeneity, while uncertainty per clinic was small because of large numbers (median 654 cycles). This resulted in a substantial rankability with only 10% of the observed differences between the clinics attributed to chance.[20] Compared with this research, rankability in our data was much lower. In the Dutch outcome indicators, not only were the total numbers of patients sometimes small (median between 90 and 277) but also the outcome was frequently low. Simple rankings based on fixed effects of hospital performance disregard both the magnitude and the uncertainty of the differences between hospitals.[21] An illustrative example is the cholecystectomy indicator, where the number of cases was too low to detect any differences between hospitals.

Small samples and low event rates limit the statistical power of the comparison between hospitals.[22]

This raises questions about minimal power calculations or combining indicators to provide sufficient sample size to decrease measurement error. Classical power calculation or estimating minimal cases and total numbers might be performed using Cohen's d, where d is defined as the difference between two means divided by the SD. Effect sizes are commonly defined as small, d=0.2, medium, d=0.5, and large, d=0.8. A variant of Cohen's d may be used for event rate. The population size for d=0.5 then is at least 200, and at least 800 for d=0.2 for indicators with sufficient event rates.[23] These numbers can be used as 'a rule of thumb' to assess the reliability of ranking hospitals. Actual calculations of required sample sizes for random effect models are much more complex and theoretical work on this topic is needed. Looking at the sample sizes for the pressure ulcer indicator (59−548) in the Dutch hospitals, it is questionable if this indicator will ever be suitable for ranking hospitals. The maximal sample size is limited by the number of beds in a hospital. In case of inadequate numbers, presentation of the results for a specific indicator could be done using funnel plots because this would show the differences between hospitals in relation to random variation.[24] In addition, crude random effect estimates including a measure for rankability might be informative for stakeholders who are able to interpret them, for example, hospitals or the government. Realistic presentation is important to avoid gaming and truly encourage actions to improve the quality of care.[25]

A categorisation for rankability is still arbitrary. Lingsma *et al* suggested that over 70% rankability should be fair to rank hospitals.[20] Higgins *et al* assigned adjectives of low, moderate and high to the I^2 values of 25%, 50% and 75%.[26] I^2 is used to measure heterogeneity in meta-analyses[27] and is similar in nature to our rankability measure. I^2 can be interpreted as the percentage of the total variability in a set of effect sizes due to heterogeneity, that is, between-study variability. Adopting this categorisation, the authors found that none of the outcome indicators had a high rankability. It could be argued that in case of moderate rankability, 'expected ranks' should be used that take into account random variability.[13−15] This requires statistical knowledge and access to advanced statistical programs. No ranking attempt should be made when rankability is low. It would also be interesting to identify subsets of hospitals that meet or exceed a standard, fall below a standard, and a subset that cannot be classified due to sample size limitations. The random effect estimates with CIs shows if a hospital significantly differs from the mean beyond statistical uncertainty. Therefore, random effect estimates can be used to identify subsets, and funnel plots used as a graphical display of these subsets.

Reliability of ratings depends on sample size and heterogeneity, but also on biases. A conceptual framework can be drawn to summarise the elements of between-hospital differences (figure 1).[20] The observed differences can be divided into unexplained differences and chance. By using random effect models, chance can be corrected for, leaving patients characteristics, registration bias, quality of care and residual confounding as elements of the unexplained differences. Consequently, ranking reflects the total of unexplained differences between hospitals and not true differences in the quality of care. This is a limitation of this study, but the

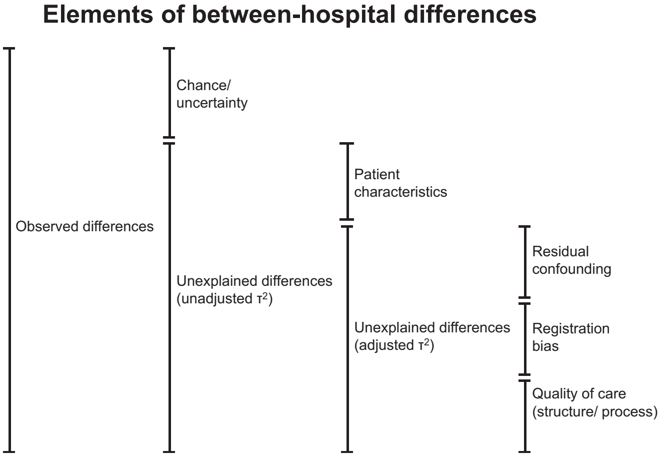
Elements of between-hospital differences
Observed differences
Chance/ uncertainty
Unexplained differences (unadjusted τ²)
Patient characteristics
Unexplained differences (adjusted τ²)
Residual confounding
Registration bias
Quality of care (structure/ process)

Figure 1 Conceptual framework of between-hospital differences. Observed differences can be divided into random variation and unexplained differences, which can be further attributed to patient characteristics that were not adjusted for, residual confounding because of imperfect case-mix correction, registration bias. Differences in quality of care remain the explanation for the final part of between-hospital differences.

publicly reported data do not provide any additional information.

The authors conclude that none of the currently used Dutch outcome indicators are suitable for ranking hospitals. When judging hospital quality the influence of random variation must be accounted for to avoid over-interpretation of the numbers in the quest for more transparency in healthcare. Adequate sample size is a prerequisite when attempting reliable ranking.

Funding Internal Erasmus MC grant for healthcare research (Mrace).

Competing interests None to declare.

Contributors JPM and EWS had the original idea for the study, EWS developed the study design. AMvD gathered and analysed the data with HL. AMvD wrote a first draft of the paper. All authors contributed to further drafts.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Jencks SF, Cuerdon T, Burwen DR, *et al.* Quality of medical care delivered to Medicare beneficiaries: a profile at state and national levels. *JAMA* 2000;284:1670−6.
2. Berg M, Meijerink Y, Gras M, *et al.* Feasibility first: developing public performance indicators on patient safety and clinical effectiveness for Dutch hospitals. *Health Policy* 2005;75:59−73.
3. Halasyamani LK, Davis MM. Conflicting measures of hospital quality: ratings from 'Hospital Compare' versus 'Best Hospitals'. *J Hosp Med* 2007;2:128−34.
4. Lemmers O, Kremer JA, Borm GF. Incorporating natural variation into IVF clinic league tables. *Hum Reprod* 2007;22:1359−62.
5. Lilford R, Mohammed MA, Spiegelhalter D, *et al.* Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;363:1147−54.
6. Mohammed MA, Mant J, Bentham L, *et al.* Comparing processes of stroke care in high- and low-mortality hospitals in the West Midlands, UK. *Int J Qual Health Care* 2005;17:31−6.
7. Ranstam J, Wagner P, Robertsson O, *et al.* Health-care quality registers: outcome-orientated ranking of hospitals is unreliable. *J Bone Joint Surg Br* 2008;90:1558−61.
8. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? *Med Care* 2005;43:1177−84.
9. Spiegelhalter D. Ranking institutions. *J Thorac Cardiovasc Surg* 2003;125:1171−3; author reply 3.
10. Anderson J, Hackman M, Burnich J, *et al.* Determining hospital performance based on rank ordering: is it appropriate? *Am J Med Qual* 2007;22:177−85.
11. Adab P, Rouse AM, Mohammed MA, *et al.* Performance league tables: the NHS deserves better. *BMJ* 2002;324:95−8.
12. Diehr P, Cain K, Connell F, *et al.* What is too much variation? The null hypothesis in small-area analysis. *Health Serv Res* 1990;24:741−71.
13. Robertsson O, Ranstam J, Lidgren L. Variation in outcome and ranking of hospitals: an analysis from the Swedish knee arthroplasty register. *Acta Orthop* 2006;77:487−93.
14. Davidson G, Moscovice I, Remus D. Hospital size, uncertainty, and pay-for-performance. *Health Care Financ Rev* 2007;29:45−57.
15. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital profiling. *Stat Sci* 2007;22:206−26.
16. Houwelingen van JC, Brand R, Louis TA. *Empirical Bayes methods for monitoring health care quality*, 2005. http://www.msbi.nl/dnn/People/Houwelingen/Publications/tabid/158/Default.aspx.
17. Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Health Care* 2005;14:347−51.
18. Spiegelhalter D, Abrahams KR, Myles JP. *Bayesian approaches to clinical trials and health care evaluation*, John Wiley & Sons Ltd; Chichester: 2004.
19. de Reuver PR, Gouma DJ. [Bile leakage. A performance indicator with markedly different consequences for the patient, specialist and care insurer]. (In Dutch). *Ned Tijdschr Geneeskd* 2007;151:1709−12.
20. Lingsma HF, Eijkemans MJ, Steyerberg EW. Incorporating natural variation into IVF clinic league tables: the Expected Rank. *BMC Med Res Methodol* 2009;9:53.
21. Lingsma HF, Dippel DW, Hoeks SE, *et al.* Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry* 2008;79:888−94.
22. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA* 2004;292:847−51.
23. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd edn. Philadelphia: Lawrence Erlbaum Associates, 1988.
24. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24:1185−202.
25. Gibberd R, Hancock S, Howley P, *et al.* Using indicators to quantify the potential to improve the quality of health care. *Int J Qual Health Care* 2004;16(Suppl 1):i37−43.
26. Higgins JP, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557−60.
27. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, *et al.* Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods* 2006;11:193−206.