American Journal of Epidemiology
Vol. 169, No. 2
DOI: 10.1093/aje/kwn340
Advance Access publication December 8, 2008

Practice of Epidemiology

More Than Numbers: The Power of Graphs in Meta-Analysis

Leon Bax, Noriaki Ikeda, Naohito Fukui, Yukari Yaju, Harukazu Tsuruta, and Karel G. M. Moons

Initially submitted March 17, 2008; accepted for publication July 25, 2008.

In meta-analysis, the assessment of graphs is widely used in an attempt to identify or rule out heterogeneity and publication bias. A variety of graphs are available for this purpose. To date, however, there has been no comparative evaluation of the performance of these graphs. With the objective of assessing the reproducibility and validity of graph ratings, the authors simulated 100 meta-analyses from 4 scenarios that covered situations with and without heterogeneity and publication bias. From each meta-analysis, the authors produced 11 types of graphs (box plot, weighted box plot, standardized residual histogram, normal quantile plot, forest plot, 3 kinds of funnel plots, trim-and-fill plot, Galbraith plot, and L'Abbé plot), and 3 reviewers assessed the resulting 1,100 plots. The intraclass correlation coefficients (ICCs) for reproducibility of the graph ratings ranged from poor (ICC = 0.34) to high (ICC = 0.91). Ratings of the forest plot and the standardized residual histogram were best associated with parameter heterogeneity. Association between graph ratings and publication bias (censorship of studies) was poor. Meta-analysts should be selective in the graphs they choose for the exploration of their data.

epidemiologic methods; evaluation studies; meta-analysis; publication bias; review

Abbreviation: ICC, intraclass correlation coefficient.

Over the past few decades, systematic reviews—with meta-analyses as their quantitative and analytical core—have become a cornerstone of evidence-based medicine. As such, meta-analyses play a central role in the development of clinical guidelines and in clinical decision-making. The majority of meta-analytical endeavors involve the use of graphs, mostly forest or funnel plots, to decide which analytical approach serves the synthesis of the study data best, and notably to explore heterogeneity and possible publication bias (1–4). Heterogeneity and publication bias are perhaps the 2 major threats to the validity of a meta-analysis.

Heterogeneity refers to the variation among effect parameters across studies. In meta-analyses, one commonly tests whether the assumption of a single underlying effect (fixed-effect analysis) or similar effects (random-effects analysis) is indeed reasonable. Approaches based on the assumption of a fixed effect can be used if the between-study variation is not excessive. Approaches based on the assumption of random effects are used to incorporate the heterogeneity if any is present. In the latter case, sources of heterogeneity can be investigated and incorporated by means of stratified meta-analyses or meta-regression (5). Publication bias can be viewed as a systematic error in a meta-analysis that occurs because not all of the evidence is properly represented. Although there is both over- and underrepresentation of evidence, underrepresentation is more prominent and often refers to ''negative'' evidence—that is, evidence that either is not statistically significant or conflicts with the prevailing beliefs about the association under investigation. Underrepresentation occurs when researchers do not submit their study results to a journal or submit only part of their results (selective reporting), journals decide not to publish certain studies (selective publication), or the study retrieval and selection procedures of a meta-analysis do not include a publication (selective inclusion) (6). The bias caused by these phenomena is often referred to as *publication bias*, although this is literally a ''pars pro toto'' and the term *dissemination bias* may be more accurate (7).

Graphical assessments of heterogeneity and publication bias are numerous (as we will explain below). Some types of

Correspondence to Dr. Leon Bax, Kitasato Clinical Research Center, Kitasato University, 1-15-1 Kitasato, Sagamihara, 228-8555 Kanagawa, Japan (e-mail: leonbax@kitasato-crc.org).

plots, such as the funnel plot and L'Abbé plot, have been evaluated individually to some extent (1, 2, 8–17). However, little is known about the relative performance of the majority of graphs in terms of reproducibility (interrater agreement) and the validity of judgments regarding evidence of heterogeneity and publication bias. Our objective was to conduct a comprehensive evaluation of the reproducibility and validity of such judgments with simulated data sets (based on empirical data) with varying heterogeneity and publication bias.

MATERIALS AND METHODS

Frequently used graphs in meta-analysis

Figure 1 shows the most commonly used types of graphs in meta-analyses. Each graph is constructed from the same meta-analytical data set—that is, the meta-analysis by Colditz et al. (18) on the efficacy of Bacillus Calmette-Guérin vaccine in preventing tuberculosis.

The *forest plot* (Figure 1, part A) is the most common graph in meta-analysis reports. It shows the estimate (often a risk ratio or odds ratio) and confidence interval for each study and, commonly at the bottom, the corresponding summary estimate. The forest plot dates back to at least the 1970s (19) and was used for the first time in a meta-analysis context in 1982 (20). In early use of the plot, studies and their confidence intervals were indicated by bars and studies with the largest variance (and thus the largest confidence intervals and bars) were most prominent, even though they were least important. This problem was solved by displaying the confidence intervals with thin lines and marking a study's effect estimate with a square that was proportional to the study's weight (and inversely proportional to the variance). The summary or meta-analysis estimate evolved into a diamond, with its center at the summary estimate and its outer edges at the confidence limits. Forest plots may be useful for showing how the effect estimates from individual studies accumulate to a meta-analysis result. It has also been suggested that the plots provide a visual representation of the amount of variation between study estimates (2, 21) and that one can ''eyeball'' the overlap of the confidence intervals of the effect estimates to judge the presence of between-study heterogeneity (22).

The *funnel plot* was introduced by Light and Pillemer in 1985 (4). It typically has a measure for effect size on the *x*-axis and a measure related to the within-study variance (e.g., inverse standard error) on the *y*-axis. Each study is represented by a single equal-sized dot (Figure 1, part B). Under most circumstances, there are relatively more small studies (with larger variance) than big studies in a meta-analytical data set, and the smaller studies have estimates that are more scattered and further removed from the summary estimate. This creates a funnel-like distribution of the dots in the plot that, without publication bias, is assumed to be symmetrical. If studies are relatively more ''missing'' on one side of the plot—commonly studies with low sample size—the missingness and subsequent asymmetry can be attributed to publication bias. Because funnel plots are constructed in many ways (with different measures on the axes) and asymmetry can be caused by more than just publication bias (23), it has been the subject of substantial methodological scrutiny. Tang and Liu (15) assessed various funnel plots from published meta-analyses and concluded that the choice of axis measure could alter conclusions on the presence of publication bias in most studies. Sterne and Egger (14) evaluated the performance of funnel plots for binary event meta-analyses and concluded that the (inverse) standard error is the preferred measure for the *y*-axis and a log-transformed ratio measure of effect size (e.g., log risk ratio or log odds ratio) is the preferred measure for the *x*-axis, as shown in Figure 1, part B. Finally, interpretations of funnel plots may be different for different readers (10, 16).

The *trim-and-fill plot* (Figure 1, part C) is also a funnel plot, although it not merely involves a representation of raw data but also shows the results of a kind of imputation algorithm called ''trim and fill'' (24, 25). The algorithm assesses the symmetry of the plot analytically and can impute new studies to plots in which studies appear to be missing. The asymmetry assessment is performed via a rank correlation, after which the studies causing the asymmetry are trimmed on 1 side of the plot. This shifts the meta-analysis estimate, possibly causing asymmetry again. If reestimation shows residual asymmetry, the process of trimming and reestimation is repeated and usually runs for 3–4 cycles. Once no asymmetry is left, the trimmed studies are put back and their counterparts on the other side of the last symmetry axis are imputed. This is followed by a meta-analysis that includes the imputed studies. The plot itself looks like a regular funnel plot, but it contains additional dots (usually not filled), representing the imputed studies, and an additional vertical line that indicates the summary effect when these studies are included in the meta-analysis. The number of imputed studies, particularly the difference between the original summary estimate and the summary estimate after imputation, are assumed to be indicative of publication bias. Although the algorithm's performance decreases when heterogeneity increases (25–27), in the presence of publication bias meta-analyses with imputations from the trim-and-fill algorithm are less biased than meta-analyses without imputations (26).

The *Galbraith plot* (Figure 1, part D) is designed to assess the extent of heterogeneity between studies in a meta-analysis (1). The *y*-axis shows the (log-transformed) effect size divided by its standard error (*z* score) and the inverse of the standard error on the *x*-axis. Each study is represented by a single dot, and a regression line runs centrally through the plot. Parallel to the regression line, at a 2-standard-deviation distance, 2 lines create an interval in which most dots would be expected to fall if the studies were estimating a single fixed parameter. Galbraith originally proposed putting the *y*-axis on the right side and making it radial so that the graph would look like a speedometer (the so-called radial plot). This is not necessary for interpretation of the graph, and only a few meta-analysis programs have implemented this feature (28).

The *L'Abbé plot* (Figure 1, part E), introduced in 1987 (3), is only applicable to meta-analyses of studies with binary outcomes. It plots the risks (or odds) in the exposed or index group (*y*-axis) against those of the control group

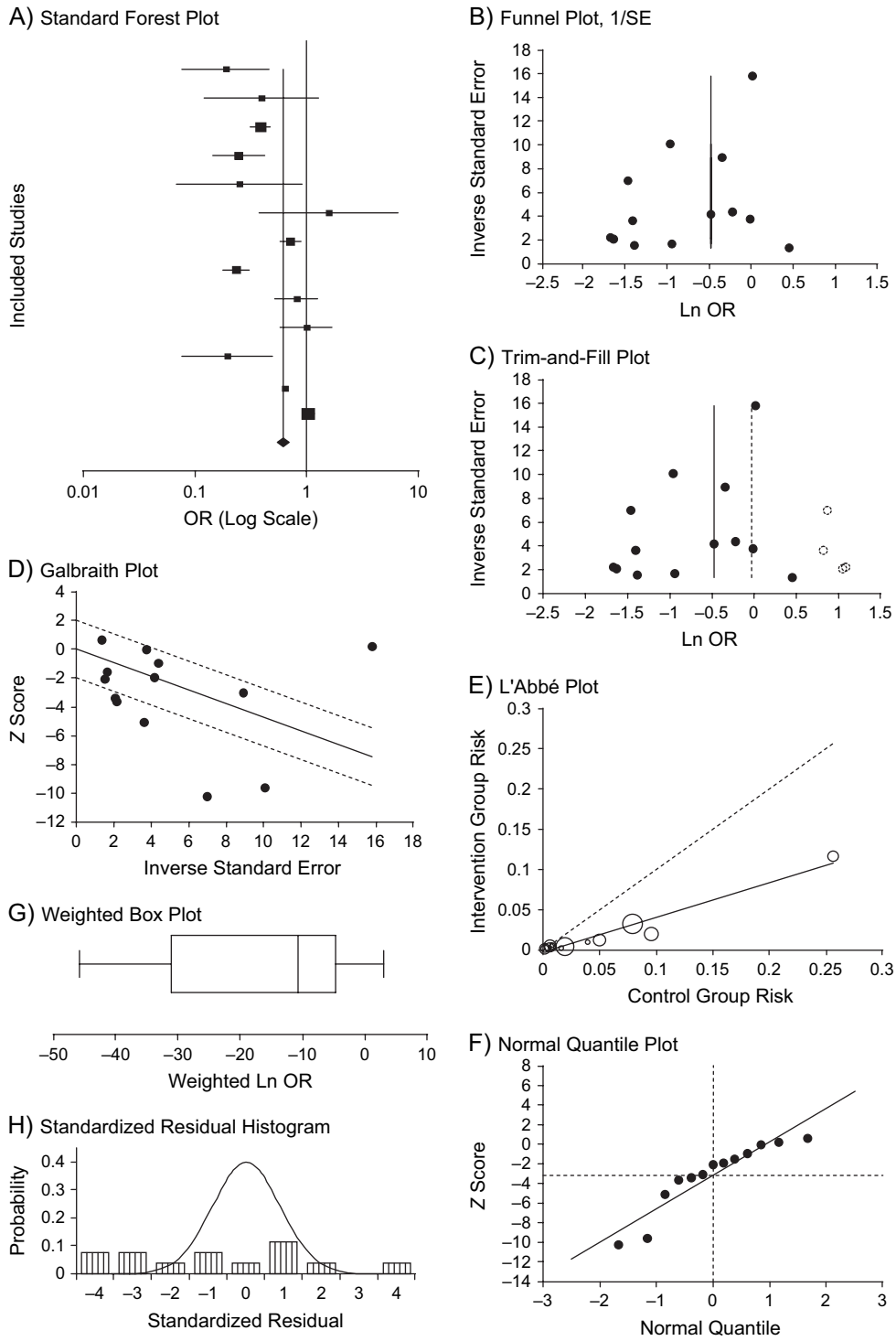

Figure 1. Graphs corresponding to 8 types of plots available in meta-analysis software for the assessment of heterogeneity and publication bias in meta-analyses. Two variations of the funnel plot and 1 variation of the box plot were included in the assessments but are not shown here. SE, standard error.

(*x*-axis) and often contains a regression line and a central diagonal line indicating identical risks in each group. The sizes of the dots are proportional to the study weights. With a great deal of between-study heterogeneity, there may be substantial spread around the regression line, although it has been reported that naive use of the distance between the

regression line and the dots as an indicator of heterogeneity may be misleading (11). Multiple clustering of control group risks along the *x*-axis indicates violation of the assumption that a single underlying baseline risk exists for all studies and thus also indicates heterogeneity.

In statistics, *normal quantile plots* are often used to assess data normality. The normal quantile plot for meta-analysis, as recommended by Wang and Bushman (17), has each individual study's *z* score on the *y*-axis and the normalized quantile of its rank on the *x*-axis (Figure 1, part F). The plot can be used in meta-analysis to check the normality of the data (dots expected on a straight line), to investigate heterogeneity (clustering of dots), and to assess the presence of publication bias (deviation of the tails from the regression line) (17).

The *box plot*, like the normal quantile plot, is also frequently used in statistics (29). Application of the box plot to meta-analytical data sets may nevertheless be suboptimal because it does not take into account the weight of the studies. Consequently, the center (median or mean) does not correspond to the meta-analysis summary estimate and is therefore misleading. There are 2 adjusted box plots that do not have these disadvantages: 1) a standard box plot to which the meta-analysis summary measure has been added (not shown) and 2) a box plot of weighted estimates (Figure 1, part G) (30).

The *standardized residual histogram* (Figure 1, part H) is briefly explained in a meta-analysis context by Greenland (31) and is further described by Sutton et al. (32). The histogram plots the fractions of categorized standardized residuals (the individual estimate minus the summary estimate, divided by the standard error of the individual estimate) in vertical bars. An overlay of a normal distribution can then be used to assess heterogeneity and departures from normality.

Other plots worth mentioning are the Egger regression asymmetry plot (9), the exclusion sensitivity plot (30), and the Baujat plot (8). The Egger plot is essentially a Galbraith plot with the intercept of the regression line unconstrained, and therefore it was not added to the lineup of plots included in our assessments. The other 2 plots assess the influence of (excluding) a study on the total analysis; although they are useful for checking which studies are causing most of the heterogeneity and the largest shifts in summary estimates, they are not meant to quantify the heterogeneity itself and are also not further addressed here.

Reproducibility and validity of meta-analytical graphs

Assessment of graphs is inherently a subjective task and thus suffers from reproducibility issues. Reproducibility refers to the extent that a procedure, measurement, or judgment is replicated either over time by the same persons or equipment or by different persons or under different circumstances. Here, reproducibility of the judgment of meta-analytical graphs refers to different judgments by different observers or raters (interobserver or interrater agreement). Validity, in this context, refers to how well a graph measures what it is supposed to measure. It reflects both how well the judgment of a graph is guided by changes in heterogeneity and publication bias and the ability of the raters to properly interpret what is shown in the graph.

Simulation study

We designed a simulation study in which we created 100 meta-analysis data sets for each of 4 scenarios with different heterogeneity and publication bias (Table 1). The simulation parameters were based on those used in previous meta-analytical research (26, 33, 34) and on data from the 50 most recent (July 2007) English-language meta-analyses of clinical studies reporting odds ratios in PubMed (National Library of Medicine). Publication bias was induced by censoring studies with a commonly used exponential selection function (26, 34, 35) that gives studies with high *P* values a higher chance to be censored. We used the MIX meta-analysis software (30, 36), incorporating a Visual Basic for Applications version of the Mersenne Twister random number generator (37), to produce the data sets with randomly varying characteristics described in Table 1. Scenarios 1 and 2 produced data sets without publication bias, and scenarios 3 and 4 produced data sets with publication bias. Data sets from scenarios 1 and 3 exhibited little or no statistical between-study variation (heterogeneity), and scenarios 2 and 4 produced data sets with substantial heterogeneity. For each simulated data set, we created the 8 graphs displayed in Figure 1 plus 2 additional funnel plots (with the standard error or sample size on the *y*-axis) and an additional box plot, described above. The 11 graphs were created for each of the 100 simulated data sets, resulting in 1,100 graphs that needed to be assessed. Numerical analyses, such as Mantel-Haenszel meta-analyses of odds ratios, were performed simultaneously.

Graphical assessments

A special Visual Basic for Applications program (available upon request) was written in which all graphs were presented to 3 of the 6 authors (L. B., N. F., Y. Y.) in random order and in a blinded fashion, meaning that they were kept unaware of the source data (meta-analytical studies) of the graphs. These researchers, with considerable experience in meta-analysis, rated the heterogeneity and publication bias shown by the graphs from ''none'' to ''extensive'' with a continuous rating instrument (a scrollbar sliding from 0 to 100) inside the program. The ratings were performed over a period of 3 weeks.

Data analysis

We had 2 primary outcomes of interest in our study: the interrater reproducibility and the validity of the graphical assessments in judging the presence of heterogeneity and publication bias. Reproducibility was evaluated by means of intraclass correlation coefficients (ICCs) in SPSS 14.0 (38). The ICC can be viewed as a measure of correlation, consistency, or conformity for a data set when it has multiple groups (39), and we used a 2-way random-effects ICC for consistency of individual measurements (39). ICCs range from 0 to 1, typically with the following

Table 1. Simulation Parameters Used in an Evaluation of the Reproducibility and Validity of Graphs for Assessing Heterogeneity and Publication Bias in Meta-Analyses

Parameter	Distribution of Parameter	Scenario 1 (Heterogeneity Absent/ Publication Bias Absent)	Scenario 2 (Heterogeneity Present/ Publication Bias Absent)	Scenario 3 (Heterogeneity Absent/ Publication Bias Present)	Scenario 4 (Heterogeneity Present/ Publication Bias Present)
Baseline risk	Uniform (minimum, maximum)	Minimum = 0.2	Minimum = 0.2	Minimum = 0.2	Minimum = 0.2
		Maximum = 0.4	Maximum = 0.4	Maximum = 0.4	Maximum = 0.4
Odds ratio	Lognormal (μ, τ)	$\exp(\mu) = 0.85$	$\exp(\mu) = 0.85$	$\exp(\mu) = 0.85$	$\exp(\mu) = 0.85$
		$\tau = 0$	$\tau = 0.45$	$\tau = 0$	$\tau = 0.45$
No. of studies	Uniform (minimum, maximum)	Minimum = 5	Minimum = 5	Minimum = 5	Minimum = 5
		Maximum = 25	Maximum = 25	Maximum = 25	Maximum = 25
Study arm size	Uniform (minimum, maximum)	Minimum = 25	Minimum = 25	Minimum = 25	Minimum = 25
		Maximum = 500	Maximum = 500	Maximum = 500	Maximum = 500
Study selection	Bernoulli (*ps*) $ps = \exp(-1b \times p_i^a)$	$a = 0$	$a = 0$	$a = 3$	$a = 3$
		$b = 0$	$b = 0$	$b = 4$	$b = 4$

classification: ICC $<$ 0.75 = poor agreement; ICC 0.75–0.90 = moderate agreement; and ICC $>$ 0.90 = high agreement (40).

The validity of the graphical judgments of heterogeneity and publication bias was evaluated by the association between the simulation parameter settings and the graph scores from the raters in regression analyses. The dependent variables were the presence or absence of between-study variability (heterogeneity) and the counts of censored studies (publication bias). The average score of the 3 raters was used as an independent (predictor) variable, and the logistic and Poisson regression analyses were performed in R (41).

RESULTS

Reproducibility

The overall ICC (across all raters and all 1,100 graphs) was 0.58 (95% confidence interval: 0.54, 0.62) for assessments of heterogeneity and 0.65 (95% confidence interval: 0.62, 0.69) for assessments of publication bias. With regard to heterogeneity, the forest plot and the standardized residual histogram had the highest reproducibility (ICCs of 0.87 and 0.69, respectively). The weighted box plot ratings were the least reproducible. For judgment of publication bias, the trim-and-fill plot and the weighted box plot had the best reproducibility (ICCs of 0.91 and 0.71) and the standard error funnel plot and the normal quantile plot had the lowest ICCs. Details are provided in Table 2.

Validity of graphs

For assessment of heterogeneity, the scores for the forest plot, standardized residual histogram, Galbraith plot, and L'Abbé plot showed significant associations with the presence of heterogeneity (Table 3). When their ICCs were also taken into consideration, the standardized residual histogram and the forest plot appeared to be the best candidates if multiple graphs were to be used. For the assessment of publication bias, validity was low in general (Table 4). None of the funnel plots correlated well with the underlying parameters of publication bias.

DISCUSSION

We examined the interrater reproducibility and validity of 11 types of graphs that are frequently used in assessments of heterogeneity and publication bias in meta-analysis. One hundred data sets with varying heterogeneity and publication bias were simulated, and the resulting 1,100 graphs were judged in random order by 3 raters on the degree of heterogeneity and/or publication bias. The reproducibility of heterogeneity assessments was highest for the forest plot and the standardized residual histogram, and the ratings of these graphs were also well associated with simulated actual

Table 2. Reproducibility of Assessments of Heterogeneity and Publication Bias With Graphs in Meta-Analyses

Graph	Heterogeneity		Publication Bias	
	ICC[a]	CI	ICC[a]	CI
Box plot	0.61	0.506, 0.703	0.59	0.488, 0.689
Weighted box plot	0.34	0.211, 0.462	0.71	0.622, 0.783
Standardized residual histogram	0.69	0.600, 0.768		
Normal quantile plot	0.47	0.353, 0.585	0.50	0.387, 0.613
Forest plot	0.87	0.823, 0.905		
Standard error funnel plot			0.51	0.397, 0.621
Inverse standard error funnel plot			0.53	0.412, 0.632
Sample size funnel plot			0.58	0.473, 0.679
Galbraith plot	0.63	0.527, 0.718		
L'Abbé plot	0.55	0.436, 0.651		
Trim-and-fill plot			0.91	0.874, 0.934
Overall	0.58	0.544, 0.621	0.65	0.617, 0.686

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient.

[a] Two-way, random-effects ICC.

Table 3. Validity: Associations Between Graph Scores and the Heterogeneity Simulation Parameters in Logistic Regression Analysis

Graph	Odds Ratio[a]	95% Confidence Interval	*P* Value
Box plot	1.01	0.992, 1.027	0.29
Weighted box plot	0.99	0.960, 1.009	0.22
Standardized residual histogram	1.12	1.076, 1.170	<0.001
Normal quantile plot	1.00	0.980, 1.023	0.9
Forest plot	1.10	1.065, 1.137	<0.001
Galbraith plot	1.10	1.062, 1.135	<0.001
L'Abbé plot	1.05	1.023, 1.073	<0.001

[a] Exponentiated coefficients from logistic regression.

heterogeneity. For assessment of publication bias, the ratings of the trim-and-fill plot and the weighted box plot had the best reproducibility properties. Association between graph ratings and publication bias (censorship of studies) was poor in general.

The issue of poor reproducibility of graph assessments in meta-analyses has been raised by other authors (10, 15, 16), particularly for the funnel plot. However, to our knowledge, it had not been formally investigated or quantified. Our results underline the need to use multiple raters and produce composite or consensus scores. We acknowledge that our results are likely to be sensitive to the experience and training of the raters and to some extent the number of raters. The reproducibility and validity of graph assessments will probably decrease when the reviewers or meta-analysts are less experienced and perhaps increase when they are all experts. We consequently decided to use raters that had experience and a background in meta-analysis that is common for authors of systematic reviews (1 experienced rater (L. B.) and 2 raters with moderate experience in meta-analysis (N. F., Y. Y.)). We used 3 raters because this is a common number for systematic review teams. Both aspects should make our findings generalizable to the average practice of meta-analysis.

We found that the forest plot and the standardized residual histogram (plots with the best reproducibility) indicated the presence or absence of heterogeneity well. Although eyeballing of the forest plot is traditionally not recommend for investigating heterogeneity (5), we found in an additional explorative analysis (data available upon request) that ratings of the forest plot as well as the histogram correlated very strongly with the results of the I^2 test (42, 43), which is often used to quantify heterogeneity in meta-analyses.

To simulate publication bias, we used a common selection approach based on an exponential function of the *P* value (26, 34, 35). This assumes that censoring is related to (absence of) significance, whereas the funnel plots assume censorship based on (absence of) extremeness of small study estimates. These premises may in many situations assume censoring of different studies, which could explain the relatively poor performance of these plots in our validity evaluation. Although our approach is common and likely to be close to reality, future simulation studies might explore alternative censorship mechanisms, including possible differences in these mechanisms for experimental and observational studies.

To put the analytics of this paper into perspective, we stress that prospective registries of studies are essential in order to gain knowledge on why certain studies are published and others are not. Tracking of study reporting via these registries of protocols is the only way to prevent or properly correct for publication biases. Time will tell whether this is a utopian vision or a future reality.

There is a well-known expression that says ''A picture is worth a thousand words.'' We would like to add that, in meta-analysis, a picture may be worth more than a million numbers. Interpretation of graphs requires care, however, because reproducibility and validity depend heavily on the type of graph and the construct it is meant to visualize. Meta-analysts should be selective in the graphs they choose for the exploration of their data.

Table 4. Validity: Associations Between Graph Scores and the Publication Bias Simulation Parameters in Poisson Regression Analysis

Graph	Rate Ratio[a]	95% Confidence Interval	*P* Value
Box plot	1.00	0.993, 1.007	0.96
Weighted box plot	0.99	0.988, 1.001	0.10
Normal quantile plot	1.00	0.991, 1.006	0.74
Standard error funnel plot	1.00	0.994, 1.009	0.69
Inverse standard error funnel plot	1.00	0.994, 1.008	0.78
Sample size funnel plot	1.00	0.994, 1.009	0.71
Trim-and-fill plot	1.00	0.995, 1.006	0.83

[a] Exponentiated coefficients from Poisson regression.

ACKNOWLEDGMENTS

Author affiliations: Kitasato Clinical Research Center, Kitasato University, Sagamihara, Japan (Leon Bax); Department of Medical Informatics, Kitasato University, Sagamihara, Japan (Leon Bax, Noriaki Ikeda, Harukazu Tsuruta); Japan Clinical Research Support Unit, Tokyo, Japan (Naohito Fukui); Department of Epidemiology and Biostatistics, School of Health Sciences and Nursing, University of Tokyo, Tokyo, Japan (Yukari Yaju); and Julius Center for Health Sciences and Primary Care, Universitaire Medisch Centrum Utrecht, Utrecht, The Netherlands (Leon Bax, Karel G. M. Moons).

This study was supported by research grant 3084 from the Graduate School of Medical Sciences of Kitasato University. Kitasato University played no role in any aspect of the study.

Conflict of interest: none declared.

REFERENCES

1. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med*. 1988;7(8):889–894.
2. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ*. 2001;322(7300):1479–1480.
3. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med.* 1987;107(2):224–233.
4. Light RJ, Pillemer DB. *Summing Up*. Cambridge, MA: Harvard University Press; 1985.
5. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*. 1994;309(6965):1351–1355.
6. Rothstein H, Sutton AJ, Borenstein M. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, United Kingdom: John Wiley & Sons Ltd; 2005.
7. Song F, Eastwood AJ, Gilbody S, et al. Publication and related biases. *Health Technol Assess*. 2000;4(10):1–115.
8. Baujat B, Mahé C, Pignon JP, et al. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med.* 2002;21(18):2641–2652.
9. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629–634.
10. Lau J, Ioannidis JP, Terrin N, et al. The case of the misleading funnel plot. *BMJ*. 2006;333(7568):597–600.
11. Song F. Exploring heterogeneity in meta-analysis: is the L'Abbé plot useful? *J Clin Epidemiol.* 1999;52(8):725–730.
12. Song F, Khan KS, Dinnes J, et al. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol.* 2002;31(1):88–95.
13. Souza JP, Pileggi C, Cecatti JG. Assessment of funnel plot asymmetry and publication bias in reproductive health meta-analyses: an analytic survey [electronic article]. *Reprod Health*. 2007;4:3.
14. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol*. 2001;54(10):1046–1055.
15. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol*. 2000;53(5):477–484.
16. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol*. 2005;58(9):894–901.
17. Wang MC, Bushman BJ. Using the normal quantile plot to explore meta-analytic data sets. *Psychol Methods*. 1998;3(1):46–54.
18. Colditz GA, Brewer TF, Berkey CS, et al. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA*. 1994;271(9):698–702.
19. Freiman JA, Chalmers TC, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 ''negative'' trials. *N Engl J Med*. 1978;299(13):690–694.
20. Lewis JA, Ellis SH. A statistical appraisal of post-infarction beta-blocker trials. *Primary Cardiol*. 1982(suppl 1):31–37.
21. Moja L, Moschetti I, Liberati A, et al. Understanding systematic reviews: the meta-analysis graph (also called ''forest plot''). *Intern Emerg Med*. 2007;2(2):140–142.
22. Ried K. Interpreting and understanding meta-analysis graphs—a practical guide. *Aust Fam Physician*. 2006;35(8):635–638.
23. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol*. 2000;53(11):1119–1129.
24. Duval S, Tweedie R. A non-parametric ''trim and fill'' method of accounting for publication bias in meta-analysis. *J Am Stat Assoc.* 2000;95:89–98.
25. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*. 2000;56(2):455–463.
26. Peters JL, Sutton AJ, Jones DR, et al. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Stat Med*. 2007;26(25):4544–4562.
27. Sterne JA. High false-positive rate for trim and fill method [letter]. *BMJ*. 2000;320:1574–1577. (http://www.bmj.com/cgi/eletters/320/7249/1574#EL1).
28. Bax L, Yu LM, Ikeda N, et al. A systematic comparison of software dedicated to meta-analysis of causal studies [electronic article]. *BMC Med Res Methodol*. 2007;7:40.
29. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
30. Bax L, Yu LM, Ikeda N, et al. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data [electronic article]. *BMC Med Res Methodol*. 2006;6:50.
31. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987;9:1–30.
32. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for Meta-Analysis in Medical Research*. Chichester, United Kingdom: John Wiley & Sons Ltd; 2000.
33. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006;25(20):3443–3457.
34. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med*. 2001;20(4):641–654.
35. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994;50(4):1088–1101.
36. Bax L, Yu LM, Ikeda N, et al. *MIX: Comprehensive Free Software for Meta-Analysis of Causal Research Data. Version 1.7*. Sagamihara, Japan: Kitasato Clinical Research Center; 2008. (http://mix-for-meta-analysis.info). (Accessed May 20, 2008).
37. Matsumoto M, Nishimura T. *Mersenne Twister Home Page: A Very Fast Random Number Generator of Period 2^{19937}-1*. Hiroshima, Japan: Department of Mathematics, Hiroshima University; 2002. (http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html). (Accessed February 12, 2008).
38. SPSS, Inc. *SPSS for Windows*. Chicago, IL: SPSS, Inc; 2005.
39. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30–46.
40. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 1999.
41. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2007.
42. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539–1558.
43. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560.