In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias

Norma Terrin*, Christopher H. Schmid, Joseph Lau

Institute for Clinical Research and Health Policy Studies, Tufts–New England Medical Center, 750 Washington Street, Box 63, Boston, MA 02111, USA

Accepted 30 January 2005

Abstract

Background and Objective: Publication bias and related biases can lead to overly optimistic conclusions in systematic reviews. The funnel plot, which is frequently used to detect such biases, has not yet been subjected to empirical evaluation as a visual tool. We sought to determine whether researchers can correctly identify publication bias from visual inspection of funnel plots in typical-size systematic reviews.

Methods: A questionnaire with funnel plots containing 10 studies each (the median number in medical meta-analyses) was completed by 41 medical researchers, including clinical research fellows in a meta-analysis class, faculty in clinical care research, and experienced systematic reviewers.

Results: On average, participants correctly identified 52.5% (95% CI 50.6–54.4%) of the plots as being affected or unaffected by publication bias. The weighted mean percent correct, which adjusted for the fact that asymmetric plots are more likely to occur in the presence of publication bias, was also low (48.3 to 62.8%, depending on the presence or absence of publication bias and heterogeneous study effects).

Conclusion: Researchers who assess for publication bias using the funnel plot may be misled by its shape. Authors and readers of systematic reviews need to be aware of the limitations of the funnel plot.

Keywords: Publication bias; Funnel plot; Systematic review; Meta-analysis; Heterogeneity

1. Introduction

Health care providers and policy makers rely on systematic reviews, including meta-analyses, for important decisions concerning clinical care. Yet the conclusions of a review may be invalid if the available studies form a biased sample. Bias occurs because statistically significant research results are more widely disseminated, and therefore easier to retrieve. Unpublished work is the most difficult to access, but selective reporting, frequency of citation, duplicate publication, and language of publication also influence accessibility [1–6].

Researchers are becoming more aware of the potential for publication and related biases. Many have turned to the funnel plot, a simple graphic technique for detecting these biases [7]. The funnel plot is a scatter plot of the component studies in a review or meta-analysis (quantitative summary), with the treatment effect on the horizontal axis, and a weight, such as the inverse standard error, or sample size, on the vertical axis. Larger weights correspond to more precise estimates of treatment effect. A common interpretation is that a symmetric, inverted funnel shape implies no publication bias, but if the funnel appears to be missing points in the lower corner of the plot associated with ineffectiveness of treatment, there is potential bias (Fig. 1). The funnel shape occurs because the larger, more precise studies tend to be closer to the true effect, whereas the smaller ones are more variable. The simplicity of this method is appealing, and the subjectivity of visual assessment for asymmetry can be overcome by statistically testing for a relation between precision and effect using either regression [8,9] or rank correlation [10]. However, funnel plots and related statistical methods are problematic because there are several possible causes of funnel plot asymmetry other than publication bias. These include heterogeneity, chance, choice of effect measure, and choice of precision measure [11–14].

Heterogeneity occurs when results vary from study to study because of differences in study protocol, study quality, illness severity, and patient characteristics. When multiple treatment effects are estimated, there is no reason to expect a funnel shape (Fig. 2).

Although funnel plots are often published or reported for heterogeneous studies [15–27], the impact of heterogeneity...
on the shape of the funnel plot is frequently overlooked. For example, in a systematic review of intra-articular hyaluronic acid in treatment of knee osteoarthritis [26], the authors interpreted the funnel plot and regression test as supporting the presence of publication bias (Fig. 3). They recognized molecular weight as a source of heterogeneity, and even plotted the highest molecular weight studies with a different symbol, but they did not take heterogeneity into consideration when assessing for publication bias. Without the high molecular studies, the plot appears more symmetric.

A funnel plot with a small number of studies may appear asymmetric simply due to chance. More than 50% of meta-analyses have 10 studies or fewer [28,29], and it is not uncommon for researchers to publish interpretations of funnel plots with 10 or fewer points [17,18,30–38].

Inappropriate applications of the funnel plot are not surprising in light of expert advice to assess for publication bias by examining funnel plots for asymmetry [10,39]. Caveats [11–14,28,39] have generally gone unheeded. One commentary in a subspecialty journal ignored all ambiguity, advising that funnel plots “always” be used, and that a relation between treatment effect and sample size is “indicative of publication bias” [40].

We found previously [14] that a quantitative method of filling in the sparse corner of the funnel plot [41] spuriously adjusted for nonexistent bias if the studies were heterogeneous, or if there were only 10 studies per meta-analysis. We conjectured that visual inspection would also be inadequate for separating the effects of publication bias, heterogeneity, and chance. Many published reports display and visually interpret the funnel plot. The study reported here used a questionnaire to determine whether researchers can correctly identify publication bias from funnel plots, when they are shown both symmetric and asymmetric plots with and without publication bias.

2. Methods

2.1. The participants

A questionnaire on funnel plot interpretation was completed in the spring and summer of 2001 by a convenience sample with three groups and a total of 41 participants.
Seventeen were taking a meta-analysis course at Tufts University School of Medicine. These were mostly fellows in the Clinical Research Graduate Program of the Sackler School of Graduate Biomedical Sciences. The questionnaires were completed close to the end of the course, but before a class on publication bias. Three more research fellows, all with some knowledge of meta-analysis, were added to this group. Eleven faculty members of the Division of Clinical Care Research at Tufts–New England Medical Center participated. This was a diverse group, with some having systematic review experience, but the majority having none. The group included both clinical and health services researchers. Ten clinical researchers outside Tufts University were recruited. These individuals had all either published or used systematic reviews. All had some knowledge of the funnel plot, although none had published methodologic research on publication bias. The three groups will be referred to as “fellows,” “faculty,” and “reviewers.” The Tufts-New England Medical Center Institutional Review Board certified the exempt status of the research.

2.2. The questionnaire

The questionnaire contained an introduction to the funnel plot quoted from a standard source [42]:

Publication bias occurs when studies reporting statistically significant results are published and studies reporting less significant results are not … A funnel plot is a scatterplot of sample size [or other measure of precision] versus estimated effect size for a group of studies. [Effect size is a measure of how effective the intervention is.] Since small studies will typically show more variability among the effect sizes than larger studies, and there will typically be fewer of the latter, the plot should look like a funnel, hence its name. When there is publication bias … a bite will be taken out of a part of [the funnel] plot.

This quote was followed by two examples of funnel plots from the literature, with brief explanations ([7] p. 68–69 and [43]). Then 22 funnel plots were shown (Fig. 4), with instructions:

On the following pages are 22 examples of funnel plots from both simulated and actual meta-analyses. Because the outcome is dichotomous (disease or no disease) the effect size is the odds ratio. The lower the odds ratio, the more effective the treatment, so the missing studies (if any are missing) would be expected on the right, where the odds ratio is near 1.0 or larger. The measure of precision on the vertical axis is the inverse of the standard error (1/se). The larger 1/se, the greater the weight of the study in the meta-analysis.
For each of the following plots, please place an “X” in front of the response that best fits your interpretation of whether there is evidence for publication bias: “yes,” “no,” or “maybe.” Feel free to make comments at the end.

The meta-analyses in the questionnaire were selected from simulated meta-analyses previously generated for a related project [14]. Each had 10 studies, a typical number [28,29]. Study sample size was generated randomly, on the log scale from 50 to 500 (median size, \( n \approx 158 \)). Homogeneous and heterogeneous meta-analyses were generated by fixed and random effects models, respectively. Parameters of the models were chosen based on actual data [29]. To create publication bias, we allowed studies with lower \( P \)-value or larger sample size to have greater likelihood of inclusion. Four levels of asymmetry were represented. The characteristics of 16 of the plots are in Table 1. In addition, four plots of published meta-analyses were included, to see whether participants would respond to them differently. The questionnaire also included two extremely asymmetric, computer-generated plots without publication bias. To avoid penalizing the funnel plot, we excluded from our analyses these examples of a rare type of meta-analysis (see Appendix for details).

2.3. The analyses

We analyzed the 16 plots in Table 1. The percentage of correctly identified plots was calculated for each participant. “Maybe” answers counted as 50% correct. Also, the percentage of correctly identified plots, excluding the plots answered “maybe,” was calculated for each participant. “Yes,” “no,” and “maybe” answers were tabulated at each level of asymmetry. The 16 plots were compared with the plots of the published meta-analyses on percentage of “yes,” “no,” and “maybe” answers. To avoid penalizing the funnel plot, we excluded from our analyses these examples of a rare type of meta-analysis (see Appendix for details).

3. Results

The mean percentage of plots correctly identified was 52.5% (95% CI, 50.6–54.4%). The means for fellows, faculty, and reviewers (51.5, 53.4, and 53.4%, respectively) were not significantly different. The mean percentage correct, excluding the plots answered “maybe,” was 54.5% (95% CI, 50.9–58.0%) (Fig. 5). Two faculty members are not represented in Fig. 5 because they answered “maybe” to all 16 plots. The variability in the scores decreased as the number answered “yes” or “no” increased. The higher scores were not necessarily indicative of ability, but may have occurred simply due to chance.

The percentages of “yes” responses for plots with low, moderate, high, and very high levels of asymmetry were 15, 47, 42, and 57%, respectively (Fig. 6). Plots with only moderate asymmetry that were bias free (#6 and #9, Fig. 4) got a “yes” response 45% of the time, indicating that even moderate asymmetry can be misleading. Plots with low asymmetry that were affected by bias (#11 and #21), got a “no” response 49% of the time, indicating that lack of asymmetry can also be misleading.

The percentages of “yes,” “no,” and “maybe” responses for computer generated plots were 40, 23, and 36%, respectively, comparable to the frequencies for published data (36, 31, and 33%, respectively).

The plots included one at each level of asymmetry for each of the four heterogeneity-bias categories (Table 1). In the real world, frequency of asymmetry varies, depending on presence of publication bias and heterogeneity. In particular, asymmetric plots are more likely to occur in the presence of publication bias. Therefore, we calculated the weighted mean percent correct in each of the four heterogeneity-bias categories. The weight for each plot was the frequency of its level of asymmetry in a simulation study [14]. Plots in the questionnaire that would be unusual in the real world received less weight, whereas more typical plots received more weight. The plots in the simulation study were generated with the same parameters for size, odds ratio, and prevalence as the questionnaire plots. We also used the simulation study to calculate Type I error rates and power for regression [8] and rank correlation tests [10] and to analyze the amount of bias by asymmetry category.

![Fig. 5. The score was the percentage of correctly identified plots among those judged (answered “yes” or “no”). “Fellows” were clinical research fellows and students in a meta-analysis class. “Faculty” were clinical and health services researchers, and “Reviewers” were clinical researchers with systematic review experience. The horizontal line is at the score 50.](image)
Table 2 shows the frequencies of funnel plot asymmetry that were found in the simulation study for the four different heterogeneity-bias groups. Using the entries in Table 2 as weights for the corresponding questionnaire plots listed in Table 1, we calculated the weighted mean percent correct in each of the four heterogeneity-bias categories (Table 3).

Type I error rates for regression and rank correlation were within 0.02 of the nominal $\alpha = 0.05$, but both had very low power, less than 30%.

When there was publication bias, the degree of asymmetry in the funnel plot had only a slight relation to the degree of bias in the point estimate. The relation existed only when there was homogeneity. The bias in the point estimate was 0.061, 0.068, 0.070, and 0.072 for homogeneous meta-analyses with low, moderate, high, and very high asymmetry, respectively. For heterogeneous meta-analyses, the bias was 0.102, 0.091, 0.113, and 0.105.

4. Discussion

Medical researchers who participated in the survey tended to interpret funnel plot asymmetry in plots of 10 studies as evidence of publication bias. On average, they identified about 50% of the plots correctly, because at each level of asymmetry, the same number of plots with and without publication bias was included. Although it is not surprising that participants could not distinguish between plots of similar shape with and without publication bias, the results are important because many authors of systematic reviews continue to assess for publication bias with the funnel plot.

To evaluate the implications of these results, it is critical to understand how frequently funnel plot asymmetry actually occurs with and without publication bias. We used simulations from a previous project to provide the necessary context [14]. In funnel plots with 10 studies, we found that moderate and high levels of visual asymmetry were common, even in the absence of publication bias, and that many plots with publication bias had approximately symmetric appearance (Table 2). Furthermore, the degree of asymmetry had only a slight relation to the degree of bias, when there was publication bias. Thus, researchers may frequently be misled by the shape of the plot, interpreting asymmetry as publication bias when none exists, and symmetry as absence of publication bias, when actually it is present. The weighted mean percent correct, which adjusted for the fact that asymmetric plots are more likely to occur in the presence of publication bias, was low (Table 3).

Participants appeared to respond similarly to the computer-generated and published data, answering “yes,” “no,” and “maybe” with comparable frequency. The large number of “maybe” responses made by a few of the participants perhaps indicates recognition that the plots contained inadequate information for deciding whether there was evidence of publication bias. One participant showed awareness that chance and heterogeneity may have played a role, commenting “The problem is that getting a 6:4 or even a 7:3 distribution isn’t all that unlikely in a simulation with only 10 observations,” and “There was a problem when there was an outlier, because I had to decide whether to ignore it or not.” None of the other participants addressed the issues of chance or heterogeneity, though a few wrote comments.

A convenience sample was adequate for this study. Each of the three groups of participants did barely any better, on average, than they would have by flipping a coin. Experienced reviewers performed similarly to researchers with little or no meta-analysis experience. Given this result, it would be unlikely that a random sample of researchers would find the funnel plot useful for discriminating among different

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of funnel plot asymmetry within each of the four heterogeneity-bias categories, shown as percent of 1,000 simulated meta-analyses</td>
</tr>
<tr>
<td><strong>Asymmetry</strong></td>
</tr>
<tr>
<td><strong>Heterogeneity</strong></td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>

The meta-analyses had the same parameters for size, odds ratio, and prevalence as the 16 plots in Table 1.
causes of asymmetry. Before the questionnaire was administered, it was tested on the two coauthors who did not participate in the selection of plots. They scored 44 and 47, suggesting that meta-analysis experts may be no better able than other researchers to determine when asymmetry is due to bias.

The instructions to the study participants were intended to represent the degree of knowledge generally displayed by authors of reviews, and were taken from a standard meta-analysis book [42]. Therefore, the various factors other than bias that can affect funnel plot shape were not described. However, even those with greater awareness of the subtleties, including the participant who commented on chance and heterogeneity, the experienced reviewers, and the authors of this article, were unable to distinguish plots with bias from those with none, suggesting that further instruction to the participants would not have helped.

The implication of this research is that, for 10 studies or fewer (the size of more than 50% of medical meta-analyses), the funnel plot is inconclusive as a tool for detecting publication bias. For the type of data used in the questionnaire (moderate trial size and prevalence, and no underlying relation between sample size and treatment benefit), the funnel plot tests [8,10] had low Type I error rates that were much more acceptable than Type I error rates for visual inspection. However, the tests had very low power. Therefore, for this type of data, the tests may help reviewers avoid incorrectly inferring bias when it is absent, but they frequently cannot pick up existing bias.

The tests cannot prevent incorrect inference of bias when there is an underlying relation between sample size and treatment effect. For example, trials of low-sodium diets that achieve greater sodium reduction also achieve greater blood pressure change [22,44]. If the most extreme sodium-reduction diets and the most severely hypertensive subjects were more likely to be studied in small trials, then small trials would be more likely to achieve greater sodium reduction, and hence, greater blood pressure changes. The underlying relation between sample size and blood pressure change would cause asymmetry that might be picked up by funnel plot tests, even in the absence of bias.

Any assessment for publication bias in the presence of heterogeneity requires caution. Investigation of the sources of heterogeneity may help explain asymmetry in the funnel plot. In some cases, it may be possible to remove sources of heterogeneity before assessing for publication bias, through meta-regression, control-rate regression, stratification, or selection modeling with covariates [44–49]. However, it is frequently not possible to identify a source for the existing heterogeneity [50], or to remove all of the heterogeneity even when a source can be identified, and it is unclear how useful these methods are for meta-analyses with 10 or fewer studies. Some authors refer to tests for funnel plot asymmetry as tests for “small study effects,” described as “biases which result in an association between treatment effects and sample size” [28]. However, heterogeneity among treatment effects can cause funnel plot asymmetry without creating bias of any kind.

Funnel plot tests should be used only to check for asymmetry, and only when they have adequate power. They should not be referred to as tests for publication bias.

Our results add to a growing body of research indicating that the funnel plot and methods based on it are problematic. A previous study [14] found that a quantitative method of filling in the sparse corner of the plot [41] spuriously imputed studies. Other studies have found that funnel plot shape is sensitive to the choice of effect measure and precision measure [11], that variation in quality can affect the shape, with smaller, lower quality studies showing greater benefit of treatment [13], and that the funnel plot tests have high Type I error rates [9,12] and low power [9,28]. A number of alternative methods, none without drawbacks, are discussed elsewhere [51,52].

This study was limited by the number of factors we could feasibly vary in the questionnaire plots. We did not want the questionnaire to be prohibitively long. All meta-analyses had 10 studies and true mean odds ratio = 0.50.

In conclusion, the funnel plot and statistical tests based on it have a number of problems. The specific contribution of the study reported here is that researchers may be misled by the shape of the plot, interpreting asymmetry as publication bias when none exists, and symmetry as absence of publication bias, when actually it is present.

Acknowledgments

This work was supported by the Agency for Healthcare Research and Quality (AHRQ), grant number R01 HS10254.

Appendix

Models for generating the data

The numbers of events were drawn from binomial distributions. Homogeneous meta-analyses assumed true odds ratio (OR) 0.5 and true control group prevalence 0.15. Heterogeneous meta-analyses assumed normally distributed true log OR, mean = −0.768, variance = 0.15. Consequently, mean true OR = 0.50. True control group prevalence for heterogeneous meta-analyses was normally distributed, mean = 0.15, variance = 0.005.

Generating publication bias

The chance of selection in $P$-value intervals 0–0.01, 0.01–0.05, 0.05–0.06, 0.06–0.10, 0.10–0.50, 0.50–1.0 was 1.0, 0.9, 0.6, 0.5, 0.2, and 0, respectively. Studies not included after selection based on $P$-value had additional chance to be selected that was equal to sample size divided by 1,000.
Categories of asymmetry

Low: five or fewer points to the left of the most precise study, out of 10. Moderate: six or seven to the left. High: eight or nine to the left and nonsignificant (α = 0.05) regression test [8]. Very high: eight or nine to the left and significant regression test.

Study sample size

Sample size for the 16 plots was drawn uniformly on log 50 to log 500, then exponentiated.

The excluded plots

Four funnel plots of actual meta-analyses were included in the questionnaire, one at each level of asymmetry (#2, #5, #10, #15, Fig. 4). It is not known whether they were affected by publication bias. Plots #14 and #18 (Fig. 4) were created by giving each study 80% power to detect the true effect. These plots are extremely asymmetric, because a large number of subjects are required to achieve statistical significance when the true odds ratio is near 1.

References