11. PUBLICATION BIAS

11.1. Overview

Publication bias refers to the concern that studies which report relatively large effects are more likely to be published than studies which report smaller effects.

More generally, the term publication bias is sometimes called reporting bias or dissemination bias, and refers to the following chain of events. As compared with studies that are *not* statistically significant, those that *are* statistically significant are more likely to be published at all, to be published sooner, to be published more than once, to be published in journals with higher profiles, to be cited, and/or to be published in English. Additionally, if a study includes multiple outcomes, the published papers are more likely to report and/or highlight the outcomes that showed statistically significant results (J. Sterne, Egger, & Moher, 2008).

Studies that are statistically significant tend to be ones that report larger effect sizes, and so the studies that are promoted at each link in the chain tend to be the ones that exaggerate the size of the effect. The elements in this chain tend to build on each other, and as a group can severely impact what we see in a systematic review (Carroll, 2018; Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Dwan et al., 2014; J. P. A. Ioannidis & T. A. Trikalinos, 2007; Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; J. A. Sterne et al., 2011).

11.1.1. Example | Second-hand smoke and lung cancer

While there are many techniques that are employed to address publication bias, almost all of them are based on a common set of assumptions, which I will illustrate using an analysis that assessed the relationship between exposure to second-hand smoking and lung cancer (Hackshaw et al., 1997). The analysis included 37 studies, which all followed the same design. In each study, researchers recruited couples where neither partner smoked, and designated one partner as the subject. They also recruited couples where one partner smoked and designated the non-smoker as the subject. They then computed the risk ratio for the subject in the second group vs. the first.

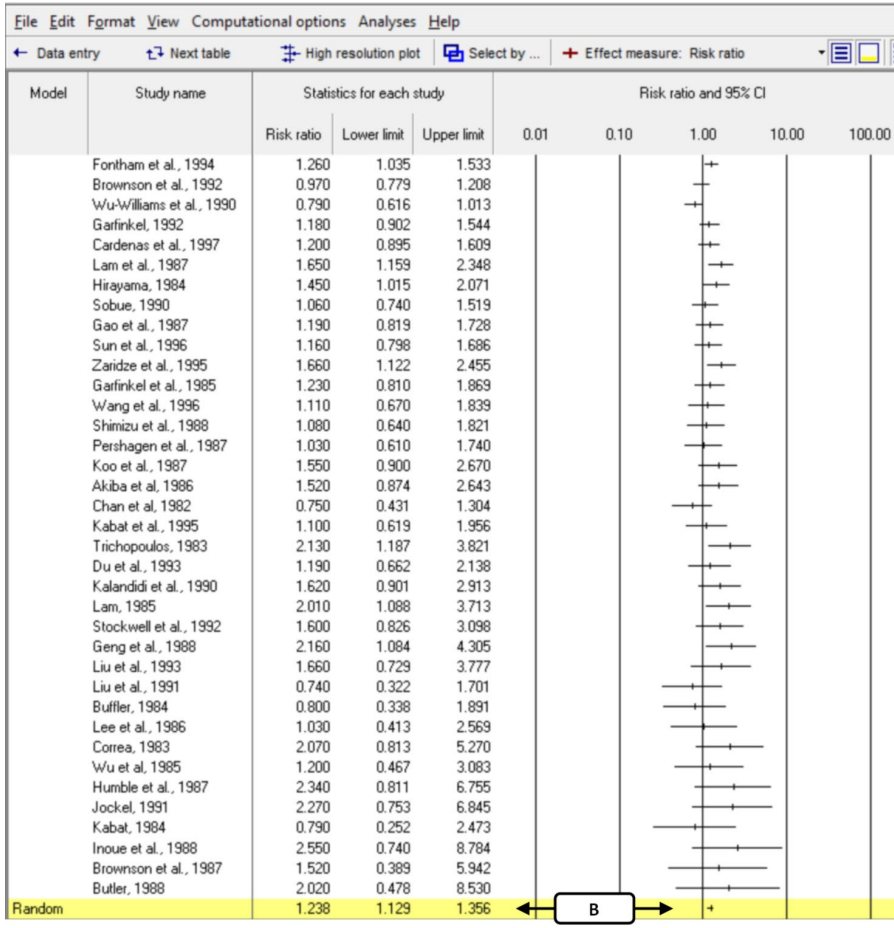

Figure 60 | Forest plot on ratio scale | Risk ratio > 1 indicates increased risk

The results are shown in Figure 60. The summary risk ratio [B] is 1.238, which indicates that non-smokers living with a smoker were 24% more likely to develop lung cancer, as compared with those living with a non-smoker. The confidence interval is 1.129 to 1.356, which tells us that the mean risk ratio probably falls somewhere in this range. The *Z*-value for a test of the null hypothesis is 4.526, with a corresponding *p*-value of < 0.001. We can reject the null hypothesis, and conclude that the mean risk ratio in the universe of comparable studies is greater than 1.0.

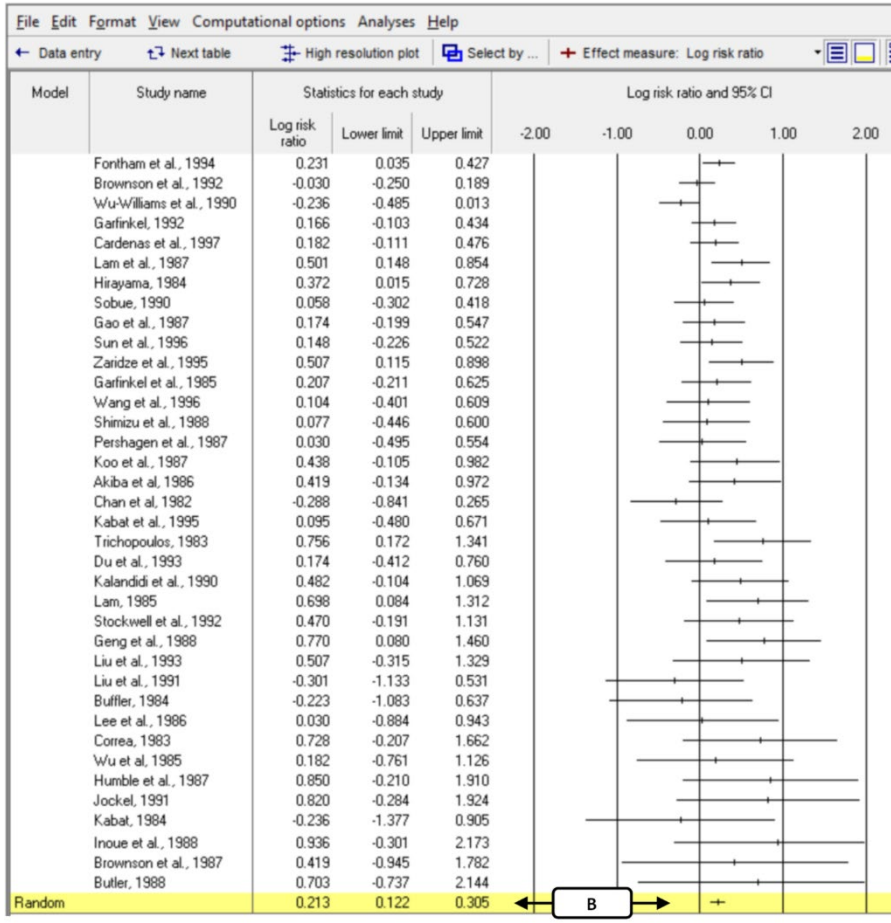

Figure 61 | Forest plot on log scale | Log risk ratio > 0 indicates increased risk

The procedure to assess publication bias employs a log scale rather than a risk ratio scale. For that reason, I also display Figure 61, which shows the same data on a log scale. In log units, the mean risk ratio [B] is 0.213 with a confidence interval of 0.122 to 0.305. The other statistics (the *Z*-value and the *p*-value) remain the same.

It is clear that in this set of studies second-hand smoke was associated with a 24% increased risk of lung cancer. However, it is not clear that this set of studies is representative of all studies that were actually performed. Specifically, there is reason to believe that these 37 studies may be a biased subset of all actual studies.

The reason for this possibility can be explained with reference to Figure 62. This plot shows the same 37 studies as the prior figure, but in a different

format. The X-axis corresponds to the log risk ratio, as in Figure 61. The Y-axis corresponds to the standard error of each study. In general, this means that large studies (small standard error) appear toward the top while small studies (large standard error) appear toward the bottom. For example, Fontham et al. (1994) was one of the largest studies in the analysis. It has a standard error of 0.100 and appears near the top [A]. By contrast, Butler (1988) was one of the smallest studies in the analysis. It has a standard error of 0.735 and appears near the bottom [B].

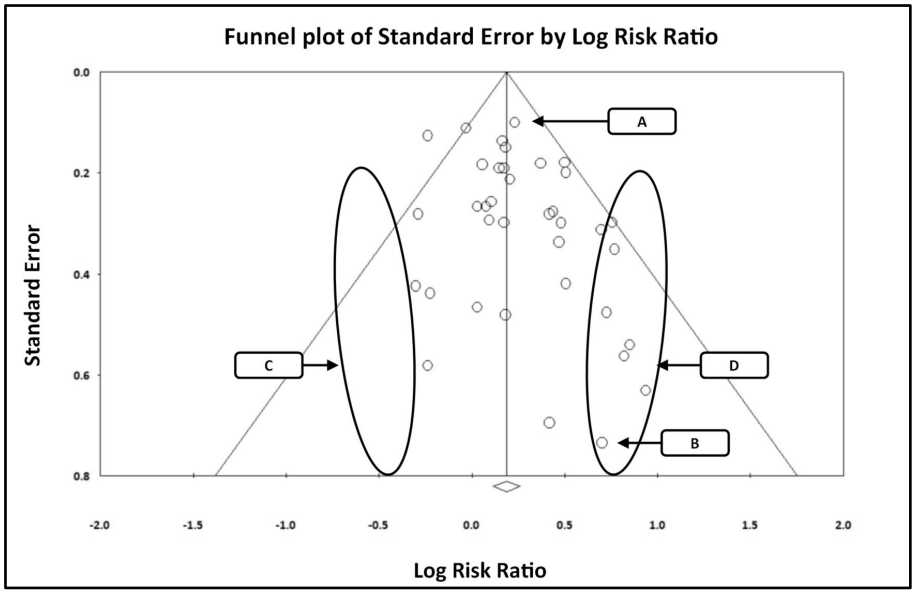

Figure 62 | Second-hand smoking and lung cancer |Observed studies only

Consider what this plot would look like if the mean effect size is actually a risk ratio of 1.238 (log value of 0.213) and we had included all studies that were performed. We have drawn a line at the mean, and statistical theory tells us that roughly 50% of all effects should fall to the left of this line, and roughly 50% should fall to the right of this line.

On the other hand, consider what the plot would look like if we were missing some studies due to publication bias. Specifically, consider what would happen if studies that are statistically significant were more likely to be published, while studies that were not statistically significant were less likely to be published.

Toward the top of the plot (where the studies are large) most studies will be statistically significant based on their sample size, even if the risk ratio

falls to the left of the mean. Additionally, large studies tend to be published even if they are not statistically significant, since a relatively large cadre of people will have responsibility for the project and a vested interest in seeing it published (Egger & Smith, 1995). Therefore, toward the top of the plot we would expect that almost all studies would be published, and they would appear in equal numbers on either side of the line.

By contrast, toward the bottom of the plot (where the sample size for each study is relatively small) studies with effects to the left of the mean might not be statistically significant. Indeed, studies *at* the mean or slightly to the right of the mean might not be statistically significant. Rather, only studies with effect sizes toward the extreme right on the X-axis will be statistically significant. Therefore, toward the bottom of the plot, we would expect to see more studies toward the right, and relatively few toward the left.

Put another way, it is plausible that the studies which were *actually conducted* were equally distributed on both sides of the mean, corresponding to areas [C] and [D] in Figure 62. That is, studies were conducted that fall into [C], but many of these studies were not published. In that case, the fact that the area labeled [C] has substantially fewer studies than the one labeled [D] could be due to publication bias.

While there are several mechanisms for addressing publication bias, most are based on this same core idea – that there will be a relationship between the size (or precision) of the study and the size of the effect. Concretely, as the sample size gets smaller, the mean effect size will get larger. Equivalently, as we move from top to bottom on the plot, the effects will shift toward the right (J. A. Sterne et al., 2011).

One method for assessing publication bias was proposed by Begg and Mazumdar (1994). They suggested that we compute the rank correlation between precision and effect size. A statistically significant correlation tells us that the mean effect size is larger in the smaller studies. In this example Kendall's tau=0.143 and p=0.107. Note that Kendall's tau is no relation to the standard deviation of true effects, also called tau.

A similar procedure was proposed by Egger, Davey Smith, Schneider, and Minder (1997) . This procedure is similar to the Begg and Mazumdar approach in that it looks for a relationship between precision and effect size. However, rather than using a rank correlation it uses a regression, which tends to have better statistical power. A statistically significant intercept in the regression tells us that the mean effect size is larger in the smaller studies. In this example the intercept is 0.892 and p=0.012.

The problem with both procedures is that they may alert us to the possibility of publication bias, but do not tell us what to do with that

information. (The Egger procedure can estimate the extent of bias but is rarely used to do so.)

One approach that does address this issue is the Trim and Fill procedure (Duval & Tweedie, 2000), which is displayed in Figure 63. This procedure looks for asymmetry in the plot. In this example, there seem to be more small studies on the right, and relatively few on the left. The procedure "Trims" the plot by removing the studies [D] that are responsible for the asymmetry. It creates a mirror image for each of these studies, and then "Fills" the plot by re-inserting each of the studies that had been removed [D], along with its imputed counterpart [C]. The idea is that we have "located" the missing studies and included them in the analysis. It then computes the mean and variance for the "full" set of studies. These new values are taken to be the values that we would have seen if all studies had been included in the analysis.

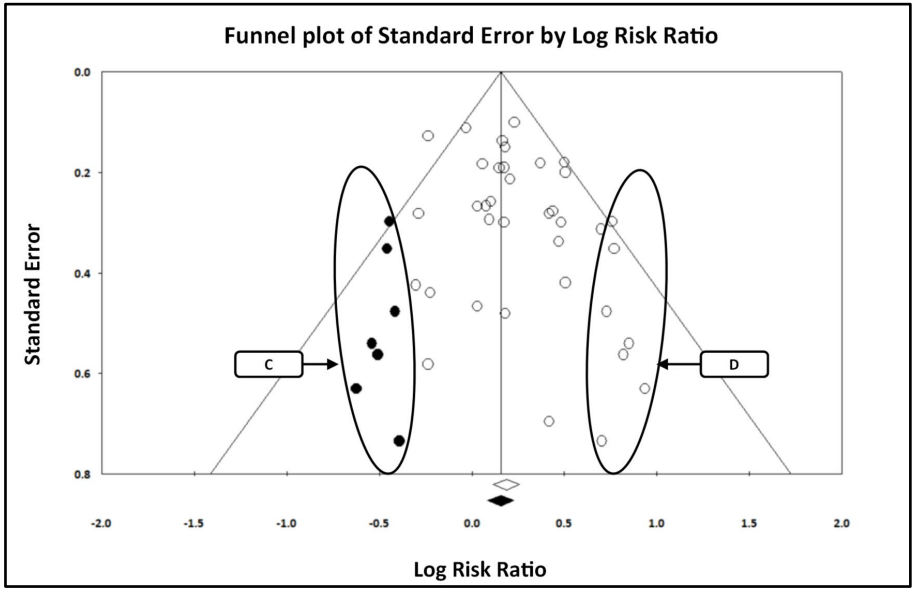

Figure 63 | Observed studies, and studies imputed by Trim and Fill

In the current analysis, the observed risk ratio was 1.238, indicating that second-hand smoke was associated with a 24% increase in the risk of lung cancer. The adjusted risk ratio was 1.189, indicating that the increased risk (after adjusting for bias) was roughly 19%. One could argue that in this context, the difference between the two is not of substantive import. That is, most people who would be concerned about a 24% increase in risk would also be concerned about a 19% increase in risk. On that basis we could conclude

that while publication bias may have shifted the effect size upward, the impact of the shift was not clinically important, and the basic conclusion (that second-hand smoke increases the risk of lung cancer) remains unchanged (Borenstein, Hedges, Higgins, & Rothstein, 2009; Givens, Smith, & Tweedie, 1997; Takagi, Sekino, Kato, Matsuno, & Umemoto, 2006).

On the pages that follow I address various issues including the following

- Researchers sometimes conflate publication bias with a small-study effect.
- Researchers sometimes over-interpret the results of the tests for publication bias.
- Researchers sometimes apply tests for bias indiscriminately.

11.2. Conflating bias with the small-study effect

11.2.1. Mistake

When a test for publication bias is statistically significant or (in the case of Trim and Fill) indicates that studies are missing, the researcher typically concludes that this is evidence of publication bias. The reality is more complicated.

11.2.2. Details

The idea that we can identify publication bias by looking for a larger effect size in smaller studies works well in cases where the fixed-effect model is called for – that is, when all studies are estimating a common parameter. In that case, when the effect size is larger in smaller studies, the possible reasons are –

A. Random sampling error
B. Publication bias

Here, if the test for asymmetry is statistically significant, we can rule out (A), and so the correlation between sample size and effect size is probably due to publication bias (B).

The situation becomes more complicated when the true effect size varies from study to study. In this case, if the effect size is larger in smaller studies, the possible reasons are –

A. Random sampling error
B. Publication bias
C. The effect size really is larger in the smaller studies

Here, if the test for asymmetry is statistically significant, we can rule out (A), but we cannot distinguish between (B) and (C). The fact that the effect size is larger in smaller studies *could be* due to publication bias. However, it is also possible that the effect size actually *is larger* in smaller studies, for reasons having nothing to do with bias. There are any number of reasons why the effect size could be larger in smaller studies. Consider the following examples.

(1) Suppose that a new intervention is being studied. The initial trials are small, enroll patients who are very ill, and show large benefits from the treatment. Later trials are larger, enroll patients who are only moderately ill, and show more modest benefits. The effect size actually is larger in the smaller studies because the patients in these studies have more room to improve than those in the larger studies (P. P. Glasziou & Irwig, 1995; Smith & Egger, 1994; Stuck, Siu, Wieland, Adams, & Rubenstein, 1993).
(2) Suppose that a new intervention is being studied. The initial trials are small, and run by people who ensure that the patients take the medication as prescribed. Later trials are large, and run by staff who are not able to track the patients as carefully. The effect size is larger in the smaller studies because the treatment in these studies is applied more consistently (Stuck, Rubenstein, & Wieland, 1998).
(3) Suppose that an intervention is tested in a series of studies which vary in size. Large studies tend to be run by professionals who employ methods to minimize the risk of bias. Smaller studies are run by researchers with less experience, and methodological flaws in these studies (for example patients with a better prognosis being pushed into the treatment group) yield larger effects (Egger, Juni, Bartlett, Holenstein, & Sterne, 2003; Ioannidis, 2008b; Linde et al., 1999; Terrin, Schmid, Lau, & Olkin, 2003; Wood et al., 2008).
(4) Suppose that a meta-analysis includes studies which employed unique variants of an intervention. Those which employed weaker variants included large sample sizes to yield adequate power, while those which employed stronger variants included smaller sample sizes to yield the same power. The effect size will be larger in the smaller studies because these are the studies with the more effective variants of the treatment (Linde et al., 1997; Terrin et al., 2003).

In these examples, what we are calling a "small-study effect" is simply a special case of heterogeneity.

The various tests outlined earlier (rank correlation, regression, Trim and Fill) can be used to rule out (A), but they are not able to distinguish between (B) and (C). For this reason, when we do find evidence that the effect size is larger in the smaller studies, it is generally a good idea to refer to this as a "small-study effect" rather than publication bias. Rather than saying "there *was* publication bias and therefore the true effect size is smaller than our estimate" we would say "*if* the small-study effect was due to publication bias, then the true effect size would be smaller than our estimate" (J. P. Ioannidis & T. A. Trikalinos, 2007; Jaime L Peters et al., 2010; J. Sterne et al., 2008; J.

A. Sterne et al., 2011; J.A.C. Sterne, Egger, & Davey Smith, 2001; Jonathan A. C. Sterne, Gavaghan, & Egger, 2000).

11.2.3. Use logic in trying to disentangle bias from small study

Once we have ruled out random sampling error, we might be able to argue that the small-study effect is (or is not) probably due to publication bias (J. Sterne et al., 2008; J. A. Sterne et al., 2011).

For example, if most studies in the analysis are statistically significant, and the effect size is larger in the smaller studies, publication bias is a plausible explanation. By contrast, if only a small proportion of studies are statistically significant, it is less likely that publication bias had a substantial impact on which studies were included.

The way studies were located might also be relevant in this context. If we are pulling studies from the literature that assessed the impact of drugs for treating depression, it is plausible to expect some bias based on selective publication and reporting. By contrast, in a prospective meta-analysis (where a set of primary studies had been planned in advance by a group of researchers and are now being included in a meta-analysis) we know that we have included all the trials, and so a small-study effect cannot be due to publication bias.

These examples are not intended to be exhaustive, but to provide a framework for thinking about possible causes of asymmetry in the funnel plot.

Summary

When we use a random-effects analysis, if the effect size tends to be larger in small studies this *could* be due to publication bias, but *alternatively* could reflect the fact that the effect size actually is larger in small studies. The procedures outlined here cannot distinguish between these two possibilities.

11.3. Publication bias does not invalidate the analysis

11.3.1. Mistake

When a test for publication bias is statistically significant, researchers sometimes conclude that the meta-analysis is not useful. The reality is more nuanced.

11.3.2. Details

Almost any meta-analysis where studies are pulled from the literature will be affected by publication bias. Fortunately, that does not invalidate the analysis. The key issue is not whether *any* bias exists, but rather *how much* of an impact this bias might have caused. In many cases we will be able to say that while bias may have inflated the effect size, the basic conclusions of the analysis are robust.

In this context it is important to keep in mind that publication bias is only one of many types of bias that could have an impact on the analysis. Publication bias refers to the fact that some studies may be missing from the analysis, but there are other types of bias that could affect the validity of the studies that are included in the analysis. These include selective reporting of outcomes, poor methods for randomizing patients, poor allocation concealment, among others.

For perspective, it is also helpful to keep in mind that the mean effect size in the analysis will depend on the particular mix of populations that we happen to include in the analysis. If we happen to include populations where the effect size is relatively large, the mean will shift upward. If we happen to include populations where the effect size is relatively small, the mean will shift downward. As such, publication bias is not operating in a pristine environment. Rather, it is one more source of noise in an environment that is somewhat noisy to begin with.

Summary

When we consider the validity of the results, we need to consider the potential impact of publication bias. However, the presence of bias does not automatically invalidate the results.

11.4. Tests to detect bias may be over-interpreted

11.4.1. Mistake

The first two approaches discussed in section 11.1, the rank correlation test and the regression test, both pose the null hypothesis that there is no evidence of publication bias, and then attempt to disprove that hypothesis. There are two problems with this approach, as follows.

11.4.2. Details

First, researchers tend to assume that the test's *p*-value is an index of the amount of bias. If the test for bias yields a *p*-value of 0.001 the researcher might assume that there is substantial bias. Conversely, if the test for bias yields a *p*-value of 0.200 the researcher might assume that there is no bias. Neither assumption is necessarily correct. The *p*-value for a test of bias is a function both the strength of the relationship and the number of studies in the analysis. Therefore, a significant *p*-value *could* reflect a strong correlation between sample size and effect size, but *could also* reflect the fact that the analysis includes many studies. Similarly, a non-significant *p*-value *could* reflect the fact that sample size and effect size are not correlated, but *could also* reflect the fact that there are a small number of studies.

Second, this approach assumes that if the effect size tends to be larger in small studies, this is evidence of publication bias. In fact, the effect size could to be larger in small studies for reasons that have nothing to do with bias, as explained in the discussion on *small-study effects* (section 11.2).

Summary

A significant *p*-value for the rank correlation test and the regression test tells us that the effect size tends to be larger in the smaller studies. However, the *p*-value should not be used as a surrogate for the magnitude of this relationship. Additionally, while the relationship may be due to publication bias, it could also be due to other factors.

11.5. Trim and fill

11.5.1. Mistake

As discussed in the prior section, one of the key problems with some procedures is that they test for the *presence* of bias, rather than estimating (and correcting for) the *extent* of bias. One procedure that does estimate and adjust is the Trim and Fill approach. This approach is useful when employed correctly, but is sometimes over-interpreted.

11.5.2. Details

The Trim and Fill procedure looks for asymmetry in the plot as a method for locating missing studies. Figure 64 is a funnel plot for the analysis of second-hand smoke and lung cancer introduced in section 11.1.1. In the absence of bias, we would expect the effects to be evenly distributed on either side of the mean effect. Concretely, we would expect to find roughly the same number of studies in area [C] as in area [D], but we do not.

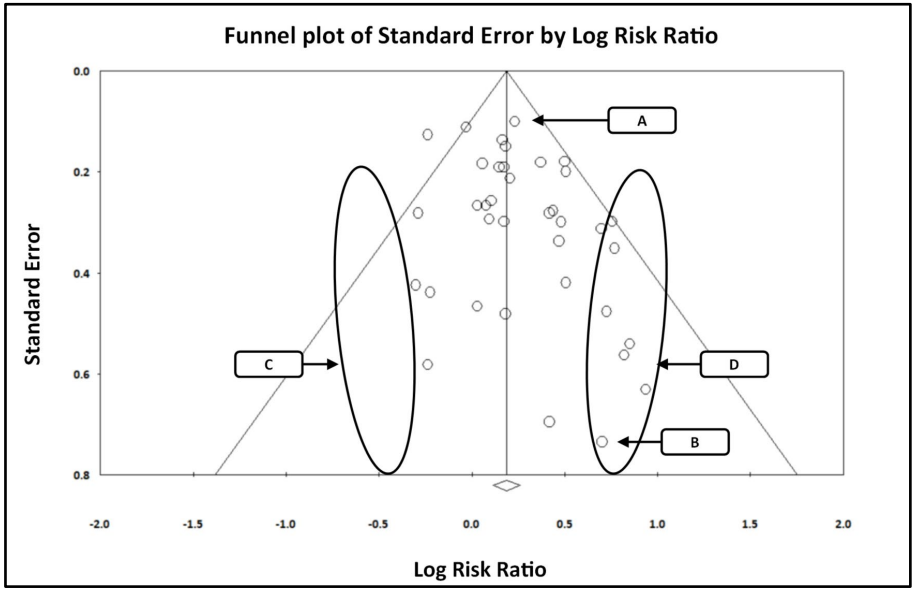

Figure 64 | Observed studies only

The Trim and Fill method assumes that the area [C] studies are missing due to publication bias. It creates these studies, inserts them into the analysis, and runs the analysis using the original and imputed studies to yield an adjusted mean (Figure 65).

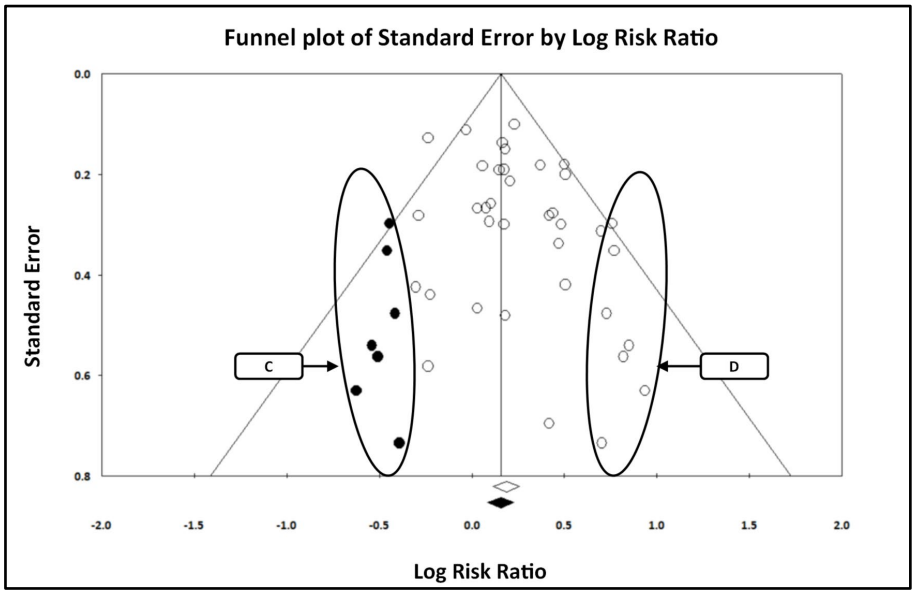

Figure 65 | Observed studies, and studies imputed by Trim and Fill

The advantage of this approach is that it actually provides an adjusted effect size. If the asymmetry is due to publication bias, this procedure allows us to estimate what the true effect size would be with the bias removed.

In the *smoking* example, the observed effect size was a risk ratio of 1.238 while the adjusted effect size is a risk ratio of 1.189. This tells us that while publication bias may have led us to overestimate the magnitude of the risk, we would arrive at basically the same conclusion if we were able to eliminate the bias. In other words, if someone is concerned about the initial estimate (that risk is increased by 24%), they would also be concerned about the adjusted estimate (that risk is increased by 19%). By comparison, if the adjusted estimate had been that risk is increased by only 1%, we would conclude that the substantive meaning of the results could be due primarily to publication bias.

This procedure avoids the first problem associated with the other tests, in that it tells us *how much* impact the bias might have had on the effect-size estimate. However, it is still subject to the second problem. If there are fewer

studies in the lower left quadrant, we assume that those studies were actually performed, and are missing due to publication bias. In fact, it is also possible that small studies actually do have larger effects, as discussed in the section on *small study effects* (section 11.2). In that case, these studies never existed and when we add them to the analysis, we are actually *introducing* bias, rather than removing it.

For this reason, when we use this method, we need to be very clear about the assumptions. Consider the following two options for reporting the results of this analysis.

- "The Trim and Fill method tells us that the plot is asymmetric. Specifically, the smaller studies tend to be clustered toward the right side of the plot (the lower right-hand quadrant) with relatively few studies toward the left (the lower left-hand quadrant). If we impute these studies and include them in the analysis, the adjusted effect size is a risk ratio of 1.189. *This is the effect size that we would have seen in the absence of publication bias*."
- "The Trim and Fill method tells us that the plot is asymmetric. Specifically, the smaller studies tend to be clustered toward the right side of the plot (the lower right-hand quadrant) with relatively few studies toward the left (the lower left-hand quadrant). *There are various reasons why this might be the case, one of which is publication bias. If publication bias is indeed the reason, it makes sense to impute the missing studies and compute an adjusted effect size*. The adjusted effect size is a risk ratio of 1.189."

The first approach assumes that publication bias is responsible for the small number of studies in the lower-left quadrant. The second approach acknowledges that this might not be the case, and is therefore more appropriate.

In general, this procedure should be seen as a kind of sensitivity analysis that tells us whether the essential conclusion is robust to publication bias, rather than an attempt to yield adjusted numbers (J. L. Peters, Sutton, Jones, Abrams, & Rushton, 2007).

11.5.3. Additional problems associated with Trim and Fill

As discussed earlier (section 11.2) all common procedures to address publication bias are intended for the case when the true effect size is the same in all studies, and have serious limitations when the true effect size varies

across studies. This applies to the Trim and Fill procedure as well (J. L. Peters et al., 2007; Terrin et al., 2003).

Additionally, the Trim and Fill method is not always robust. In some cases, adding or removing one or two studies can substantially change the number of studies that are imputed. This follows from the method's algorithm, which uses statistical tests to look for asymmetry. A small change in the pattern of results will sometimes have this effect.

When one does apply the Trim and Fill method, there is some confusion about what computational options to use. There are two parts to the method – first we impute the missing studies. Then we re-run the analysis with the original studies plus the imputed ones. The first part can be performed using either a fixed-effect or random-effects model. Sutton (2005) and J. L. Peters et al. (2007) recommend using a fixed-effect model for the first part. One should always use a random-effects model for the second part.

Summary

The Trim and Fill approach is more informative than some other approaches in that it provides an estimate of the adjusted effect size. However, it looks for the same pattern of effects as the other methods (the effect size is larger in small studies) and assumes that this pattern is due to bias. In fact, the pattern could be due to other factors.

11.6. The tests only work under certain conditions

11.6.1. Mistake

All the procedures outlined above are based on the idea that we can look for a relationship between the size of the study and the size of the effect. For these analyses to work, it is necessary that we have sufficient data to apply the procedures properly. Sometimes, researchers apply these procedures when the data does not allow for a meaningful analysis.

11.6.2. Details

In order to apply any of the procedures for publication bias being discussed here, several conditions must apply (J. P. Ioannidis & T. A. Trikalinos, 2007; J. A. Sterne et al., 2011).

- We need to have a reasonable number of studies. There is a consensus that we should use 10 studies as a minimum, but that many more studies would be needed in the presence of substantial heterogeneity (J. P. T. Higgins & Green, 2008; J. P. Ioannidis & T. A. Trikalinos, 2007; J. A. Sterne et al., 2011).
- We need to have a reasonable amount of variation in the sample size. The procedures all look for a relationship between effect size and sample size. If all studies have approximately the same sample size, then (by definition) there can be no correlation between sample size and effect size. When all studies in the analysis have been performed by drug companies, there is a distinct possibility that all studies will have a similar sample size, since drug companies often use a standard sample size for a particular type of study.
- We need to have a reasonable amount of variation in the effect size. If all studies have approximately the same effect size, then (by definition) there can be no correlation between sample size and effect size.
- There must be at least one study in the analysis that is statistically significant. If no studies are statistically significant, it makes no sense to suggest that our sample was biased by the preferential inclusion of statistically significant studies (J. P. Ioannidis & T. A. Trikalinos, 2007).

Summary

Before applying the procedures to assess publication bias, we need to be sure that we have a sufficient number of studies, with sufficient variation in sample size and effect size, for the analysis to be meaningful.

11.7. Procedures do not apply to studies of prevalence

11.7.1. Mistake

The common procedures to address publication bias are based on the idea that studies which are statistically significant are more likely to be published than studies which are not statistically significant. Sometimes, researchers apply these methods in cases where this framework is not applicable.

11.7.2. Details

The idea that studies are more likely to be published when they are statistically significant only makes sense when the studies in the analysis test for statistical significance. As such, it applies to the vast majority of studies, including virtually all studies that assess the impact of an intervention, or that assess the relationship between two variables. However, it does not apply to studies that report the prevalence of a condition. When we report a prevalence, we simply report the prevalence. We do not test it to see whether it is significantly different from any specific value. The possibility that a study will not be published because it is not statistically significant simply does not apply when there is no test of significance.

One could make an argument that studies which report higher estimates of prevalence are more likely to be published than studies which report lower estimates of prevalence. For example, one could suggest that a study reporting the prevalence of PTSD as 50% is more likely to be published than one reporting the prevalence of PTSD as 10%. However, it is not clear that this is generally the case. And, even if this was the case, the bias would need to be limited to the smaller studies before we could apply the model (since the algorithms work by looking for a relationship between the size of the study and the size of the effect).

In sum, it is always a mistake to apply these (or any) tests by rote. We need to think about the logic of the test and whether this logic applies in any given case. When the goal of the analysis is to assess the impact of an intervention, the default would be to test for publication bias using the tools outlined earlier. But in other cases, we might not want to test for publication bias at all. And if we do, we would want to think about how bias would manifest itself. We would not assume that the usual model applies (J. A. Sterne et al., 2011).

Summary

The procedures for publication bias should only be applied when the likelihood that a study will be published is affected by a finding of statistical significance.

11.8. The model for publication bias is simplistic

11.8.1. Mistake

The procedures being discussed are all based on a model that assumes studies which are statistically significant are more likely to find their way into an analysis than studies which are not statistically significant. In many cases, the publication process is more complicated than that (Bax & Moons, 2011).

11.8.2. Details

The model that underlies publication bias is that if a study is not statistically significant the researcher is less likely to submit it for publication (within any given time frame) as compared with a paper that is statistically significant. And, if the researcher does submit the paper for publication, the journal is less likely to accept it for publication as compared with a paper that is statistically significant. Research shows that these assumptions do mirror the true state of affairs, in general. However, it is important to recognize that this is a simplistic view of the overall situation, and may not apply in any given case.

For example, one could imagine a scenario where the *first* studies looking at the impact of a new intervention are more likely to be published if they are statistically significant. After that, studies that confirm the original findings might be *harder* to publish, since they (merely) confirm what we already know, while studies that are *not* statistically significant might be published more readily, since they challenge the current state of information.

Additionally, the basic idea that statistically significant studies are more likely to be published than non-significant studies varies by trends and by journals. Recent awareness of the problems with lack of replication may shift editorial priorities, and some journals will agree to publish any study provided that the protocol is of high value and is followed correctly.

In this volume I have focused exclusively on procedures that are in common use and are relatively simple to use. More advanced procedures are discussed by (Bayarri, 1988; DuMouchel, 1988; Hedges, 1988, 1992; Iyengar & Greenhouse, 1988a, 1988b; Keith & Begg, 1992; Laird, Patil, & Taillie, 1988; Rao, 1988; Rosenthal & Rubin, 1988).

Summary

The procedures for bias outlined here are based on a model wherein studies that are statistically significant are more likely to be published than studies which are not statistically significant. The reality is often more complicated.

11.9. Publication bias and the grey literature

11.9.1. Mistake

Some researchers think that publication bias refers to a distinction between studies that were published in a journal vs. those that were published as technical reports, dissertations, or abstracts. This is incorrect.

11.9.2. Details

The term *grey literature* refers to technical reports, dissertations, abstracts, and so on. Researchers who locate some studies in journals and other studies in the grey literature sometimes think that the term publication bias refers to the difference between these two. This is incorrect. Publication bias (or more generally, retrieval bias) refers to fact that we can locate some studies and not others. In this context, all studies that we can locate are considered published (or retrieved), without regard to where they were located.

It is true that there may be a difference between published studies and the grey literature, with the latter reporting smaller effects. In some cases, we might want to explore that as a separate issue. However, once we locate a study, we no longer have a concern that the study is missing. For our purpose it does not matter if the study was located in a top-tier journal or in a file drawer.

Summary

The term *publication bias* refers to the concern that we might fail to retrieve some studies. It does not refer to the fact that some of the studies in the analysis had been published and others had been located in other places.

11.10. Lines on funnel plot

11.10.1. Mistake

Some researchers think that studies outside the triangle on a funnel plot are evidence of bias. This is incorrect.

11.10.2. Details

The funnel plot is typically drawn with a pair of lines that start at the mean effect size and extend two standard error on either side (Figure 66). At the top, where the standard error is zero, the lines converge. Toward the bottom, where (in this example) the standard error is 0.8, the lines extend 1.6 standard units on either side of the mean. These lines make it easier to identify studies that fall more than two standard error from the mean. The fact that a study happens to fall outside these lines [A] says nothing about publication bias.

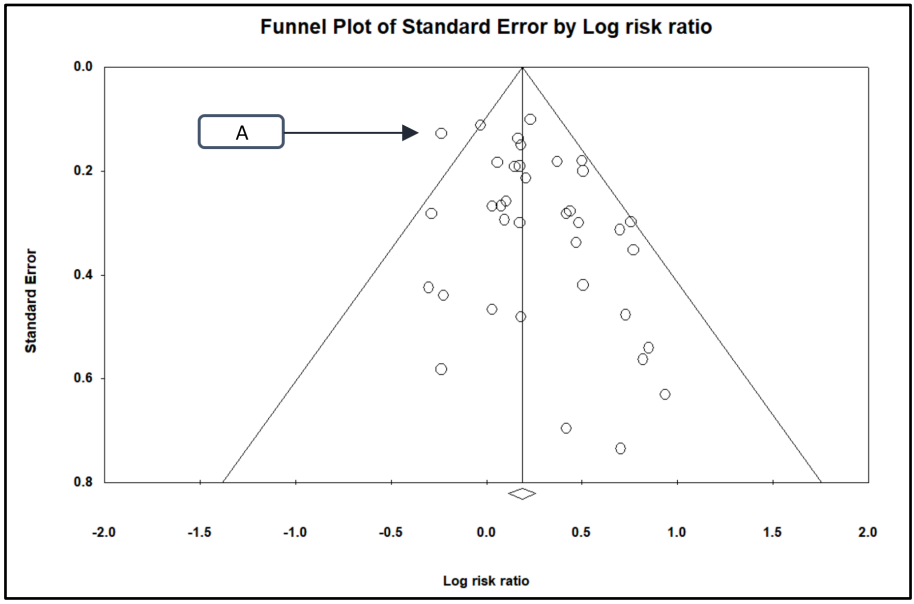

Figure 66 | Funnel plot for smoking data set

Summary

The lines on a funnel plot indicate two standard error on either side of the mean effect. The fact that a study falls outside this range does not indicate publication bias.

11.11. Fail-Safe *N*

11.11.1. Mistake

Researchers sometimes apply a method known as Fail-Safe N to address publication bias. This approach should be avoided.

11.11.2. Details

The Fail-Safe *N* is an idea developed by Rosenthal in the late 1970s (Rosenthal, 1979). Rosenthal wanted to address the concern that the analysis yields a statistically significant result only because it is based on a biased subset of all studies that had actually been performed. Specifically, he wanted to address the possibility that if all missing studies were somehow located and then included in the analysis, the results would no longer be statistically significant. He suggested that we compute the number of missing studies with a nil effect (here, a risk ratio of 1.0) that we would need to add to the analysis, to shift the *p*-value above 0.05.

This number is called the Fail-Safe *N*. If the number is low, relative to the number of studies in the analysis, then there may be cause for concern. If the number is high, relative to the number of studies in the analysis, then there is less cause for concern. Harris Cooper coined the term File-Drawer problem to refer the presumed location of these missing studies.

The smoking analysis includes 37 studies, and so it seems plausible in this case that there may be twenty, or forty, or perhaps sixty studies that were performed and not published. But the Fail-Safe *N* is 269, and it is not likely that this number of studies were performed and then simply filed away without being published. Therefore, we would conclude that the true risk ratio for all studies is not 1.0. There is no clear line between what number is acceptable and what would be a source of concern. This would depend on the reviewer's judgment, informed by knowledge of the field.

This approach was appropriate when Rosenthal proposed it, but has little utility for meta-analysis as we practice it today. There are several reasons for this.

When Rosenthal proposed the use of Fail-Safe *N*, he was working with meta-analyses where the primary goal was to test the null hypothesis of no effect. In that context, if we rejected the null hypothesis, it would be useful to know that the basic conclusion was robust. Today, our focus is not on a test of the null hypothesis, but rather on estimating the mean effect size. The relevant question today would be "How many studies do we need to add to

the analysis before the mean effect size is no longer of substantive import," and this question is not addressed by the Fail-Safe *N*.

Even if we did want to focus on a test of the null hypothesis, there is another problem with using the Fail-Safe *N* for this purpose. The Fail-Safe *N* deals with the null hypothesis that the true effect size in *all* studies is zero. By contrast, the null hypothesis addressed by current methods is that the *mean* effect size is zero. These two hypotheses are not the same, and it is possible (for example) for a test of the first to yield a *p*-value of 0.02 while a test of the second yields a *p*-value of 0.10. In this case the Fail-Safe *N* might tell us that we would need to add twenty studies to make the *p*-value non-significant, when the *p*-value is *already* non-significant.

For these reasons, the Fail-Safe *N* should be avoided. For additional discussion see J. Sterne et al. (2008) and Orwin (1983). Copas and Jackson (2004) discuss an advanced method that uses the same basic idea but can be applied to the kinds of analyses we perform today.

Summary

The Fail-Safe *N* is intended to provide assurance that the results are not entirely an artifact of publication bias. This approach was developed in another era, when the goal of a meta-analysis was primarily to test a specific null hypothesis. It has little relevance today, when the goal is to estimate the mean effect size and to test a different null hypothesis.

11.12. Using cumulative analysis

11.12.1. Mistake

All the approaches discussed above assume that a small-study effect is due to publication bias. There is another procedure that allows us to assess the small-study effect without making any assumptions about the reason for this effect. This procedure is rarely applied.

11.12.2. Details

The procedures for addressing publication bias are based on the idea that as the studies get smaller, the effect size gets larger. Some procedures merely test to see if this relationship exists. One procedure (Trim and Fill) tried to gauge the size of the relationship and adjust for it.

There is another way to approach the same issue. This is to use a cumulative meta-analysis as shown in Figure 67. We sort the studies from most precise to least precise (based on the standard error of each study), and then we run a sequence of 37 separate analyses. In this figure, the effect size shown on the row for any given study is *not* the effect size for that study. Rather, it is the effect size for an analysis that includes all studies in the plot up to and including that one. The first is based on the first study alone, the next is based on the first two studies, the next is based on the first three studies, and so on (Borenstein et al., 2009; Rothstein, Sutton, & Borenstein, 2005).

This plot offers a useful perspective on the issue of a small-study effect. By looking at the mean effect size as we move from the top toward the bottom of the plot, we can detect a shift toward the right. For example, if we pick the analysis based on the top 18 studies (that is, halfway down the plot, on the line for Chan et. al.) the mean effect size is 1.169 [X]. From that point on the mean shifts consistently toward the right. In the last analysis (based on all 37 studies) the mean effect size is 1.238 [B].

We can interpret this analysis the same way we interpret Trim and Fill. In the bottom half of the plot (the studies below Chan et al), the studies tend to show higher effects, which shift the mean from 1.169 to 1.238. The rightward shift tells us that the effect size tends to be larger in the smaller studies, but the reason remains unknown. If this shift reflects the fact that the effect size actually is larger in smaller studies, the preferred estimate of the mean effect size would be the one based on all thirty-seven studies (a risk ratio of 1.238). On the other hand, if this shift is due to publication bias, the

preferred estimate of the mean effect size would be the one based on the top half of the plot (1.169). Importantly, the shift from 1.169 to 1.238 does not substantially alter the conclusions of the analysis (that second-hand smoke is associated with a clinically important risk of lung cancer).

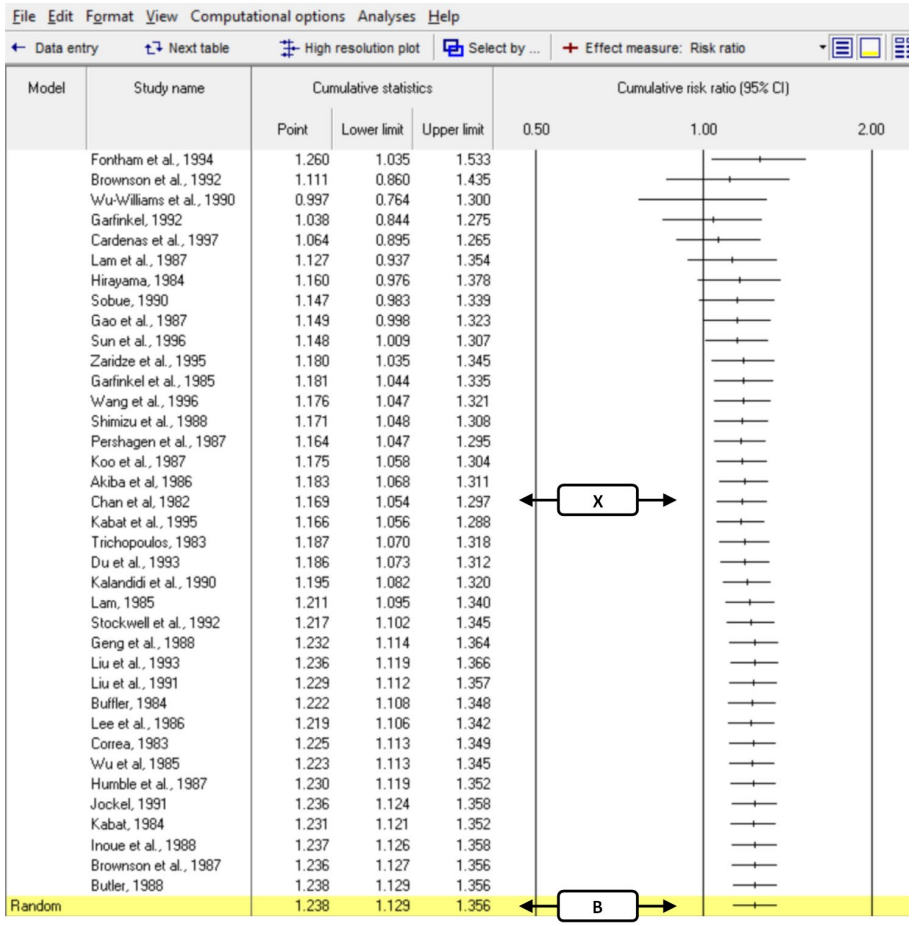

Figure 67 | Cumulative analysis | Risk ratio > 1 indicates increased risk

In a sense, this approach has the same benefits as Trim and Fill without some of the drawbacks. Like Trim and Fill, it shows the extent of possible bias, and provides an adjusted value. Unlike Trim and Fill, this procedure is relatively robust, in the sense that the results will not shift dramatically based on the effect size in one or two studies.

Importantly, neither procedure allows us to determine whether the shift is due to bias or a small-study effect. The difference between the two

procedures is that Trim and Fill only makes sense if the asymmetry is due to publication bias. By contrast, the cumulative procedure makes no assumption about the reason for asymmetry – it simply reports how much the mean shifts as small studies are added. This works whether the shift is due to an increase in bias, or because the effect size actually is larger in small studies.

In this example I chose to divide the top and bottom of the plot at the eighteenth study since that is the midpoint, but there is nothing special about this point. One could divide the plot into those with sample sizes of *N* patients or less. One could divide the plot into thirds. This approach is not intended to be a statistical procedure that yields clear-cut distinctions. Rather, it is intended as a vehicle that people can use to get a sense of how bias (or a small-study effect) may have impacted the analysis. To be clear, the Trim and Fill procedure should be seen in this light as well. The algorithm that it uses to identify the number of missing studies was never intended to yield definitive estimates.

Note.

If the cumulative effect shifts to the right, we assume this is because the additional studies tend to have larger effect sizes. It is also possible that the shift is (partly) due to a re-estimation of tau-squared, and the consequent re-weighting of the larger studies. To apply this approach carefully, we would look at the weights and ensure that this is not the case.

Summary

It is possible to use a cumulative analysis, starting with the larger studies and sequentially adding smaller studies, to get a sense of how much the effect size shifts as the smaller studies are added to the analysis.

11.13. The focus on publication bias ignores other types of bias

11.13.1. Mistake

Researchers sometimes focus on publication bias to the exclusion of other types of bias. This is a mistake, since there are other types of bias that could pose a substantial threat to the validity of the analysis.

11.13.2. Details

Any meta-analysis is subject to various types of bias. In addition to publication bias, which looks at the possibility that some studies have been excluded from the analysis, we also need to be concerned about the risk of bias *within* the studies that are included in the analysis. One approach to dealing with these biases is to address them using the "Risk of bias" table that should accompany every analysis.

Among these potential biases is one called selective reporting bias. This refers to the situation where studies are included in the analysis, but data within those studies is cherry-picked. For example, if someone tested an intervention using several outcomes, they might report the outcome that showed a large effect while ignoring ones that showed smaller effects. This type of bias can be especially pernicious and may have a substantially larger impact than publication bias.

Publication bias is generally addressed separately from the other types of bias. This follows from the fact that publication bias deals with the issue of whether studies are missing from the analysis, whereas the others deal with the potential for bias inside the studies that we have included. As such, this is a reasonable approach. However, we need to assess the potential for all these biases. The absence of publication bias should not lull the researcher into a false sense of security.

Summary

Publication bias addresses the fact that the studies included in the analysis may be a biased subset of all studies that had been performed. We also need to address the fact that the studies included in the analysis also suffer from various types of bias.

11.14. Putting it all together

Publication bias refers to the fact that studies which overestimate the impact of an intervention are more likely to be included in meta-analyses than studies which accurately estimate or underestimate the impact of that intervention. Researchers have developed methods that are intended to identify the presence of publication bias. While these methods are sometimes helpful, they have some serious limitations.

The tests for publication bias can only work under certain conditions. Specifically, they require a reasonable number of studies, and a reasonable amount of variance in the sample size and in the effect size. The tests only make sense if the studies report testing the effect size for statistical significance. If any of these conditions is missing, the procedures outlined in this chapter are pointless.

The procedures to address publication bias were developed primarily for cases where the true effect size is basically the same in all studies, and so a finding that the effect size is larger in the smaller studies would most likely be due to bias. In cases where the true effect size varies across studies, which is true for the vast majority of meta-analyses, these methods are problematic. If the effect size tends to go up as the sample size goes down, this could be evidence of publication bias. However, this could also reflect the fact that the effect size actually is larger in smaller studies for reasons that are unrelated to bias.

Additionally, when the effect size varies across studies we are working in an environment where the observed mean will vary depending on the mix of populations that happen to be included in the analysis. In this context, the possibility of publication bias is one source of noise among many.

In any event, we need to be careful not to over-interpret the tests. Even if we can establish with reasonable certainty that there is some publication bias, the results of the analysis may still be very useful. Conversely, even if we can establish with reasonable certainty that there is no publication bias, the results of the analysis may be seriously tainted by other kinds of bias.