6

Malleability of Meta-Analysis

6.1 Meta-Analysis

Medical scientists are faced with a daunting volume and diversity of evidence for many hypotheses regarding the effectiveness and harmfulness of medical interventions. This avalanche of evidence contributed to the formation of groups dedicated to the systematic review of evidence (such as the Cochrane Collaboration), to journals that publish reviews of existing evidence rather than evidence from original research, and to methods of amalgamating evidence, including social methods, such as consensus conferences, and quantitative methods, such as meta-analysis. My focus in this chapter is on meta-analysis, widely thought to be among the most reliable methods in medical research (see Chapter 5). I describe the purported merits of meta-analysis and the aims that analysts set out to achieve with this method, critically assess the details of the method, and argue that meta-analysis does not generally have the merits that many claim for it. Meta-analysis is malleable.

Meta-analysis is a method that combines evidence from individual studies into summary measures of the beneficial and harmful effects of medical interventions. Many claim that meta-analysis is an especially reliable method (§6.2). I articulate two purported methodological principles that many seem to think meta-analysis is especially good at satisfying: *constraint*—the use of meta-analysis should constrain assessments of medical interventions—and *objectivity*—meta-analysis should be performed in a way that limits the influence of subjective biases and idiosyncrasies of researchers.

I show that the use of meta-analysis often fails to achieve *constraint* (§6.3). Meta-analysis fails to constrain assessments of medical interventions because numerous decisions must be made when performing a meta-analysis, which allow wide latitude for subjective idiosyncrasies to influence the results of a meta-analysis. My argument involves a close examination of these details (§6.4). Meta-analysis is performed by selecting which primary studies are to be included in the meta-analysis, calculating the magnitude of the effect attributed to an intervention for each study, assigning a weight to each study, and then calculating a weighted average of the effect sizes. To help describe the methodology of meta-analysis I draw on the published guidance of the Cochrane Collaboration, an institution of evidence-based medicine that commissions a large number of meta-analyses. Finally, I end by discussing an alternative, older, and

arguably better strategy for assessing a large volume and diversity of evidence (§6.5), associated with the epidemiologist Sir Bradford Hill (1897–1991).

There is a debate about whether or not randomized trials are the gold standard of evidence to assess medical interventions.[1] However, it is in fact meta-analysis (or systematic reviews that typically include meta-analyses) that is at the top of the most prominent evidence hierarchies in medicine (see Chapter 5). In what follows I criticize this assumed status of meta-analysis. Meta-analyses, like randomized trials, are malleable, and liable to be influenced by numerous biases. This fact, together with a broader concern about biases in medical research (see Chapter 10), provides support to one of the central premises of the master argument for medical nihilism described in Chapter 11.

6.2 Constraint and Objectivity

The first comprehensive meta-analysis was about extra-sensory perception.[2] Meta-analysis later became the platinum standard of evidence in medicine for several reasons. The sheer volume of available evidence meant that most users of evidence (for example, physicians and policy-makers) could not be aware of all relevant evidence. A proposed solution was to produce systematic reviews of the available evidence. By the 1990s, hundreds of meta-analyses were being published every year, and now thousands are published every year.

Meta-analysis became a prominent method in part due to its purported rigor compared with qualitative and unstructured methods of amalgamating evidence. In contrast with qualitative literature reviews and consensus conferences, meta-analyses have a constrained structure and a quantitative output. The importance of using systematic methods of amalgamating evidence became apparent by the 1970s, when scientists began to review a plethora of evidence with what some took to be personal idiosyncrasies.[3] A recent textbook on meta-analysis worries that unstructured reviews "come to opposite conclusions, with one reporting that a treatment is effective while the other reports that it is not"—the solution to this problem, according to the authors, is to use meta-analysis, a more structured method that (goes this suggestion) can

[1] See (Worrall, 2002), (Worrall, 2007), (Borgerson, 2008), (Cartwright, 2007), and (Cartwright, 2009).

[2] (Rhine, Pratt, Stuart, Smith, & Greenwood, 1940). This is a nice historical accident, because Hacking (1988) showed that the practice of randomizing subjects into different groups also began in psychical research—thus both our alleged gold standard of evidence and our alleged platinum standard of evidence first arose in research about paranormal psychology.

[3] An early defender of meta-analysis claimed that "A common method for integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students or friends" (Glass, 1976). An example is (Pauling, 1986), in which the Nobel Laureate cited dozens of his own studies supporting his hypothesis that vitamin C reduces the risk of catching a cold, and yet he did not cite studies contradicting this hypothesis, though several had been published (Knipschild, 1994).

constrain assessments of medical interventions.[4] Likewise, a statistics textbook emphasizes a worry regarding reviewers' idiosyncrasies—"the conclusions of one reviewer are often partly subjective, perhaps weighing studies that support the author's preferences more heavily than studies with opposing views"—and the authors suggest that meta-analysis can mitigate this concern.[5]

The scientific basis of meta-analysis is simple. Many purported causes in medicine have a small effect, and so when analyzing data from a single trial on an intervention with a small effect, there might be no statistically significant difference between the experimental group and the control group of the trial. But by pooling data from multiple trials the sample size of the analysis increases, thereby rendering estimates of the magnitude of an intervention's effects more precise, and perhaps statistically significant. A key feature of meta-analysis, then, is quantitative precision, which is especially important for detecting small effects (as I argue in Chapters 8 and 11, many medical interventions have tiny effects).

In short, meta-analysis is a method to amalgamate evidence from multiple studies. Relative to other methods of amalgamating evidence, such as informal reviews or consensus conferences, meta-analysis is said to have the virtues of constraining estimates of the effectiveness and harmfulness of medical interventions and doing so in a way that is not influenced by subjective idiosyncrasies of analysts. The purported rigor, transparency, quantitative precision, and freedom from personal bias can be summarized by these two principles:

CONSTRAINT: Meta-analysis should constrain estimates of the effectiveness and harmfulness of medical interventions.

OBJECTIVITY: Meta-analysis should not be sensitive to idiosyncratic or personal biases.

A straightforward way of construing the relation between these two norms is that *OBJECTIVITY* is in the service of *CONSTRAINT*: meta-analysis can constrain estimates of the effectiveness and harmfulness of medical interventions only if it is not sensitive to analysts' idiosyncratic or personal biases.[6] Defenders of meta-analysis claim that, compared with other methods of amalgamating a large volume of evidence, meta-analysis best satisfies these principles. This is the basis of the alleged status of meta-analysis at the top of evidence hierarchies.

However, in the following sections I argue that meta-analysis, unfortunately, generally fails to satisfy these principles. The details of the methodology of a meta-analysis require many decisions at multiple stages, which allow wide latitude for an analyst's idiosyncrasies to affect its outcome. Meta-analysis is malleable.

[4] (Borenstein, Hedges, Higgins, & Rothstein, 2009).

[5] Since "it is extremely difficult to balance multiple studies by intuition alone without quantitative tools" (Whitlock & Schluter, 2009), the authors claim meta-analysis should be used.

[6] For a recent historical account of objectivity see (Daston & Galison, 2007), and for a recent philosophical account see (Douglas, 2004).

6.3 Failure of Constraint

Medical scientists have recently noted that multiple meta-analyses about the same medical interventions can reach contradictory conclusions. For example, there have been numerous inconsistent studies on the benefits and harms of synthetic dialysis membrane versus cellulose membrane for patients with acute renal failure: one meta-analysis of these studies found greater survival of such patients using the synthetic membrane compared with those using the cellulose membranes, while another meta-analysis reached the opposite conclusion. Here is another example. Two meta-analyses published in the same issue of *BMJ* came to contradictory conclusions regarding whether or not an association exists between the use of selective serotonin reuptake inhibitors (SSRIs, a class of antidepressants) and suicide attempts. In one, there was no association found between the use of these antidepressants and suicide attempts, and only a weak association between antidepressant use and risk of self-harm, while in the other there was a strong association between antidepressant use and suicide attempts.[7] Contradictory conclusions have been reached from meta-analyses on the benefits of acupuncture and homeopathy, mammography for women under fifty, and the use of antibiotics to treat otitis, to name a few other examples.

Differential outcomes between contradictory meta-analyses can be associated with the analysts' professional or financial affiliations. For example, several meta-analyses have investigated a potential causal relation between formaldehyde and leukemia. Two meta-analyses concluded that formaldehyde exposure does not cause leukemia. In contrast, a third found a modest elevation of risk of developing leukemia in professionals who work with formaldehyde, such as pathologists and embalmers. A fourth found an even higher risk.[8] The meta-analyses that concluded that formaldehyde exposure does not cause leukemia were performed by employees of private consulting and industrial companies. In contrast, the authors of the two meta-analyses that found some evidence for a causal relation between formaldehyde exposure and leukemia worked in academic and government institutions.

Barnes and Bero (1998) performed a quantitative second-order assessment of multiple reviews that reached contradictory conclusions regarding the same hypothesis, and found a very strong correlation between the outcomes of the meta-analyses and the analysts' relationships to industry. They analyzed 106 review papers on the health effects of passive smoking ('secondhand smoke'): thirty-nine of these reviews concluded that passive smoking is not harmful to health, and the remaining sixty-seven concluded that there is some adverse health effect from

[7] The four meta-analyses cited here are, respectively: (Subramanian, Venkataraman, & Kellum, 2002), (Jaber et al., 2002), (Gunnell, Saperia, & Ashby, 2005), and (Fergusson et al., 2005).

[8] The citations are: (Bachand, Mundt, Mundt, & Montgomery, 2010), (Collins & Lineker, 2004), (Bosetti, McLaughlin, Tarone, Pira, & La Vecchia, 2008), and (Zhang, Steinmaus, Eastmond, Xin, & Smith, 2009). Formaldehyde exists in products that account for more than 5 percent of the U.S. gross national product (Zhang et al. 2009).

passive smoking. Of the variables investigated, the only significant difference between the analyses that showed adverse health effects versus those that did not was the analysts' relationship to the tobacco industry: analysts who had received funding from the tobacco industry were eighty-eight times more likely to conclude that passive smoking has no adverse health effects compared with analysts who had not received tobacco funding.

Here is another example. Antihypertensive drugs have been tested by hundreds of studies, and as of 2007 there had been 124 meta-analyses on such drugs. Meta-analyses of these drugs were five times more likely to reach positive conclusions regarding the drugs if the reviewer had financial ties to a drug company. Or consider the second-order review of meta-analyses of studies on spinal manipulation as a treatment for lower back pain: some meta-analyses have reached positive conclusions regarding the intervention while other meta-analyses have reached negative conclusions, and a factor associated with positive meta-analyses was the presence of a spinal manipulator on the review team.[9]

Such examples could easily be multiplied. About one third of meta-analyses in medicine are published by employees of the company that manufactures the drug that is assessed by the meta-analysis, and these meta-analyses are twenty times less likely to make negative claims about that drug.[10] The above examples illustrate the fact that multiple meta-analyses of the same primary set of evidence can reach contradictory conclusions. The examples suggest that idiosyncratic features of analysts influence the results of meta-analyses. Moreover, the features of meta-analysis that explain its occasional failure to attain *CONSTRAINT* are shared by all meta-analyses. That is, the conditions under which multiple meta-analyses of the same primary evidence can reach contradictory conclusions are inherent features of all meta-analyses. To show this I turn to a detailed examination of the method.

6.4 Meta-Analysis is Malleable

The failure of *CONSTRAINT* in the above cases is at least partially a consequence of the failure of *OBJECTIVITY*: constraint on assessments of medical interventions was not met by the meta-analyses in §6.3 because the meta-analyses were not sufficiently objective. Subjectivity is infused at many levels of a meta-analysis: when designing and performing a meta-analysis, decisions must be made—based on judgment, expertise, and personal preferences—at each step of a meta-analysis, which include the choice of primary evidence, outcome measure, quality assessment tool, and averaging technique. I examine each choice in turn.

[9] The two second-order reviews cited in this paragraph are (Yank, Rennie, & Bero, 2007) and (Assendelft, Koes, Knipschild, & Bouter, 1995).

[10] (Ebrahim, Bance, Athale, Malachowski, & Ioannidis, 2016).

6.4.1 Choice of primary evidence

Multiple decisions must be made regarding what primary evidence to include in a meta-analysis. The dominant view in evidence-based medicine is to include only evidence from randomized trials in a meta-analysis.[11] Such a view excludes other common kinds of evidence, including that from cohort studies and case-control studies, as well as other kinds of evidence that are not in the domain of meta-analyses, such as pathophysiological evidence, evidence from animal experiments, mathematical models, and clinical expertise.

When assessing a medical intervention one should use all available evidence. Consider: an effect size of 2.0x from three randomized trials testing a particular medical intervention should have a different impact on one's assessment of the intervention when considered in the background of fifty case-control studies on the same intervention that show an effect size of 2.2x, versus fifty case-control studies that show an effect size of −0.8x. If one's assessment of the intervention were not different in the two scenarios, one would be making an unreliable inference. One's assessment of a hypothesis after observing new evidence should be guided by all of one's previous evidence (this general norm is called the 'principle of total evidence'), and if it is not then one is liable to make an unreliable inference about the probability that the hypothesis is true in light of the new evidence.

Consider the following guidance from the Cochrane collaboration: "review authors should not make any attempt to combine evidence from randomized trials and [non-randomized studies]" (Cochrane Handbook 13.2.1.1). Such a practice could limit the external validity of a meta-analysis, since randomized trials are typically performed with relatively narrow study parameters while other kinds of evidence—including evidence from non-randomized human studies, studies on animals, and experiments designed to elucidate causal mechanisms, which are often performed on tissue and cell cultures—can have diverse experimental parameters and aid in causal inference.[12]

Even if we grant that randomized trials provide the most reliable evidence, that would not mean that evidence from non-randomized studies is negligible. Indeed, some of our best medical interventions were supported by evidence from non-randomized studies (such as insulin for type 1 diabetes, discussed in Chapter 4), and for many medical interventions we only have evidence from non-randomized studies. A joke in such discussions is that there has never been a randomized trial that has tested the effectiveness of parachutes.

The exclusive use of a narrow range of evidence is purportedly justified by the garbage-in-garbage-out argument: if low-quality evidence is included in a meta-analysis, then the output of the meta-analysis will be low quality. Some take this to entail that

[11] For instance, when performing a meta-analysis, Egger, Smith, and Phillips (1997) claim that "researchers should consider including only controlled trials with proper randomisation."

[12] I discuss this further in Chapters 8 and 9. See also (Illari, 2011), (Leuridan & Weber, 2011), (Russo & Williamson, 2007), and (Howick, 2011a).

rather than including all available evidence, meta-analyses should only include the best evidence (Slavin, 1995).

There is something correct about the garbage-in-garbage-out argument, but there are several problems with it. First, as above: if we ignore evidence, even if it comes from a lower-quality method, we violate the principle of total evidence. Second, Worrall (2002), Cartwright (2007), and others have argued that there is no gold standard of evidence; it follows that we ought to take into account evidence of multiple kinds when it is available. Third, the possibility of defeating evidence should compel us to consider all available evidence.[13] Fourth, there is no reason why an analyst cannot assess lower-quality evidence appropriately, simply by assigning a lower weight to such evidence when calculating a weighted average. Fifth, and finally, the veiled premise of the garbage-in-garbage-out argument—that only randomized trials are reliable while non-randomized studies are unreliable—is false. All inductive methods are potentially unreliable.

This last point hints at what is right about the garbage-in-garbage-out argument. If the available evidence from randomized trials suffers from shared systematic biases, then a meta-analysis on that evidence will be systematically biased. Entire domains of medical research suffer from the same systematic biases. For example, all randomized trials on the effectiveness of antidepressants use one of very few scales for measuring the severity of depression, and I argue in Chapters 8 and 9 that such scales are systematically biased toward overestimating the benefits and underestimating the harms of antidepressants—thus any meta-analysis in this domain will be biased. Similarly, any domain in which publication bias is rampant will render meta-analyses in that domain systematically biased.[14] Unfortunately, publication bias is rampant in medicine (Chapter 10).

In short, although all evidence is inductively risky, there are good reasons for including as much evidence as possible in a meta-analysis, though one must be wary of systematically biased evidence. Regardless, when performing a meta-analysis one must make a decision regarding the breadth of methodological quality to include, and this decision can be made differently by different analysts—this is one feature that makes meta-analyses malleable.

Besides methodological quality, there are other properties of medical studies that can vary, and when performing a meta-analysis one must determine the heterogeneity of such properties that one is willing to accept. Some limitation of the diversity of evidence that gets included in a meta-analysis is justifiable. The Cochrane group gives the following proviso: "Meta-analysis should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes"

[13] For example, if Tamara, a specialist in ocean geography, tells me that Kiribati is an island in the Atlantic, then I have some evidence that Kiribati is indeed an island in the Atlantic; but if I later get evidence that Tamara is a compulsive liar then I have lost my reason to believe that Kiribati is an island in the Atlantic. Attending to some of my evidence (Tamara's claim) and ignoring other evidence (about Tamara's honesty) would lead me to believe something false.

[14] See (Jukola, 2015) for a criticism of meta-analysis that focuses on these issues.

(Cochrane Handbook 9.5.1). 'Outcomes' here refers to which parameters are measured; for example, if one study tests the effect of a drug on lowering blood pressure, and another study tests the effect of the same drug on the rate of heart attacks, then there is no shared outcome on which to calculate an average. If multiple studies do not measure the same parameters then there is no sense in calculating an average value of those parameters.

However, sometimes analysts assess heterogeneity among study designs by assessing the statistical variability of the data between studies: high statistical variability, according to this approach, suggests substantive heterogeneity in study designs (the questionable assumption seems to be that a single type of causal relation should generate relatively homogeneous data among subjects in different trials). It would be odd to decide to not perform a meta-analysis simply because of the variability of data between studies, because such data could be produced by a single causal relation that in fact generated variable data. In any case, as the Cochrane group rightly states, deciding whether or not a meta-analysis should be performed requires a judgment regarding the substantive or statistical homogeneity of the relevant studies. Analysts can demarcate the boundary between those studies that are deemed homogeneous and those outside the homogeneous set in a relatively unconstrained manner.

A similar consideration applies to assessing homogeneity of participants and interventions. If we are interested in the effect of a given intervention, we must be consistent with what that intervention is—although a narrow range of intervention diversity (say, using a single dose of an experimental drug) will narrow the range of conclusions one can draw about the intervention. Likewise for the use of a narrow range of participants: before we can know if an intervention works in a broad demographic range, it is reasonable to try to determine if it works in a narrow demographic range. Moreover, some interventions only have a specific effect in a narrow range of subject diversity.[15] Thus, there can be good reasons for limiting the diversity of participants, interventions, and kinds of outcomes to be included in a meta-analysis. In any case, such parameters of meta-analyses are decision points that can influence the outcomes of a meta-analysis.

In sum: there are a plurality of relatively unconstrained decisions regarding what evidence to include that analysts must make when performing a meta-analysis. The worry is that such choices can vary between analysts, and such differences can affect the outcome of a meta-analysis.

Another choice that must be made regarding which primary evidence to include in a meta-analysis is the degree of discordance—that is, the degree to which evidence from different primary studies disagree or contradict each other—that the analyst is willing to accept amongst the primary set of evidence. The Cochrane Handbook has a section that discusses strategies for dealing with discordant evidence (9.5.3). An examination of

[15] On the other hand, Epstein (2007) argues that our knowledge of the effectiveness and harm profile of many medical interventions is limited because these interventions have been tested on a narrow demographic range of subjects.

these strategies is revealing. One strategy is to 'explore' the discordance: discordance might be due to systematic differences between studies, and so an analysis can be done to determine if differences between studies are related to differences in outcomes. Another strategy is to exclude studies from the meta-analysis: the handbook claims that discordance might be a result of several outlying studies, and if a factor can be found that might explain the discordance, then those outliers can be excluded. The handbook notes, however, that "Since usually at least one characteristic can be found for any study in any meta-analysis which makes it different from the others, this criterion is unreliable because it is all too easy to fulfill." Indeed, a study can be similar or dissimilar to another in an infinite number of ways, and so if one had sufficient data and resources, one could always find a potential difference-maker about a study. Each of these strategies for dealing with discordance can be pursued in a multitude of ways, with varying amounts of time and energy devoted to the particular strategies. The extent of discordance deemed acceptable in a meta-analysis is something that can be freely decided upon. Differing approaches to discordance can have a direct effect on the outcomes of meta-analyses.

Decisions regarding what primary evidence to include in a meta-analysis are constrained by what primary evidence is available. A well-known problem in medical research is publication bias: studies that show positive findings are more likely to be published than studies that have null or negative findings (Chapter 10). An illustrative example is provided by Whittington et al. (2004), who shows that the risk-benefit profile of some SSRIs for the treatment of childhood depression is positive when considering only published studies but negative when both published and unpublished studies are evaluated. Reviewers performing a meta-analysis often have less access to evidence that suggests that an intervention is ineffective or harmful (because it is unpublished) than they do to evidence that suggests the intervention is effective, and this can influence the results of a meta-analysis because publication bias systematically favors medical interventions.[16]

In sum, a number of decisions must be made regarding which studies to include in a meta-analysis, including the acceptable range of methodological quality of studies, the acceptable range of study parameter diversity, whether or not to exclude studies with outlying data, and if publication bias is severe or not. In terms of the principles described in §6.2, the plurality of required decisions regarding which studies to include in a meta-analysis threatens *objectivity*, and thereby *constraint*. Decisions regarding the choice of primary evidence to be included in a meta-analysis must be based on judgment, thereby inviting idiosyncrasy and allowing a degree of latitude in the results of a meta-analysis. This renders meta-analysis malleable.

[16] The issue of which primary studies to include in a meta-analysis is appealed to by analysts when explaining contradictory outcomes between their own meta-analysis and other meta-analyses. For instance, in the report by Bachand et al. (2010)—one of the meta-analyses that tested if formaldehyde exposure causes leukemia, discussed in §6.3—the authors claimed that their finding contradicted that of an earlier meta-analysis because of a difference in selection of primary studies.

6.4.2 Choice of outcome measure

Data from primary studies must be summarized quantitatively by an outcome measure before being amalgamated into a weighted average. An outcome measure is used to compute an effect size, which is an estimate of the magnitude of the purported strength of the causal relation under investigation. Multiple outcome measures can be used for this—including the risk difference, relative risk, and relative risk reduction (I give examples of these below, and discuss outcome measures in detail in Chapter 8, where I criticize the use of relative measures such as the risk ratio, and argue that absolute measures such as risk difference should always be employed). The choice of outcome measure can influence the degree to which the primary evidence appears concordant or discordant, and so ultimately the choice of outcome measure influences the results of meta-analysis.

As discussed above, the Cochrane group gives several strategies for dealing with discordant primary evidence. One of these strategies is to change the outcome measure when faced with discordance. Because of the mathematical relationship between ratios and differences, discordant relative effect sizes can entail concordant absolute effect sizes, and vice versa. A hypothetical case will help me illustrate this.

Consider two studies (1 and 2), each with two experimental groups (E and C), and each with a binary outcome (Y and N). Table 6.1 indicates the possible outcomes for each study, where the letters (a–d) are the numbers for each outcome in each group.

The risk ratio (RR) is defined as:

$$RR = \left[a/(a+b)\right]/\left[c/(c+d)\right]$$

The risk difference (RD) is defined as:

$$RD = a/(a+b) - c/(c+d)$$

Now, suppose for Study 1 the numbers for the two outcomes in each group are a = 1, b = 1, c = 1, d = 3 and for Study 2 they are a = 6, b = 2, c = 3, d = 5. This would give the following effect sizes for the two studies:

RR of study 1 = 2

RR of study 2 = 2

RD of study 1 = 0.25

RD of study 2 = 0.375

Table 6.1. A 2 × 2 table for defining binary outcome measures

	Group	**Outcome**
	Y	N
E	a	b
C	c	d

Thus, these two studies, using risk difference as the outcome measure, have discordant effect sizes (0.25 and 0.375); but by switching the outcome measure to risk ratios the studies have concordant effect sizes (2 and 2). Although the Cochrane group advises changing the outcome measure if the primary studies have discordant effect sizes, choosing between outcome measures on the basis of trying to avoid discordance is ad hoc. Although it may be true that evidence from multiple studies appears discordant only because one outcome measure is used rather than another, it might not be true: discordance might simply be due to a lack of systematic effect by the intervention.

More to the point, the choice of outcome measure is another decision in which personal judgment is required, and the fact that there are multiple outcome measures allows a range of possible outputs for any meta-analysis. Again, this threatens *objectivity*, since some analysts might choose to change their outcome measure when the primary evidence appears discordant using the originally chosen outcome measure, while other analysts might resist such switching. One's choice of outcome measure has a direct influence on the outcome of a meta-analysis, and thus differing choices of outcome measures directly threatens *constraint*.

6.4.3 Choice of quality assessment tool

Analysts often attempt to account for differences in the size and methodological quality of studies included in a meta-analysis by weighing the studies with a quality assessment tool (QAT).[17] The conclusion of a meta-analysis depends on how the primary evidence is weighed, because the weights are used as a multiplier when the effect sizes are averaged.

There are many features of evidence that should influence how primary evidence is weighed, including features that are relevant to both the internal validity and the external validity of studies. In Chapter 7 I argue that scientists lack principles to determine how these features should be weighed relative to each other. The trouble is that different weighing schemes can give contradictory results when evidence is amalgamated. An empirical demonstration of this was given by a research group that amalgamated data from seventeen trials testing a particular intervention, using twenty-five different tools to assess study quality (thereby effectively performing twenty-five meta-analyses). These quality assessment tools varied in the number of assessed study attributes, from a low of three attributes to a high of thirty-four, and varied in the weight given to the various study attributes. The results were troubling: the amalgamated effect sizes between these twenty-five meta-analyses differed by up to 117 percent —*using exactly the same primary evidence*.[18]

[17] In Chapter 5 I argue that QATs are superior to evidence hierarchies for assessing evidence, but in Chapter 7 I focus on QATs and argue that they face their own fundamental problems.

[18] Reported in (Juni, Witschi, Bloch, & Egger, 1999). The authors concluded that "the type of scale used to assess trial quality can dramatically influence the interpretation of meta-analytic studies." In Chapter 7 I note more demonstrations of low inter-tool reliability of QATs.

Not only does the choice of quality assessment tool dramatically influence the results of meta-analysis, but so does the choice of analyst using these tools. A quality assessment tool known as the 'risk of bias tool' was devised by the Cochrane group. To test this tool, Hartling et al. (2009) distributed 163 manuscripts of randomized trials among five reviewers, who assessed the quality of the trials with this tool. They found the inter-rater agreement of quality assessments to be very low. In other words, even when given a single quality assessment tool, and a narrow range of methodological diversity, there was a wide variability in assessments of trial quality.

In short, when performing a meta-analysis, analysts must choose a quality assessment tool and apply the tool to the assessment of particular primary-level studies. The choice of quality assessment tool and variations in the assessments of quality by different analysts violate *OBJECTIVITY*, and this threatens *CONSTRAINT*: differing decisions regarding one's quality assessment tool lead to contradictory outcomes of a meta-analysis.

6.4.4 Choice of averaging technique

Once effect sizes are calculated for each study, two common ways to determine the average effect size are possible: sub-group averages and pooled averages. In a pooled average, all subjects from the included studies are merged in the analysis as if they were part of one large study with no distinct demographic sub-groups. A problem with the pooled average approach is that different demographic groups might respond differently to an intervention. For example, a drug might, on average, have a large benefit to males and a small harm to females, and if data from these groups were combined in a pooled average we would erroneously conclude that the drug has, on average, a small benefit to all people, including females.

Maintaining distinct sub-groups in a meta-analysis, which the Cochrane group rightly advises, is an attempt to avoid such problems. However, to determine a sub-group average, either the sub-groups must be consistently demarcated amongst primary studies, or the patient-level data necessary to demarcate sub-groups, such as age and gender, must be available to the analyst. The former is often not the case and the latter is often not available. However, if patient-level demographic data is available, then the analyst can demarcate sub-groups any way she wishes until she finds something interesting, but of course such retrospective data-dredging is liable to support spurious findings (see Chapter 10 for a discussion of this practice, sometimes called p-hacking). More to the point: the choice of average type—pooled or sub-group, and if the latter, the choice of sub-groups—is another decision point in meta-analysis that threatens *OBJECTIVITY* and *CONSTRAINT*. It is another feature that makes meta-analysis malleable.

6.5 The Hill Strategy

An older tradition of evidence in medicine, associated with the epidemiologist Sir Bradford Hill, provides a more compelling way to consider the variety of evidence

in medical research. Hill was one of the epidemiologists involved in the first large case-control studies during the 1950s that showed a correlation between smoking and lung cancer.[19] The statistician Ronald Fisher had noted the absence of controlled experimental evidence on the association between smoking and lung cancer. Fisher's infamous criticism was that the smoking-cancer correlation could be explained by a common cause of smoking and cancer: he postulated a genetic predisposition that could be a cause of both smoking and cancer, and so he argued that the correlation between smoking and cancer did not show that smoking caused lung cancer. The only way to show this, according to Fisher, was to perform a controlled experiment; of course, for ethical reasons no such experiment could be performed. Hill responded by appealing to a plurality of kinds of evidence that, he argued, when taken together made a compelling case that the association was truly causal.

The evidence that Hill cited as supporting this causal inference was: strength of association between measured parameters; consistency of results between studies; specificity of causes (a specific cause has a specific effect); temporality (causes precede effects); a dose-response gradient of associations between parameters; a plausible biological mechanism that can explain a correlation; coherence with other relevant knowledge, including evidence from laboratory experiments; evidence from controlled experiments; and analogies with other well-established causal relations.[20] Hill considered these as inferential clues, or as epistemic desiderata for discovering causal relations. Although Hill granted that no single desideratum was necessary or sufficient to demonstrate causality, he claimed that jointly the desiderata could make for a good argument for the presence of a causal relation.[21] The important point for the purpose of contrast with meta-analysis is the plurality of reasons and sources of evidence that Hill appealed to.

The desiderata appealed to by Hill depend on diverse kinds of evidence, which lack a shared quantitative measure, so that the evidence cannot be combined by a simple weighted average of numerical effect sizes. Versions of the problems I raised for meta-analysis apply to Hill's approach—especially the choice of primary evidence to include, the choice of measures to quantify the evidence (at least, the evidence that can be quantified), the choice of a quality assessment scale to assess or weigh the evidence, and the choice of averaging technique—these are troublesome for the Hill strategy. But this strategy can at least be used to consider all available evidence, at least in principle.

One can have evidence that satisfies only some of the desiderata while still having ample justification for causal inference. Moreover, unlike meta-analysis, there is no

[19] (Doll & Hill, 1950, 1954).

[20] Meta-analysis can be thought of as a formal technique to assess the 'consistency' criterion. Framing meta-analysis this way shows just how much meta-analysis neglects, but also shows that it can be a useful technique nevertheless.

[21] See (Doll, 2003). Howick, Glasziou, and Aronson (2009) restructure Hill's desiderata, and Rothman and Greenland (2005) offer a brief discussion of each of the desiderata. Woodward (2010) more thoroughly discusses the specificity desideratum. Of these desiderata, temporality is plausibly a necessary condition for a causal relation.

simple algorithm to amalgamate the diverse forms of evidence in Hill's approach. There is, then, malleability in the Hill strategy. As I argued above, meta-analysis itself is malleable. The complexity of assessing and amalgamating a large volume and diversity of evidence might inevitably require malleable methods. But the Hill strategy is more constraining than meta-analysis in some respects. If a meta-analysis supports a hypothesis while most of Hill's desiderata provide evidence against the hypothesis, this ought to warrant serious reservation in this hypothesis. Conversely, if most of the desiderata coherently support a particular hypothesis, this is suggestive that the hypothesis is roughly correct.[22] Endorsing the Hill strategy, then, does not necessarily mean endorsing a more tolerant or relaxed attitude toward amalgamating evidence. However, nothing very general can be said regarding when the satisfaction of the desiderata is sufficient to infer causality—the Hill approach requires judgment, just as meta-analysis requires. Both approaches to amalgamating evidence are malleable.

6.6 Conclusion

I have argued that meta-analyses fail to adequately constrain assessments of the effectiveness of medical interventions. This is because the numerous decisions that must be made when designing and performing a meta-analysis require judgment and expertise, and allow biases and idiosyncrasies of reviewers to influence the outcome of the meta-analysis. The failure of *objectivity* at least partly explains the failure of *constraint*: the many judgments required for meta-analysis explain how multiple meta-analyses of the same primary evidence can reach contradictory conclusions.

There are better and worse ways to perform a meta-analysis. Though I have used the published guidance from the Cochrane group to frame my criticisms, this group has worked to improve the quality of meta-analyses.[23] I appeal to meta-analyses throughout this book.[24] Meta-analysis, when done well, is a valuable method in medical research.

Nevertheless, the general epistemic importance given to meta-analysis is unjustified, since it is so malleable: meta-analysis allows unconstrained choices to influence its results, which in turn explains why the results of meta-analyses are unconstrained. The upshot, one might claim, is merely to urge the improvement of the quality of meta-analyses in ways similar to that already proposed by evidence-based medicine

[22] For instance, in §6.3 I discussed meta-analyses that tested whether formaldehyde exposure causes leukemia. One of these (Zhang et al., 2009) concluded that formaldehyde exposure is indeed associated with leukemia, and in addition to this analysis the authors proposed possible causal mechanisms meant to undergird the outcome of their meta-analysis, thereby appealing to the coherence and plausibility desiderata.

[23] Meta-analyses that are not performed by Cochrane collaborators are twice as likely to have positive conclusions compared with meta-analyses performed by Cochrane collaborators (Tricco, Tetzlaff, Pham, Brehaut, & Moher, 2009). Assuming that Cochrane meta-analyses were higher quality than non-Cochrane meta-analyses (a generally safe assumption), it follows that better meta-analyses are less likely to have a positive conclusion regarding a medical intervention.

[24] For instance, in Chapter 9 I cite a prominent meta-analysis which shows that the drug rosiglitazone causes serious harms (Nissen & Wolski, 2007), and in Chapter 11 I cite another prominent meta-analysis which shows that SSRIs are nearly ineffective for treating depression (Kirsch et al., 2008).

methodologists, in order to achieve more constraint. However, my discussion of the many particular decisions that must be made when performing a meta-analysis indicates that such improvements can only go so far. For at least some of these decisions, the choice between available options is arbitrary; the various proposals to enhance the transparency of reporting of meta-analyses are unable, in principle, to referee between these arbitrary choices (in Chapter 7 I argue that this is the case for many aspects of medical research generally).

One of the criticisms I raised against meta-analysis is its reliance on a narrow range of evidential diversity. An older tradition of evidence in medicine, associated with Sir Bradford Hill, is in this respect superior. However, there is no structured method for assessing, quantifying, and amalgamating the very disparate kinds of evidence that Hill considered. Thus the Hill strategy lacks the apparent objectivity, methodological simplicity, and quantitative output of meta-analysis. But given the central argument of this chapter, the fact that the Hill strategy lacks a simple method of objectively amalgamating diverse evidence is not a strike against it relative to meta-analysis, since I have argued that the objectivity of the latter is a chimera. Both approaches to amalgamating evidence in medicine are malleable.