Is Meta-Analysis the Platinum Standard of Evidence?

Jacob Stegenga

March 24, 2011
Forthcoming in *Studies in History and Philosophy of Biological and Biomedical Sciences*

Abstract

An astonishing volume and diversity of evidence is available for many hypotheses in the biomedical and social sciences. Some of this evidence – usually from randomized controlled trials (RCTs) – is amalgamated by meta-analysis. Despite the ongoing debate regarding whether or not RCTs are the 'gold-standard' of evidence, it is usually meta-analysis which is considered the best source of evidence: meta-analysis is thought by many to be *the platinum standard of evidence*. However, I argue that meta-analysis falls far short of that standard. Different meta-analyses of the same evidence can reach contradictory conclusions. Meta-analysis fails to provide objective grounds for intersubjective assessments of hypotheses because numerous decisions must be made when performing a meta-analysis which allow wide latitude for subjective idiosyncrasies to influence its outcome. I end by suggesting that an older tradition of evidence in medicine – the plurality of reasoning strategies appealed to by the epidemiologist Sir Bradford Hill – is a superior strategy for assessing a large volume and diversity of evidence.

Keywords

Meta-analysis, evidence, medicine, randomized controlled trial (RCT), Sir Bradford Hill, epidemiology

1. Introduction

Biomedical and social scientists are faced with a daunting volume of evidence for many hypotheses of interest. For example, by 1985 there had been over 700 studies on the relationship between class size and academic achievement, over 800 studies on the effectiveness of psychotherapy, and 120 studies testing if the phase of the moon affects human behavior.[1] The diversity of evidence available for many hypotheses in medicine and the social sciences is also daunting. Standard hypotheses regarding contemporary pharmaceutical interventions, for example, have evidence from computational models of toxicity, cell-based studies, experiments on multiple animal species (murine and canine, and sometimes primate and porcine) investigating multiple organ systems, and multiple kinds of study designs on humans. This avalanche of a large volume and diversity of evidence contributed to the formation of groups dedicated to the systematic review of evidence (such as the Cochrane Collaboration), to journals which publish reviews of existing evidence rather than evidence from original research (e.g. *Annual Review of Genetics* or *Epidemiologic Reviews*), and to methods of amalgamating evidence, including social methods, such as consensus conferences, and formal methods, such as meta-analysis. My focus in this paper is on meta-analysis. I describe the purported virtues 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, contrary to the standard view regarding the epistemic status of meta-analysis, meta-analysis does not have the virtues that many claim for it.

Here is the definition from the U.K. National Health Service:

> **Meta-analysis**: a mathematical technique that combines the results of individual studies to arrive at one overall measure of the effect of a treatment.

A frequent goal of using meta-analysis is to discover causal relationships and to determine the magnitude of an effect for a particular magnitude of a purported cause. To

[1] See, e.g., Smith and Glass (1977), Glass and Smith (1979), and Rotton and Kelly (1985).

achieve this end when faced with a huge volume and diversity of evidence, many claim that, given its methodological virtues, meta-analysis is an especially good method (§2). I identify these methodological virtues as two general norms for any method of amalgamating evidence: Constraint – the use of meta-analysis should constrain intersubjective assessments of hypotheses – and Objectivity – meta-analysis should be performed in a way which limits the influence of subjective biases and idiosyncrasies of particular researchers.

I describe several cases to show that the use of meta-analysis often fails to achieve Constraint (§3). Meta-analysis fails to constrain intersubjective assessments of hypotheses 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. Some of these decisions are required for any method of amalgamating evidence while others are particular to the technical details of meta-analysis. The bulk of my argument involves a close examination of these decisions involved in the methodological details of meta-analysis (§4). Meta-analysis is performed by i) selecting which primary studies are to be included in the meta-analysis, ii) calculating the magnitude of the effect due to a purported cause for each study, iii) assigning a weight to each study, which is often determined by the size and the quality of the study, and then iv) calculating a weighted average of the effect magnitudes. Although meta-analysis is often used in the biological, human, and social sciences, my focus is on medical research. I draw on the published guidance of the Cochrane Collaboration, a primary institution of the so-called 'evidence-based medicine' movement which commissions a large number of meta-analyses, to help describe the methodology of meta-analysis. Finally, I end by discussing an alternative, older, and arguably better strategy for assessing a large volume and diversity of evidence (§5), associated with the epidemiologist Sir Bradford Hill (1897 - 1991).

Many arguments have been proposed debating whether or not randomized controlled trials (RCTs) provide the best evidence for causal hypotheses in medicine and

the social sciences.[2] Nancy Cartwright (2007), for instance, asks "Are RCTs the gold standard?" to which she answers 'no'. However, despite the debates surrounding the gold-standard status of RCTs, it is in fact meta-analysis which is at the top of the most prominent evidence hierarchies in medicine and social policy.[3] Coining a neologism analogous to the metaphor of the gold-standard, it is widely thought that *meta-analysis is the platinum standard of evidence*. In what follows I criticize the purported platinum standard status of meta-analysis.

2. Constraint and Objectivity

The first comprehensive meta-analysis performed on a single hypothesis with evidence from multiple sources was about extra-sensory perception (Rhine et al. 1940).[4] Meta-analysis later became the platinum standard of evidence in medicine and the social sciences for several reasons. The sheer volume of available evidence meant that most users of evidence (e.g. physicians or 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 recently the number of published meta-analyses has exceeded two thousand per year (Sutton and Higgins 2008).

Meta-analysis became a prominent method in part due to the purported rigor of meta-analyses compared with qualitative methods of amalgamating evidence. In contrast with qualitative literature reviews and social methods of amalgamating evidence such as consensus conferences, meta-analyses have both quantitative inputs and outputs. The

---

[2] See, e.g., Worrall (2002), Worrall (2007), Borgenson (2008), Banerjee and Duflo (manuscript), Duflo and Kremer (manuscript), Deaton (2008), Cartwright (2007), and Cartwright (2010).

[3] Meta-analysis is that the top of the evidence hierarchies in the evidence ranking schemes of the Oxford Centre for Evidence-Based Medicine, the Scottish Intercollegiate Guidelines Network, and the Australian National Health and Medical Research Council). As I discuss below, however, those meta-analyses which are usually considered to be the best are those which include only RCTs.

[4] This is a nice historical accident, because Ian Hacking (1988) showed that the practice of randomizing subjects into different groups also began in psychical research – thus both our gold standard of evidence and our platinum standard of evidence come from research in paranormal psychology.

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: "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 of such a review is (Pauling 1986), in which the Nobel Laureate cited dozens of his own studies supporting his pet hypothesis that large doses of vitamin C can reduce the risk of catching a cold, and yet he did not cite any studies contradicting this hypothesis, though several had been published (Knipschild 1994). Similarly, a recent textbook on meta-analysis worries that unsystematic reviews (sometimes called 'narrative reviews') can fail to constrain intersubjective assessments of hypotheses: "there are examples in the literature where two narrative reviews come to opposite conclusions, with one reporting that a treatment is effective while the other reports that it is not" (Borenstein et al. 2009). The solution to this problem, according to the authors of this textbook, is to use meta-analysis, a more formal method which (it is claimed) can constrain intersubjective assessments of hypotheses. Likewise, a recent statistics textbook emphasizes a worry regarding reviewers' personal 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." These authors suggest that meta-analysis is superior in this regard, since "it is extremely difficult to balance multiple studies by intuition alone without quantitative tools" (Whitlock and Schluter 2009). The quantitative tool most often used to achieve such a 'balance' of multiple studies in medicine (and the social sciences) is meta-analysis.

The best account of the scientific value of meta-analysis is rather simpler than one might suppose. One might think that an aim of meta-analysis is to satisfy a principle stipulating the consideration of all available evidence for a hypothesis (such as Carnap's "Principle of Total Evidence"). However, as I argue below, meta-analyses violate such a principle because they normally include only a small fraction of available evidence. Alternatively, one might think that an aim of meta-analysis is to satisfy a principle of robustness: hypotheses are often said to be more likely to be true if they are supported by

evidence from multiple independent sources.[5] However, because meta-analyses usually include only evidence from a narrow range of methodological diversity (such as RCTs), such evidence typically fails to be methodologically independent, which is often said to be a requirement of robustness arguments. One proposal to amalgamate diverse evidence is to use the evidence to build causal models, or models of a network of interconnected causal relations (Danks 2005; Cartwright and Stegenga 2011). Accordingly, one might think that an aim of meta-analysis is to construct causal models. But meta-analyses amalgamate evidence on a single causal relation, not on a network of interconnected causal relations.

Instead, the best justification or explanation of the value of meta-analysis is statistical: many purported causes in medicine and the social sciences have a small observable effect, and so when analyzing data from a single study on an intervention with a small effect, there might be no statistically significant difference between the experimental group and the control group. But by pooling data from multiple studies the sample size of the analysis increases, which tends to decrease the width of confidence intervals, thereby potentially rendering estimates of the magnitude of an intervention effect more precise, and perhaps statistically significant. One aim of meta-analysis, then, is quantitative precision. Such quantitative precision is perhaps best construed as a means to the end of constraint on intersubjective assessments of hypotheses.

In short, meta-analysis is a method to assess and amalgamate evidence from multiple studies. Relative to other methods of amalgamating evidence, such as informal literature reviews or social methods like consensus conferences, meta-analysis is said to have the virtues of constraining intersubjective assessments of hypotheses and doing so in a way which is not infused with the subjective idiosyncrasies of the analysts. The purported rigor, transparency, quantitative precision, and freedom from personal bias can be summarized by these two general norms for any method of amalgamating evidence:

> Constraint: An evidence amalgamation method should constrain intersubjective assessment of hypotheses.

[5] See, e.g., Wimsatt (1981), Trout (1995), Thagard (1998), Douglas (2004), and Stegenga (2009).

> Objectivity: An evidence amalgamation method 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: an evidence amalgamation method can constrain intersubjective assessments of hypotheses only if it is not sensitive to analysts' idiosyncratic or personal biases. It is beyond the scope of this paper to provide a full explication and assessment of these two norms.[6] Nevertheless, they are, *prima facie*, worthwhile norms for any method of amalgamating evidence. The important point for my present purpose is that statisticians, institutions of evidence-based medicine, and other defenders of meta-analysis claim that, compared with other methods of assessing and amalgamating a large volume of evidence, meta-analysis best satisfies these norms. This is the basis of the purported platinum standard status of meta-analysis.

However, in the following section I argue that meta-analysis, unfortunately, often fails to satisfy these norms (§3). In §4 I argue that 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.

3. Failure of Constraint

Epidemiologists have recently noted that multiple meta-analyses on the same hypotheses, performed by different analysts, can reach contradictory conclusions. For example, there have been numerous inconsistent studies on the benefits and harms of a newer synthetic dialysis membrane versus an older cellulose membrane for patients with acute renal failure: one recent meta-analysis of these studies found greater survival of such patients using the newer synthetic membrane compared with those using the older cellulose membranes (Subramanian et al. 2002), while another meta-analysis reached the opposite conclusion (Jaber et al. 2002). Here is another example. Two meta-analyses

[6] Recent excellent scholarship has investigated the notion of objectivity, both from a historical perspective (e.g., Daston and Galison 2007) and from a philosophical perspective (e.g., Douglas 2004).

published in the same issue of the *British Medical Journal* came to contradictory conclusions regarding whether or not an association exists between the use of selective serotonin reuptake inhibitors (SSRI, a common class of antidepressant) and suicide attempts. In the meta-analysis reported by Gunnell et al. (2005), there was no association between SSRI use and suicide attempts, and only a weak association between SSRI use and risk of self harm. In contrast, in the meta-analysis reported by Fergusson et al. (2005), there was a relatively strong association between SSRI use and the suicide attempts. Similarly, 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 (see e.g. Linde and Willich 2003).

There is good reason to think that differential outcomes between contradictory meta-analyses are associated with the analysts' professional or financial affiliations. Several meta-analyses have recently been published which amalgamate evidence testing if formaldehyde exposure causes leukemia. Bachand et al. (2010) and Collins and Lineker (2004) conclude that formaldehyde exposure does not cause leukemia. In contrast, Bosetti et al. (2008) found a modest elevation of risk of developing leukemia in professionals who work with formaldehyde, such as pathologists and embalmers. Zhang et al. (2009) found an even higher risk of developing leukemia among professionals who work with formaldehyde. The meta-analyses which concluded that formaldehyde exposure is not associated with leukemia were performed by employees of private consulting companies.[7] In contrast, the authors of the two meta-analyses that found some evidence for a causal link between formaldehyde exposure and leukemia worked in academic and government institutions.[8] Lest readers think this is a crude *ad hominem* anecdote regarding an isolated example, consider the following similar cases.

Barnes and Bero (1998) performed a quantitative assessment of multiple meta-analyses which reached contradictory conclusions regarding the same hypothesis, and

[7] In the case of Collins and Lineker (2004) one of the authors was an employee of The Dow Chemical Company. An organization representing the chemical industry estimates that formaldehyde exists in products which account for more than 5% of the U.S. gross national product (cited in Zhang et al. 2009).

[8] I am grateful to Heather Douglas for bringing this example to my attention. She should not, of course, be held responsible for my interpretation of the case.

found a 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: thirty-nine of these reviews concluded that passive smoking is not harmful to health, and the remaining 67 concluded that there is at least some adverse health effect associated with 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 88 times more likely to conclude that passive smoking has no adverse health effects compared with analysts who had not received tobacco funding.

Here is yet 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 (Yank et al. 2007). Or consider the meta-meta review of meta-analyses of studies on spinal manipulation as a treatment for lower back pain: some meta-analyses of this intervention 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 (Assendelft et al. 1995).

Such examples could easily be multiplied. I have made no attempt to comprehensively document the cases in which multiple meta-analyses on the same hypothesis reach contradictory conclusions. These examples are merely meant to show that multiple meta-analyses of the same primary set of evidence can reach contradictory conclusions, not that they must, or even often do, reach contradictory conclusions. The examples suggest that idiosyncratic features of analysts influence the results of meta-analyses. Moreover, the features of meta-analysis which 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 the methodology common to all meta-analyses. I now turn to a detailed examination of the methodology of meta-analysis.

4. Inherent Subjectivity

The failure of Constraint in the above cases is at least partially a consequence of the failure of Objectivity: constraint on intersubjective assessments of hypotheses is not met by the meta-analyses in §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 most importantly include the:

i. Choice of primary evidence
ii. Choice of effect measure
iii. Choice of quality assessment scale
iv. Choice of averaging technique

Some of these choices are not specific to meta-analysis (i and perhaps iii), but are nevertheless relevant to explaining the shortcomings of meta-analysis, while others are particular to the technicalities of meta-analysis (ii and perhaps iv). The general principles of meta-analysis are simple and are not unique to the biomedical or social sciences. For example, a common method of combining multiple expert probability forecasts (say, for sunshine in three days, or for a stock price increase in the next fiscal quarter, or for a victory for a presidential candidate) is to calculate a statistical average: when multiple experts give probability forecasts, a standard way to combine these multiple forecasts into a single forecast is to simply calculate an average of the probabilities.
However straightforward a weighted average may seem, the subtleties of meta-analysis are complex. In what follows I consider each class of choices required in the steps of a meta-analysis.

4.1 Choice of Primary Evidence

Multiple decisions must be made regarding what primary evidence to include in a meta-analysis. I survey some of these decisions, and critically evaluate arguments for particular strategies to these decisions.

4.1.1 Methodological Quality

The dominant view in evidence-based medicine is to include only evidence from RCTs in a meta-analysis; according to a statement of leaders in evidence-based medicine, in a meta-analysis "researchers should consider including only controlled trials with proper randomisation" (Egger, Smith, and Phillips 1997). Such a view excludes other common kinds of statistical evidence, including that from cohort studies and case-control studies, as well as non-statistical evidence which is not in the domain of usual technical meta-analyses, such as pathophysiological evidence, and evidence from animal experiments, mathematical models, and clinical expertise.

In contrast, others argue that an evidence amalgamation method should use all available evidence. Glass (1976), for instance, claims that an effect size of 2.0 x from 3 RCTs testing a purported causal relation should have a different impact on one's assessment of the causal hypothesis when considered in the light of (i) 50 matched case-control studies, purportedly testing the same causal relation as the RCTs, that show an effect size of 2.2 x, versus (ii) 50 matched case-control studies, purportedly testing the same causal relation as the RCTs, that show an effect size of -0.8 x. A standard argument supports Glass's contention: if one's assessment of the causal hypothesis were *not* different in the two scenarios, one would effectively be committing the base-rate fallacy: one's assessment of a hypothesis after observing new evidence should also be guided by all of one's previous evidence, and if it is not then one is liable to make an ill-formed judgment of the probability that the hypothesis is true in light of the new evidence.

Here is another argument to support Glass's contention. In (i) there is concordance between the new evidence (from RCTs) and the previous evidence (from case-control studies), which might suggest that the two kinds of studies are converging on a true effect size (but such concordance can occur for other reasons). In (ii) there is *dis*cordance between the new evidence (from RCTs) and the previous evidence (from case-control studies), which might suggest (a) that there is a systematic problem with the case-control studies, given the known potential biases with case-control studies compared with RCTs (this is a typical response in the evidence-based medicine community when faced with discordance between RCTs and case-control studies), (b) that there is a systematic problem with the RCTs, given the low number of them compared with the large number of case-control studies, (c) that the two kinds of studies were not similar

enough in all important parameters, including the causal structure of the study populations, (d) that the purported cause is spurious, (e) that a highly unlikely series of events has occurred. In other words, in (ii) there is no *general* reason to assume (a) as an explanation of the discordance, and if one blindly does assume (a) as an explanation then one is liable to be wrong.

Another way to put this consideration is that even if RCTs are justifiably the gold standard of evidence, that would not mean that evidence from non-randomized studies is negligible. Indeed, some of our most believable causal hypotheses were first supported by evidence from non-randomized studies, and for many hypotheses we only have evidence from non-randomized studies. A joke in such discussions is that there has never been a carefully performed RCT which has tested the causal efficacy of parachutes (e.g. Smith and Pell 2003).

The exclusive use of a narrow range of evidence is purportedly justified on the grounds that the methods of meta-analysis are only valid for homogeneous evidence (I discuss this below), and 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 also be low quality, and so rather than including *all* available evidence, meta-analyses should only include the 'best evidence' (e.g. Slavin (1995), who argues that meta-analysis should be limited to 'best evidence synthesis'). There are numerous problems with this argument, one of which is outlined above: if we ignore some evidence, even if it comes from a method deemed to be of low quality, we effectively commit the base-rate fallacy.[9] Moreover, there is no reason why an analyst cannot assess lower-quality evidence appropriately, simply by assigning a lower weight to such evidence when calculating the

[9] My appeal to the base-rate fallacy here might suggest that I am relying on Bayesian principles. But the problem with ignoring evidence should be a problem for everyone. Worrall (2002) and Cartwright (2007) have forcefully argued that there is no single 'gold standard' of evidence and thus we ought to take into account evidence of all kinds when available. Moreover, the possibility of 'defeating' evidence provides further reason why one ought to consider all available evidence. For example, if Beth, a specialist in ocean geography, tells me that Kiribati is an island nation in the Atlantic, then I have some evidence that Kiribati is indeed an island nation in the Atlantic; but if I later get evidence that Beth is a compulsive liar then I have lost my reason to believe that Kiribati is an island nation in the Atlantic. Attending to some of my evidence (Beth's claim) and ignoring other evidence (about Beth's honesty) leads me to believe something false.

weighted average. Finally, the veiled premise of the garbage-in-garbage-out argument – that all and only non-randomized studies require problematic background assumptions in order for evidence from such studies to be truth-conducive – is false. *All* methods presuppose background assumptions that must be met for the evidence from such methods to be considered truth-conducive, and such assumptions may or may not be problematic, but this depends on specific features of the study design, both in the abstract and in relation to one's hypothesis of interest. In short, although all evidence is inductively risky, there are good reasons for including as much evidence as possible in a meta-analysis. Regardless, when performing a meta-analysis one must make a decision regarding the breadth of methodological quality to include, and this decision might be made differently by different analysts.

4.1.2 Methodological Diversity

Another justification for only including evidence from select methods is the possibility of variable treatment effects among different subjects or different experimental circumstances. Consider the following guidance from the Cochrane Collaboration:

> you have to be confident that clinical and methodological diversity is not so great that we should not be combining studies at all. This is a judgement, based on evidence, about how we think the treatment effect might vary in different circumstances.[10]

For the Cochrane Collaboration, the standard for what counts as methodological diversity is low; these meta-analyses only include a narrow range of study designs in any given review. Some limitation to the diversity of primary evidence which 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" (Cochrane Handbook 9.5.1). Including only studies with homogenous outcomes is fine if by 'outcome' they mean *kind* of outcome; for example, if one study tests the effect of a drug on lowering blood

[10] Cochrane website http://www.cochrane-net.org/openlearning/html/mod13-4.htm, accessed Oct 20, 2009.

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. More generally, a meta-analysis is only meaningful if the data from multiple studies is generated from a single kind of causal relation. But even when multiple studies are purported to measure the same causal relation, the only evidence that analysts have to assess this (besides the substantive features of the study designs) is by the statistical variability between the data from the studies. As the Cochrane group rightly states, this is a 'judgement' regarding whether or not a meta-analysis is even meaningful in the first place.

Homogeneity of participants and interventions might be similarly justifiable. 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, it is reasonable to try to determine if it works in a narrower demographic.[11] (But of course, if we already have evidence from a broader population of subjects, including non-human subjects, then we should not ignore such evidence.) Moreover, some interventions only have a specific effect in a narrow range of subject diversity. Thus, there can be good reasons for limiting the diversity of participants, interventions, and kinds of outcomes to be included in a meta-analysis. Nevertheless, though, such parameters of meta-analyses are decision points which can influence the outcomes of a meta-analysis.

Other limitations to the primary evidence included in a meta-analysis are more troublesome. Consider the following Cochrane guidance: "we strongly recommend that review authors should not make any attempt to combine evidence from randomized trials and NRS [non-randomized studies]" (13.2.1.1). No justification is provided for this limitation; not only is evidence from non-randomized studies not to be amalgamated with evidence from RCTs, but neither is evidence from pathophysiological knowledge, background considerations of underlying mechanisms, animal experiments, and results

[11] I give short shrift to a growing debate: Epstein (2007) argues that our knowledge of the safety and efficacy of many biomedical interventions is limited because for too many years these interventions were tested on a narrow demographic range of subjects.

from mathematical models. Such a practice could limit the external validity of a meta-analysis, since RCTs on humans 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 study parameters at lower cost. Moreover, as discussed above, this practice violates a principle of total evidence, which comes with possibly significant epistemic risk: neglecting other kinds of evidence risks making an uninformed judgment (or, the base-rate fallacy) on a hypothesis.

Methods of amalgamating evidence from multiple studies, but which systematically exclude all evidence but that from a single *kind* of study, are not limited to medicine. A non-medical example is in 'driving under the influence' (DUI) cases. In most jurisdictions in the United States there are at least three kinds of evidence that can be used to detect intoxication of drivers: (1) a police officer's subjective assessment of the driver;[12] (2) the driver's blood alcohol concentration as extrapolated from a portable breath test machine in the officer's car; (3) the driver's blood alcohol concentration as extrapolated from a more reliable breath test machine in a police station (Mnookin 2008). The use of breath test machines is meant to mitigate officers' subjective assessments; to use a term of Daston and Galison (2007), the 'mechanical objectivity' of breath test machines are thought to counter the subjectivity of officers. In many jurisdictions, evidence from (3) trumps evidence from (2) or (1): if a driver is suspected of being intoxicated according to (1), and fails the breath test in (2), but gets to the station and then passes the breath test in (3), the driver is released with no charges. In short, in such cases a single kind of evidence trumps other available kinds of evidence. Thus medicine is not the only domain in which one kind of evidence trumps all other kinds of evidence. However, to the extent that one is committed to the principle of total evidence, one will find such practices dissatisfying.

[12] This subjective assessment is itself comprised of various kinds of evidence, including the driver's ability to perform behavioral tasks, the driver's conversational capability, and the driver's outward appearance and smell.

The obvious worry about the plurality of unconstrained decisions regarding the methodological diversity to be included in a meta-analysis is that such choices can vary between analysts, and if so, such differences might affect the outcome of a meta-analysis.

4.1.3 Discordance

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 Collaboration Handbook has a section which discusses strategies for dealing with discordant primary evidence (9.5.3). An examination of these strategies is revealing. One strategy is to "explore" the discordance: discordance might be due to systematic differences between studies, and so a post-hoc meta-study can be done to determine if systematic differences between studies are related to systematic 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 some factor can be found that might explain the discordance between these outlying studies and the remainder of the studies, 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 study on an infinite number of features, and so if one had sufficient data and resources, one could always find a potential difference-maker about a study that would purportedly justify its exclusion. Finally, when faced with discordant primary evidence, the Cochrane group suggests that a meta-analysis may not be meaningful – "If you have clinical, methodological or statistical heterogeneity it may be better to present your review as a systematic review using a more qualitative approach to combining results..."[13] This is because, as discussed above, the primary evidence might be discordant not because of random variations of measures from a single causal relation, but rather because the multiple primary studies were measuring multiple causal relations.

[13] http://www.cochrane-net.org/openlearning/html/mod13-4.htm Accessed May 13, 2009.

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. There is no reason to think that different analysts will follow these strategies in the same way. Differing approaches to discordance have a direct affect on the outcomes of meta-analyses.

4.1.4 Data Access

Decisions regarding what primary evidence to include in a meta-analysis are constrained by what primary evidence is available. The internet has improved access to primary evidence. Nevertheless, a well-known problem in medical research is publication bias: papers which show statistically significant positive findings are more likely to be published than papers that have null or negative findings (especially when the research is funded by private companies – see Brown (2008)). An illustrative example is provided by Whittington et al. (2004), who showed that the risk-benefit profile of some SSRIs for the treatment of childhood depression is positive when considering only published studies and negative when both published and unpublished studies are evaluated. A corollary of publication bias has its own name: the File Drawer Problem.[14] In short, reviewers performing a meta-analysis often have less access to null or negative evidence (because it is sitting in file drawers or on hard drives) than they do to published positive evidence, and this is likely to influence the results of a meta-analysis (often, it seems, such influence is in the favor of the medical intervention under study).

A related problem is faced by analysts who want to do a meta-analysis with patient-level outcomes (which has several advantages over published study-level outcomes which I do not discuss here): often patient-level data is confidential or is protected by corporate interests. Other practical problems regarding access to primary evidence include studies published in languages foreign to the analyst, and evidence available only in the 'gray literature' of conference proceedings and dissertations; evidence from 'gray literature' tends to have lower estimates of medical interventions than does evidence published in mainstream literature (McAuley et al. 2000). How

[14] That this is not called the Hard Drive Problem suggests that it has been with us for some time.

intensely an analyst grapples with these practical problems of data access can influence the results of a meta-analysis.

4.1.5 Summary

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, how hard to look through the gray literature, if the File Drawer Problem is severe or not, and whether or not a meta-analysis is even feasible in the first place. In the words of a critic of meta-analysis: "It is precisely in those areas where there is most disagreement that these methods [meta-analysis] are least applicable" (Eysenck 1984). In terms of the norms described in §2, the plurality of required decisions regarding which studies to include in a meta-analysis threatens Objectivity, and thereby Constraint. Regardless of how justified the decisions regarding choice of primary evidence are for any particular meta-analysis, they must be based on expertise and judgment, thereby inviting idiosyncrasy, and allowing a degree of latitude in the possible results of a meta-analysis.[15]

4.2 Choice of Effect Measure

Data from primary studies must be summarized quantitatively by a standardized measure, usually referred to as an 'effect measure', before being amalgamated into a weighted average. An effect measure (sometimes also called an outcome measure) is used to summarize data into an 'effect size', which is an estimate of the magnitude of the purported strength of the cause-effect relationship under investigation. Multiple effect

[15] The issue of which primary studies to include in a meta-analysis is often appealed to by analysts when explaining contradictory outcomes between their own meta-analysis and previous meta-analyses. For instance, in the report by Bachand et al. (2010) – one of the meta-analyses testing if formaldehyde exposure causes leukemia, discussed in §3 – the authors claimed that the apparently contradictory outcome of their meta-analysis with the outcome of an earlier meta-analysis was due to a difference in selection of primary studies: "Zhang et al. (2009) identified all relevant epidemiological studies published on formaldehyde and lymphohematopoietic cancer, but due to lack of case-control studies meeting their inclusion criteria, restricted their analysis to cohort and PMR studies."

measures can be used for this – frequent choices include the odds ratio, the risk difference, and the correlation coefficient (I give examples of these below). The choice of effect measure can influence the degree to which the primary evidence appears concordant or discordant, and so ultimately the choice of effect measure influences the results of meta-analysis, and can even influence whether or not an analyst thinks a meta-analysis is worth doing in the first place. The guidance from the Cochrane group will again help me to explain this.

As discussed above, the Cochrane group gives several strategies for dealing with discordant primary evidence. One of these strategies is to "change the effect measure", because discordance "may be an artificial consequence of an inappropriate choice of effect measure." The Cochrane Handbook is correct to claim that "when control group risks vary, homogeneous odds ratios or risk ratios will necessarily lead to heterogeneous risk differences, and vice versa." This is simply due to the mathematical relationship between ratios and differences. However, although it may be true that evidence from multiple studies appears discordant only because one effect measure is used rather than another, it might not be true: heterogeneity might simply be due to a lack of systematic effect by the intervention. A hypothetical case will help me illustrate the trouble with choosing between effect measures based on discordance between primary studies.

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

Group	Outcome	
	Y	N
E	a	b
C	c	d

The risk ratio (RR) is defined as:

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

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:

$RR_1 = 2$; $RR_2 = 2$; $RD_1 = 0.25$; $RD_2 = 0.375$

Thus a meta-analysis on just these two studies, using risk difference as the effect measure, would have discordant primary effect sizes to amalgamate (0.25 and 0.375); but by switching the effect measure to risk ratios the meta-analysis would have concordant primary results to amalgamate (2 and 2). Although the Cochrane Collaboration advises changing the effect measure if the primary studies have discordant results, choosing between effect measures on the basis of trying to avoid discordance is ad hoc. More to the point, the choice of effect measure is another decision in which personal judgment is required, and the fact that there are multiple effect measures allows a range of outputs for any meta-analysis. Again, this threatens Objectivity, since some analysts might choose to change their effect measure when the primary evidence appears discordant using the originally chosen effect measure, while other analysts might resist such switching given that such switching seems ad hoc. Regardless of one's view of whether or not such switching is ad hoc, one's choice of effect measure has a direct influence on the outcome of a meta-analysis, and thus differing choices of effect measures directly threatens what I have been calling Constraint.

4.3 Choice of Quality Assessment Scale

Analysts often attempt to account for differences in the size and methodological quality of primary studies included in a meta-analysis by weighing the primary studies with a formalized quality assessment scale. The conclusion of a meta-analysis depends on how the primary evidence is weighed, because the weights are used as a multiplier when the primary effect sizes are averaged. There are many features of evidence that should influence how primary evidence is weighed, including multiple features that influence the internal validity of a study (e.g. freedom from numerous potential biases) and the external validity of a study (i.e. the relevance of the evidence to one's general hypothesis of interest). Scientists lack principles to precisely determine how these numerous 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 Jüni and his colleagues (1999). They amalgamated data from 17 trials testing a particular medical intervention, using 25 different scales to assess study quality (thereby effectively performing 25 meta-analyses).[16] These quality assessment scales varied in the number of assessed study attributes, from a low of three attributes to a high of 34, and varied in the weight given to the various study attributes; however, Jüni and his colleagues note that "most of these scoring systems lack a focused theoretical basis." Their results were troubling: the amalgamated effect sizes between these 25 meta-analyses differed by up to 117% – *using exactly the same primary evidence*. The authors concluded that "the type of scale used to assess trial quality can dramatically influence the interpretation of meta-analytic studies."

Not only does the choice of quality assessment scale dramatically influence the results of meta-analysis, but so does the choice of analyst. A quality assessment scale known as the 'risk of bias tool' was devised by the Cochrane group to assess the degree to which the results of a study "should be believed." Alberta researchers distributed 163 manuscripts of RCTs among five reviewers, who assessed the RCTs with this tool, and they found the inter-rater agreement of the quality assessments to be very low (Hartling et al. 2009). In other words, even when given a *single* quality assessment tool, and training on how to use it, and a narrow range of methodological diversity, there was a wide variability in assessments of study quality.

Much evidence suggests that personal differences in the assessment of the quality of scientific studies is a deeply rooted phenomenon. Kunda (1990) presents psychological research on what she calls "motivated reasoning", in which subjects assess evidence differentially depending on subjective idiosyncrasies.[17] For example, after reading a scientific article which concludes that consuming caffeine is risky for females, female caffeine consumers were less convinced by the article than were females who do not consume caffeine. In another study, subjects were presented with mixed evidence about the efficacy of capital punishment, and both supporters and opponents of capital

[16] These quality assessment scales were summarized and described in Moher et al. (1995).

[17] I am grateful to Boaz Miller for bringing these findings to my attention, and to the discussion of them in Miller (2010).

punishment subsequently became more polarized in their respective views, which is perhaps best explained by a differential assessment of the mixed evidence.[18]

In short, when performing a meta-analysis, analysts must choose a quality assessment scale and apply the scale to the assessment of particular primary-level studies. The choice of a quality assessment scale, and variations in the assessments of quality by different analysts, violates what I have been calling Objectivity, and the above examples show that such a violation of Objectivity straightforwardly threatens Constraint: differing decisions regarding one's quality assessment scale lead to contradictory outcomes of a meta-analysis.

4.4 Choice of Averaging Technique

Once effect measures are calculated for each primary study, two common ways to determine the average effect measure 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. One problem with the pooled average approach is Simpson's paradox: the comparative success rate of two groups can be reversed in all of their respective sub-groups, so if a meta-analysis simply pooled all participants into an analysis of overall groups then the calculated effect of the intervention could be the opposite of what one would find in every sub-group. Another 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, maintaining sub-groups does not avoid Simpson's paradox unless there is a principled way to demarcate sub-groups such that the 'true' result one is interested in is relative to those sub-groups and

[18] Although these examples suggest that differential assessments of the quality of scientific studies is influenced by non-epistemic features of the subjects involved in the assessment, such differential assessment of the quality of scientific studies can also arise by subjects variably weighing relevant epistemic considerations.

these exact sub-groups were used in the primary analyses. Moreover, 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 to the analyst, then the analyst can re-group individual sub-groups any way she wishes until she finds something interesting, but of course such retrospective data-dredging is liable to support spurious findings. More to the point, once again: the choice of average type – pooled or sub-group (and if the latter, the choice of appropriate sub-groups) – is another decision point in the methodology of meta-analysis which threatens Objectivity and Constraint.

4.5 Summary

Let me recap. I am not the first to note difficulties with meta-analysis. Others have claimed that formal methods of amalgamating evidence "bury under a series of assumptions many value judgments" (Lomas et al. 2003). I have attempted to identify those specific aspects of meta-analysis in which such "value judgments" have an influence on the results of a meta-analysis.

5. The Hill Strategy

A long-time critic of meta-analysis has argued that subjective knowledge is necessary to properly assess a large volume and diversity of evidence:

> A good review is based on intimate personal knowledge of the field, the participants, the problems that arise, the reputation of different laboratories, the likely trustworthiness of individual scientists, and other partly subjective but extremely relevant considerations. Meta-analysis rules out any such subjective factors. (Eysenck 1994)

While I concur that meta-analysis has a primary aim of ruling out subjective factors when amalgamating evidence (which is another way of stating the Objectivity norm), if my arguments in §4 are correct, then meta-analysis is not successful at reaching this aim. Others have urged that in situations in which a large volume of primary-level evidence

which is discordant, we do not have (and likely will not find) a satisfactory "formula or set of principles designed to provide decision-making rules" (Klein and Williams 2000). Such pessimism is perhaps most acutely justified when the discordant primary evidence comes from very different kinds of experiments. Nevertheless, there is, at least at first glance, a tension between the purported objectivity and quantificational simplicity of meta-analyses and the subjectivity and qualitative complexity required to assess and interpret the relevant aspects of all available evidence.

A consideration of an older tradition of evidence in medicine, associated with the epidemiologist Sir Bradford Hill (1897 - 1991), might go some way toward resolving this tension. Hill was one of the leading epidemiologists involved in the first large-scale case-control studies during the 1950s which showed a correlation between smoking and lung cancer (Doll and Hill 1950, 1954). Hill's statistician nemesis Ronald Fisher (1890 - 1962) noted the absence of controlled experimental evidence required to prove that the smoking-cancer association was indeed causal. Fisher's now infamous criticism was that the smoking-cancer correlation could be explained by a confounding variable, or common cause of the smoking and cancer. Fisher postulated a genetic predisposition which could be a common cause of both smoking and cancer, and so the observed association between smoking and cancer could be spurious. The only way to determine a true causal relation, according to Fisher, was to perform a controlled experiment; of course, for ethical reasons no such experiment could be performed. Hill, at the time an epidemiologist at the London School of Hygiene and Tropical Medicine, responded by appealing to a plurality of reasoning strategies which, he claimed, when taken together made a compelling case that the observed association was truly a causal relation (Hill 1965).

These reasoning strategies were as follows:

1. strength of associations between variables: strong associations between variables are more likely to be causal than weak associations

2. consistency of results between studies: an association between variables which is observed in multiple studies is more likely to be causal[19]
3. specificity of variables: a single specific cause has a single specific effect; correlations between coarse-grained or non-specific variables are less-compelling evidence for a true causal relation
4. temporality: a cause must precede its effect
5. biological gradient: a dose-response pattern of associations between variables suggests a true causal relation
6. plausibility: a plausible biological mechanism which can explain a correlation suggests that the association is a true causal relation[20]
7. coherence: a causal interpretation of an association should not conflict with other relevant knowledge, and epidemiological evidence should cohere with evidence from laboratory experiments
8. experimental evidence: despite criticisms from Fisher, Hill of course recognized the value, when available, of evidence from controlled experiments
9. analogy: analogies with other known causal relations can aid in causal inference; that is, if the purported cause and purported effect are similar in important respects to a known cause and its effect, then there is at least some reason to think that the purported causal relation is real

Although some have erroneously called these considerations 'causal criteria', Hill considered them only as guidelines rather than necessary or sufficient conditions or 'criteria' (except perhaps for temporality, which is plausibly a necessary condition for a causal relation). Since Hill seems to have intended these as epistemic desiderata for discovering causal relations, I will simply call them 'desiderata'.[21] Although Hill granted that no single desideratum was necessary or sufficient to demonstrate causality, he

[19] It is worth noting that 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.
[20] For an interesting study of an eighteenth century case in which the search for a causal mechanism led the researchers astray, see De Vreese (2008).
[21] I am grateful to an anonymous referee for this suggestion.

claimed that jointly the desiderata could make for a good argument for the presence of a causal relation (Doll 2003).[22] Each particular desideratum could use philosophical critique, but the important point for the purpose of contrast with meta-analysis is the plurality of reasons and sources of evidence that Hill appealed to.[23]

The desiderata appealed to by Hill depend on diverse kinds of evidence, which lack a shared quantitative measure – like that of evidence solely from RCTs – such that the evidence can be combined by a simple weighted average. The four specific problems I raised for meta-analysis – the choice of primary evidence to include, the choice of a metric or effect size to quantify the evidence, the choice of a quality assessment scale to assess or weigh the evidence, and the choice of averaging technique– are even more troublesome for the Hill strategy. Thus one might think: meta-analysis has the virtue of amalgamating evidence with objectivity and quantitative simplicity, yet has the vice of amalgamating only a narrow range of evidence, while the Hill strategy has the virtue of considering all available evidence, yet has the vice of qualitative subjectivity. But given my arguments in §3 and §4, the purported virtues of meta-analysis – objectivity and constraint – are less apparent than many have thought.

Since Hill's desiderata are not individually necessary (with the exception, noted above, of the temporality desideratum) for inferring causal relations, one can have evidence which satisfies only some of the desiderata while still having ample justification for causal inference. There is, then, some malleability in the Hill strategy. Defenders of formal methods of amalgamating evidence, such as meta-analysis, might object to such malleability. Such an objection could appeal to the Objectivity and Constraint norms: if the Hill strategy is so malleable, then different analysts could apply the Hill strategy in a variety of ways which reach contradictory conclusions. This objection would misfire twice over. First, I have already shown that meta-analysis is also highly malleable. This is not a mere *tu quoque*. The complexity of assessing and amalgamating a large volume and

---

[22] The Hill strategy could perhaps be understood as part of a shift in epidemiological concepts of cause and disease from a monocausal to a multifactorial model; for a discussion of concepts of cause and disease in epidemiology, see Broadbent (2009).

[23] See Howick et al. (2009) for a recent analysis and restructuring of Hill's desiderata, and Rothman and Greenland (2005) for a brief discussion of each of the desiderata. Woodward (2010) provides a careful analysis of the specificity desideratum.

diversity of evidence might inevitably require malleable techniques, in which case malleability *per se* could hardly be a criticism of such a technique. Second, when properly applied the desiderata are constraining. If a meta-analysis supports a hypothesis while most of Hill's desiderata provide reasons against belief in the hypothesis, this ought to sustain serious reservation in this hypothesis. For example, Hodge (2007) reports a meta-analysis which concludes that intercessory prayer (praying on behalf of others) has a small but significant effect on the well-being of those prayed for. Such a claim, of course, fares poorly on at least several of Hill's desiderata.[24] Endorsing the Hill strategy, then, does not mean endorsing a more tolerant or relaxed attitude toward amalgamating evidence compared with purportedly rigorous and quantitative methods of amalgamating evidence. Conversely, if most of the desiderata coherently support a particular hypothesis, this is suggestive that the hypothesis is roughly correct. For instance, in §3 I discussed meta-analyses which 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 conclusion the authors proposed possible causal mechanisms meant to undergird the outcome of their meta-analysis, thereby appealing to the coherence and plausibility desiderata.[25]

Some epidemiologists now argue that desiderata such as those used by Hill should be employed more often (Weed 1997), whereas others argue that such criteria should not be used to assess causal relations (Charlton 1996). At the very least, the Hill strategy of dealing with a huge volume and diversity of evidence might, given the problems with meta-analysis discussed in §3 and §4, be more virtuous than meta-analysis.

6. Conclusion

I have argued that meta-analyses fail to adequately constrain intersubjective assessments of hypotheses. This is because the numerous decisions that must be made when designing and performing a meta-analysis require personal judgment and expertise,

[24] Moreover, this is another example in which multiple meta-analyses reach contradictory conclusions. Masters and Spielmans (2007) and Roberts et al. (2009) both report meta-analyses which conclude that intercessory prayer has no effect.

[25] However, it should be clear that nothing very general can be said regarding when the satisfaction of the desiderata are sufficient to infer causality.

and allow personal 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: that is, the subjectivity required for meta-analysis explains how multiple meta-analyses of the same primary evidence can reach contradictory conclusions regarding the same hypothesis.

Defenders of meta-analysis have noted that although my critique shows that there are better and worse ways to perform a meta-analysis, it does not follow that we ought to discard the technique altogether. I agree. Although I have used the published guidance from the Cochrane group as a foil to frame my criticisms, the Cochrane group has been active in working to improve the quality of meta-analyses. There have been multiple attempts at formulating the features that a report of a meta-analysis should include, prominently including that of the Quality of Reporting of Meta-analyses (QUORUM) group (Moher et al. 1999). This response from defenders of meta-analysis does not, however, directly address my central argument, namely that the epistemic prominence given to meta-analysis is unjustified, since meta-analysis allows idiosyncratic biases to influence its results, which in turn explains why the results of meta-analyses are unconstrained. The upshot to this critique, one might claim, is merely to urge the improvement of the quality of meta-analyses in ways similar to that already proposed by the QUORUM and Cochrane group, in order to achieve some higher degree of constraint.[26] However, my discussion of the many particular decisions that must be made when performing a meta-analysis suggests that such improvements can only go so far. For at least some of these decisions, the choice between available options is entirely arbitrary; the various proposals to enhance the transparency of reporting of meta-analyses are unable, in principle, to referee between these arbitrary choices. More generally, this rejoinder from the defenders of meta-analysis – that we ought not altogether discard the technique – over-states the strength of the conclusion I have argued for, which is not that

[26] For an illustration of the variable quality of meta-analyses, consider this: meta-analyses which were not performed by Cochrane collaborators were twice as likely to have positive conclusion statements compared with meta-analyses performed by Cochrane collaborators (Tricco et al. 2009). Assuming that Cochrane meta-analyses were higher quality than non-Cochrane meta-analyses (surely a safe assumption), it follows that better meta-analyses are less likely to have a positive conclusion regarding a medical intervention.

meta-analysis is entirely a bad method of amalgamating evidence, but rather is that meta-analysis ought not be considered the best kind of evidence for assessing causal hypotheses in medicine and the social sciences. I have not argued that meta-analysis cannot provide *any* compelling evidence, but rather, contrary to the standard view, I have argued that meta-analysis is not the *platinum* standard of evidence.

One of the primary 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 the epidemiologist Sir Bradford Hill, is in this respect superior. Moreover, the Hill strategy can accommodate the response from defenders of meta-analysis considered immediately above: the 'consistency' desideratum can be tested by meta-analysis, and so even if one were to use the Hill strategy, one could still use meta-analysis as part of one's assessment of a hypothesis of interest. Meta-analysis, then, would be one of many kinds of evidence appealed to when amalgamating available evidence for some hypothesis. However, there is no formal method for assessing, quantifying, and amalgamating the very disparate kinds of evidence that Hill considered. Thus the Hill strategy lacks the apparent objectivity and quantificational simplicity of meta-analysis. But given the central argument of this paper, 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 quantitative simplicity and objectivity of the latter is a chimera. Despite the ubiquitous view that meta-analysis is the platinum standard of evidence in medicine, meta-analysis is not, in the end, very shiny.

Acknowledgments

Nancy Cartwright, Heather Douglas, Miriam Solomon, Eric Martin, and Boaz Miller gave detailed feedback on earlier drafts of this paper, and I am grateful for discussions with audiences at University of Toronto, Michigan State University, University of Western Ontario, the American Association for the Advancement of Science, participants in my seminar at Virginia Tech, and members of the UCSD Philosophy of Science Reading Group. Three anonymous reviewers suggested many valuable improvements.

References

Assendelft, W. J., Koes, B.W., Knipschild, P.G., Bouter, L.M. (1995). The relationship between methodological quality and conclusions in reviews of spinal manipulation. *The Journal of the American Medical Association*, 274, 1942-1948.

Bachand, A.M., Mundt, K.A., Mundt, D.J., Montgomery, R.R. (2010). Epidemiological studies of formaldehyde exposure and risk of leukemia and nasopharyngeal cancer: A meta-analysis. *Critical Reviews in Toxicology*, 40(2), 85-100.

Banerjee, A., & Duflo, E. (Unpublished m.s.). The experimental approach to development economics. MIT JPAL manuscript. Accessed at http://www.povertyactionlab.org/methodology (March 15, 2011).

Barnes, D., & Bero, L. (1998). Why review articles on the health effects of passive smoking reach different conclusions. *The Journal of the American Medical Association*, 279(19), 1566-70.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley and Sons.

Borgenson, K. (2008). *Valuing and evaluating evidence in medicine.* PhD diss., University of Toronto.

Bosetti, C., McLaughlin, J.K., Tarone, R.E., Pira, E., La Vecchia, C. (2008). Formaldehyde and cancer risk: a quantitative review of cohort studies through 2006. *Annals of Oncology*, 19, 29-43.

Broadbent, A. (2009). Causation and models of disease in epidemiology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 40, 302-311.

Brown, J. (2008). "The Community of Science." In (Eds.) Carrier, Howard, and Kourany, *The challenge of the social and the pressure of practice: Science and values revisited*. Pittsburgh: University of Pittsburgh Press.

Cartwright, N., & Stegenga, J. (2011). A theory of evidence for evidence-based policy. In (Eds.) Dawid, Twining, and Vasilaki, *Evidence, inference and enquiry.* Oxford University Press.

Cartwright, N. (2007). Are RCTs the gold standard? *Biosocieties*, 2, 11-20.

Cartwright, N. (2010). The long road from 'it works somewhere' to 'it will work for us'. *Philosophy of Science Association, Presidential Address*.

Charlton, B.G. (1996). Attribution of causation in epidemiology: chain or mosaic? *Journal of Clinical Epidemiology*, 49, 105-107.

Cochrane Handbook. Available online at http://www.cochrane.org/resources/handbook

Collins, J.J., & Lineker, G.A. (2004). A review and meta-analysis of formaldehyde exposure and leukemia. *Regulatory Toxicology and Pharmacology*, 40, 81-91.

Danks, D. (2005). Scientific coherence and the fusion of experimental results. *The British Journal for the Philosophy of Science*, 56, 791-807.

Daston, L., & Galison, P. (2007). *Objectivity*. Cambridge: Zone Books.

De Vreese, L. 2008. Causal (mis)understanding and the search for scientific explanations: a case study from the history of medicine. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 39, 14-24.

Deaton, A. (2008). Instruments of development: randomisation in the tropics, and the search for the elusive keys to economic development. *Proceedings of the British Academy*, 162, 123-160.

Doll, R., & Hill, A.B. (1950). Smoking and carcinoma of the lung: Preliminary report. *British Medical Journal*, 2(4682), 739-748.

Doll, R., & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits. *British Medical Journal*, 1(4877), 1451-5.

Doll, R. (2003). Fisher and Bradford Hill: Their personal impact. *International Journal of Epidemiology*, 32, 929-931.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, 138(3), 453-73.

Duflo, E., & Kremer, M. (2003)' The use of randomization in the evaluation of development effectiveness. World Bank manuscript, accessed at http://www.povertyactionlab.org/methodology (March 15, 2011).

Egger, M., Smith, G. D., Phillips, A.N. (1997). Meta-analysis: principles and procedures *British Medical Journal*, 315, 1533-37.

Epstein, S. (2007). *Inclusion: The politics of difference in medical research*. Chicago: Chicago University Press.

Eysenck, H. (1984). Meta-analysis: an abuse of research integration. *Journal of Special Education*, 18(1), 41-59.

Eysenck, H. (1994). Systematic reviews: meta-analysis and its problems" *British Medical Journal*, 309, 789-792.

Fergusson, D., Doucette, S., Glass, K.C., Shapiro, S., Healy, D., Hebert, P., Hutton, B. (2005). Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. *British Medical Journal*, 330, 396-9.

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 10, 3-8.

Glass, G.V. & Smith, M.L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2-16.

Gunnell, D. Saperia, J., Ashby, D. (2005). Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. *British Medical Journal*, 330, 385-8.

Hacking, I. (1988). Telepathy: origins of randomization in experimental design. *Isis,* 79(3), 427-51.

Hartling, L., Ospina, M., Liang, Y., Dryden, D., Hooten, N., Seida, J., Klassen, T. (2009). Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *British Medical Journal*, 339:b4012.

Hill, B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295-300.

Hodge, D.R. (2007). A systematic review of the empirical literature on intercessory prayer. *Research on Social Work Practice*, 17, 174-187.

Howick, J., Glasziou, P., Aronson, J.K. (2009). The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *J R Soc Med*, 102, 186-194.

Jaber, B.L., Lau, J., Schmid, C.H., Karsou, S.A., Levey, A.S., Pereira, B.J. (2002). Effect of biocompatibility of hemodialysis membranes on mortality in acute renal failure: a meta-analysis. *Clinical Nephrology*, 57(4), 274-82.

Jüni, P., Witschi, A., Bloch, R., Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *The Journal of the American Medical Association*, 282(11), 1054-60.

Knipschild, P. (1994). Systematic reviews: Some examples. *British Medical Journal*, 309, 719-721.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

Linde, K. & Willich, S. (2003). How objective are systematic reviews? Differences between reviews on complementary medicine. *Journal of the Royal Society of Medicine*, 96, 17-22.

Lomas J., Fulop, N., Gagnon, D., Allen, P. (2003). On being a good listener: setting priorities for applied health services research. *Milbank Quarterly*, 81(3), 363-388.

Masters, K.S., Spielmans, G.I. (2007). Prayer and health: review, meta-analysis, and research agenda. *Journal of Behavioral Medicine*, 30(4), 329-338.

McAuley, L., Pham, B., Tugwell, P., Moher, D. (2000) Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet*, 356(9237), 1228-31.

Miller, B. (2010). *A social theory of knowledge*. PhD diss., University of Toronto.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. and the QUORUM Group (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet*, 354, 1896-900.

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P., Walsh, S. (1995). Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials*, 16, 62-73.

Mnookin, J. (2008). Under the influence of technology: DUI and the legal production of objectivity. UCSD Science Studies Colloquium, April 21 2008.

Pauling, L. (1986). *How to live longer and feel better*. New York: W.H. Freeman.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Rhine, J.B., Pratt, J.G., Stuart, C.E., Smith, B.M., and Greenwood, J.A. (1940). *Extrasensory perception after sixty years.* New York: Holt.

Roberts, L., Ahmed, I., Hall, S., Davison, A. (2009). Intercessory prayer for the alleviation of ill health. *Cochrane Database of Systematic Reviews*, Apr 15;(2):CD000368.

Rothman K.J., & Greenland S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, 95, S144-S150.

Rotton, J. & Kelly, I.W. (1985). Much ado about the full moon: A meta-analysis of lunar-lunacy research. *Psychological Bulletin*, 97, 286-306.

Slavin, R. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, 48(1), 9-18.

Smith, G.C. & Pell, J.P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematice review of randomised controlled trials. *British Medical Journal*, 327(7429), 1459-1461.

Smith, M.L., & Glass, G.V. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-60.

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76, 650-661.

Subramanian, S., Venkataraman, R., Kellum, J. A. (2002). Influence of dialysis membranes on outcomes in acute renal failure: A meta-analysis. *Kidney International*, 62, 1819–23.

Sutton, A.J. & Higgins, J.P.T. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, 27, 625-50.

Thagard, P. (1998). Ulcers and bacteria I: Discovery and acceptance. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 29, 107-136.

Tricco, A.C., Tetzlaff, J., Pham, B., Brehaut, J., Moher, D. (2009). Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. *Journal of Clinical Epidemiology*, 62(4), 380-386.

Trout, J. D. (1995). Diverse tests on an independent world. *Studies in History and Philosophy of Science*, 26(3), 407-29.

Weed, D. (1997). On the use of causal criteria. *International Journal of Epidemiology*, 26, 1137-1141.

Whitlock, M. & Schluter, D. (2009). *The analysis of biological data*. Greenwood Village: Roberts and Company Publishers.

Whittington, C.J., Kendall, T., Fonagy, P., Cottrell, D., Cotgrove, A., Boddington, E. (2004). Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet*, 363(9418), 1341-5.

Wimsatt, W. (1981). Robustness, reliability, and overdetermination. In (Eds.) Brewer & Collins, *Scientific inquiry and the social sciences*, San Francisco: Jossey-Bass.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, 25, 287-318.

Worrall, J. (2002). *What* evidence in evidence-based medicine? *Philosophy of Science*, 69, S316-30.

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, 58, 451-88.

Yank, V., Rennie, D., Bero, L.A. (2007). Financial ties and concordance between results and conclusions in meta-analyses: A retrospective cohort study. *British Medical Journal*, 335, 1202-5.

Zhang, L., Steinmaus, C., Eastmond, D.A., Xin, X.K., Smith, M.T. (2009). Formaldehyde exposure and leukemia: A new meta-analysis and potential mechanisms. *Mutation Research*, 681, 150-168.