Herding QATs: Quality Assessment Tools for Evidence in Medicine

Jacob Stegenga

Forthcoming in *Classification, Disease and Evidence: New Essays in Philosophy of Medicine* (Huneman, Silberstein, Lambert, eds.), Springer

Abstract

Medical scientists employ 'quality assessment tools' (QATs) to measure the quality of evidence from clinical studies, especially randomized controlled trials (RCTs). These tools are designed to take into account various methodological details of clinical studies, including randomization, blinding, and other features of studies deemed relevant to minimizing bias and error. There are now dozens available. The various QATs on offer differ widely from each other, and second-order empirical studies show that QATs have low inter-rater reliability and low inter-tool reliability. This is an instance of a more general problem I call the underdetermination of evidential significance. Disagreements about the strength of a particular piece of evidence can be due to different—but in principle equally good—weightings of the fine-grained methodological features which constitute QATs.

Jacob Stegenga

Banting Postdoctoral Fellow
Institute for the History and Philosophy of Science and Technology
University of Toronto
91 Charles Street West
Toronto, Ontario
Canada M5S 1K7

and

Assistant Professor
Department of Philosophy
University of Utah
215 South Central Campus Drive
Carolyn Tanner Irish Humanities Building
4th Floor
Salt Lake City, Utah 84112

jacob.stegenga@utoronto.ca
http://individual.utoronto.ca/jstegenga

Acknowledgements
This paper has benefited from discussion with Nancy Cartwright, Eran Tal, Jonah Schupbach, and audiences at the University of Utah, University of Toronto, and the Canadian Society for the History and Philosophy of Science. Medical scientists Ken Bond and David Moher provided detailed written commentary, as did Philippe Huneman and Gérard Lambert. For the title I thank Frédéric Bouchard. I am grateful for financial support from the Social Sciences and Humanities Research Council of Canada.

1 Introduction

The diversity of evidence in modern medicine is amazing. Many causal hypotheses in medicine, for instance, have evidence generated from experiments on cell and tissue cultures, experiments on laboratory animals (alive at first, then dead, dissected, and analyzed), results of mathematical models, data from epidemiological studies of human populations, data from controlled clinical trials, and meta-level summaries from systematic reviews based on techniques such as meta-analysis and social processes such as consensus conferences. Moreover, each of these kinds of evidence has many variations. Epidemiological studies on humans, for instance, include case-control studies, retrospective cohort studies, and prospective cohort studies.

Evidence from each of these diverse kinds of methods has varying degrees of credibility and relevance for a hypothesis of interest. It is crucial, in order to determine how compelling the available kinds of evidence are, and to make a well-informed assessment of a causal hypothesis, that one take into account substantive details of the methods that generated the available evidence for that hypothesis. Methodological quality, in medical research at least, is typically defined as the extent to which the design, conduct, analysis, and report of a medical trial minimizes potential bias and error. Medical scientists attempt to account for the various dimensions of quality of evidence in a number of ways.

Methodological quality is a complex multi-dimensional property that one cannot simply intuit, and so formalized tools have been developed to aid in the assessment of the quality of medical evidence. Medical evidence is often assessed rather crudely by rank-ordering the types of methods according to an 'evidence hierarchy'. Systematic reviews and specifically meta-analyses are typically at the top of such hierarchies, randomized controlled trials are near the top, non-randomized cohort and case-control studies are lower, and near the bottom are laboratory studies and anecdotal case reports.[1] Evidence from methods at the top of this hierarchy, especially evidence from clinical trials, is often assessed with more fine-grained tools that I call quality assessment tools (QATs). There are many such tools now on

[1] I discuss evidence hierarchies in more detail in §6. Such evidence hierarchies are commonly employed in evidence-based medicine. Examples include those of the Oxford Centre for Evidence-Based Medicine, the Scottish Intercollegiate Guidelines Network (SIGN), and The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. These evidence hierarchies have recently received much criticism. See, for example, Bluhm (2005), Upshur (2005), Borgerson (2008), and La Caze (2011), and for a specific critique of placing meta-analysis at the top of such hierarchies, see Stegenga (2011). In footnote 4 below I cite several recent criticisms of the assumption that RCTs ought to be necessarily near the top of such hierarchies.

offer—QATs are quickly becoming an important tool of epidemiologists and other medical scientists. QATs are used to assess the primary-level evidence amalgamated by a systematic review, and since most causal hypotheses in medicine are assessed by evidence generated from systematic reviews, much of what we think we know about causal hypotheses in medicine is influenced by QATs.

A widely accepted norm holds that when determining the plausibility of a hypothesis one should take into account all of the available evidence. For hypotheses about medical interventions this principle stipulates that one ought to take into account the range of diverse kinds of evidence which are available for that hypothesis.[2] A similar norm states that when determining the plausibility of a hypothesis one should take into account how compelling the various kinds of evidence available for that hypothesis are, by considering detailed qualitative features of the methods used to generate that evidence. The purpose of using a QAT is to evaluate the quality of evidence from medical trials and observational studies in such a fine-grained way. Their domain of application is relatively focused, therefore, since they do not apply to other kinds of evidence that is typically available for causal hypotheses in medicine (such as mechanistic evidence generated by basic science research, or results from experiments on animals), but for better or worse it is usually only RCTs, meta-analyses of RCTs, and sometimes observational studies that are considered when assessing causal hypotheses in medicine, and it is these types of methods that QATs are designed for.

A burgeoning literature has investigated the strategies that scientists employ when generating and assessing evidence. In what follows I examine the use of QATs as codified tools for assessing evidence in medical research. Although there has been some criticism of QATs in the medical literature, they have received little philosophical critique.[3] I begin by describing general properties of QATs, including the methodological features that many QATs share and how QATs are typically employed (§2). I then turn to a discussion of empirical studies which test the inter-

[2] The general norm is usually called the principle of total evidence, associated with Carnap (1947). See also Good (1967). Howick (2011) invokes the principle of total evidence for systematic reviews of evidence related to medical hypotheses. A presently unpublished paper by Bert Leuridan contains a good discussion of the principle of total evidence as it applies to medicine.

[3] Although one only needs to consider the prominence of randomization in QATs to see that QATs have, in fact, been indirectly criticized by the recent literature criticizing the assumed 'gold standard' status of RCTs (see footnote 4). In the present paper I do not attempt a thorough normative evaluation of any particular QAT. Considering the role of randomization suggests what a large task a thorough normative evaluation of a particular QAT would be. But for a systematic survey of the most prominent QATs, see West et al. (2002). See also Olivo et al. (2007) for an empirical critique of QATs.

rater reliability (§3) and inter-tool reliability (§4) of QATs: most QATs are not very good at constraining intersubjective assessments of hypotheses, and more worrying, the use of different QATs to assess the same primary evidence leads to widely divergent quality assessments of that evidence. This is an instance of a more general problem I call the underdetermination of evidential significance, which holds that in a rich enough empirical situation, the strength of the evidence is underdetermined (§5). Despite this problem, I defend the use of QATs in medical research. I end by comparing QATs to the widely employed evidence hierarchies, and argue that despite the problems with QATs, they are better than evidence hierarchies for assessing evidence in medicine (§6).

2 Quality Assessment Tools

A quality assessment tool (QAT) for medical evidence can be either a scale with elements that receive a quantitative score representing the degree to which each element is satisfied by a medical trial, or else a QAT can be simply a checklist with elements that are marked as either present or absent in a medical trial. Given the emphasis on randomized controlled trials (RCTs) in medical research, most QATs are designed for the evaluation of RCTs, although there are several for observational studies and systematic reviews.[4] Most QATs share several elements, including questions about how subjects were assigned to experimental groups in a trial, whether or not the subjects and experimenters were concealed to the subjects' treatment protocol, whether or not there was a sufficient description of subject withdrawal from the trial groups, whether or not particular statistical analyses were performed, and whether or not a report of a trial disclosed financial relationships between investigators and companies.[5] Most QATs provide instructions on how to

[4] The view that RCTs are the 'gold standard' of evidence has recently been subjected to much philosophical criticism. See, for example, Worrall (2002), Worrall (2007), Cartwright (2007), and Cartwright (2010); for an assessment of the arguments for and against the gold standard status of RCTs, see Howick (2011). Observational studies also have QATs, such as QATSO (Quality Assessment Checklist for Observational Studies) and NOQAT (Newcastle-Ottawa Quality Assessment Scale – Case Control Studies).

[5] A note about terminology: sometimes the term 'trial' in the medical literature refers specifically to an experimental design (such as a randomized controlled trial) while the term 'study' refers to an observational design (such as a case control study), but this use is inconsistent. I will use both terms freely to refer to any method of generating 'human level' evidence in biomedical research, including both experimental and observational designs.

score the individual components of the QAT and how to determine an overall quality score of a trial.

A comprehensive list of QATs developed by the mid-1990s was described by Moher et al. (1995). The first scale type to be developed, known as the Chalmers scale, was published in 1981. By the mid-1990s there were over two dozen QATs, and by 2002 West et al. were able to identify 68 for RCTs or observational studies. Some are designed for the evaluation of any medical trial, while others are designed to assess specific trials, or trials from a particular medical sub-discipline. Some are designed to assess the quality of a trial itself, while others are designed to assess the quality of a report of a trial, but most assess both.

QATs are now widely used for several purposes. When performing a systematic review of the available evidence for a particular hypothesis, QATs help reviewers take the quality of medical studies into account. This is typically done in one of two ways. First, QAT scores can be used to generate a weighting factor for the technique known as meta-analysis. Meta-analysis usually involves calculating a weighted average of so-called effect sizes from individual medical studies, and the weighting of effect sizes can be determined by the score of the respective trial on a QAT.[6] Second, QAT scores can be used as an inclusion criterion for a systematic review, in which any primary-level trial that achieves a QAT score above a certain threshold would be included in the systematic review (and conversely, any trial that achieves a QAT score below such a threshold would be excluded). This application of QATs is perhaps the most common use to which they are put. Finally, QATs can be used for purposes not directly associated with a particular systematic review or meta-analysis, but rather to investigate relationships between QAT scores and other properties of medical trials. For instance, several findings suggest that there is an inverse correlation between QAT score and effect size (in other words, higher quality trials tend to have lower estimates of the efficacy of medical interventions).[7]

Why should medical scientists bother using QATs to assess evidence? Consider the following argument, similar to an argument for following the principle of total evidence, based on a concern to take into account any 'defeating' properties of one's evidence. Suppose your evidence seems to provide definitive support for some

---

[6] There are several commonly employed measures of effect size, including mean difference (for continuous variables), or odds ratio, risk ratio, or risk difference (for dichotomous variables). The weighting factor is sometimes determined by the QAT score, but a common method of determining the weight of a trial is simply based on the size of the trial (Egger, Smith, and Phillips, 1997), often by using the inverse variability of the data from a trial to measure that trial's weight (because inverse variability is correlated with trial size).

[7] See, for example, Moher et al. (1998), Balk et al. (2002), and Hempel et al. (2011).

hypothesis, H_1. But then you learn that there is a systematic error in the method which generated your evidence. Taking into account this systematic error, the evidence no longer supports H_1 (perhaps instead the evidence supports a competitor hypothesis, H_2). Had you not taken into account the fine-grained methodological information regarding the systematic error, you would have unwarranted belief in H_1. You do not want to have unwarranted belief in a hypothesis, so you ought to take into account fine-grained methodological information.

Here is a related argument: if one does not take into account all of one's evidence, including one's old evidence, then one is liable to commit the base-rate fallacy. In terms of Bayes' Theorem—p(H|e) = p(e|H)p(H)/p(e)—one commits the base-rate fallacy if one attempts to determine p(H|e) without taking into account p(H). Similarly, if one wants to determine p(H|e) then one ought to take into account the detailed methodological features which determine p(e|H) and p(e).

One need not be a Bayesian to see the importance of assessing evidence at a fine-grain with QATs. For instance, Mayo's notion of 'severe testing', broadly based on aspects of frequentist statistics, also requires taking into account fine-grained methodological details. The Severity Principle, to use Mayo's term, claims that "passing a test *T* (with *e*) counts as a good test of or good evidence for *H* just to the extent that *H* fits *e* and *T* is a *severe test* of *H*" (Mayo 1996). Attending to fine-grained methodological details to ensure that one has minimized the probability of committing an error is central to ensuring that the test in question is severe, and thus that the Severity Principle is satisfied. So, regardless of one's doctrinal commitment to Bayesianism or frequentism, the employment of tools like QATs to take into account detailed information about the methods used to generate the available evidence ought to seem reasonable.

One of the simplest QATs is the Jadad scale, first developed in the 1990s to assess clinical studies in pain research. Here it is, in full:

1. Was the study described as randomized?
2. Was the study described as double blind?
3. Was there a description of withdrawals and dropouts?

A 'yes' to question 1 and question 2 is given one point each. A 'yes' to question 3, in addition to a description of the number of withdrawals and dropouts in each of the trial sub-groups, and an explanation for the withdrawals or dropouts, receives one point. An additional point is given if the method of randomization is described in the paper, and the method is deemed appropriate. A final point is awarded if the method of blinding is described, and the method is deemed appropriate. Thus, a trial can receive between zero and five points on the Jadad scale.

The Jadad scale has been praised by some as being easy to use—it takes about ten minutes to complete for each study—which is an obvious virtue when a reviewer must assess hundreds of studies for a particular hypothesis. On the other hand, others complain that it is too simple, and that it has low inter-rater reliability (discussed in §3). I describe the tool here not to assess it but merely to provide an example of a QAT for illustration.

In contrast to the simplicity of the Jadad scale, the Chalmers scale has 30 questions in several categories, which include the trial protocol, the statistical analysis, and the presentation of results. Similarly, the QAT developed by Cho and Bero (1994) has 24 questions. At a coarse grain some of the features on the Chalmers QAT and the Cho and Bero QAT are similar to the basic elements of the Jadad QAT: these scales both include questions about randomization, blinding, and subject withdrawal. (In §5 I briefly describe how Cho and Bero developed their QAT, as an illustration of the no-best-weighting argument). In addition, these more detailed QATs include questions about statistical analyses, control subjects, and other methodological features deemed relevant to minimizing systematic error. These QATs usually take around 30 to 40 minutes to complete for each study. Despite the added complexity of these more detailed QATs, their scoring systems are kept as simple as possible. For instance, most of the questions on the Cho and Bero QAT allow only the following answers: 'yes' (2 points), 'partial' (1 point), 'no' (0 points), and 'not applicable' (0 points). This is meant to constrain the amount of subjective judgment required when generating a QAT score.

Although most QATs share at least several similar features, the relative weight of the overall score given to the various features differs widely between QATs. Table 1 lists the relative weight of three central methodological features—subject randomization, subject allocation concealment (or 'blinding'), and description of subject withdrawal—for the above QATs, in addition to three other QATs.

Table 1. Number of methodological features used in six QATs, and weight assigned to three widely shared methodological features (adapted from Jüni et al. 1999).

Scale	Number of Items	Weight of Randomization	Weight of Blinding	Weight of Withdrawal
Chalmers et al. (1981)	30	13.0	26.0	7.0
Jadad et al. (1996)	3	40.0	40.0	20.0
Cho & Bero (1994)	24	14.3	8.2	8.2
Reisch et al. (1989)	34	5.9	5.9	2.9
Spitzer et al. (1990)	32	3.1	3.1	9.4
Linde et al. (1997)	7	28.6	28.6	28.6

Note two aspects of Table 1. First, the number of items on a QAT is highly variable, from 3 to 34. Second, the weight given to particular methodological features is also highly variable. Randomization, for instance, constitutes 3.1% of the overall score on the QAT designed by Spitzer et al. (1990), whereas it constitutes 40% of the overall score on the QAT designed by Jadad et al. (1996). The differences between QATs explains the low inter-tool reliability, which I describe in §4. But first I describe the low inter-rater reliability of QATs.

3 Inter-Rater Reliability

The extent to which multiple users of the same rating system achieve similar ratings is usually referred to as 'inter-rater reliability'. Empirical evaluations of the inter-rater reliability of QATs have shown a wide disparity in the outcomes of a QAT when applied to the same primary-level study by multiple reviewers; that is, the inter-rater reliability of QATs is, usually, poor.

The typical set-up of evaluations of inter-rater reliability of a QAT is simple: give a set of manuscripts to multiple reviewers who have been trained to use the QAT, and compare the quality scores assigned by these reviewers to each other. A statistic called kappa (κ) is typically computed which provides a measure of agreement between the quality scores produced by the QAT from the multiple reviewers (although other statistics measuring agreement are also used, such as

Kendall's coefficient of concordance and the intraclass correlation coefficient).[8] Sometimes the manuscripts are blinded as to who the authors were and what journals the manuscripts were published in, but sometimes the manuscripts are not blinded, and sometimes both blinded and non-blinded manuscripts are assessed to evaluate the effect of blinding. In some cases the manuscripts all pertain to the same hypothesis, while in other cases the manuscripts pertain to various subjects within a particular medical sub-discipline.

For example, Clark et al. (1999) assessed the inter-rater reliability of the Jadad scale, using four reviewers to evaluate the quality of 76 manuscripts of RCTs. Inter-rater reliability was found to be "poor", but it increased substantially when the third item of the scale (explanation of withdrawal from study) was removed and only the remaining two questions were employed.

A QAT known as the 'risk of bias tool' was devised by the Cochrane Collaboration (a prominent organization in the so-called evidence-based medicine movement) to assess the degree to which the results of a study "should be believed." A group of medical scientists subsequently assessed the inter-rater reliability of the risk of bias tool. They distributed 163 manuscripts of RCTs among five reviewers, who assessed the RCTs with this tool, and they found the inter-rater reliability of the quality assessments to be very low (Hartling et al. 2009).

Similarly, Hartling et al. (2011) used three QATs (Risk of Bias tool, Jadad scale, Schulz allocation concealment) to assess 107 studies on a medical intervention (the use of inhaled corticosteriods for adults with persistent asthma). This group employed two independent reviewers who scored the 107 studies using the three QATs. They found that inter-rater reliability was 'moderate'. However, the claim that inter-rater reliability was moderate was based on a standard scale in which a κ measure between 0.41 - 0.6 is deemed moderate. The κ measure in this paper was 0.41, so it was just barely within the range deemed moderate. The next lower category, with a κ measure between 0.21 - 0.4, is deemed 'fair' by this standard scale. But at least in the context of measuring inter-rater reliability of QATs, a κ of 0.4 represents wide disagreement between reviewers.

---

[8] For simplicity I will describe Cohen's Kappa, which measures the agreement of two reviewers who classify items into discrete categories, and is computed as follows:

$\kappa = [p(a) - p(e)]/[1 - p(e)]$

where p(a) is the probability of agreement (based on the observed frequency of agreement) and p(e) is the probability of chance agreement (also calculated using observed frequency data). Kappa was first introduced as a statistical measure by Cohen (1960). For more than two reviewers, a measure called Fleiss' Kappa can be used. I give an example of a calculation of κ below.

Here is a toy example to illustrate the disagreement that a κ measure of 0.4 represents. Suppose two teaching assistants, Beth and Sara, are grading the same class of 100 students, and must decide whether or not each student passes or fails. Their joint distribution of grades is:

		Sara	
		Pass	Fail
Beth	Pass	40	10
	Fail	20	30

Of the 100 students, they agree on passing 40 students and failing 30 others, thus their frequency of agreement is 0.7. But the probability of random agreement is 0.5, because Beth passes 50% of the students and Sara passes 60% of the students, so the probability that Beth and Sara would agree on passing a randomly chosen student is 0.5×0.6 (= 0.3), and similarly the probability that Beth and Sara would agree on failing a randomly chosen student is 0.5×0.4 (= 0.2) (and so the overall probability of agreeing on passing or failing a randomly chosen student is 0.3 + 0.2 = 0.5). Applying the kappa formula gives:

$(0.7 - 0.5)/(1 - 0.5) = 0.4$

Importantly, Beth and Sara disagree about 30 students regarding a relatively simple property (passing). It is natural to suppose that they disagree most about 'borderline' students, and their disagreement is made stark because Beth and Sara have a blunt evaluative tool (pass/fail grades rather than, say, letter grades). But a finer-grained evaluative tool would not necessarily mitigate such disagreement, since there would be more categories about which they could disagree for each student; a finer-grained evaluative tool would increase, rather than decrease, the number of borderline cases (because there are borderline cases between each letter grade). This example is meant to illustrate that a κ measure of 0.4 represents poor agreement between two reviewers.[9] A κ score is fundamentally an arbitrary measure of disagreement, and the significance of the disagreement that a particular κ score represents presumably varies with context. This example, I nevertheless hope, helps to illustrate the extent of disagreement found in empirical assessments of the inter-rater reliability of QATs.

[9] I owe Jonah Schupbach thanks for noting that a κ measure can not only seem inappropriately low, as in the above cases of poor inter-rater reliability, but can seem inappropriately high as well. If a κ measure approaches 1, this might suggest agreement which is 'too good to be true'. Returning to my toy example, if Beth and Sara had a very high a κ measure, then one might wonder if they colluded in their grading. Thus when using a κ statistic to assess inter-rater reliability, we should hope for a κ measure above some minimal threshold (below which indicates too much disagreement) but below some maximum threshold (above which indicates too much agreement). What exactly these thresholds should be are beyond the scope of this paper (and are, I suppose, context sensitive).

In short, different users of the same QAT, when assessing the same evidence, generate diverging assessments of the strength of that evidence. In most tests of the inter-rater reliability of QATs, the evidence being assessed comes from a narrow range of study designs (usually all the studies are RCTs), and the evidence is about a narrow range of subject matter (usually all the studies are about the same causal hypothesis regarding a particular medical intervention). The poor inter-rater reliability is even more striking considering the narrow range of study designs and subject matter from which the evidence is generated.

4 Inter-Tool Reliability

The extent to which multiple instruments have correlated measurements when applied to the same property being measured is referred to as inter-tool reliability. One QAT has inter-tool reliability with respect to another if its measurement of the quality of medical studies correlates with the measurement of the quality of the same studies by the other QAT. A QAT score is a measure on a relatively arbitrary scale, and the scales between multiple QATs are incommensurable, so constructs such as 'high quality' and 'low quality' are developed for each QAT which allow the results from different QATs to be compared. That is, when testing the inter-tool reliability of multiple QATs, what is usually being compared is the extent of their agreement regarding the categorization of particular medical trials into pre-defined bins of quality. Similar to assessments of inter-rater reliability, empirical evaluations of the inter-tool reliability have shown a wide disparity in the outcomes of multiple QATs when applied to the same primary-level studies; that is, the inter-tool reliability of QATs is poor. I should note, however, that there are few such assessments available, and those published thus far have varied with respect to the particular QATs assessed, the design of the reliability assessment, and the statistical analyses employed.[10]

An extensive investigation of inter-tool reliability was performed by Jüni and colleagues (1999). They amalgamated data from 17 studies which had tested a

[10] For this latter reason I refrain from describing or illustrating the particular statistical analyses employed in tests of the inter-tool reliability of QATs, as I did in §3 on tests of the inter-rater reliability of QATs. Nearly every published test of inter-rater reliability uses a different statistic to measure agreement of quality assessment between tools. Analyses include Kendall's rank correlation coefficient (τ), Kendall's coefficient of concordance (W), and Spearman's rank correlation coefficient (ρ).

particular medical intervention (the use of low molecular weight heparin to prevent post-operative thrombosis), and they used 25 QATs to assess the quality of these 17 studies (thereby effectively performing 25 meta-analyses). The QATs that this group used were the same that Moher et al. (1995) had earlier described, which 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. Jüni and his colleagues noted 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*. They found that medical trials deemed high quality according to one QAT could be deemed low quality according to another. The authors concluded that "the type of scale used to assess trial quality can dramatically influence the interpretation of meta-analytic studies."

Perhaps the most recent evaluation of inter-tool reliability is Hartling et al. (2011), discussed above in §3. Recall that this group used three QATs (Risk of Bias tool, Jadad scale, Schulz allocation concealment) to assess 107 trials on a particular medical intervention. They also found that the inter-tool reliability was very low.

Yet another example of a test of inter-tool reliability of QATs was reported by Moher et al. (1996). This group used six QATs to evaluate 12 trials of a medical intervention. Again, the inter-tool reliability was found to be low.

Low inter-tool reliability of QATs is troubling: it is a quantitative empirical demonstration that the determination of the quality of a medical trial depends on the choice of QAT. Moreover, in §2 I noted that there are many QATs available, and between them there are substantial differences in their design. Thus the *best* tools that medical scientists have to determine the strength of evidence generated by what are typically deemed the *best* study designs (RCTs) are relatively unconstraining and liable to produce conflicting assessments. Such low inter-tool reliability of QATs has important practical consequences. Elsewhere I show that multiple meta-analyses of the same primary evidence can reach contradictory conclusions regarding particular causal hypotheses, and one of the conditions which permits such malleability of meta-analysis is the choice of QAT (Stegenga 2011).[11] The discordant results from the 25 meta-analyses performed by Moher et al. (1995) are a case in point.

[11] Low inter-tool reliability of QATs is only one of several problems with meta-analysis. Other parameters of meta-analysis that render this method malleable include the choice of primary-level studies to include in the analysis, the choice of outcome measure to employ, the choice of kind of data to amalgamated (patient-level or study-level), and the choice of averaging technique to employ. See Stegenga (2011) for a critical account of meta-analysis.

Moreover, this low inter-tool reliability has philosophical consequences, which I explore in §5.

Such low inter-tool reliability might be less troubling if the various QATs had distinct domains of application. The many biases present in medical research are pertinent to varying degrees depending on the details of the particular circumstances at hand, and so one might think that it is a mistake to expect that one QAT ought to apply to all circumstances. For some causal hypotheses, for instance, it is difficult or impossible to conceal the treatment from the experimental subjects and/or the investigators (that is, 'blinding' is sometimes impossible)—hypotheses regarding chiropractic spinal manipulation are a case in point. Thus, no study relevant to such a hypothesis will score well on a QAT that gives a large weight to allocation concealment. Such a QAT would be less sensitive to the presence or absence of sources of bias other than lack of allocation concealment, relative to QATs that give little or no weight to allocation concealment. In such a case one might argue that since the absence of allocation concealment is fixed among the relevant studies, an appropriate QAT to use in this case should not give any weight to allocation concealment, and would only ask about the presence of those properties of a study that might vary among the relevant studies. On the other hand, one might argue that since we have principled reasons for thinking that the absence of allocation concealment can bias the results of a study, even among those studies that cannot possibly conceal subject allocation, an appropriate QAT to use in this case *should* evaluate the presence of allocation concealment (in which case all of the relevant studies would simply receive a zero score on allocation concealment), just as a QAT ought to evaluate the presence of allocation concealment in a scenario in which the studies in fact can conceal subject allocation. The former consideration is an appeal to determining the *relative* quality between studies, and the latter consideration is an appeal to determining the *absolute* quality of studies. The latter consideration should be more compelling in most cases, since, as discussed above, the typical use of QATs is to help estimate the true efficacy of a medical intervention, and such estimates ought to take into account the full extent of the potential for biases in the relevant evidence, regardless of whether or not it was possible for the respective studies to avoid such biases.

There are scenarios, though, in which we might have reasons to think that a property of a study that causes bias in other scenarios does not cause bias (or perhaps causes less bias) in these scenarios. For example, the placebo effect might be stronger in studies that are designed to assess the *benefits* of pharmaceuticals compared with studies that are designed to assess the *harms* of pharmaceuticals.

Such a difference could be independently and empirically tested. If this were true, then the different scenarios would indeed warrant different QATs, suitable for the particularities of the scenario at hand. If the low inter-tool reliability of QATs were merely the result of employing multiple QATs to different kinds of empirical scenarios (different kinds of studies, say, or studies of different kinds of hypotheses, such as benefits versus harms of pharmaceuticals), then such low inter-tool reliability would hardly be troubling. Indiscriminate use of QATs might lead to low inter-tool reliability, such thinking would go, but discriminate use would not.

Similarly, low inter-tool reliability of QATs would be less troubling if one could show that in principle there is only one good QAT for a given domain, or at least a small set of good ones which are similar to each other in important respects, because then one could dismiss the observed low inter-tool reliability as an artefact caused by the inclusion of poor QATs in addition to the good ones.

Unfortunately, on the whole, these considerations do not mitigate the problem of low inter-tool reliability of QATs. There are, in fact, a plurality of equally fine QATs, designed for the same kinds of scenarios (typically: assessing RCTs of the efficacy of pharmaceuticals). A systematic review by medical scientists concluded that there were numerous QATs that "represent acceptable approaches that could be used today without major modifications" (West et al. 2002). Moreover, all of the empirical demonstrations of their low inter-tool reliability involve the assessment of the quality of studies from a very narrow domain: for instance, the low inter-tool reliability of QATs shown in Jüni et al. (1999) involved assessing studies of a *single* design (RCTs) about a *single* causal hypothesis, and these QATs had been developed with the purpose of assessing the quality of that very study design. Although there are some QATs which are arguably inferior to others, at least among the reasonably good ones I argue below that we lack a theoretical basis for distinguishing among them, and so we are stuck with a panoply of acceptable QATs which disagree widely about the quality of particular medical studies and thus the strength of the evidence generated from those studies.

One might agree with the view that there is no uniquely best QAT, but be tempted to think that this is due only to the fact that the quality of a study depends on particularities of the context (e.g. the particular kind of study in question and the form of the hypothesis being tested by that study). Different QATs might, according to this thought, be optimally suited to different contexts. While this latter point is no doubt true—above I noted that some QATs are designed for assessing particular kinds of *studies*, and others are designed for assessing studies in particular *domains* of medicine—it does not explain the low inter-tool reliability of QATs. That is

because, as above, the low inter-tool reliability of QATs is demonstrated in narrowly specified contexts. Moreover, the research groups that design QATs usually claim (explicitly) that their QATs are meant to be applicable to a given study design (usually RCTs) in almost any domain of medical research. In short, QATs are intended to apply to a broad range of contexts, but regardless, the empirical demonstrations of their low inter-tool reliability are almost always constrained to a single particular context.

Despite their widespread and growing use, among medical scientists there is some debate about whether or not QATs ought to be employed at all (see, for example, Herbison et al. (2006)). Their low inter-rater and inter-tool reliability might suggest that resistance to their use is warranted. There are three reasons, however, that justify the continuing improvement and application of QATs to assessing the quality of medical evidence. First, when performing a meta-analysis, a decision to not use an instrument to differentially weight the quality of the primary-level studies is equivalent to weighting all the primary-level studies to an equal degree. So whether one wishes to or not, when performing a meta-analysis one is forced, in principle, to weight the primary-level studies, and the remaining question then is simply how arbitrary one's method of weighting is. Assigning equal weights regardless of methodological quality is maximally arbitrary. The use of QATs to differentially weight primary-level studies is an attempt to minimize such arbitrariness. Second, as argued in §2 above, one must account for fine-grained methodological features in order to guarantee that one avoids potential defeating properties of evidence, and QATs can help with this. Third—but closely related to the second point—there is some empirical evidence which suggests that studies of lower quality have a tendency to over-estimate the efficacy of medical interventions (see footnote 7), and thus the use of QATs helps to accurately estimate the efficacy of medical interventions. In short, despite their low inter-rater and inter-tool reliability, QATs are an important component of medical research, and should be employed when performing a systematic review or meta-analysis.

5 Underdetermination of Evidential Significance

The primary use of QATs is to estimate the quality of evidence from particular medical studies, and the primary use of such evidence is to estimate the strength (if any) of causal relations in relevant domains. The relata in these purported causal relations are, of course, the medical intervention under investigation and the change

in value of one or more parameters of a group of subjects. The best available QATs appropriate to a given domain differ substantially in the weight assigned to various methodological properties (§2), and thus generate discordant estimates of evidential quality when applied to the same evidence (§4). The differences between the best available QATs are fundamentally arbitrary. Although I assume that there must be a unique value (if at all) to the strength of purported causal relations in the domains in which these tools are employed, the low inter-tool reliability of QATs—together with the fundamentally arbitrary differences of their content—suggests that, in such domains and for such relations, there is no uniquely correct estimate of the quality of evidence. This is an instance of the general problem I call the underdetermination of evidential significance.

Disagreement regarding the strength of evidence in particular scientific domains has been frequently documented with historical case studies. One virtue of examining the disagreement generated by the use of QATs is that such disagreements occur in highly controlled settings, are quantifiable using measures such as the κ statistic, and are about subjects of great importance. Such disagreements do not necessarily represent shortcoming on the part of the disagreeing scientists, and nor do such disagreements necessarily suggest a crude relativism. Two scientists who disagree about the strength of a particular piece of evidence can both be rational because their differing assessments of the strength of the same evidence can be due to their different weightings of fine-grained features of the methods which generated the evidence. This explains (at least in part) the low inter-rater and inter-tool reliability of QATs.

Concluding that there is no uniquely correct determination of the epistemic significance of some piece of evidence by appealing to the poor inter-rater and inter-tool reliability of QATs is not merely an argument from disagreement. If it were, then the standard objection would simply note that the mere fact of disagreement about a particular subject does not imply that there is no correct or uniquely best view on the subject. Although different QATs disagree about the strength of evidence from a particular trial, this does not imply that there is no true or best view regarding the strength of evidence from this particular trial—goes the standard objection—since the best QATs might agree with each other about the evidence from this trial, and even more ambitiously, agreement or disagreement among QATs would be irrelevant if we just took into account the quality assessment of this particular trial by the uniquely best QAT. The burden that this objection faces is the identification of the single best QAT or at least the set of good ones (and then hope that multiple users of the best QAT will have high inter-rater reliability, or that the

set of good QATs will have high inter-tool reliability). As noted in §4, medical scientists involved in the development and assessment of QATs claim that there are simply a plurality of decent QATs that differ from one another in arbitrary respects. More fundamentally, we lack a theory of scientific inference that would allow us to referee between the most sophisticated QATs. Recall the different weightings of the particular methodological features assessed in QATs, noted in Table 1. Another way to state the burden of the 'mere argument by disagreement' objection is that to identify the best QATs, one would have to possess a principled method of determining the optimal weights for the methodological features included on a QAT. That we do not presently have such a principled method is an understatement.

Consider this compelling illustration of the arbitrariness involved in the assignment of weights to methodological features in QATs. Cho and Bero (1994) employed three different algorithms for weighting the methodological features of their QAT (discussed in §2). Then they tested the three weighting algorithms for their effect on quality scores of medical trials, and their effect on the inter-rater reliability of such scores. They selected for further use—*with no principled basis*—the weighting algorithm that had the highest inter-rater reliability. Cho and Bero explicitly admitted that nothing beyond the higher inter-rater reliability warranted the choice of this weighting algorithm, and they rightfully claimed that such arbitrariness was justified because "there is little empiric [sic] evidence on the relative importance of the individual quality criteria to the control of systematic bias."[12] Medical scientists have no principled foundation for developing a uniquely good QAT, and so resort to a relatively arbitrary basis for their development.

One could press the standard objection by noting that while it is true that we *presently* lack an inductive theory that could provide warrant for a unique system for weighting the various methodological features, it is overly pessimistic to think that we will *never* have a principled basis for identifying a uniquely best weighting system. It is plausible, this objection goes, to think that someday we will have a uniquely best QAT, or perhaps uniquely best QATs for particular kinds of epistemic scenarios, and we could thereby achieve agreement regarding the strength of evidence from medical studies. To this one would have to forgive those medical scientists, dissatisfied with this response, who are concerned with assessing evidence today. But there is another, deeper reason why such a response is not compelling.

---

[12] There is a tendency among medical scientists to suppose that the relative importance of various methodological features is merely an empirical matter. One need not entirely sympathize with such methodological naturalism to agree with the point expressed by Cho and Bero here: we lack reasons to prefer one weighting of methodological features over another, regardless of whether one thinks of these reasons as empirical or principled.

It is not a mere argument from present disagreement—I reiterate—to claim that the poor inter-tool reliability of QATs implies that the strength of evidence from particular medical studies is underdetermined. That is because, as the example of the Cho and Bero QAT suggests, the disagreements between QATs are due to arbitrary differences in how the particular methodological features are weighed in the various QATs. There are, to be sure, better and worse QATs. But that is about as good as one can do when it comes to distinguishing between QATs. Of those that account for the majority of relevant methodological features, some weight those features in a slightly different manner than others, and we have no principled grounds for preferring one weighting over another. We do not possess a theory of scientific inference that could help determine the weights of the methodological features in QATs. If one really wanted to, one could sustain the objection by claiming that it is possible that in the future we will develop a theory of inference which would allow us to identify a uniquely best QAT. There is a point at which one can no longer argue against philosophical optimism. The underdetermination of evidential significance is a hard problem; like other hard philosophical problems, it does not preclude optimism.

One could put aside the aim of finding a *principled* basis for selecting among the available QATs, and instead perform a selection based on their *historical* performance. Call this a 'naturalist' selection of QATs.[13] Since QATs are employed to estimate the quality of evidence from medical studies, and such evidence is used to estimate the strength of causal relations, the naturalist approach would involve selecting QATs based on a parameter determined by the 'fit' between (i) the strength of presently known causal relations and (ii) the quality of the evidence for such causal relations available at a particular time, as determined in retrospect by currently available QATs. The best QAT would be the one with the best average fit between (i) and (ii). Such an assessment of QATs would be of some value. It would be fundamentally limited, though, given an epistemic circularity. In the domains in which QATs are employed, the best epistemic access to the strength of causal relations is the total evidence from all the available medical studies, summarized by a careful systematic review (which, in this domain, usually takes the form of a meta-analysis), appropriately weighted to take into account relevant methodological features of those studies. But of course, those very weightings are generated by QATs. The naturalist approach to assessing QATs, then, itself requires the employment of QATs.

[13] Such an approach was first suggested to me by Jim Tabery.

The underdetermination of evidential significance is *not* the same problem that is often associated with Duhem and Quine. One formulation of the standard underdetermination problem—underdetermination of theory by evidence—holds that there are multiple theories compatible with a given body of evidence. The underdetermination of evidential significance is the prior problem of settling on the strength of a given piece of evidence in the first place. Indeed, one may wish to say that an appropriate name for the present problem is just the inverse of the Quinean locution: *underdetermination of evidence by theory*. Our best theories of inference underdetermine the strength of evidence, exemplified by tools such as QATs.

6 QATs and Hierarchies

The most frequently used tools for assessing the quality of medical studies are not QATs, but rather evidence hierarchies. An evidence hierarchy is a rank-ordering of kinds of methods according to the potential for bias in that kind of method. The potential for bias is usually based on one or very few parameters of study designs, most prominently randomization. QATs and evidence hierarchies are not mutually exclusive, since an evidence hierarchy can be employed to generate a rank-ordering of types of methods, and then QATs can be employed to evaluate the quality of tokens of those methods. However, judicious use of QATs should replace evidence hierarchies altogether. The best defense of evidence hierarchies that I know of is given by Howick (2011), who promotes a sophisticated version of hierarchies in which the rank-ordering of a particular study can increase or decrease depending on parameters distinct from the parameter first used to generate the ranking. Howick's suggestion, and any evidence hierarchy consistent with his suggestion (such as that of GRADE), ultimately amounts to an outright abandonment of evidence hierarchies. Howick gives conditions for when mechanistic evidence and evidence from non-randomized studies should be considered, and also suggests that sometimes evidence from RCTs should be doubted. If one takes into account methodological nuances of medical research, in the ways that Howick suggests or otherwise, then the metaphor of a hierarchy of evidence and its utility in assessing quality of evidence seem less compelling than more quantitative tools like QATs.

For instance, the GRADE evidence hierarchy employs more than one property to rank methods. GRADE starts with a quality assignment based on one property and takes other properties into account by subsequent modifications of the quality

assignment (shifting the assignment up or down). Formally, the use of *n* properties to rank methods is equivalent to a scoring system based on *n* properties which discards any information that exceeds what is required to generate a ranking. QATs generate scores that are measured on scales more informative than ordinal scales (such as interval, ratio, or absolute scales). From any measure on one of these supra-ordinal scales, a ranking can be inferred on an ordinal scale, but not vice versa (from a ranking on an ordinal scale it is impossible to infer measures on supra-ordinal scales). Thus hierarchies (including the more sophisticated ones such as GRADE) provide evaluations of evidence which are *necessarily less informative* than evaluations provided by QATs.

Moreover, because these sophisticated hierarchies begin with a quality assignment based on one methodological property and then shift the quality assignment by taking other properties into account, the weights that can be assigned to various methodological properties are constrained. With QATs, on the other hand, the weight assigned to any methodological property is completely open, and can be determined based on rational arguments regarding the respective importance of the various properties, without arbitrary constraints imposed by the structure of the scoring system. In short, despite the widespread use of evidence hierarchies and the defense of such use by Howick (2011), and despite the problems that I raise for QATs above, QATs are superior to evidence hierarchies for assessing the great volume of evidence in contemporary medical research.

7 Conclusion

An examination of QATs suggests that coarse-grained features of evidence in medicine, like freedom from systematic error, are themselves amalgams of a complex set of considerations; that is why QATs take into account a plurality of methodological features such as randomization and blinding. The various aspects of a specific empirical situation which can influence an assessment of a coarse-grained evidential feature are numerous, often difficult to identify and articulate, and if they can be identified and articulated (as one attempts to do with QATs), they can be evaluated by different scientists to varying degrees and by different quality assessment tools to various degrees. In short, there are a variety of features of evidence that must be considered when assessing evidence, and there are numerous and potentially contradictory ways to do so. Our best theories of scientific inference

provide little guidance on how to weigh the relevant methodological features included in tools like QATs.

A group of medical scientists prominent in the literature on QATs notes that "the quality of controlled trials is of obvious relevance to systematic reviews" but that "the methodology for both the assessment of quality and its incorporation into systematic reviews are a matter of ongoing debate" (Jüni, Altman, and Egger, 2001). I have argued that the use of QATs are important to minimize arbitrariness when assessing medical evidence and to accurately estimate probabilities associated with measures of confirmation. However, available QATs vary in their constitutions, and when medical evidence is assessed using QATs their inter-rater reliability and inter-tool reliability is low. This, in turn, is a compelling illustration of a more general problem: the underdetermination of evidential significance. Disagreements about the strength of evidence are, of course, ubiquitous in science. Such disagreement is especially striking, however, when it results from the employment of carefully codified tools designed to quantitatively assess the strength of evidence. QATs are currently the *best* instruments available to medical scientists to assess the strength of evidence, yet when applied to what is purported to be the *best* quality evidence in medicine (namely, evidence from RCTs), different users of the same QAT, and different QATs applied to the same evidence, lead to widely discordant assessments of the strength of evidence.

References

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J. (2002) "Correlation of Quality Measures with Estimates of Treatment Effect in Meta-Analyses of Randomized Controlled Trials" *JAMA* 287(22):2973-82.

Bluhm, R. (2005). "From Hierarchy to Network: A Richer View of Evidence For Evidence-Based Medicine" *Perspectives in Biology and Medicine* 48(4):535–547.

Borgerson, K. (2008) *Valuing and Evaluating Evidence in Medicine*. PhD diss., University of Toronto.

Carnap, R. (1947) "On the Application of Inductive Logic" *Philosophy and Phenomenological Research* 8: 133-148.

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*.

Chalmers TC, Smith H, Blackburn B, et al. (1981) "A Method for Assessing the Quality of a Randomized Control Trial" *Control Clin Trials* 2: 31-49.

Cho MK, Bero LA. (1994) "Instruments for Assessing the Quality of Drug Studies Published in the Medical Literature" *JAMA* 272: 101-104.

Cohen, J. (1960) "A Coefficient of Agreement for Nominal Scales" *Educational and Psychological Measurement* 20(1): 37-46.

Clark, H.D., Wells, G.A., Huët, C., McAlister, F.A., Salmi, L.R., Fergusson, D., Laupacis, A. (1999) "Assessing the Quality of Randomized Trials: Reliability of the Jadad Scale" *Controlled Clinical Trials* 20: 448-452.

Egger, M., Smith, G. D., Phillips, A.N. (1997) "Meta-Analysis: Principles and Procedures" *British Medical Journal* 315: 1533-37.

Good, I.J. (1967) "On the Principle of Total Evidence" *The British Journal for the Philosophy of Science* 17(4): 319-321.

Hartling, L., Bond, K., Vandermeer, B., Seida, J., Dryen, D.M., Rowe, B.H. (2011) "Applying the Risk of Bias Tool in a Systematic Review of Combination Long-Acting Beta-Agonists and Inhaled Corticosteroids for Persistent Asthma" *PLoS One* 6(2): 1-6. e17242

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.

Hawthorne, J. (2011) "Bayesian Confirmation Theory" in *The Continuum Companion to the Philosophy of Science* (French and Saatsi, eds.). Continuum Press.

Hempel S, Suttorp MJ, Miles JNV, Wang Z, Maglione M, Morton S, Johnsen B, Valentine D, Shekelle PG. (2011) "Empirical Evidence of Associations Between Trial Quality and Effect Sizes" *Methods Research Report*, AHRQ Publication No. 11-EHC045-EF. Available at: http://effectivehealthcare.ahrq.gov.

Herbison P, Hay-Smith J, Gillespie WJ. (2006) "Adjustment of Meta-Analyses on the Basis of Quality Scores Should be Abandoned" *J Clin Epidemiol* 59: 1249-56.

Howick, J. (2011) *The Philosophy of Evidence-Based Medicine*. Wiley-Blackwell.

Jadad AR, Moore RA, Carroll D, et al. (1996) "Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?" *Control Clin Trials* 17: 1-12.

Jüni, P., Altman, D.G., Egger, M. (2001) "Assessing the Quality of Randomised Controlled Trials" in *Systematic Reviews in Health Care: Meta-Analysis in Context* (Egger, Smith, and Altman, eds.). London: BMJ Publishing Group.

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.

La Caze, A. (2011) "The Role of Basic Science in Evidence-Based Medicine" *Biology and Philosophy* 26(1): 81-98.

Linde K, Clausius N, Ramirez G, et al. (1997) "Are the Clinical Effects of Homoeopathy Placebo Effects?" *Lancet* 350: 834-843.

Maher, C.G., Sherrington, C., Herbert, R.D., Moseley, A.M., & Elkins, M. (2003) "Reliability of the PEDro Scale for Rating Quality of Randomized Controlled Trials" *Physical Therapy* 83: 713-721.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*. University of Chicago Press.

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.

Moher, D., Jadad, A.R., Tugwell, P. (1996) "Assessing the Quality of Randomized Controlled Trials. Current Issues and Future Directions" *Int J Technol Assess Health Care* 12(2): 195-208.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. (1998) "Does Quality of Reports of Randomised Trials Affect Estimates of Intervention Efficacy Reported In Meta-Analyses?" *Lancet* 352(9128):609-13.

Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ (2007) "Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review" *Physicial Therapy* 88(2): 156-175.

Reisch JS, Tyson JE, Mize SG. (1989) "Aid to the Evaluation of Therapeutic Studies" *Pediatrics* 84: 815-827.

Spitzer WO, Lawrence V, Dales R, et al. (1990) "Links Between Passive Smoking and Disease: A Best-Evidence Synthesis. A Report of the Working Group on Passive Smoking" *Clin Invest Med* 13: 17-42.

Stegenga, J. (2011) "Is Meta-Analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Biological and Biomedical Sciences* 42: 497-507.

Upshur, R. (2005) "Looking for Rules in a World of Exceptions: Reflections on Evidence-Based Practice" *Perspectives in Biology and Medicine* 48(4): 477-489.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F., Lux, L. (2002) "Systems to Rate the Strength of Scientific Evidence" *Evidence Report/Technology Assessment Number* 47, AHRQ Publication No. 02-E016.

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.