Philosophical Controversies in the Evaluation of Medical Treatments: With a Focus on the Evidential Roles of Randomization and Mechanisms in Evidence-Based Medicine

ALEXANDER MEBIUS

Doctoral Thesis in Philosophy
Stockholm, Sweden 2015

This compilation thesis consists of an introduction and the following papers:

I Mebius, A. (2014) Corroborating evidence-based medicine. *Journal of Evaluation in Clinical Practice*, 20:915–920. (Reproduced by permission of John Wiley and Sons)

II Howick, J., Mebius, A. (2016) Randomized trials and observational studies: the current philosophical controversy. In Schramme T and Edwards S (eds.) *Handbook of the Philosophy of Medicine*. Springer.

III Howick, J., Mebius, A. (2014) In search of justification for the unpredictability paradox. *Trials*, 15:480.

IV Mebius, A. (2014) A weakened mechanism is still a mechanism: on the causal role of absences in mechanistic explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 45:43–48.

V Mebius, A. (submitted for publication) An omitted hallmark of mechanism function.

VI Mebius, A. (submitted for publication) Ascribing functions to medical technologies.

KTH Royal Institute of Technology
Architecture and the Built Environment
Department of Philosophy and History
SE-100 44 Stockholm, Sweden

Typeset with LaTeX.

Printed by Universitetsservice US-AB, Stockholm.

ISBN 978-91-7595-489-9
ISSN 1650-8831

Abstract

This thesis examines philosophical controversies surrounding the evaluation of medical treatments, with a focus on the evidential roles of randomised trials and mechanisms in Evidence-Based Medicine. Current 'best practice' usually involves excluding non-randomised trial evidence from systematic reviews in cases where randomised trials are available for inclusion in the reviews. The first paper challenges this practice and evaluates whether adding of evidence from non-randomised trials might improve the quality and precision of some systematic reviews. The second paper compares the alleged methodological benefits of randomised trials over observational studies for investigating treatment benefits. It suggests that claims about the superiority of well-conducted randomised controlled trials over well-conducted observational studies are justified, especially when results from the two methods are contradictory. The third paper argues that postulating the unpredictability paradox in systematic reviews when no detectable empirical differences can be found requires further justification. The fourth paper examines the problem of absence causation in the context of explaining causal mechanisms and argues that a recent solution (Barros 2013) is incomplete and requires further justification. Solving the problem by describing absences as causes of 'mechanism failure' fails to take into account the effects of absences that lead to vacillating levels of mechanism functionality (i.e. differences in effectiveness or efficiency). The fifth paper criticises literature that has emphasised functioning versus 'broken' or 'non-functioning' mechanisms emphasising that many diseases result from increased or decreased mechanism function, rather than complete loss of function. Mechanistic explanations must account for differences in the effectiveness of performed functions, yet current philosophical mechanistic explanations do not achieve this. The last paper argues that the standard of evidence embodied in the ICE theory of technological function (i.e. testimonial evidence and evidence of mechanisms) is too permissive for evaluating whether the proposed functions of medical technologies have been adequately assessed and correctly ascribed. It argues that high-quality evidence from clinical studies is necessary to justify functional ascriptions to health care technologies.

Keywords: Evidence; randomized controlled trials; observational studies; systematic reviews; meta-analysis; methodology; process assessment; outcome assessment; medical care; randomization; evidence-based medicine; selection bias; philosophy of medicine; philosophy of science; mechanisms; quality of evidence; animal studies; treatment effect; causation by absence; medical technology.

Acknowledgements

I thank with deep gratitude my supervisors Professor John Cantwell, Dr. Jeremy Howick, and Professor Sven Ove Hansson for their prudence, good judgement and wisdom.

My grateful thanks are extended to my friends and colleagues at the Royal Institute of Technology (KTH) and University of Oxford. I also wish to thank Ann-Charlotte and Karl Sundén for their unfailing support and interest in the development of this work. Above all I thank my wife Eva-Charlotta for her total support, and my parents for their generosity and help.

Contents

Papers

1 Introduction

This thesis examines philosophical controversies surrounding the evaluation of medical treatments, with a focus on the evidential roles of randomised trials and mechanisms in Evidence-Based Medicine (EBM). Identifying issues with current evidential comparisons and evaluations of medical interventions, and indicating how they can be improved is philosophically interesting and also has the potential to improve the scientific basis upon which health services are based.

The first major issue evaluated in this thesis is whether it is advisable to combine evidence from observational studies alongside evidence from randomised trials within systematic reviews. Observational studies tend to suffer from problems that are believed to increase the risk of producing inflated effects in randomised trials. Hence it seems reasonable to:

(1) side with randomised trials in cases where randomised trials and observational studies have conflicting results,

(2) excluding observational studies alongside randomised trials in systematic reviews and meta-analysis, and

(3) pre-rank evidence from randomized trials higher than observational studies.

However, differences between results from randomised trials and observational studies have proven difficult to detect empirically. Based on the failure to detect an average difference, the authors of a recent Cochrane review concluded that: "there is little evidence for significant effect estimate difference between observational studies and RCTs [randomised controlled trials], regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmacological interventions" (Anglemyer et al. 2014, p. 2). The consequent conclusion that the theoretical benefits of randomised trials are not borne out in practice is nothing short of groundbreaking and (if acceptable) would lead to a revolution in medical practice; and provide empirical warrant to philosophical critiques of the EBM epistemology of evidence (e.g., Worrall 2002; Cartwright 2007).

The second major issue investigated in this thesis concerns the epistemic status of mechanisms in biomedical research. Knowledge of causal mechanisms is considered essential by many prominent philosophers of science for adequately guiding and planning controlled experiments. However, a number of serious concerns raised in the methodological literature regarding the epistemic significance of mechanisms have received scant attention in the philosophical literature to date. Many of these considerations are crucial because they pose more general problem for philosophical accounts of mechanisms in that they question their ability to (1) distinguish differences in the causal productivity of mechanisms—especially by not recognising the importance of effect size in research studies—and (2) to explain contradicting results from mechanistic findings. In this thesis I investigate these two issues and conclude both that (a) the special privilege of randomised trials (supported by EBM

proponents), and (b) the alleged benefits of mechanistic evidence (supported by EBM critics) require further justification, some of which I provide.

1.1 How this thesis is organised

Paper I, "Corroborating evidence-based medicine", published in *Journal of Evaluation in Clinical Practice*, discusses whether the addition of non-randomised studies might improve the quality of some systematic reviews by enhancing the evidence available from clinical trials; Paper II (co-authored by Dr. Jeremy Howick) compares the ability of randomised controlled trials and observational studies to reduce confounding and draws ethical implications (forthcoming in *Handbook of the Philosophy of Medicine*); Paper III is a short editorial/commentary published in *Trials* (co-authored by Dr. Howick) on an important issue in synthesizing evidence that has the potential to modify the Cochrane Collaboration's methodology for undertaking systematic reviews.

The epidemiology and methodology of medicine connects to important philosophical questions about the nature of mechanisms and mechanistic evidence. Paper IV, "A weakened mechanism is still a mechanism: on the causal role of absences in mechanistic explanation", published in *Studies in History and Philosophy of Biological and Biomedical Sciences*, discusses the problem of causation by absence in the context of explaining the productive continuity of causal mechanisms. Paper V (submitted for publication) proposes that establishing differences in causal effect size is often necessary to understand corresponding differences in the performance of mechanism functions. Paper VI (submitted for publication) argues that mechanistic evidence and expert judgment are inadequate for detecting small or moderate effects of drug interventions and specific medical technologies, and thus inappropriate for assessing whether a clinical function can be safely ascribed to interventions.

These articles are organised around two overlapping themes: epistemological and methodological issues in evidence-based medicine in papers I, II, and III (outlined in section 2, below); and philosophical accounts of mechanisms in the biomedical sciences in papers IV, V, and VI (outlined in section 3, below). In addition, Section 2.2.1 considers the methodological value of blinding and placebo control in clinical research. Section 3.3.1 considers the potential of mechanistic evidence to advance therapeutic knowledge and inform disease management.

2 The best study design for evaluating medical treatments

The following section discusses the features of observational and randomised studies and evaluates whether randomised studies are intrinsically more reliable. It also discusses how evidence from different types of studies might be compared in practice. Section 2.2.1 addresses problems surrounding the ethics and implementation of blinding and placebo

controls (e.g. as related to sham or placebo surgery) where the methodological importance risks being overlooked in the debate in papers I–III.

2.1 Corroborating evidence-based medicine

It is generally accepted in the medical community that randomised studies are superior to non-randomised studies for investigating treatment efficacy (see, for example, Higgins and Green 2008; Straus et al. 2011). This standpoint is used to justify the inclusion criteria for systematic reviews. These tend to encourage using randomised studies rather than observational ones, give greater weight to the results of randomised studies, and discourage pooling the results from these two types of study. The assumed superiority of randomised studies also informs judgments about the risk of bias (Guyatt et al. 2008; OCEBM 2011). However, in practice, significant differences between the results of randomised and non-randomised studies in testing the efficacy of clinical treatments have proven difficult to detect empirically, calling into question the very basis on which we make decisions about which clinical treatments to recommend.

For example, five meta-analyses contrasting the two types of study did not find any large systematic differences between randomised trials and observational studies (Benson and Hartz 2000; Concato et al. 2000; MacLehose et al. 2000; Ioannidis et al. 2001). Benson and Hartz (2000) found no differences between the two types of study. Concato et al. (2000) found that randomised studies showed larger variation. MacLehose et al. (2000) found no differences in higher quality studies, suggesting that well-conducted studies of both types gave consistent and accurate results. Ioannidis et al. (2001) found a high correlation coefficient between the results of the two types of study, although observational studies tended to make slightly higher effect estimates and have greater heterogeneity. More recently a Cochrane Review on the differences between randomised and observational studies found no statistically significant difference between the two study types (Anglemyer et al. 2014). The reviewers found that some randomised studies reported larger effect sizes than observational studies of the same treatment, while others had smaller, but often similar, effect sizes. The authors concluded that, on average, there was very little difference between randomised controlled trials and observational studies. An earlier Cochrane Review found that some adequately randomised studies reported larger effect sizes than inadequately randomised studies, while others reported smaller (but again often similar) effect sizes (Odgaard-Jensen et al. 2011). The authors concluded that whether studies are adequately randomised would likely affect the chances of detecting a significant result, but that the direction of the effect cannot be determined with certainty.

These conclusions have important implications for clinicians and clinical practice: they call into question the basis for judging what evidence to accept when there are conflicting results from different interventions or between observational and randomised studies. Accepting the equal validity of observational studies implies, for example, that we do not know whether vitamin C reduces the risk of cardiovascular disease (Knekt et al. 1994),

whether homoeopathy (pace physiological rational) reduces the risk of depression (Oberai et al. 2013), or whether metformin reduces the risk of cancer among patients with diabetes (DeCensi et al. 2010). In each case, observational studies suggest these interventions are effective, while randomised trials do not support this conclusion.

These conclusions also suggest that systematic reviews should include both observational and randomised studies (Higgins and Green 2008). This might improve the quality of some systematic reviews, since many draw on very few studies (Shrier et al. 2007). In some cases, the addition of observational studies might augment the data available and provide a more complete picture. This might be particularly pertinent if the randomised studies were of poor quality, for example, if the randomisation process was easily subverted, or the studies failed to control adequately for other forms of bias. Many thus argue that observational study designs eliminate the effect of confounders. Paper I notes, however, that including observational studies in systematic reviews would require rewriting the rules of evidence ranking systems such the Grading Assessment Development and Evaluation (GRADE) system (Guyatt et al. 2008) that currently assume that evidence from randomised controlled trials is always of higher quality than evidence from observational studies.

Evidence ranking systems generally allow studies to be up- or down-graded based on quality. However, this option is seldom used, not least because most ranking systems start from the assumption that randomised trials have higher quality, which effectively skews the baseline. Paper I (Mebius 2014b) suggests that the Cochrane Handbook (Higgins and Green 2008) is internally inconsistent: it cautions against combining results from randomised and non-randomised studies in meta-analyses, but subsequently states that meta-analysis should be considered when the studies are sufficiently homogeneous in "participants, interventions and outcomes" (Higgins and Green 2008, p. 137). This, I argue, implies that observational studies can and should be included in meta-analyses if they meet the homogeneity criteria.

The solution is clearly more nuanced than simply giving observational studies equal weight. Shrier et al. (2007) suggest that when carrying out meta-analyses or deciding what evidence to rely upon, researchers can choose between (1) taking the time to assess the probability of bias from lack of randomisation in an observational study to determine whether to include its data, or (2) simply deciding that data from observational studies should be excluded—the implication being, perhaps, that it is rather lazy to exclude observational studies without further thought. Indeed, observational studies may enhance the data available to the meta-analyst, but they should be included with caution, and only after a careful assessment of the quality of each study and the evidence that it provides.

If the results of the two different types of study are consistent or homogenous, as is often the case, there is no reason not to accept evidence from non-randomised studies. In that case, it seems reasonable to pool results from the two types of studies when carrying out meta-analysis, especially where there is a shortage of randomised studies on a particular clinical treatment, because doing so may improve the quality of the meta-analysis and

systematic review.

2.2 Randomised trials and observational studies: the current philosophical controversy

As noted in the previous section, the issue is often not the type of study but how well it is conducted. Paper II (Howick and Mebius 2016) suggests that carefully controlled observational study with a large effect could provide stronger evidence than a confounded randomised trial with a small effect. As others have noted, many randomised trials are poorly conducted (Altman 2002). In contrast, some observational studies are extremely well conducted (Guyatt et al. 2008), to the degree that the relative advantages of randomised trials can sometimes be outweighed by the quality of observational studies.

The main issue is that empirical evidence indicates that observational studies tend to suffer from problems that are known to increase the risks of producing inflated effect estimates in clinical trials. There are problems with inadequate random allocation of participants to treatment groups (Savović et al. 2012), unsuccessful concealment of allocation sequence from those assigning participants to intervention groups (allocation concealment) (Schulz and Grimes 2002a), and lack of adequate double-blinding of investigators assessing the outcome and/or doctors and participants in the study.

Random allocation, allocation concealment and double-blinding are all used to reduce bias in randomised studies (Jüni et al. 2001) but are much harder if not impossible to implement in observational studies. Since observational studies cannot use such measures, they should be more likely, on average, to overestimate treatment effects because they are less likely to reduce bias.[1]

Although arguments for the methodological value of double-blinding in randomised trials are important to the philosophical discussion surrounding the evaluation of medical treatments, they are independent from the arguments for the value of randomisation. Nevertheless, they have important epistemological and ethical implications that deserve attention. The rest of this section outlines philosophical/scientific problems surrounding the use of blinding and placebo controls in clinical studies (for a comprehensive survey and discussion of related issues, see Howick 2011).

2.2.1 Blinding and placebos

The importance of blinding is widely recognised, and is emphasised when judging a study's risk of bias or methodological quality (Guyatt et al. 2008). Yet many crucial aspects of blinding are not well understood. This section describes some features of blinding and placebo controls that are particularly relevant to the discussion in papers I–III (for extensive

[1] It is important to note that single blinding can be achieved in observational studies by ensuring that the outcome assessors are blinded to exposure status, even if the participants and their caregivers are aware of the treatment.

philosophical discussions of blinding and placebos, see Howick 2009; 2011; Turner 2012). I propose that the justification for blinding is, on the whole, sound, but that many ethical and epistemological issues need to be further addressed.

A double-blind trial is normally a trial in which neither the participants nor the caregivers are aware of who receives the experimental intervention. The blinding of outcome assessors is often implicit in a double-blind design (Schulz, Chalmers, and Altman 2002). Double-blinding is considered an important feature of randomised trials to minimise bias caused by beliefs, attitudes, and expectations (Higgins and Green 2008).

Blinding can protect against expectation bias introduced by the presumed benefit of the therapy. If caregivers are not blinded, or if they become unblinded during the performance phase of a trial, their beliefs, attitudes, and expectations about the efficacy of the experimental treatment can be transferred to the trial's participants. Caregivers' attitudes about an intervention might also affect the amount of supplementary care they provide to participants. For example, a caregiver who believes that the experimental treatment is insufficiently effective might be tempted to provide additional care (a co-intervention), increase the dose of the experimental medication, or even withdraw the patient from the trial (Schulz and Grimes 2002b). Participants who believe that they are receiving the treatment may entertain favourable expectations, thinking that they are receiving the 'latest and greatest' therapy. Conversely, participants may have increased anxiety about the new treatment. Study participants may consequently experience changes in health status caused by their beliefs and those of their caregivers rather than because of the treatment.

Blinding can also be implemented to protect against bias in outcome assessment. This is usually called 'ascertainment bias' or 'sampling bias' in the non-medical literature. Ascertainment bias refers to the systematic distortion of results that may arise when those assessing the outcome know which group received the experimental treatment. For example, when assessing a randomised trial of patients with multiple sclerosis, unblinded neurologists found a clear benefit to the experimental treatment, whereas blinded neurologists found that the treatment had no benefit over placebo (Noseworthy et al. 1994).

The underlying rationale for blinding is that, in principle, if the people involved in the trial are ignorant of who is in the treatment and control groups, the effects of expectations will be balanced across all groups. However, there is debate about the extent that this is achieved in practice in trials described as 'double-blind' (Hróbjartsson et al. 2007).

Evidence from meta-epidemiological studies suggests that blinding, when successful, reduces bias (Schulz and Grimes 2002a), especially in trials where there is a risk of subjectivity in expectations or assessment influencing the results (Day and Altman 2000). Inadequate blinding has been associated with inflated effect estimates in studies of drug interventions (Schulz and Grimes 2002b) and with underreporting statistically significant harmful effects of drug treatments (Nieto et al. 2007). Inadequate blinding of outcome assessors is closely associated with the risk of ascertainment bias, which introduces systematic differences between groups in how outcomes are assessed (Higgins and Green 2008). Schulz et al. (1995) found that odds ratios were overestimated by 17% in clinical trials

lacking blind outcome assessment. Hróbjartsson et al. (2007) similarly found that trials not described as double-blind reported 14% larger treatment effects on average compared to similar studies described as double-blind.

Methodological criteria to protect against bias associated with inadequate blinding have been developed and advocated (Hróbjartsson and Boutron 2011), but implementing these criteria has proven difficult to achieve in practice. The empirical studies suggesting that blinding is beneficial also indicate that blinding procedures are seldom tested, and that when tested they are not very successful (Hróbjartsson et al. 2007). The difficulty of maintaining double blinding in clinical trials may in part be explained by the presence or absence of side effects that differentiate participants receiving the active treatment from the placebo. Double-blind conditions can be particularly challenging to maintain in trials with dramatic treatment effects because obvious or apparent effects are difficult to disguise (Howick 2011). However, dramatic effects appear to be rare in clinical trials (Pereira, Horwitz, Ioannidis 2012), and many of the challenges in implementing and maintaining double-blind conditions can be traced to the use of inadequate placebo controls.

To achieve double-blinding, study participants, clinicians, and other trial personnel must be kept ignorant of which participants receive which treatment. This usually requires the use of placebo controls rather than 'no treatment' groups. Very roughly (see Howick 2011, for a detailed discussion), a placebo control is an intervention capable of making people believe that it is, or could be, the experimental intervention, but that is not the experimental intervention. Although there are different forms of blinding, the most common procedure is to provide a placebo to participants in the control group. Researchers commonly believe that blinding necessitates placebos, including in trials comparing new treatments with existing treatments (Schulz and Grimes 2002b).

To prevent those involved in the trial from distinguishing the placebo control from the experimental intervention, researchers may choose to use 'active placebos'. These are placebos designed to induce side effects that resemble those of the experimental treatment (Howick 2011). To avoid being identified as a placebo, active placebos are also designed to imitate the smell, weight, taste, and outward appearance of the experimental intervention. If the active placebo properly mimics the relevant properties of the experimental treatment, those involved in the trial should be unable to guess who is in the treatment group. Consequently, active placebos are frequently advocated as a means of improving the effectiveness of blinding procedures in clinical trials (Feys et al. 2014).

Reproducing features such as taste can be a formidable challenge, but reproducing side effects can be an intractable problem for three reasons. First, the type and degree of a treatment's harmful effects are often unknown or unexpected until a trial is underway. Second, even if all possible side effects were to be identified before the trial, reproducing them is a separate challenge. Developing a compound that induces specific side effects is similar to developing a drug, warranting a development plan and validation process similar to that required for therapeutic compounds. Several additional studies may be required to confirm both that the substance added to the placebo can produce the identified side

effects and that the added substance has no effect on the experimental outcome (Howick 2011). Third, inducing harmful effects with a control treatment might be unethical for certain clinical conditions.

The many difficulties involved in using active placebos may call into question the possibility of reducing bias through blinding. As noted, empirical studies suggest that successful blinding protects against bias and that blinding is an important measure to implement. However, because blinding is frequently unsuccessful, it is necessary to consider why double-blinding is so difficult to achieve. What are the limitations of current blinding procedures? Can these procedures be strengthened to achieve greater success? How much evidential weight, if any, should be assigned to blinding as an indicator of trial quality compared with other methodological features such as adequate randomisation?

Judgments about the utility of blinding will likely have further implications for many related philosophical issues in discussions about evidence. Philosophical consideration of blinding connects to questions about (i) the epistemic basis for judging the assumed superiority of randomised over observational studies for investigating treatment efficacy, (ii) the ethics involved in blinding with active placebos (as in placebo surgery), (iii) the criteria used to make judgments about risk of bias in studies, and (iv) the criteria used to justify the inclusion of studies in systematic reviews.

If blinding reduces the risk of bias and can also be implemented successfully in randomised trials, then blinding might arguably provide randomised trials with an additional advantage over observational studies by producing less biased estimates of treatment effects. However, it is necessary to consider under what circumstances administering active placebos is ethically acceptable, because study participants randomised to the control arm of a trial would be exposed to possible adverse effects, but would receive no potential benefit from the experimental treatment (Howick 2011).

2.3 In search of justification for the unpredictability paradox

As indicated, differences between randomised and non-randomised studies have proven difficult to detect empirically, which has led to the suggestion that non-randomised studies may have equal merit. This has serious implications for clinical practice, especially when deciding which evidence to accept in the face of conflicting results.

Concluding that evidence from observational studies may sometimes be equally reliable means that in the event of a conflict randomised trial results should not necessarily be presumed to have (epistemic) priority. This conclusion also has a direct effect on meta-analysis, as it means that data from both types of studies may be considered side by side in systematic reviews. The data may even be pooled if conditions are suitable—that is, if the direction of the effect is the same across all studies, both observational and randomised. The Cochrane Handbook (Higgins and Green 2008) cautions against pooling results if the direction of effects varies:

> A systematic review need not contain any meta-analyses...particularly if there

> is inconsistency in the direction of effect, it may be misleading to quote an average value for the intervention effect. (Higgins and Green 2008, p. 279)

An issue consequently arises when some individual results show an effect in one direction and others show an effect in the other direction.

Paper III (Howick and Mebius 2014) notes that the decision of whether to pool data in such cases can produce different conclusions. If conflicting data are pooled, the conclusion drawn is likely to be that the intervention produces no statistically significant effect (see Anglemyer et al. 2014). If data are not pooled, the conclusion is more likely to be that the intervention produces a statistically significant effect, but the direction of this effect is indeterminable (see Odgaard-Jensen et al. 2011). This is an important difference. The choices made in different studies about whether to pool results also reveal a worrying inconsistency in systematic review methods. An outside observer might suspect that the decision of whether to pool results is sometimes made to match the authors' desired conclusions. For example, authors comparing studies with and without adequate randomisation may have prior beliefs about the value of randomisation, leading them to fail to report a non-significant effect. Alternatively, they may have assumed that randomisation did not make a difference and therefore reported the pooled value, showing no significant effect.

Although the Cochrane guidance (Higgins and Green 2008) is helpful in deciding when to pool, the warning against pooling when the effect direction is inconsistent should be made more emphatic (see Mebius et al. forthcoming, for a more complete discussion). In short, when the results of individual studies included in a systematic review differ by a statistically significant margin in both directions, either the results should not be pooled, or the pooled difference should be reported alongside the clarification that the direction of effect for the final result is unclear. This difference in direction may result from the difference between randomised and observational studies or it may result from how well the studies have been conducted.

Note that the problem here is not simply the heterogeneity of the results. It is not misleading to pool results in a systematic review in which there is a very high degree of heterogeneity if the direction of the effect is consistent between the individual studies used, regardless of whether those studies are randomised or observational. However, in such a case, the pooled result will not provide an accurate measure of the real effect.

The issue of heterogeneity also has practical implications. As long as clinicians are confident of the direction of the effect, they can prescribe a clinical treatment secure in the knowledge that at worst it will have very little effect in some patients, and that in some patients it is likely to be very effective. However, if the direction of the effect varies between studies, then the treatment cannot be given without potentially causing harm in some patients, even though it would be very effective in others. Although such treatments are clearly effective for some patients, they warrant further investigation to establish the reasons for the variation before they are prescribed and used widely.

3 The role of mechanisms in evaluating causal effects

Philosophers of science are increasingly interested in looking at how attention to mechanisms can assist thinking about causality and causal explanation. In their landmark paper 'Thinking about mechanisms', Machamer, Darden, and Craver (MDC 2000) present the idea that knowledge of causal mechanisms is often necessary for understanding and explaining biological phenomena. The authors characterise mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (MDC 2000, p. 3).[2]

In more recent work, Craver and Darden (2014) state that the value of reasoning from evidence of mechanism in research studies is clear, for instance in guiding controlled experimentation:

> [T]o design an experiment that rigorously tests a claim about the active organization of the mechanism, one often has to know or presuppose a great deal about what the parts of the mechanisms are likely to be and how they are likely to be (and not be) organized. Meaningful experimentation (with useful interventions and detections) can take place only against a wealth of background knowledge about the active organization of the system under study. (p. 125)

In other words, knowledge of mechanisms is considered by these authors (and by many others) to be essential for adequately planning and conducting controlled experiments. However, a number of general concerns regarding mechanistic evidence have received scant attention in the philosophical literature. Recent developments in methodological work and discussions about mechanistic research could be of interest to the philosopher of science because they raise a number of questions about the epistemic significance of mechanistic evidence. In section 3.3.1, I briefly cover some methodological considerations about the relative advantages and disadvantages of mechanistic evidence from basic science research. I argue that we have little warrant for having confidence in mechanistic evidence.

Many of these considerations are connected to a more general problem for philosophical accounts of mechanisms in that they are neither set up to distinguish nor very good at distinguishing differences in the causal productivity of mechanisms, especially by not recognising the importance of effect size in research studies. These problems are coupled with those covered in some detail in papers IV and V. Paper VI covers issues that arise in evidential accounts of technological functions that rely on mechanisms for knowledge about medical functions.

[2] There are considerable dissimilarities between the MDC definition and those proposed in competing accounts of mechanisms (see Bechtel and Abrahamsen 2005; Glennan 2002; Woodward 2011) but these differences do not affect the argument developed in this thesis.

3.1 On the causal role of absences in mechanistic explanation

Much work in basic medical science involves understanding the causes and effects of absences or absent causes. This is because mechanisms typically involve absences as causes or effects. Causation by absence is an especially interesting type of causal relation where the cause (or the effect) is the absence or 'non-presence' of something causing something else. Finding these causes and understanding the impact they have is an important part of understanding impaired functions associated with different forms of disease.

An oft-cited instance of absence causation in the philosophical literature is an individual's house-plant dying because of that individual's failure (or, to be more specific since this case involves agency, his or her *omission*) to water the plant. Other well-known examples from the same literature include being late for a meeting because of missing a bus (again, a case of omission), the absence of oxygen in the bloodstream causing death, the absence of insulin causing type 1 diabetes, and the absence of vitamin C causing scurvy.

Paper IV (Mebius 2014a) notes that causal effects caused by absences present an interesting problem for the notion of productive (causal) continuity. Such a notion is currently adopted in one of the most prominent theories of mechanisms, advanced by MDC (see above). Productive continuity entails a continuous connection or link (or 'oomph') between the entities involved in mechanistic processes. The idea of an absent cause becomes problematic in such processes because absences, if thought of as causes, cannot be physically linked to their effect. Absences are nowhere to be found. But how, then, are they able to produce physical effects?

Philosophers have tried to solve this problem by viewing absences as causes of mechanism failures:

> The standard response of the mechanist to this problem of causation between absences is to say that the causal claim is made true by the fact that the expected mechanism didn't operate. (Williamson 2013, p. 263)

The solution recently proposed by Benjamin Barros (2013) accordingly suggests that we interpret causal absences as causes of mechanism *failures*. This solution is interesting because it deals with the problem of causal continuity directly—no longer does locating and explaining the effects of absences in causal process present a problem because absences are causes that make mechanism productivity grind to a halt (that is, they cause an interruption in mechanism production). It is thought that many diseases result from failures caused by absences (see section 3.2 below).

Paper IV challenges this view and questions the value of solving the problem by characterising absences as causes of mechanism failure. It argues that this description seems inadequate because the effects of absences rarely cause mechanism failure. Instead, causal absences typically produce less dramatic effects. For example, unlike in type 1 diabetes, where the body fails to produce insulin, in type 2 diabetes (which is considerably more common) the body produces insufficient insulin to meet increased cellular requirements; consequently, the mechanism productivity in type 2 diabetes is not disrupted.

Broadly, the argument presented is that the causal effects attributed to absences will depend on the size of the effect and that most effects of absences are small or moderate. As such, these effects do not generally cause mechanisms to fail. In practice, absences tend to produce small to moderate effects on mechanisms that often cause *attenuating* effects on the mechanism production. Additionally, this paper suggests that the effects of absences are best measured with continuous effect size estimates since they may well vary both in direction (that is, in having a negative or positive outcome) and in effect size.

3.2 An omitted hallmark of mechanism function

Paper V suggests that the notion of causal function presented in much recent philosophical work on mechanisms omits an account of the varying effectiveness and efficiency with which functions can be performed. I argue that contemporary theories of mechanism function have failed to recognise the importance of accounting for these differences. Paper V suggests that researching biomedical functions often requires describing the continuum (i.e. undisrupted operation) of vacillating levels of efficacy and efficiency by which mechanistic entities and activities work.

When it comes to understanding a mechanism's function it is often misleading or unsatisfactory to ask 'Does it work?' rather than 'How much does it work?' or 'To what extent does it work?'; other relevant questions include 'How well does it work?' and 'How likely is it to work?' Answers to the second category of questions require more quantitative precision than is typically present in contemporary accounts of mechanism function, which have traditionally been more concerned with qualitative features of mechanisms. One overlooked but important point about mechanism functions is made clear in cases such as declining cognitive function in Alzheimer's disease, where functioning is not simply present or absent. Rather, functioning may be more or less present in gradations ranging from highly effective cognitive performance to highly ineffective cognitive performance. Degrees of effectiveness are important in theories of mechanisms because if the mechanism that serves the function has some variation in the size of its produced effect, it is necessary to be able to account for any resulting differences in functional performance.

Hence, an under-emphasised feature of mechanisms is the vacillating efficiency or differing effectiveness with which a mechanism's functions can be performed. Functional effectiveness or efficiency can be viewed as continuous variable rather than a dichotomous one. Nevertheless, at certain points along the continuum, mechanism operation becomes increasingly or decreasingly pathological, with correspondingly low levels of functional efficiency and effectiveness. This incremental variability in mechanism function is similar to the ways that pharmacological treatments and other medical technologies can be more or less effective or harmful in patients.

An underlying assumption here is that mechanisms' functions need to be established because their effects cannot simply be assumed before being established empirically. For example, it would be a misuse of language to say that a pharmacologic substance has the

function of curing a disease if it kills the patient in the process of doing so. Accordingly, a mechanism can only have a function once a useful (i.e. functional) effect has been established.

3.3 Ascribing functions to medical technologies

Most breakthroughs in modern medicine would not have been possible without technological innovations. Medical technologies have significantly improved our ability to diagnose and treat patients and have consequently led to important advances in the treatment and prevention of disease, saving countless lives in the process. However, many episodes in clinical medicine reveal technologies that have been ineffective or harmful because they were implemented on the basis of scant or unreliable evidence. A special point of concern, therefore, is how to assure that technologies are properly functional before adopting them.

Paper VI argues that a recent method developed by philosophers of technology for ascribing and evaluating technological functions requires further improvement because it fails to provide adequate evidential justification. The influential ICE theory of functions (Vermaas and Houkes 2006) exemplifies an excessively permissive theory of evidence that relies too extensively on expert testimony and knowledge of mechanisms as both evidence for and means of evaluating technological functions (cf. Hansson 2006). Expert testimony is problematic because some experts may claim that a technological intervention is effective while other experts may disagree. The key question then is how to determine the relative value of a medical intervention given contradictory claims about the safety and effectiveness of that intervention. I argue that the ICE epistemology of function ascription proves inadequate for ascribing clinically relevant functions to medical technologies because of the very high risk of bias generally inherent in evidence of expertise and evidence of mechanisms.

3.3.1 Mechanisms as evidence

A common critique of the EBM epistemology of evidence is that it advises against considering mechanistic evidence alongside evidence from clinical studies. For example, in a recent paper by Clarke, Gilles, Illari, Russo, and Williamson (2014), the authors criticise EBM proponents for not trusting mechanistic evidence when evaluating medical interventions. The authors argue that mechanistic evidence "should be treated alongside, rather than as inferior to, the evidence of correlation provided by statistical trials" (Clarke et al. 2014, p. 340, cf. Steele 2008). Clarke et al. support their claim with several compelling case studies where evidence of mechanisms derived from animal experiments or in vitro studies was used successfully in conjunction with evidence of correlation and led to satisfactory treatments, as in the case of streptomycin therapy for treating tuberculosis. It should be noted that most mechanistic knowledge is obtained from experiments on animal models (in vivo) or cells (in vitro).

Many critics of the EBM view of mechanisms emphasise the role of mechanistic evi-

dence derived from laboratory experiments (mostly using animal models) in guiding clinical research and in extrapolating results to individual patients. As observed by Craver and Darden in their recent book, "the use of model systems is an indispensable part of biological practice" (2014, p. 138). La Caze (2011) likewise notes that:

> What is known or supposed at the mechanistic level plays a vital role in specifying both the models of experiment and the models of data. And once specified, the statistical findings based on the observed data provide information on the clinical applicability of the mechanism provided by basic science only if the assumptions made in the experimental and data model hold. (p. 94)

Bluhm (2013) agrees, arguing that "knowledge of mechanisms should be used to design epidemiological studies...epidemiological research informed by knowledge of physiological mechanisms will provide a better evidence base for medicine" (p. 426). Anderson (2012) considers that the most important role of mechanistic evidence is "generating ideas for novel treatment possibilities in the first place, such that subsequent controlled trials can test" (p. 993).

However, no evidence to date supports many of the arguments that clinical research produces high levels of clinical benefit (Pound and Bracken 2014). Recent empirical evidence suggests, to the contrary, that very few findings successfully translate from animal research to effective treatments in clinical practice (Ioannidis et al. 2014). This is not to say that developing *in vivo* or *in vitro* models is a useless methodological approach and that successful clinical applications of discoveries in animal studies are never obtained. Rather, recent empirical studies seem to suggest that the positive clinical impact of basic science research is far less frequent than is commonly assumed (Landis et al. 2014). The substantial epistemological and methodological challenges facing the translation of animal models of human disease to successful treatments in clinical practice has been almost entirely overlooked in the philosophical discussion about the role of mechanisms in EBM.

Many researchers rightly lament the poor quality of study design, execution, and reporting of results that currently prevails in animal research (Macleod 2014; Perrin 2014). Ioannidis et al. (2014), for example, note a clear lack of methodological quality in animal studies that can be interpreted as a sign of sloppy research,[3] including inadequate randomisation of animals to experimental and control groups (Hirst et al. 2014). Ioannidis et al. also report a problem with inadequate blinding of experimenters; proper blinding promotes impartial handling of measurements (see section 2.2.1). The poor methodological quality of animal studies appears to contribute to the irreproducibility of most of the evidence for mechanisms (Begley and Ioannidis 2015):

> Over the recent years, there has been an increasing recognition of the weaknesses that pervade our current system of basic and preclinical research. This has

[3]Note that some of these concerns can be expressed for certain observational studies and randomised trials, while others are unique to animal research.

> been highlighted empirically in preclinical research by the inability to replicate the majority of findings presented in high-profile journals. The estimates for irreproducibility based on these empirical observations range from 75% to 90%. (p. 116)

Similarly, Begley and Ellis (2012) suggest that the low validity of preclinical findings is neither sustainable nor acceptable, and that researchers must reassess the ability of current approaches to translate preclinical findings into clinical benefit, noting that "[t]he lack of rigour that currently exists around generation and analysis of preclinical data is reminiscent of the situation in clinical research about 50 years ago" (p. 531).

Researchers have highlighted major issues related to the use of animal models (e.g. rodents) to predict clinical efficacy in biomedical research (Sorge et al. 2014). However, it is possible that the failure of animal models to predict human outcomes might result less from the poor quality of basic research than from the incomplete nature of the models (Pound and Bracken 2014; Kokolus et al. 2013).

One major concern in clinical research is the effect of bias caused by baseline differences between experimental and control groups. Human test subjects are especially susceptible to bias due to significant interspecies variance and varying living conditions. Methodological measures such as random allocation of treatment and double-blinding can help to minimise this bias.

A related problem caused by the poor quality and poor reporting of animal findings is the common occurrence of false positive results (Perrin 2014; Prinz et al. 2011). False positive results should be of great concern because they suggest that interventions are effective when they are not. Perhaps more importantly, they may lead researchers to conduct several unnecessary and unethical trials where study participants will be exposed to ineffective and potentially harmful interventions.

As mentioned previously, many philosophers propose using mechanisms to solve the problem of external validity when generalising from randomised controlled trial results:

> Evidence of mechanisms helps assess the external validity of an RCT [randomised controlled trial] (or indeed of any study) because it adds precious knowledge about the similarities between the test and target populations. (Clarke et al. 2013, p. 746)

However, while clinical studies are performed on human patients, mechanistic evidence is inferred largely from controlled experiments or from trials on *non-human animals*. The artificial environment created in the lab is arguably likely to be less externally valid than the environment of a clinical trial with actual patients in a hospital. As Howick (2011b) correctly points out, why should results from animal studies with severely limited external validity be used as evidence to support extrapolating the results of randomised trials? Evidence from in vitro experiments (e.g. cell cultures and test tubes) are even less likely to generalise because these tests are not conducted on whole animals (Ioanidis 2006).

This brings us to another issue in the philosophical discussion of mechanistic evidence. Clearly, many authors believe that mechanistic evidence is different from statistical evidence. For example, Clarke et al. (2014) state that:

> For a wide variety of reasons, non-statistical evidence of mechanisms should often be used in conjunction with, rather than viewed as inferior to, statistical evidence of correlation. (p. 358)

However, there is no reason to assume that the mechanistic evidence that matters here is non-statistical or qualitative. First, talk about (causal) effects, as in 'therapy C has an effect E', implies that the effect can be quantified. Talking about effects entails that the effect size can be estimated. Consequently, mechanisms that support effects will provide quantitative or statistical evidence, because there is no way for qualitative studies to provide a similar estimate of effect.[4] Second, as in human trials, animal studies are conducted on a sample and measure the overall effect size. Most animal studies can therefore in principle be regarded as clinical trials on non-human animals. As such, similar methodological considerations to those that apply in controlled clinical trials with human participants should apply when conducting medical experiments on animals.

One potential benefit of animal studies is a reduced risk of self-selection bias, since animals do not choose to be experimented upon. This confers an advantage to animal designs compared with observational designs, because observational studies are more likely to suffer from self-selection bias (see section 2.2.1, above). However, unlike observational studies, animal studies do not provide data about patients in healthcare settings, but rather data on animals in controlled laboratory environments. This raises serious doubts about their external validity—that is, whether results from animal trials can be extrapolated to humans or to other animals outside the trial environment. In fact, given the results of meta-epidemiologic studies investigating the quality of animal research, there is no reason to suspect results from observational studies and human trials to be more inconsistent than those between animal studies and randomised trials (Ioannidis 2006).

4 Summary

In this thesis I evaluated controversies surrounding the evidential roles of randomised trials and mechanistic evidence. I found that the special privilege given to randomised trials—most vociferously by proponents of EBM—lacks a sound justification. Specifically, the decision to exclude evidence from observational studies from clinical decision making in general and systematic reviews of 'best evidence' in particular cannot be defended. On the other hand, the solution to some of the problems with randomised trials put forth by many critics of EBM—namely the use of mechanistic evidence—is also problematic.

[4] That is, claiming that there is a causal relation between C and E assumes that C has an effect on E by which the relation can be qualified; one might express this as a difference in 'oomph' between E and C.

Not only is mechanistic evidence both intrinsically and extrinsically unreliable as evidence for the clinical benefits of medical treatments, but there are a variety of other problems with mechanistic evidence such as their inability to incorporate effects of absences, and the heterogeneous ways in which functions of mechanisms should be interpreted. All of the arguments in this thesis therefore add to the growing body of work in the philosophy of medicine, and how it can be applied to medical treatments and technologies.

References

Altman DG (2002) Poor-quality medical research: what can journals do? *Journal of the American Medical Association*, 287:2765–2767.

Andersen H (2012) Mechanisms: what are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice*, 18:992–999.

Anglemyer A, Horvath HT, Bero L (2014) Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews*, 4:MR000034.

Barros DB (2013) Negative causation in causal and mechanistic explanation. *Synthese*, 190:449–469.

Bechtel W, Abrahamsen, A (2005) Explanation: a mechanist alternative. *Studies in History and Philosophy of Science*, 36:421–441.

Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. *Nature*, 483:531–533.

Begley CG, Ioannidis JPA (2015) Improving the standard for basic and preclinical research. *Circulation Research*, 116:116–126.

Benson K, Hartz AJ (2000) A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342:1878–1886.

Bluhm R (2013) Physiological mechanisms and epidemiological research. *Journal of Evaluation in Clinical Practice*, 19:422–426.

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

Clarke B, Gillies D, Illari P, Russo F, Williamson J (2014) Mechanisms and the evidence hierarchy. *Topoi*, 33:339–360.

Concato J, Shah N, Horwitz RI (2000) Randomized controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342:1887–1892.

Craver CF, Darden L (2014) *In Search of Mechanisms*. Chicago: University of Chicago Press.

Day SJ, Altman DG (2000) Statistics notes: blinding in clinical trials and other studies. *British Medical Journal*, 321:504.

DeCensi A, Puntoni M, Goodwin P et al. (2010) Metformin and cancer risk in diabetic patients: a systematic review and meta-analysis. *Cancer Prevention Research*, 3:1451–

1461.

Feys F, Bekkering GE, Singh K, Devroey D (2014) Do randomized clinical trials with inadequate blinding report enhanced placebo effects for intervention groups and nocebo effects for placebo groups? *Systematic Reviews*, 3:14.

Glennan S (2002) Rethinking mechanistic explanation. *Philosophy of Science*, 69:S342–S353.

Guyatt GH, Oxman AD, Vist GE et al. (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336:924–926.

Hansson SO (2006) Defining technical function. *Studies in History and Philosophy of Science*, 37:19–22.

Higgins JJ, Green S (2008) *The Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]. Chichester: Wiley–Blackwell.

Hirst JA, Howick J, Aronson JK et al. (2014) The need for randomization in animal trials: an overview of systematic reviews. *PLoS ONE*, 9:e98856.

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

Howick J (2011b) Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making. *Philosophy of Science*, 75:926–940.

Howick J (2009) Questioning the methodologic superiority of 'placebo'over 'active'controlled trials. *American Journal of Bioethics*, 9:34–48.

Howick J, Mebius A (2016) "Randomized trials and observational studies: the current philosophical controversy". In Schramme T and Edwards S (eds.) *Handbook of the Philosophy of Medicine*. Springer.

Howick J, Kennedy AG, Mebius A (2015) "Philosophy of evidence-based medicine". In Pritchard D (ed.) *Oxford Bibliographies in Philosophy*. Oxford: Oxford University Press. Available at: http://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0253.xml (Accessed 10-03-2015).

Howick J, Mebius A (2014) In search of justification for the unpredictability paradox. *Trials*, 15:480.

Hróbjartsson A, Forfang E, Haahr MT et al. (2007) Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology*, 36:654–663.

Hróbjartsson A, Boutron I (2011) Blinding in randomized clinical trials: imposed impartiality. *Clinical Pharmacology and Therapeutics*, 90:732–736.

Ioannidis JPA (2006) Evolution and translation of research findings: from bench to where? *PLoS Clinical Trials*, 1:e36.

Ioannidis JPA, Greenland S, Hlatky MA et al. (2014) Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, 383:166–175.

Ioannidis JPA, Haidich AB, Pappa M et al. (2001) Comparison of evidence of treatment effects in randomized and non–randomized studies. *Journal of the American Medical Association*, 286:821–830.

Jüni P, Altman DG, Egger M (2001) Systematic reviews in health care: assessing the quality of controlled clinical trials. *British Medical Journal*, 323:42–46.

Knekt P, Reunanen A, Jarvinen R et al (1994) Antioxidant vitamin intake and coronary mortality in a longitudinal population study. *American Journal of Epidemiology*, 139:1180–1189.

Kokolus KM, Capitano ML, Lee CT et al. (2013) Baseline tumor growth and immune control in laboratory mice are significantly influenced by subthermoneutral housing temperatures. *Proceedings of the National Academy of Sciences*, 110:20176–20181.

La Caze A (2011) The role of basic science in evidence-based medicine. *Biology and Philosophy*, 26:81–98.

Landis SC, Amara SG, Austin CP et al. (2012) A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, 490:187–189.

Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. *Philosophy of Science*, 60:1–25.

MacLehose RR, Reeves BC, Harvey IM et al. (2000) A systematic review of comparisons of effect sizes derived from randomised and non–randomised studies. *Health Technology Assessment*. 4:1–154.

Mebius A (2014a) A weakened mechanism is still a mechanism: on the causal role of absences in mechanistic explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 45:43–48.

Mebius A (2014b) Corroborating evidence-based medicine. *Journal of Evaluation in Clinical Practice*, 20:915–920.

Mebius A, Aronson JK, Howick J et al. (forthcoming) Illegitimate pooling in systematic reviews of randomized trials versus observational studies.

Mebius A, Kennedy AG, Howick J (forthcoming) Research gaps in the philosophy of evidence-based medicine.

Nieto A1, Mazon A, Pamies R et al. (2007) Adverse effects of inhaled corticosteroids in funded and nonfunded studies. *Archives of Internal Medicine*, 167:2047–2053.

Noseworthy JH, Ebers GC, Vandervoort MK et al. (1994) The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*, 44:16-20.

Oberai P, Balachandran I, Janardhanan N et al (2013) Homoeopathic management in depressive episodes: a prospective, unicentric, non–comparative, open–label observational study. *Indian Journal of Research in Homoeopathy*, 7:116–125.

OCEBM Levels of Evidence Working Group (2011): The Oxford 2011 Levels of Evidence, Oxford Centre for Evidence-Based Medicine. Available at: http://www.cebm.net/index.aspx?o=5653 (Accessed 10-03-2015).

Odgaard–Jensen J, Vist GE, Timmer A et al (2011) Randomization to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews*, 4:MR000012.

Pereira TV, Horwitz RI, Ioannidis JPA (2012) Empirical evaluation of very large treatment effects of medical interventions. *Journal of the American Medical Association*,

308:1676–1684.

Perrin S (2014) Make mouse studies work. *Nature*, 507:423–425.

Pound P, Bracken MB (2014) Is animal research sufficiently evidence based to be a cornerstone of biomedical research? *British Medical Journal*, 348:g3387.

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10:712.

Savović, J, Jones HE, Altman DG et al. (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Annals of Internal Medicine*, 157:429–438.

Schulz KF, Chalmers I, Altman DG (2002) The landscape and lexicon of blinding in randomized trials, *Annals of Internal Medicine*, 136:254–259.

Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, 273:408–412.

Schulz KF, Grimes DA (2002a) Allocation concealment in randomised trials: defending against deciphering. *Lancet*, 359:614–618.

Schulz KF, Grimes DA (2002b) Blinding in randomised trials: hiding who got what. *Lancet*, 359:696–700.

Shrier I, Boivin JF, Steele RJ et al. (2007) Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology*, 166:1203–1209.

Sorge RE, Martin LJ, Isbester KA et al. (2014) Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods*, 11:629–632.

Steel D (2008) *Across the Boundaries. Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Straus SE, Glasziou P, Richardson WS, Haynes RB (2011). *Evidence-Based Medicine: How to Practice and Teach EBM*. 4th edn. Edinburg: Churchill Livingston.

Turner A (2012) 'Placebos' and the logic of placebo comparison. *Biology and Philosophy*, 27:419–432.

Vermaas PE, Houkes W (2006) Technical functions: a drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science*, 37:5–18.

Williamson J (2013) How can causal explanations explain? *Erkenntnis*, 78:257–275.

Woodward J (2011) Mechanisms revisited. *Synthese*, 183:409–427.

Worrall J (2002) What evidence in evidence-based medicine? *Philosophy of Science*, 69:S316–S330.