Evidence Based Medicine: Evolution, Revolution, or Illusion?

A philosophical examination of the foundations of Evidence Based Medicine

Adam La Caze

A thesis submitted for the degree of Doctor of Philosophy
Department of Philosophy, The University of Sydney, 2009

Preface

Declaration of Originality. The content of this thesis represents my own original contribution. Where the work of others is discussed it is clearly referenced. No part of this thesis has been submitted for another degree.

Published Papers. The contents of Chapter 2, *Evidence based medicine can't be ...*, is to be published in *Social Epistemology* (La Caze 2008a).

Abstract

Evidence based medicine (EBM) has become the dominant model for informing and justifying therapeutic decisions. There is increasing philosophical interest in whether the central claims of EBM can be justified. EBM puts forward a methodological hierarchy as a criterion for justifying therapeutic decisions. This hierarchy privileges evidence from randomised interventional studies as a basis for inference over alternative forms of evidence such as observational studies and basic medical science. Proponents of EBM have provided surprisingly little justification for their methodological claims. The early chapters of the thesis examine the justification and interpretation of the hierarchy provided by proponents of EBM. The only sustainable justification of the hierarchy is as a hierarchy of comparative internal validity. Study designs higher up the hierarchy have the capacity to rule out more sources of systematic error than study designs lower down. While there are good reasons for preferring randomised interventional studies for testing the efficacy of drugs, high internal validity is not sufficient for informing therapeutic decisions. In the later sections of the thesis, I turn to the question of external validity. The crucial role that observational studies and basic science play in the application of clinical research is demonstrated. In the final chapters, I argue that some of the frequentist methods currently employed to analyse clinical data are ill suited to the task of informing therapeutic decisions. EBM is promoted as a rationalist turn in medicine. If EBM is to fulfil this promise, more attention is needed on the foundations of the approach. This thesis examines the foundational arguments of EBM, and observes the limits of these arguments in informing therapeutic decisions.

Acknowledgements

I have been fortunate to receive much support and advice in completing this work. I especially thank my supervisors. From the project's inception to its current status, Mark Colyvan, my principle supervisor, has been an exceptional guide. I have gained much from witnessing Mark's approach to philosophy, both as an enjoyable activity of intrinsic merit and as a profession. I am a grateful beneficiary of Mark's generosity, optimism and wisdom.

Jason Grossman, my associate supervisor, has been an tireless source of helpful advice and optimism. I am indebted to Jason for his passionate and patient guidance through the challenging terrain of philosophy of statistics. Jason's expertise has benefited the thesis immensely.

Katie Steele has been my unofficial third supervisor and confidant. Katie graciously took me under her wing early in my PhD. Our many conversations on matters of formal and informal philosophy have enlightened me as much as I have enjoyed them.

For coffee, conversations, ideas and inspiration I thank Neil Cottrell, Damian Cox, Chris Cutts, Steve Duffull, Fiona Fidler, Christopher Grannell, Paul Griffiths, Margeurite La Caze, Aidan Lyon, Frank May, John Matthewson, Hatha McDivitt, Fabien Medvecky, Lisa Nissen, Debra Rowett, Nick Shaw, Sue Tett, Neil Thomason, Scott Tyler, Karl Winckel, Bonnie Wintle, and John Worrall.

I have worked with a number of excellent pharmacists and other health professionals who have shaped the questions broached in this thesis. In particular, I thank Chris Cutts, Neil Cottrell, Ian Coombes, Steve Duffull, Debra Rowett and Frank May. Through previous work I had the opportunity to meet regularly with a group of general practitioners working in rural Queensland, I thank them for sharing their insights on the challenges of therapeutic decision making.

Fabien Medvecky, Karl Winckel and Scott Tyler kindly read and com-

mented on sections of the thesis in its current form.

For long-term support, and for supplying all the preconditions that are impossible to list, I thank David, Maureen, Megan, Damien and Sarah. And, for her unconditional and unwavering support, I thank Sanam.

Prologue

Rofecoxib was a heavily marketed, and widely used, anti-inflammatory agent that was withdrawn from the worldwide market in September 2004. Now, rofecoxib is a well recognised, and widely used, example of the challenges of regulatory and clinical decision making.

The rofecoxib story is striking. Rofecoxib was withdrawn following post-marketing evidence that it increased the risk of heart attacks and strokes. Prior to this, rofecoxib had progressed through the phases of drug development without significant hitch. That is, its risks were not uncovered until after it had passed what we accept to be the 'best tests' of a drug's efficacy and safety. But this is not what is especially striking about rofecoxib. Despite the rigours of drug development, some rare, though catastrophic, adverse effects can only be detected once a drug is used by a very large number of people. What is especially striking in this case is that the risks of rofecoxib are relatively common, both in terms of presentation and frequency. The cardiovascular effects of rofecoxib are not entirely idiosyncratic. Indeed, the possibility of an increased risk of blood clots (leading to heart attacks or strokes) is foreseeable on pharmacological grounds. And, based on current data, the frequency of heart attacks and strokes caused by rofecoxib is in the range of 30 to 80 additional events per 10,000 patients per year; by contrast, a 'rare' adverse effect, say, for example penicillin anaphylaxis, occurs in about one patient in every 10,000.

Even more disturbing is that rofecoxib was developed in anticipation of its *safety* benefits. In many ways rofecoxib, and the class of drugs it belongs to, had all the hallmarks of being a success story for biomedical science. Traditional anti-inflammatory agents, despite their effectiveness in relieving pain and inflammation, pose significant public health risks due to their propensity to cause gastrointestinal damage. The COX-2 inhibitor class, of which rofecoxib is a member, were developed out of an increased pharmacological

and physiological understanding of how anti-inflammatories work. COX-2 inhibitors are pharmacologically targeted to work just as well as traditional anti-inflammatories while minimising gastrointestinal risks. On this basis, these drugs were eagerly awaited, and once available, heavily prescribed. The anticipated improvements in gastrointestinal safety were eventually established. Unfortunately, the cardiovascular risks identified a couple of years later are of a similar magnitude to the gastrointestinal benefits.

Rofecoxib was approved, and spent four and a half years on the market, before its propensity for grave side effects was established. Clearly, this is an undesirable state of affairs. A number of questions arise immediately, perhaps the most urgent among these is: *what went wrong?* The answer is unsettling; in a strong sense, nothing went wrong. Rofecoxib was tested according to the methods we accept to be our 'gold standard'. Specifically, a significant number of reasonably sized randomised controlled trails conducted prior to and soon after market approval failed to demonstrate the cardiovascular risks. The VIGOR study (Bombardier et al. 2000) provided some (disputed) evidence that rofecoxib may increase the risk of heart attacks and strokes, but it was not until Bresalier et al. (2005) reported the results of the APPROVe trial, that rofecoxib was withdrawn from the market. Prior to APPROVe, a meta-analysis that combined the results of the pre- and post-marketing randomised controlled trials supported the cardiovascular safety of rofecoxib (Weir et al. 2003). Perhaps it should be noted that this meta-analysis was conducted by the sponsor, and that questions have been raised as to whether the reporting and publication of rofecoxib's key trials were of the highest clinical and ethical standards (see Krumholz et al. 2007, and Ross et al. 2008). But issues such as these are considerably easier to pick up once we are sifting through the wreckage under the glare of flood lights. And, despite their significance, these transgressions are not sufficient to explain the amount of time it took to establish rofecoxib's safety profile—especially when the magnitude of rofecoxib's risks are approximately equal to its benefits. It is important to reflect on what should have been done better, but we should also acknowledge an important underlying point: our *best methods*—or, more accurately, what we currently consider to be our best methods—led us astray.

Regulatory and clinical decisions rely heavily on evidence from randomised controlled trials. This study design is seen to provide the most reliable evidence about medicines (the only evidence thought to be more reliable is

the results of a number of randomised controlled trials combined in a meta-analysis). From a regulatory perspective, drugs will only be put on the market once their efficacy is established in a randomised controlled trial. For medical decision making, Evidence Based Medicine is the main game in town. And Evidence Based Medicine advises clinicians to base their decisions, as much as possible, on evidence from randomised controlled trials.

Rofecoxib raises methodological questions. Do we place too much confidence in the results of randomised controlled trials? On what basis are randomised controlled trials our best source of evidence? Which questions are especially well suited to being tested in randomised controlled trials? And, which questions are poorly tested by this study design? Clearly these are questions for medicine and for public health policy. But just as clearly, these are questions of, and for, philosophy.

This thesis approaches the questions raised by Evidence Based Medicine more generally, as opposed to specific issues surrounding rofecoxib. The questions will be approached from the perspective of contemporary philosophy of science. According to this approach, the questions and methods of philosophy of science are seen to be continuous with the questions and methods of science. Rather than develop prior and abstract truths by which the methods of clinical research can be judged, the inherent complexities of medical research and clinical decision making take centre stage. The epistemological strengths and weakness of the methods we employ to understand and assess treatments are considered in light of these uncertainties.

Contents

Chapter Synopses

Chapter 1 Evidence based medicine (EBM) and the approach it suggests for therapeutic decisions is introduced. An overview of the thesis is provided. Primers on study designs used in clinical epidemiology, and the frequentist statistical analysis of clinical trials are supplied.

Chapter 2 EBM puts forward a hierarchy of evidence for informing therapeutic decisions. An unambiguous interpretation of how to apply EBM's hierarchy has not been provided in the clinical literature. However, as much as an interpretation is provided, proponents suggest a categorical interpretation. Most of the critical replies to EBM react to this interpretation. While proponents of EBM can avoid some of the problems raised by critics by suitably limiting the claims made on behalf of the hierarchy, further problems arise. If EBM is to inform therapeutic decisions then a considerably more restricted, and context dependent interpretation of EBM's hierarchy is needed.

Chapter 3 EBM's hierarchy of evidence places randomised interventional studies (and systematic reviews of such studies) higher in the hierarchy than observational studies, unsystematic clinical experience, and basic science. Recent philosophical work has questioned whether EBM's special weighting of evidence from randomised interventional studies can be justified. Following the critical literature, and in particular the work of John Worrall, I agree that many of the arguments put forward by advocates of EBM do not justify the ambitious claims that are often made on its behalf. However, in contrast to the recent philosophical work, I argue that a justification for EBM's hierarchy of evidence can be provided. The hierarchy should be viewed as a hierarchy of comparative internal validity. While this justification is defensible, the claims that EBM's hierarchy substantiates when viewed in this way are

considerably more circumscribed than some claims found in the EBM literature.

Chapter 4 A number of arguments have shown that randomisation is not *essential* in experimental design. Scientific conclusions can be drawn on data from experimental designs that do not involve randomisation. John Worrall, among others, has argued that randomising provides no guarantee that experimental groups are evenly distribution for all possible confounding factors on any particular allocation. In doing so, however, Worrall makes an additional claim: randomised interventional studies are epistemologically equivalent to observational studies providing the experimental groups are comparable according to background knowledge. I argue against this claim. In the context of testing the efficacy of drug therapies, well-designed interventional studies are epistemologically superior to well-designed observational studies because they have the capacity avoid an additional source of bias. Randomisation in interventional studies is defended on Bayesian grounds.

Chapter 5 The challenge of judging the external validity of clinical research is outlined. Contradicting EBM's claim that therapeutic decisions should be informed by evidence from randomised interventional studies rather than observational studies, I show that evidence from observational studies can play a vital role in making judgements of the external validity of clinical research. Therapeutic decision-makers need evidence from both observational and randomised interventional studies.

Chapter 6 Therapeutic decisions, according to proponents of EBM, should be informed by evidence from randomised interventional studies (and systematic reviews of randomised interventional studies) rather than basic science. Patrick Suppes' hierarchy of data models provides a framework that explicates the link between the theory of basic science, experimental inquiry, and observed data. Relying on the hierarchy of data models I show that basic science is vital not only for specifying experiments, but for analysing and interpreting the data that is provided. Further, and contradicting what is implied in EBM's hierarchy of evidence, basic science is integral to the application of clinical research to therapeutic questions.

Chapter 7 Therapeutic decision makers require data on which patients are especially likely to benefit from the treatment, and which patients are especially prone to adverse effects. But gaining reliable data of this kind is difficult. This is the problem of subgroup data. I outline two responses that are found in the literature. Neither provides a completely adequate reply. What is needed is a more explicit framework for incorporating the theory of basic science into the interpretation and application of clinical research. Patrick Suppes' hierarchy of data models provides such a framework.

Chapter 8 The standard view within clinical trial analysis is that power is irrelevant to the interpretation of the results of a statistical test once they have been observed. I argue against this view. In particular, I show the warrant associated with frequentist statistical inferences depends on the epistemic context of the statistical test. The withdrawal of rofecoxib from the market following the Adenomatous Polyp Prevention Trial (APPROVe) is used as a case study to illustrate this point. Understanding how the warrant for statistical tests differs according to the epistemic context of the test is important for drawing appropriate inferences from clinical trials.

Chapter 9 The main arguments of the thesis are summarised.

Chapter 1

Evidence Based Medicine: A 'paradigm' for therapeutic decisions?

> Most of us with rationalist pretensions presumably aspire to live in a society in which decisions about matters of substance with significant potential social or personal implications are taken on the basis of the best available evidence, rather than on the basis of irrelevant evidence or no evidence at all. Of course, the nature of what constitutes evidence in any particular instance could be a matter for significant debate. But, modulo such debate, most of us have the aspiration to live in a society which is more, rather than less, 'evidence based'.
>
> *The Address of the President, Adrian F. M. Smith, to* The Royal Statistical Society *on Wednesday, June 12th, 1996*[1]

Medical decisions should be based on evidence, indeed, the *best available* evidence. This is the motto of a movement in medicine that has become known as 'Evidence Based Medicine', or 'EBM' for short. As mottos go, it is hardly contentious. EBM has become medical orthodoxy in a relatively short space of time. And yet, despite EBM's status and its indisputable slogan, EBM has engendered considerable debate. This is because the details matter; or, as Adrian Smith understates, 'the nature of what constitutes evidence in any particular instance could be a matter for significant debate'. It is easy

[1]Smith 1996, p. 369

to assert that decisions should be based on the best available evidence. It is considerably harder to be precise about *what* the 'best available evidence' is, *why* it is better than the alternatives, and *how* it should be collected and applied. That medical evidence comes from a range of sources makes these questions especially challenging. How should we *compare* or *integrate* the evidence supplied by basic sciences such as physiology and pharmacology with the evidence supplied by applied clinical research? This thesis examines EBM's reply to these questions.

EBM developed out of clinical epidemiology, which is itself a relatively young discipline. Clinical epidemiology amalgamates 'clinical medicine' and 'epidemiology'. It approaches the questions that arise in the care of individual patients with methods that detect and quantify the influence of particular treatments, or risks, in populations. Clinical epidemiology has developed a wide range of methods to reduce the influence of systematic error and the play of chance in clinical studies. The development of these methods has had a large effect on the kind of research that is conducted in medicine. Prior to clinical epidemiology, 'laboratory research' conducted in the 'basic' medical sciences of physiology, pharmacology and pathophysiology dominated medical science. Now, *clinical* trials—trials which directly assess the effects of a treatment in a sample of patients—play a prominent role. Indeed, some consider clinical epidemiology as an additional basic science in medicine (Fletcher et al. 1996, p. 3).

Where clinical epidemiology has influenced the type of research considered important to medical science, EBM's influence has been in convincing clinicians of the importance of clinical epidemiology to the *decisions* they make in the care of their patients. EBM, in essence, is the application of the principles of clinical epidemiology to medical decisions. EBM developed as a distinct approach during the 1980's and 1990's in the Department of Medicine and Clinical Epidemiology and the Department of Biostatistics at McMaster University, Canada. Gordon Guyatt, who first used the term in a one-page editorial of the *American College of Physcian's Journal Club* (1991), suggests that 'evidence-based medicine' is an extension of David Sackett's notion of 'bringing critical appraisal to the bedside' (Guyatt and Rennie 2002). 'Critical appraisal' is the process of appraising clinical research according to the methods of clinical epidemiology.

Advocates of EBM, such as Guyatt and Sackett, provide 'tools' for bringing the skills of clinical epidemiology to the bedside. The most important

of these tools is EBM's 'hierarchy of evidence'. The hierarchy of evidence provides a hierarchy of study designs, and other sources of evidence. Advocates of EBM suggest that medical decisions should be based on evidence from as high up the hierarchy as possible. A number of different hierarchies have been developed for different aspects of medical decision making. For instance, The Oxford Centre for Evidence-based Medicine Levels of Evidence (see Phillips et al. 2001) provides evidence hierarchies for therapeutic decisions, prognosis, diagnosis, differential diagnosis, and economic and decision analyses. By far the most influential of these hierarchies is the hierarchy provided for *therapeutic decisions*. The Oxford Centre for Evidence-based Medicine hierarchy for therapeutic decisions is provided in Table 1.1. The two well-known and influential EBM guidebooks, Guyatt and Rennie (2002) and Straus et al. (2005), which teach busy clinician's the skills of EBM (and hence provide the account of EBM 'as it is to be practised'), utilise a hierarchy of evidence for therapeutic decisions that is similar to that provided by The Oxford Centre for Evidence-based Medicine. (The guidebook by Straus and colleagues does not explicitly provide a 'hierarchy', nevertheless they rely on a similar ranking of evidence for therapeutic decisions.)

In providing advice on how the hierarchy of evidence should be interpreted and applied, EBM becomes more than mere motto. The main components of the hierarchy of evidence for therapeutic decisions follow. The top of the hierarchy is typically reserved for meta-analyses and systematic reviews of a number of randomised controlled trials.[2,3] Next in the hierarchy is a single randomised controlled trial; then, systematic reviews of observational studies, which are placed above a single observational study; and, finally, the lower tiers of the hierarchy are filled by studies in basic science, and unsystematic clinical observation.

Of key importance is the distinction made between randomised controlled studies (or RCTs), and observational studies. I prefer the term 'randomised interventional study' to randomised controlled trials because it is more descriptive. It is the interventional nature of these trials that is of primary

[2]Some hierarchies place N-of-1 studies above systematic reviews, but N-of-1 studies, which allocate a patient to successive periods of treatment and control to decide whether the treatment is effective *in that patient*, raise a different set of questions and problems. I will leave these studies to one side for now.

[3]Systematic reviews and meta-analyses combine the results of individual studies, for more discussion see page 19.

Level	**Therapy/Prevention, Aetiology/Harm**
1a	Systematic review (with homogeneity*) of RCTs
1b	Individual RCT (with narrow confidence interval)
1c	All or none**
2a	Systematic review (with homogeniety) of cohort studies
2b	Individual cohort study (including low quality RCT; for example $< 80\%$ follow-up)
2c	'Outcomes' research; Ecological studies
3a	Systematic review (with homogeniety) of case-control studies
3b	Individual case-control study
4	Case series (and poor quality cohort and case-control studies)
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or 'first principles'

Table 1.1: Oxford Centre for Evidence-based Medicine Levels of Evidence for Therapeutic Decisions (Phillips et al. 2001).
Notes: *The authors of this hierarchy note that homogeneity of systematic reviews is important. On their view, substantial variation in the results of the individual studies *may* undermine the results of the systematic review. ** 'All or none' is 'met when *all* patients died before the treatment became available, but some now survive on it; or when some patients died before the treatment was available, but *none* now die on it.'

importance (Chapter 3, *EBM must be ...*, provides a detailed argument for this claim, and illustrates some of the ambiguity that has arisen from the term 'randomised controlled trial'). Randomised interventional studies recruit a sample of participants, and then randomly allocate this sample into a group that receives the experimental treatment, and a group that receives either placebo, or some other form of 'control', such as the treatment that is currently recommended for the condition. The treatment and control groups are then monitored to ascertain the affects of the treatment. Observational studies, by contrast, do not *allocate* patients into treatment or control, rather, in observational studies, it is the participants going about their day-to-day life who choose whether or not to take the treatment under investigation, or expose themselves to the risk being assessed. EBM's central claim is that therapeutic decisions should be informed by evidence from randomised interventional studies (or systematic reviews of evidence from a number of randomised interventional studies), rather than evidence from observational studies (or other sources of evidence from lower down EBM's hierarchy).

Proponents see EBM as *revolutionary*. Proponents of EBM explicitly invoke the work of Thomas Kuhn, and label EBM a 'paradigm shift' in medical practice (see Evidence-Based Medicine Working Group 1992, p. 2420, and Guyatt and Rennie 2002, p. 8). That EBM is a new paradigm (in Kuhn's sense) is hardly defensible. For a start, there is no incommensurability between the 'new' and 'old' approaches. EBM is more accurately seen as a shift in the type of evidence seen as optimal for therapeutic decisions. Along with evidence from basic science and the observations of experienced clinicians, clinical studies have long played a part in informing and justifying therapeutic decisions. What is new with EBM, is that it recommends evidence from clinical studies—especially randomised interventional studies—be explicitly privileged over other facets of therapeutic decision making. Presumably, what proponents of EBM are trying to emphasize in their talk of revolution is that explicitly privileging evidence from randomised interventional studies represents a *significant* shift in focus.[4]

And significant this shift is, in more ways than one. While the shift in what is considered *good* evidence for therapeutic decisions is certainly

[4]Some may argue that the different *weighing* of evidence put forward by proponents of EBM is enough to classify EBM as a paradigm shift in the Kuhn's sense. Perhaps a case could be made. The point I emphasise is that evidence from clinical studies was seen as important to medical decision making prior to EBM.

significant, of even greater significance is the epistemological promise that accompanies this shift. Proponents of EBM promise that basing therapeutic decisions on evidence from randomised interventional studies will improve the care of patients. Evidence from randomised interventional studies provides better evidence for therapeutic decisions, and hence, better care. This promise is made because it is thought that in practising EBM, medicine becomes more 'scientific', more 'rational'. As Brendan Reilly (2004) puts it, 'clinical medicine, long more art than science, is becoming the opposite'.

Whether or not these claims of EBM can be sustained—the details of which are the primary focus of this thesis—EBM has had a significant influence not only within medicine, but also more broadly. One prominent example is 'evidence based policy'. Adrian Smith (1996), in his Presidential Address to the Royal Statistical Society quoted above, uses medicine as an example to argue for a similar 'revolution' in public policy. While continuing to avoid debate about what 'constitutes evidence in any particular instance', Smith calls for an 'evidence-based society'.[5]

However, enthusiasm for EBM is not universal. Some argue that EBM is more *illusion* than revolution. The attack comes from a range of directions, I'll discuss three of the more prominent reactions. While more criticisms have been aired than are countenanced here, many (if not most) can be put into one of these three general categories.

First, some argue that EBM provides nothing new (Sehon and Stanley 2003; Shahar 1997). Medicine, after all, has always been based on evidence; *non*-evidence based medicine, whatever that may be, is not what was being practised prior to EBM. This criticism is strongest when it is directed against EBM's motto, or the most commonly cited 'definition' of EBM provided by Sackett (1996):

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.

Both EBM's motto, and this 'definition' are far too broad; they sacrifice

[5]If anything can be judged by how frequently 'evidence-based' is used to prefix something the speaker wishes to represent as the 'right' approach, then many have headed Adrian Smith's call—if not his intention.

descriptive detail for aspirational tones. A much clearer idea of what EBM is can be gained by paying attention to proponents advice on how EBM should be put into practice. But, even here proponents of EBM are rather ambiguous about some important details (as will be discussed in Chapter 2, *EBM can't be ...*).

Another avenue of criticism comes from those who view medical decisions as more 'art' than science (Henry 2006; Tanenbaum 1993). Medical decisions, according to these critics, rely heavily on the tacit knowledge that is accumulated by clinicians observing patients. EBM is seen to call for a reduction of medicine to statistics. On this criticism, medicine is the care of individual patients by their physician, and this relationship is too unique and complex to be reducible to the results of clinical studies. This criticism is provided in a range of ways, often these critics view EBM as a dangerous threat to medical autonomy, and this criticism is part of broader attempt to re-assert that autonomy.

The third, and by far most common criticism of EBM, is that it promises far more on behalf of its methodological distinctions than it manages to deliver (see Anonymous 1995 for an early example). Proponents of EBM promise that EBM makes medical decisions more scientific. The advantages of randomised interventional studies over observational studies is taken as the primary basis for this claim. But, somewhat surprisingly, proponents of EBM spend very little time justifying this view. Instead, the idea that randomised interventional studies provide superior evidence for therapeutic decisions is seen to be beyond reproach; it is the starting assumption of EBM, rather than a conclusion that is explicitly argued for. The guidebook from Straus et al. (2005, p. xii) refers those interested in the 'theoretical and methodological bases' of EBM to textbooks in clinical epidemiology. However, despite their shared methods, the aims of EBM are distinct from clinical epidemiology.

Many of the methods of clinical epidemiology are put to the task of improving the reliability of clinical research, especially with regard to *internal validity*. Forming correct inferences on the basis of good clinical data is clearly important for EBM, but EBM requires an additional step (or a series of additional steps). EBM is about medical *decisions*, especially therapeutic decisions. Properly interpreting data from the relevant clinical studies is only the first step in informing a therapeutic decision. The practice of EBM requires *applying* the results of clinical studies to individual patients (who may, or may not, resemble the patients included in the clinical studies). This

process requires consideration of not only clinical data, but also the preferences of the patient, the theories of basic science, and the experience of the clinician. Clinical epidemiology provides little advice on how evidence from well conducted studies should be applied to patients that present to the clinic. And, because EBM relies on clinical epidemiology for its theoretical basis, it too provides precious little on precisely how 'evidence' should be applied. In this regard EBM, as elucidated by proponents, does not dig too much deeper than the bald stipulation given in Sackett's description of EBM: other facets of therapeutic decision making, such as patient preferences, basic science and clinical experience, should be 'integrated' with the best available external clinical evidence. Much more detail is required.

Can EBM overcome these criticisms? Yes, in principle at least (or, so I will argue). Any reply will contain a considerable number of caveats. Whether or not proponents or critics will want to call the resulting view 'EBM' will depend on how closely participants in the debate tie the label 'EBM' to EBM's central claim *as it is most commonly stated by proponents.* If EBM *is* the claim that therapeutic decisions informed by evidence higher up EBM's hierarchy provides improved care compared to (or rather than) evidence from lower down the hierarchy, then EBM is in trouble. If, however, EBM's central claim can be modified, and the restricted scope of its methodological claims made explicit without rejecting the label, then there is some room to move.

I have some sympathy for the latter view. Clarifying the foundations of EBM allows the possibility of keeping what is what is good about EBM while at the same time exposing any over-extended claims. The resulting view provides a stronger foundation for therapeutic decision making, and it seems reasonable to continue to label this view EBM. This thesis focusses on the foundations of EBM: What claims can, and cannot, be substantiated on the basis of the hierarchy of evidence? What types of evidence are needed for therapeutic decisions? And, how should these different types of evidence be appealed to? Labelling is a concern only for the purposes of clarity. I will be clear whether I am discussing EBM's central claim—that therapeutic decisions are best informed by evidence higher up the hierarchy of evidence—or EBM more generally, which places this central claim within a broader framework for judging the evidence for therapeutic decisions (and permits clarifying EBM's central claim without losing the 'EBM' label).

To reply to the first criticism of EBM, the justification, interpretation

and application of EBM's hierarchy must be clearly articulated. This will clarify what it is to practice EBM, and distinguish it from alternative models of medical decision making. Also, in more clearly articulating how EBM's hierarchy of evidence is to be applied, and by explicitly delineating the limitations of research data in forming medical decisions, a reply can be formulated to the second general criticism of EBM. Putting clinical research into a richer framework for therapeutic decision-making, a framework that can incorporate the theories of basic science, clinical experience and patient preferences, helps EBM to avoid the charge of reducing decision-making to merely acting on the results of clinical studies. (Of course, there may be some who wish to defend the view that the complexity of each clinician-patient encounter is so specific to that encounter that *no* rational account of medical decision making can be provided. But this seems far too strong a view; in any case, I'll leave it for the defender of such a view to make their case.) Finally, assuming modification of EBM's central claim is permissible (without dropping the label), then conducting much-needed work on the foundations of EBM permits a reply to the third type of criticism. Once a justification for EBM's hierarchy of evidence is provided, the limits of EBM's methodological claims can be made clear.

If the sketched reply is adequate, then, rather than revolutionary or illusory, EBM may be seen as part of an *evolution* in therapeutic decision making. A step toward a rational framework for making and justifying therapeutic decisions. A modest step. A step with considerable qualification and further work to be done, but a step forward nonetheless.

The reluctance of proponents of EBM to clearly elucidate the foundations of the approach makes it difficult to pin down the view they currently hold with any precision. It is also likely that the view has shifted in response to early criticism. For instance, recent attempts to improve the grading of evidence for guidelines seem to incorporate replies to some of the early criticisms of EBM (see The GRADE Working Group 2004). However, even in their replies to critics, there is little focus on the theoretical matters at issue. Therefore it is difficult to judge whether proponents of EBM have changed their view, say, with regard to what is considered to constitute 'best evidence', or whether they have simply clarified their original view in response to concerns; whichever it is, the foundations of EBM remain somewhat murky. The central claim that has been consistently made, and continues to be made, is that evidence from higher up EBM's hierarchy better

informs therapeutic decisions.

This thesis examines the foundations of EBM, with special focus on this central claim. The primary aim is to better articulate what benefits (if any) arise by informing therapeutic decisions on evidence higher up EBM's hierarchy of evidence. To adequately fulfil this aim, I restrict my focus to therapeutic decisions, rather than medical decisions more generally, and to the evidential inputs into therapeutic decisions, rather than issues of patient preference, or other social or ethical factors that play important roles in making decisions regarding individual patients. Allow me to briefly elaborate.

EBM is focused on informing therapeutic decisions. Not only has EBM had its largest influence on how therapeutic decisions are made and justified, it is clear that therapeutic decisions were of primary importance in the development of EBM. When pushed, it is therapeutic decisions, as opposed to decisions about diagnosis or prognosis, that proponents suggest are best informed by consulting the hierarchy of evidence. Sackett and colleagues (1996) are especially clear on this in their early defence of EBM, 'Evidence based medicine: what it is and what it isn't'.

Under the banner of 'therapeutic decisions', I will refer to two specific types of decision, population therapeutic decisions and individual therapeutic decisions. Both decisions require balancing the expected benefits of a treatment against the expected harms, and hence can be considered within a decision theoretic framework. In decision theory, both the probability of the outcome (benefit or harm) and the utility of the outcome are important to decisions. The two therapeutic decisions differ in their application. Population therapeutic decisions involve making a judgement regarding the benefits and harms of a therapy in a defined population of patients. Such populations are typically defined according to their shared characteristics, for instance the population may suffer a shared disease, be a similar age, or have similar co-morbidities. The population may also be defined according to the inclusion and exclusion criteria of a clinical trial. Individual therapeutic decisions require the same weighing of the likely benefit and harm of the treatment, but this time the decision needs to be made for an individual patient, taking into consideration the unique characteristics of this individual. Making therapeutic decisions for individuals raises complicated questions concerning how evidence should be applied. Clinical trials, which assess the effects of a therapy on a sample population, can often provide reasonably clear evidence for population therapeutic decisions—providing the 'population' in question

is suitably defined. Applying the results of clinical trials to individual patients can be considerably more difficult. As individual therapeutic decisions are a particular focus for EBM, the challenges these decisions raise will be discussed throughout the thesis.

In discussing clinical studies, I focus especially on large drug trials. This is a consequence of focussing on therapeutic decisions. Large drug trials are the key trials that influence the clinical and regulatory decisions that I am interested in. Although this is where EBM started, its influence is now much broader. Currently, randomised interventional studies are recommended when testing a wide range of non-drug interventions, including behavioural interventions, social interventions and public health initiatives. The push for testing such interventions in randomised interventional studies is provided through funding processes and editorial policies, as much as by the recommendations of any given institution. However, different experimental contexts raise different methodological questions; the specific importance of differing contexts is too often under-recognised. Some epidemiological methods that are thought to apply to testing any intervention are actually much more specific for particular errors that occur in large drug trials. And application of these methods outside of this context is either unnecessary, or detrimental. Indeed, randomisation is a case in point (see Chapter 3, *EBM must be ...* for discussion). For these reasons I wish to emphasis that large drug trials will be the primary example of interventional studies in this thesis. Of course, by clarifying how different methods affect the reliability of large drug trials it will be possible to better identify those methods that are highly specific to this context.

The preferences of the patient play a crucial role in any therapeutic decision. In any particular instance, patient preferences often determine the specific evidential question. The benefits and harms of a treatment are only of interest if the patient agrees that the treatment is worth taking. And perhaps more importantly, patients will vary widely on what they consider a significant harm or a worthwhile benefit, and (in the clinical medicine at least) it is the patient's judgement on such issues that determines the value of the therapy to the individual. EBM's central claims, however, are epistemological. Accepting the importance of patient preference to therapeutic decisions, a philosophical question remains: how should various sources of evidence be judged in assessing whether a treatment is likely to benefit or harm an individual or group? This question arises whichever way an individual or

group defines the 'benefit' or 'harm' they are interested in.

The thesis approaches the foundational issues of EBM in the following manner. The next two chapters focus on the interpretation and justification of EBM's hierarchy of evidence. *Chapter 2: Evidence based medicine can't be ...* begins by outlining how proponents of EBM have suggested that the hierarchy should be interpreted. While an unambiguous interpretation of the hierarchy has not been provided, the clearest and most frequently provided account suggests a categorical interpretation. That is, in informing therapeutic decisions, evidence from higher up EBM's hierarchy of evidence *trumps* evidence from lower down the hierarchy. The results of a randomised interventional study, indeed, *all* of the results of a randomised interventional study, trump the results of an observational study. Often seen in conjunction with this categorical interpretation of the hierarchy is the notion that the hierarchy can be broadly applied. On this view, the methodological distinctions provided in the hierarchy give some general epistemological rules. This view fuels the idea that other areas of decision-making, such as public policy, can adopt 'evidence based practice'. In this chapter, I show that neither the categorical interpretation of EBM's hierarchy, nor the view that the hierarchy can be broadly applied, bear critical scrutiny. EBM's hierarchy of evidence is specific to therapeutic decisions. And, indeed, (I argue) a fairly narrow set of therapeutic questions at that. Only the overall result of the primary hypothesis tested in a randomised interventional study will have the warrant associated with the frequentist statistical methods used to analyse trials. Since therapeutic decisions rely on a much broader range of results from clinical studies, this undermines the view that *all* results of randomised interventional studies trump the results of observational studies. Clearly, a more nuanced view of how EBM's hierarchy should be interpreted is required.

In *Chapter 3: Evidence based medicine must be ...*, I shift focus to the justification that proponents have provided for EBM's hierarchy of evidence. Following the critical literature, and in particular the work of John Worrall, the arguments put forward to justify the ambitious claims made by proponents of EBM are shown to be inadequate. However, I argue that a justification for EBM's hierarchy of evidence can be provided. The hierarchy should be viewed as a hierarchy of comparative internal validity. *Internal validity* is the degree to which the results of a study are accurate for the participants of the study. *Comparative interval validity* refers to the capacity of different study designs to rule out, or minimise, specific sources of error. (Just how

EBM's hierarchy can be seen as a hierarchy of comparative internal validity is made clear in Chapter 3.) This justification for the methodological distinctions found in the hierarchy can be found in the epidemiological literature. While this justification is defensible, the claims that EBM's hierarchy underwrites when viewed in this way are considerably more circumscribed than the claims found in the EBM literature. Internally valid clinical research is not sufficient for informing therapeutic decisions about individual patients. The research must also be *applicable* to individuals who will be treated with the therapy, but were *not* involved in the clinical trial; that is, the research must have high *external validity* (or, at a minimum, there must be a framework for applying clinical research to individuals). Understanding that the methodological distinctions made by EBM relate *only* to the internal validity of clinical research underscores the challenge that external validity presents for therapeutic decisions.

A number of philosophers have shown that randomisation is not *essential* for experiments to provide 'scientific' data. Recent philosophical work, especially the work of John Worrall, goes one step further. Worrall (for instance, in Worrall 2007b) suggests that there is no epistemological benefit in randomisation, providing the experimental groups are similar according to background knowledge. Clearly, this view is in stark contrast with the position held by proponents of EBM. *Chapter 4: Why randomised interventional studies* provides an argument for preferring randomised interventional studies to observational studies in testing the efficacy of treatments. 'Efficacy' refers to whether the treatment works as hypothesised in an experimental situation. It can be contrasted with 'effectiveness', which refers to whether the treatment works in patients undergoing routine care. Interventional studies (whether or not they are randomised) provide better tests of the efficacy of treatments because this study design has the capacity to rule out more sources of systematic error than observational study designs. *Randomised* interventional studies, as opposed to non-randomised interventional studies, are defended on Bayesian grounds.

In chapters 5–7 attention turns to the challenge of external validity. *Chapter 5: The challenge of external validity* outlines why judgements of external validity are so difficult, as well as why they are vital for therapeutic decisions. External validity raises two types of question. First, the question of whether the overall results of the trial can be considered to accurately predict the average response in the population who will receive the treatment in routine

care. And second, the question of whether the results of the trial accurately predict the likely response in an individual; an individual who may or may not resemble some subgroup of patients within the trial, and who may or may not have been considered eligible to be enrolled in the study. Clinical studies with high internal validity do not straightforwardly provide data to answer these questions. Yet, these questions are crucial for EBM, because these are the questions that clinicians must answer to form a judgement regarding the likely benefits and possible harms of a therapy. EBM claims that therapeutic decisions should be based on evidence from randomised interventional studies *rather than* evidence from observational studies. In Chapter 5, I show that evidence from observational studies can play an important role in assisting judgements of external validity.

In a somewhat similar vein, *Chapter 6: The role of basic science in evidence based medicine* describes the role that basic science plays in informing therapeutic decisions. Basic science plays a considerably more important role in therapeutic decision making than is recognised in the EBM literature. Indeed, proponents of EBM are rather unclear about whether they see *any* role for basic science in therapeutic decisions. Patrick Suppes' hierarchy of data models is introduced as one way of spelling out the relationship between basic science and the observed results of clinical trials. Suppes explicates a hierarchy of models that links the theories of basic science with the observed data. Relying on this framework I show that basic science is vital, not only for specifying the experiments that are conducted, but also, interpreting the data that is provided. More importantly, I show that basic science is integral to the application of clinical research to therapeutic decisions.

In *Chapter 7: External validity, subgroup analysis and basic science*, I use the framework discussed in Chapter 6 to broach the 'problem of subgroup analysis'. The problem can be described as follows. Large randomised interventional studies provide data on the average effects of a treatment in the population of patients included in the trial. Ideally, however, what therapeutic decision makers require, is data regarding which patients are likely to benefit from the treatment, and which patients are prone to adverse effects. But gaining reliable data of this second kind is difficult. This problem has vexed the leading proponents of clinical epidemiology, including Austin Bradford Hill (1966), Archibald Cochrane (1971), and Alvan Feinstein (1998). Two approaches to the problem that are found in the literature are outlined, but neither is completely adequate. One group, the 'trialists', argue that reliable

subgroup data is not typically available. Trialists advise that therapeutic decisions are best informed by the overall results of clinical trials, rather than by any data observed in subgroups. Importantly, on the trialists view, basic science plays no, or very little, role in the interpretation and application of clinical research. By contrast, a second group, the 'pathophysiologists', argue that basic science should play a role in interpreting subgroup data (though, members of this group differ in their view of how reliable the data from such subgroups can be). What is needed is a more explicit framework for incorporating the theory of basic science into the interpretation and application of clinical research. I argue that the hierarchy of data models can provide such a framework.

One of the problems with the trialist's reply to the challenge of subgroup data is that they apply the frequentist statistical norms used when testing the efficacy of a treatment to questions of effectiveness. Such frequentist statistical norms, however, are both general and conservative. While good reasons can be provided for adhering to these norms in tests of efficacy, I argue they are too restrictive when it comes to assessing a therapy's effectiveness. I extend this discussion in the penultimate chapter, *Chapter 8: Power and Inference: The rofecoxib case.* Using the recent withdrawal of rofecoxib as a case study, I show that the needs of therapeutic decision makers often outstrip what the frequentist statistical tools provide. While frequentist methods provide an optimal frequentist statistical test of the primary hypothesis tested in a clinical trial, clinicians often need to make a decision on another aspect of the data; in the example of rofecoxib, decisions need to be made regarding the observed safety data. In such situations following the standard rules of frequentist analysis can be unhelpful.

The thesis concludes by summarising some of the key findings. Randomised interventional studies possess some methodological benefits over observational study designs. These benefits, however, are quite specific, and most important when it comes to rigourously testing the efficacy of new drugs. This is an important result and gives partial support to current regulatory processes, and EBM. But the support is only partial. There is much we need to know about therapies in addition to their efficacy in a defined population of patients, both from a regulatory perspective, and clinically. Two points are emphasised. The randomised interventional studies that are conducted are often less then ideal for establishing the safety of treatments. And, contrary to the advice of EBM's hierarchy of evidence, therapeutic deci-

sions rely on a range of evidential sources. These points are not entirely new, however, the lack of clarity about the foundations of EBM have hamstrung efforts to provide an adequate reply. The hegemony of proponents of EBM, and their preference for randomised interventional studies, have overshadowed alternative study designs and epistemological arguments. By clarifying the foundations of EBM, the thesis identifies the questions that randomised interventional studies answer well, and illuminates those questions that are poorly answered. The work on the foundations of EBM articulates the need for alternative approaches to questions other than efficacy, and provides a basis for approaching the questions that EBM leaves unanswered.

Before getting into the details, I provide two short primers that will prove useful for the discussion that follows. The first describes the study designs commonly employed within clinical epidemiology, and the second outlines the frequentist statistical methods used to analyse clinical data.

1.1 Study designs in clinical epidemiology

I have already briefly introduced randomised interventional studies (also known as randomised controlled trials) and observational studies, let me now provide a bit more detail.

Interventional studies start with a sample of patients. This sample is selected from some larger population. Once selected, members of the sample are *allocated* to treatment or control. Importantly, this sample is not taken randomly from any precisely defined population. Rather the sample is drawn from the population that the investigators have access to. From this population, investigators recruit patients in accordance with the study's entry criteria. Some assessment of the trial population can be made by considering the population the investigators have access to, the study's entry criteria, and the characteristics of the eventual sample.

In a *randomised interventional study* patients are allocated according to some random procedure (such as a table of random numbers, or similar). How patients progress on their allocated treatment is then monitored. A range of clinical endpoints are typically monitored, such as blood pressure, length of hospitalisation and the like. Of special interest is the 'primary endpoint', which is the focus of the trial. Typically, the active treatment is expected to have a beneficial effect on the primary clinical endpoint; the reason the trial is conducted is to observe whether the treatment under inves-

tigation realises this expectation. Statistically this expectation is stated as the 'primary hypothesis' under test in the trial (also known as the 'alternative hypothesis'); the 'null hypothesis' is the hypothesis that the treatment under test performs no better than control. The primary endpoint may be a single clinical outcome, or a combination of a number of outcomes. Contemporary clinical trials will usually also report the findings of the trial on additional clinical outcomes, some of which may be defined as 'secondary endpoints'.

Randomised interventional studies are conducted so as to ensure that the experimental groups are as similar as possible apart from their allocated treatment. One of the reasons that allocation is randomised is that it ensures that the groups are roughly equally matched according to factors that may influence how patients respond to the treatments—providing the sample is large relative to the number of patient characteristics that influence response to the treatment. (The claims that proponents of EBM make on behalf of randomisation often go much further, see Chapter 4 for discussion). After the groups have been allocated, well conducted trials will utilise a range of methods to ensure the two experimental groups are treated in the same way. Blinding the patients to their allocation, concealing the allocation from investigators, and standardising monitoring are all methods aimed at minimising differences between the experimental groups. These, as well as a number of other methods, will be discussed in more detail in subsequent chapters.

As discussed previously, the chief difference between observational studies and randomised interventional studies is that, in observational studies, no experimental intervention is imposed on the participants. Observational studies follow subjects who are going about their lives, choosing (as much as is possible) which medicines they take and to what risk factors they expose themselves. The two main forms of observational studies are *cohort*, and *case-control*. The difference between these study designs is whether or not the participants have suffered the event under investigation at the time the experimental groups are assembled. Cohort studies assemble a cohort of patients (either in 'real-time' or retrospectively) who have the potential to suffer the event under investigation but, at the time they are assembled into a group, have not suffered the event. Case-control studies, by contrast, assemble two groups, one group that has suffered the event under investigation, 'cases', and one group that has not, 'controls'.

Cohort studies partition the assembled group into those exposed to the treatment (or risk) under investigation, and those not exposed. These 'co-

horts' are compared to see if the rates of the event under investigation differ between the groups. Strictly, randomised interventional studies are a specialised form of cohort study (one in which the cohorts are formed by randomly allocating the sample into treatment and control). Throughout the thesis, however, when I refer to a cohort study I will always mean an observational cohort study. Cohort studies can be historical or prospective. In *historical cohort studies* a cohort of participants who have not suffered the event under investigation is assembled from medical records—some members of the cohort will have been taking the treatment, or been exposed to the risk, and others will not have. At the time of the study this cohort is assessed to determine whether the event has occurred. *Prospective cohort studies* assemble the cohort at the start of the study, and then follow that cohort for a period of time, monitoring participants as the study continues.

Case-control studies begin at the opposite end of the timeline to cohort studies, that is, once an event (or 'outcome') has occurred (for example, a heart attack or a diagnosis of cancer). The group for which the outcome has occurred, the 'case' group, is compared to a control group—a group for whom the outcome under investigation has not occurred. The two groups are compared according to their exposure to the risk factors (or treatments) under investigation in an attempt to isolate the cause of the event.

Historically controlled studies are something of a hybrid of the designs discussed so far. In historically controlled studies, one cohort is assembled from patients that previously received conventional treatment, and a second 'prospective' cohort is assembled from patients that will receive a different treatment, and will be monitored as the study progresses. The details of the treatment and the outcome experienced by members of the historical control group are ascertained using medical records. The outcomes of the prospective cohort are collected over the time of the study. Prior to the dominance of randomised interventional studies, historically controlled studies were used to assess the efficacy of treatments. They are less commonly seen in this role now, though some philosophers, such as Peter Urbach (1993) and Worrall (2002), argue that they could play a more prominent role.

The historically controlled studies that are conducted today are observational studies, in that neither the historical cohort (obviously) nor the prospective cohort are allocated an experimental intervention. Participants in the prospective cohort receive the therapy under investigation as part of their routine care. However, a historically controlled study could also be a

quasi-interventional study. Here the prospective cohort would be allocated a new 'unproven' experimental treatment; the historical control by necessity would continue to be a cohort of patients who had chosen conventional treatment. The distinction may seem slight, but there are important ethical considerations. Because much of the medical community now believe that randomised interventional studies provide the best methods to test whether a treatment is efficacious, it is considered unethical to test a new treatment in a historically controlled trial (unless there is some reason a randomised interventional study could not be conducted). As Worrall (2008, p. 422) notes, in clinical medicine, epistemology and ethics are 'closely intertwined'. It is only because randomised interventional studies are considered epistemologically superior that it is considered unethical to test 'unproven' treatments in a historically controlled trial. On the current view, which accepts the claims of EBM, patients in the 'active treatment' cohort are put at risk from an 'unproven' treatment without the compensatory 'benefit' that the study being conducted will provide what is considered to be conclusive evidence of the treatment's efficacy. If historically controlled trials are considered epistemologically adequate for tests of new treatments, then such trials would be ethically permissible.

Systematic reviews combine multiple studies according to explicit criteria (by contrast, *narrative reviews* provide a summation of the literature without employing a systematic criteria for choosing studies). *Meta-analysis* is a type of systematic review in which the quantitative analysis of the combined results is also conducted according to explicit criteria. Meta-analysis can be conducted on groups of randomised interventional studies or observational studies. Meta-analyses of randomised interventional studies are the highest form of evidence in EBM's hierarchy. A range of statistical methods have been developed to combine and contrast the results of the included studies—the sophistication of these methods has improved over the past decade.[6] Some meta-analyses use only the published data from the included studies, while others use patient-level data provided by the original investigators. Because meta-analyses pool data across studies of varying entry criteria, aims and analytical techniques, the 'homogeneity' of the included studies is important to the validity of the results of the meta-analysis. Validity of meta-analyses are improved when the included trials observe similar patients studied in similar ways.

[6]See Greenland and O'Rourke (2008) for an overview.

Each of these study designs possess advantages and disadvantages for assessing the benefits and harms of therapies. Each are prone to specific systematic errors and carry with them particular practical benefits or disadvantages. The forthcoming chapters provide an opportunity to discuss the pros and cons of the these study designs in informing therapeutic decisions.

1.2 Frequentist analysis of clinical trials

Contemporary clinical trials are analysed according to the methods of frequentist statistics. Particularly influential are the methods of Jerzy Neyman and Egon S. Pearson (see Neyman and Pearson 1933, and Neyman 1937). The terms and concepts of frequentist statistics plays a role in each of the subsequent chapters. Here I outline the frequentist approach to hypothesis testing and estimation—each of which play a central role in the inferences drawn from clinical studies.

Both hypothesis testing and estimation aim to provide information on an unknown parameter in a clinical study. In clinical research most of the focus is on a single unknown parameter, which is usually a function of the primary clinical endpoint. The primary clinical endpoint is the variable (or variables) that the trial is set up to observe. For instance, the unknown parameter might be the difference in the rate of the primary endpoint in the experimental groups, $\theta_{Dif} = X_T - X_C$, where θ_{Dif} is the unknown parameter, X_T is the rate of the primary endpoint in the treatment group and X_C is the rate of the primary endpoint in the control group. Alternatively, the unknown parameter might be expressed as a quotient, $\theta_{Quot} = X_T/X_C$, or some other function of the primary endpoint. In hypothesis testing, two hypotheses are considered regarding the unknown parameter, the null hypothesis and the alternative hypothesis. The null hypothesis typically holds that no difference exists between the treatment and control groups for the unknown parameter ($\theta_{Dif} = 0$, $\theta_{Quot} = 1$). And the alternative hypothesis holds that some defined difference exists in the rate of the primary endpoint between the experimental groups (for example, $\theta_{Dif} > 0$).

Hypothesis testing uses the observed data to infer which hypothesis will be 'accepted' according to the dictates of frequentist statistics. Some statisticians and philosophers bristle at the thought of accepting a hypothesis and prefer 'not reject'—hence the scare quotes. 'Accepting' or 'not rejecting' a hypothesis has a specific meaning in the context of frequentist methods; it

is this meaning I aim to clarify here. Once the meaning is clear, little hangs on the terminology selected.

Rather than specify hypotheses about the unknown parameter, the frequentist approach to estimation uses the observed data to directly infer a range of values for the unknown parameter. This range is called a confidence interval. Confidence intervals and p values—p values are reported as part of hypothesis tests in epidemiology—are widely recognised and entrenched parts clinical analysis. (A formal definition of these concepts is provided shortly).

Estimation has some benefits over hypothesis testing in the clinical setting. Therapeutic decision makers are typically more interested in the *value* of the unknown parameter, than whether or not they should accept a specific statistical hypothesis (see Ware et al. 1992 for discussion). However, both estimation and hypothesis testing are utilised in the contemporary analysis of clinical trials. Clinical trials report both p values and confidence intervals; often, focus turns to estimation once the null hypothesis is rejected. More significantly, both of these approaches rely on the same conceptual framework—that of frequentist statistics. Whatever the relative advantages of estimation over hypothesis testing, a range of criticisms have been made against the entire framework. Howson and Urbach (2006), for instance, find the warrant provided for frequentist inferences less than compelling.

Despite the philosophical criticisms, data from clinical trials are analysed with frequentist methods, for now at least. Hence, I will focus on the interpretation of data provided by these methods. Clearly elucidating the warrants provided by frequentist methods will help to achieve the aims of this thesis. Once these warrants are clear, how well these methods meet the needs of therapeutic decision makers can be assessed. I argue, especially Chapter 7 and 8, that frequentist methods are not ideal for informing therapeutic decisions. While it is hard to conceive EBM without frequentist analysis, there is no necessary connection between the two. EBM could avoid some of the criticisms made in Chapter 7 and 8 by using an alternative approach to statistical inference in clinical trials. Tackling the broader debate about the benefits or otherwise of competing approaches to statistical inference in medicine will be left for another time.

Hypothesis testing. The outline that follows should be uncontroversial; it is the standard story of Neyman-Pearson statistics. I provide a four-step outline of the formal features of the Neyman-Pearson approach to hypothesis testing.

I have made a number of simplifying assumptions, the most important of which are noted.

First, the raw data that the experiment provides about the unknown parameter needs to be summarised. This summary of the raw data is called a test statistic. The test statistic is a random variable, let me represent it by X. X is an *estimator* of the unknown parameter θ. To choose an appropriate test statistic, assumptions are made about the process under investigation. Perhaps most important in the context of parametric frequentist statistics is the choice of which probability model is thought to best describe the process, and thus the data. The choice of test statistic and the conclusions warranted by the hypothesis test rely on the probability model being adequate.

'Good' estimators ideally fulfil a number of frequentist desiderata: unbiasedness, consistency, efficiency and sufficiency. These desiderata consider the features of the estimator over the entire sample space; often, this is informally referred to as the performance of the estimator *in the long run*. 'Unbiasedness' means that the mean of the test statistic is equal to θ. 'Consistency' means that as the sample size tends to infinity X converges on θ. 'Efficiency' means X has a smaller variance compared to alternative estimators. And, 'sufficiency' means that all information about θ is contained in X. Choosing between estimators requires weighing up how they perform against these criteria.

Second, the distribution of the test statistic is considered on the assumption that the null hypothesis is true. Take the null hypothesis to hold that $\theta = 1$. To completely specify the sampling distribution of X on the assumption that the null hypothesis is true, a number of further assumptions are required. For instance, X may be assumed to be normally distributed, and unbiased. In frequentist statistics the sampling distribution has a physical interpretation: assuming $\theta = 1$, if the experiment were to be repeated indefinitely, we would expect the observed values of X, x, to form a normal distribution with a high frequency of results clustering around one. (Here, and throughout, dropped case is used to refer to the observed value of a random variable). Note, once the experiment is conducted, observing a value of x at the extremities of this sampling distribution would be considered unlikely on the assumption that the null hypothesis is true. The sampling distribution of X under the assumption of the null hypothesis provides the primary conceptual tool for testing the null hypothesis.

The third step is to divide the sampling distribution for X under the

assumption that the null hypothesis is true, into two regions: accept and reject. At this stage an *alternative hypothesis* is specified. In clinical epidemiology the alternative hypothesis is typically based on available data, or the 'smallest clinically important effect'. The 'reject', or 'critical', region of the sampling distribution for X under the null hypothesis is defined by the pre-set α level. The α level provides the pre-experimental probability that the statistical test will 'reject' the null hypothesis on the assumption that the null hypothesis is true. In clinical research, as in much of science, the α level is arbitrarily set to 5% of the sampling distribution. The critical region is located at the extremes of the sampling distribution; if the alternative hypothesis is that $\theta \neq 1$, then α is distributed in each 'tail' region of the distribution (half in each tail), and if the alternative hypothesis holds that $\theta > 1$, then α will reside in the right-hand tail region. Let us assume the alternative hypothesis of interest is some specific value for θ such that $\theta > 1$, that is, assume we are dealing with a 'simple' alternative hypothesis. Simple hypotheses specify a single value for the unknown parameter θ. 'Composite' hypotheses, by contrast, specify a set of values for θ. Composite alternative hypotheses are common in clinical epidemiology, once I have outlined the Neyman-Pearson method for testing two simple hypotheses, I will extend the discussion to a one-sided composite alternative hypothesis.

Figure 1.1 provides a diagrammatic representation of a Neyman-Pearson hypothesis test of two simple hypotheses. The sampling distribution to the left represents the distribution of X under the assumption that the null hypothesis is true, and the sampling distribution to the right represents the distribution of X under the assumption that the alternative hypothesis is true.

Neyman and Pearson (1933, p. 291) do not hope to know whether any given hypothesis is true or false; they advise what inference can be drawn on the basis of the observed data in light of the sampling distribution for X on the assumption that the null hypothesis, or the alternative hypothesis, is the true hypothesis. The experiment is set up so as to minimise two types of error, type I and type II. A type I error is committed if the null hypothesis is rejected when the null hypothesis is assumed to be true. As seen above, α provides the pre-experimental risk of a type I error; minimising α minimises the risk of rejecting the null hypothesis on the assumption that it is the true hypothesis. A type II error is committed when the null hypothesis is accepted when the null hypothesis is assumed to be false (or, equivalently,

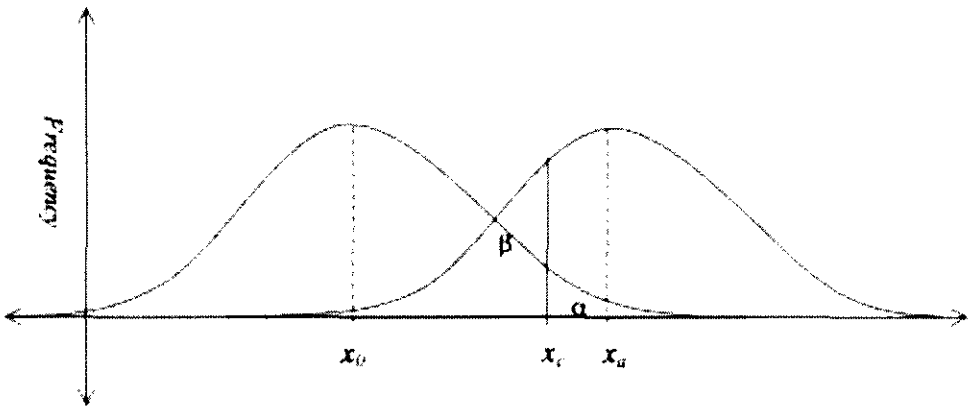

Figure 1.1: A Neyman-Pearson test of two simple hypotheses. x_0 is the value of the test statistic assuming the null hypothesis is true, x_a is the value of the test statistic assuming the alternative hypothesis is true, x_c is the critical value of the test statistic, defined by the α region. α represents the risk of rejecting the null hypothesis under the assumption that the null hypothesis is true, and β represents the risk of accepting the null hypothesis under the assumption that the alternative hypothesis is true.

when the alternative hypothesis is assumed to be true). β provides the pre-experimental risk of a type II error. Whereas α is a proportion of the sampling distribution for X under the null hypothesis, β is a proportion of the sampling distribution for X under the assumption that the alternative hypothesis is true. β is defined by the proportion of the sampling distribution of X under the alternative hypothesis that is superimposed on the 'accept' region of the sampling distribution of X under the null hypothesis (see Figure 1.1). Minimising β minimises the risk of accepting the null hypothesis on the assumption that it is false.

The alternative hypothesis is specified prospectively and β is minimised by selecting the appropriate sample size; α is arbitrarily set at 0.05. *Power*, $(1 - \beta)$, is the pre-experimental probability the test yields an α significant result, that is, the observed test statistic falls in the rejection (or, critical) region, assuming that the alternative hypothesis is true.

This set up has some benefits. For a test of two simple hypotheses, statistical tests set up in this way (i) ensure a trial designed to test a particular hypothesis has a reasonable chance of being decisive (within the terms of frequentist statistics); and (ii) provide a test that, for all values of x that fall in the critical region, the *likelihood ratio*, $P(x \mid \theta_A)/P(x \mid \theta_0) \geq c_\alpha$, where θ_A is the value for θ specified by the alternative hypothesis, θ_0 is the value for θ specified by the null hypothesis, and c_α is a constant. Informally, a test set up in this way ensures that for all of the values of the observed test statistic that fall in the critical region, the observed data favour the alternative hypothesis (providing the assumptions made in specifying the test hold). Slightly less informally, if the observed value of the test statistic falls into the critical region, the probability of the observed value assuming the alternative hypothesis is greater than the probability of the observed value assuming the null hypothesis. The likelihood ratio provides an intuitively appealing measure of what the observed data tell us about the hypotheses under examination, and plays an important role in many approaches to statistical inference. In the most common form of Bayesianism, for instance, the likelihood ratio plays an central role in 'updating' the pre-experimental information to form a post-experiment assessment of the hypotheses (see page 30 for more discussion).

For a test of two simple hypotheses, and assuming that the likelihood ratio is a continuous random variable under the null hypothesis, a best critical region can be selected such that $P(x \mid \theta_A)/P(x \mid \theta_0)$ is *maximised* for any

value of the test statistic falling in the critical region. This is Neyman and Pearson's Fundamental Lemma.

Significantly, a version of Neyman and Pearson's Fundamental Lemma holds in a number of more general situations. For instance, providing the test statistic is singly sufficient for θ—that is all the information about θ provided by the data is contained in X, and X is one dimensional—a *uniformly most powerful* test can be found for tests of a simple null hypothesis against a *one-sided composite* alternative hypothesis. Uniformly most powerful tests ensure that for all values of x within the critical region, the ratio of the *maximum likelihood* for x under the range of values that the unknown parameter θ can take on the assumption of the alternative hypothesis to the maximum likelihood for x under values for θ assuming the null hypothesis is true is maximised: $max\, P(x \mid \theta_A)/max\, P(x \mid \theta_0) \geq c_\alpha$, where $max\, P(x \mid \theta_A)$ represents the maximum likelihood for x under the range of values for θ specified by the composite alternative hypothesis; since the null hypothesis is simple $max\, P(x \mid \theta_0) = P(x \mid \theta_0)$.[7] The shift from likelihood functions in tests of two simple hypotheses, to maximal likelihood functions in tests of composite hypotheses is of importance when comparing different accounts of statistical inference (but not crucially important to the comments I wish to make about frequentist statistics later in the thesis).

Finally, the experiment is conducted. If x falls in the rejection region defined by α, the null hypothesis is 'rejected'. A p value is usually calculated for the observed test statistic. A p value is the proportion of the sampling distribution (assuming the null hypothesis is true) that corresponds to the value of the test statistic observed, or more extreme values; it tells you what proportion of the sampling distribution for X under the null hypothesis is represented by values of the test statistic equal to, or greater than, x.

While p values are a relic from R. A. Fisher's approach to frequentist statistics—an approach which focusses solely on the null hypothesis—they continue to play a role in the analysis of clinical trials. There is considerable debate regarding the differences between Fisher's and Neyman and Pearson's approach to frequentist statistics. Specifically, Fisher's interpretation of a low p value contrasts sharply with Neyman and Pearson's interpretation of a value for a test statistic which falls into the rejection region. Whereas Fisher saw p values as a somewhat informal measure of evidence, which depended on background information for appropriate interpretation, Neyman

[7]See Barnett (1999, pp. 171–174)

and Pearson provided a more explicit decision theoretic interpretation, which takes expected outcomes over a 'long run' of hypothetical trials into consideration. Goodman (1993) provides the historical context for this debate focussing on issues for epidemiology. While these interpretative differences are important and interesting they are not of direct relevance to the topics under discussion here. Recognising the basic formal similarities of the Fisherian and Neyman-Pearson methods is sufficient for the purposes of the thesis.[8]

Understanding Neyman and Pearson's Fundamental Lemma is important. It provides a dual warrant for any Neyman-Pearson test that is a uniformly most powerful test. First, there is the warrant provided for the test statistic falling into the rejection region of the sampling distribution of X under the null hypothesis. That is, if the observed test statistic falls into the rejection region, then we can infer that the observation of such a value for the test statistic is unlikely on the assumption that the null hypothesis is true, providing the assumptions made in the specification of the test hold. Second, for a test of two simple hypotheses and when the test is adequately powered, we can also infer that the likelihood ratio, $P(x \mid \theta_A)/P(x \mid \theta_0)$, for the critical region defined by the test is maximised. In this best case, for any value of the test statistic that falls in the critical region, the observed test statistic is unlikely on the assumption that the null hypothesis is true, *and* considerably more likely on the assumption that the alternative hypothesis is true than it is if the null hypothesis is assumed to be true. It is on these two warrants that Neyman-Pearson methods suggest 'rejecting' the null hypothesis, and 'accepting' the alternative hypothesis. In a test set up in this way, this is what it means to reject, or accept, a hypothesis according to frequentist methods—and, in this context, it is best not to interpret these words as meaning anything more.

Despite the same terminology being used, things are a little more complicated for a test of a simple null hypothesis against a one-sided composite alternative hypothesis. A uniformly most powerful test selects from the possible critical regions, the best critical region: that region for which the

[8]D. R. Cox (2006, p. 36) notes the formal similarities between the approaches

> There is a conceptual difference, but essentially no mathematical difference, between [Fisherian significance tests] and the treatment of testing as a two-decision problem, with control over the formal error probabilities.

maximum likelihood for x under the range of values for θ specified by the alternative hypothesis is greater than the (maximum) likelihood for x under the null hypothesis (again, since the null hypothesis is a simple hypothesis, the maximum likelihood of x under the null hypothesis is equivalent to the likelihood).

Importantly, not all Neyman-Pearson tests are uniformly most powerful tests. And even when they are, uniformly most powerful tests can arrive at counter-intuitive conclusions. Hence, some cautionary notes are required.

First, a uniformly most powerful test of the null hypothesis is not always available. A uniformly most powerful test is not possible when the alternative hypothesis is two-sided. In such situations there is no single pre-experimental likelihood ratio (nor, can a likelihood ratio be defined on the smallest clinically important value for the alternative hypothesis, which is possible when the alternative hypothesis is a one-sided composite hypothesis). Here, frequentist statisticians employ other methods (some of which can be rather arbitrary) to select from the range of admissible tests. See Barnett (1999, pp. 166–177), and Cox and Hinkley (1974, pp. 91–92) for further discussion.

Second, the likelihood ratio is of secondary importance to the set-up of Neyman-Pearson tests, and only ever considered from the pre-experimental perspective. The primary focus of Neyman-Pearson tests is α and β, which are defined in terms of the entire sample space. And because α and β are functions of the entire sample space, the acceptance and rejection properties of the test can alter *independent* of the observed data. Consider an experiment for which the data for a particular test statistic has been observed. Assume there is confusion about the rule used to decide when the experiment was over, such rules are called 'stopping rules'. Since the stopping rule influences α and β, the decision of which stopping rule on which to base the analysis can influence whether the null hypothesis is rejected or accepted (despite the observed data going unchanged).[9]

The inferential procedure for frequentist methods is focussed on the entire sample space, rather than just the observed data. This touches upon an important point of divergence between competing approaches to statistical inference. The point of divergence is whether or not the *likelihood principle* is violated. Donald Berry (1987) provides the following definition of the likelihood principle

[9]Berger and Berry (1988) provide examples.

> The likelihood function $L_x(\theta)$ [$P(x \mid H)$ in my notation] contains all the information in an experiment relevant for inferences about θ, where x stands for the observed data.[10]

Frequentist approaches violate the likelihood principle. Bayesian, likelihoodist, and a range of other approaches to statistical inference do not.

Many debates within philosophy of statistics hinge on whether (and in what circumstances) the likelihood principle should be observed. While it is important to acknowledge the philosophical criticisms of frequentist approaches to statistical inference they are not the focus of this thesis. Rather, the focus is how frequentist methods are applied in interpreting clinical research. I am especially interested in the limitations of frequentist methods for informing therapeutic decisions.[11] I show that these limitations are apparent even within the basic framework of frequentist statistics.

Estimation. Since estimation can be outlined within a conceptual framework similar to that used for hypothesis testing, I focus here on the interpretation of confidence intervals. Rather than accept or reject a pre-specified hypothesis, estimation uses the observed data to infer which values of the unknown parameter θ are supported.

Estimation is based on the properties of test statistics. The method relies on the following argument. X is an estimator of θ. An observation of the test statistic provides information about the sampling distribution of the random variable X. Assuming X is a 'good' estimator of θ (that is, unbiased, sufficient, efficient and consistent), then it is possible to infer which values of θ are supported by the observation. In the specification of the problem, estimation assumes a particular probability model best represents the process under investigation. As in hypothesis testing, the sampling distribution for X has a physical interpretation. Indefinite repetitions of the trial supply values of x normally distributed with a mean value equal to θ—given the test statistic is unbiased and sufficient, and the specification is correct.

Estimation employs critical regions to calculate the lower and upper bounds of the confidence interval. Consider the true value of θ within the parameter space Ω. The lower and upper bounds of a confidence interval are

[10] Jason Grossman (—, pp. 209–302), in a currently unpublished manuscript, provides a more precise version of the likelihood principle, but Berry's version is sufficient for our purposes.

[11] In other work, (La Caze 2008b), I discuss the limitations of frequentist methods for informing therapeutic choice from an ethical perspective.

functions of X calculated on the basis of the observed data, x. The interval created by the lower and upper bounds includes all the values for θ not rejected by a two-sided α-critical region on X.

Given the properties of X as an estimator of θ it is possible to calculate the lower and upper bounds of a confidence interval such that the pre-experimental probability of θ being within the interval can be specified, that is $P(T_\alpha < \theta < T^\alpha \mid \theta) = 1 - \alpha$ for all $\theta \in \Omega$, where T_α and T^α are the lower and upper bounds respectively.

Neyman (1937, pp. 347–350) is explicit in his interpretation of confidence intervals. θ is an unknown constant. The probability of θ falling into any interval is zero or one. However, because T_α and T^α are random variables, prior to observing any value for x, it is possible to discuss the probability of any given interval containing θ. Indeed, as mentioned in the previous paragraph, it is possible to select T_α and T^α such that the probability of θ falling into the created interval is fixed in advance. In a 95% confidence interval the probability of the interval created by these two random variables is 0.95.

The *observed data* and the assumed properties of the sampling distribution are used to calculate the confidence interval reported in clinical trials, $t_\alpha - t^\alpha$. Crucially, the 95% probability statement refers to T_α and T^α, and not t_α and t^α directly. As Vic Barnett (1999, pp. 181–182) notes, whether the *particular* interval, calculated on the basis of observed data, $t_\alpha < \theta < t^\alpha$, captures the true θ is uncertain. The confidence interval says that if the trial was repeated many times, and X is indeed a good estimator of θ, then the means of the respective distributions of values for the observed t_α and t^α would approach an interval that captures θ 95% of the time.

Confidence intervals tell us what the *trial data* says about θ. If assumptions made about X as an estimator of θ and the underlying probability model for hold, then it seems reasonable to take confidence intervals as providing *some* information about θ. Of course, if either of these conditions fail, the calculated confidence interval may be misleading.

This completes my introduction of frequentist methods. By way of counter-point I briefly outline the key components of a Bayesian approach to statistical inference. Whereas frequentist statisticians do not assign probabilities directly to scientific hypotheses, Bayesians do. Bayesians differ in their views regarding the kind of probabilities needed for Bayesian inference. (For instance, Subjective Bayesians hold that reliance on subjective

probabilities is ineliminable. Objective Bayesians, by contrast, hold that the probabilities need not be subjective). Bayesians start with a prior probability of the hypothesis in question—either a single probability or a probability distribution over the hypothesis space. Then, Bayes Theorem is applied to update their prior probability into a posterior probability. Bayes Theorem is a straightforward corollary of the product rule for probabilities.

$$P(H \mid x) = P(x \mid H) \times \frac{P(H)}{P(x)}$$

where H is the hypothesis under consideration, and x represents the observed data. For two hypothesis, H_1 and H_2, Bayes Theorem can be given in an odds form.

$$\frac{P(H_2 \mid x)}{P(H_1 \mid x)} = \frac{P(x \mid H_2)}{P(x \mid H_1)} \times \frac{P(H_2)}{P(H_1)}$$

Here, the conditional odds for the posterior probability for H_2 over H_1 based on data, x, is the product of the likelihood ratio and the prior odds for H_2 over H_1. Note, Bayesian inferences rely on the observed x and not the sampling distribution for X under assumptions about which hypothesis is the true hypothesis. Since all the information from the experiment is contained in the likelihood ratio, Bayesian methods are consistent with the likelihood principle.

Bayes Theorem is not contentious, but the *applicability* of Bayes Theorem to scientific inference is. Much debate in theoretical statistics concerns whether the required probabilities are available, and, if available, whether they are appropriate for forming scientific inferences. Frequentist statisticians argue in the negative.[12] Bayesians reply in the positive, but Bayesians

[12]Neyman (1937, pp. 343–344) makes the following remarks about Bayesian probabilities.

> It is known that, as far as we work with the conception of probability as adopted in this paper, the above theoretically perfect solution [provided by Bayes Theorem] may be applied in practice only in quite exceptional cases, and this is for two reasons:
>
> (a) It is only very rarely that the parameters $\theta_1, \theta_2, \ldots \theta_i$ are random variables. They are generally unknown constants and therefore their probability law *a priori* has no meaning.
>
> (b) Even if the parameters to be estimated $\theta_1, \theta_2, \ldots \theta_i$ could be considered as random variables, the elementary probability law *a priori*, $p(\theta_1, \theta_2, \ldots \theta_i)$, is usually unknown, and hence [Bayes Theorem] cannot

of different persuasions consider different probabilities admissible. Subjective Bayesians are the most permissive. Providing the agent's prior probabilities are consistent, Bayes Theorem can usually be applied to deduce the agent's posterior probability. Objective Bayesians narrow the appropriate application of Bayes Theorem to those situations in which objective prior probabilities are available (as defined according to the specific view of Objective Bayesianism in question). Much more could be said of Bayesian statistics specifically, and non-frequentist approaches to statistical inference more generally, but these brief points are sufficient for our purposes. For the most part, I will focus on frequentist methods.

I now shift focus to a couple of issues concerning how frequentist methods are applied in the analysis of clinical trials. First, a fairly general point. One of the main arguments provided for utilising frequentist methods to analyse clinical trials, as opposed to alternative approaches to statistical inference, is the *objectivity* promised by frequentist statistics. It is worth considering the basis for this claim.

Frequentist methods focus on the *observed data* and expectations about the sample space. Prior beliefs and cost considerations (by way of utility assessments) do not play a *formal* role in drawing inferences from data—or, more accurately, prior beliefs and utility assessments play no role in drawing inferences *once the experiment has been specified.* The personal beliefs of investigators, the conventional line of argument goes, should play no role in how clinical data are analysed. Frequentist statisticians ground their claims to objectivity on the inadmissibility of prior beliefs and cost considerations. As hinted above, it is important to realise that frequentist methods achieve their focus on the data (and expectations about the data over the sample space)—their objectivity—by incorporating the judgements that need to be made into the specification of the the trial. The form of objectivity provided by frequentist methods may have advantages, but not if the subjective judgements made in specifying the trial are overlooked.

And many subjective judgements are needed to *specify* a clinical trial. Some examples include: the selection of the test statistic, and the assumed probability model for that test statistic; how the alternative hypothesis is specified; which outcome, or outcomes, will be considered the primary clinical endpoint; how the data available prior to the study is interpreted in

be used because of the lack of the necessary data.

calculating the power of the study, and so on. None of these judgements necessarily cause problems for the frequentist analysis of a given clinical trial, but the objectivity of frequentist methods are so often touted that, at times, the importance of these judgements is under-recognised. Howson and Urbach (2006, especially Part 3) demonstrate the importance of such judgements to frequentist statistics.

The second issue is that, many statistical tests are conducted in clinical studies, but only the test of the primary hypothesis is set up to approximate the ideal according to the dictates of Neyman-Pearson theory. Most strikingly, since all therapeutic decisions require a weighing of the benefits and possible harms of a therapy, safety endpoints are typically specified as secondary endpoints. While p values and confidence intervals are provided for these endpoints, these tests are not set up in the same way as tests of primary endpoints. The appropriate interpretation of the results of secondary endpoints and subgroup analyses—analyses conducted on groups within a clinical trial—are considerably less clear cut. And yet, interpreting such results is vital for therapeutic decision making. I return to this point in the chapters that follow.

The large drug trials of primary importance to EBM are put to two different tasks. One is to provide a rigourous test of the efficacy of the drug. Often this is needed for regulatory purposes, if the drug is not seen to be efficacious, it will not be approved for marketing. The second task of these trials is to provide evidence for therapeutic decisions. Clearly there is significant overlap in these tasks. But, while a drug's efficacy is important in informing therapeutic decisions, these decisions rely on more aspects of a drug's effect than simply its efficacy. I explicitly consider aspects of this problem in Chapters 7 and 8, but the consequences of the two different tasks expected of clinical research, and the different focus of frequentist methods provided to these two tasks, are noted throughout the thesis.

Chapter 2

Evidence based medicine can't be ...

2.1 Introduction

EBM proposes that medical decisions be based on the best available evidence. While there is little to disagree with the claims of EBM at this general level—*of course* medical decisions should be based on the best evidence—the proposal is vacuous without also elucidating precisely what you mean by this evidence and how you propose that it be used. To the extent that EBM fills in these details, it does so by proposing the 'hierarchy of evidence'. EBM suggests that medical decisions be informed by evidence from as high up the methodological hierarchy as possible. These methodological claims have recently gained the attention of philosophers of science (Worrall 2002, 2007b,a; Bluhm 2005; Grossman and Mackenzie 2005; Upshur 2005). And both practitioners of EBM and philosophers, have recognised that there is much philosophical work to do within EBM (Haynes 2002; Worrall 2007a). Given the extensive practical and political influence of EBM in a wide range of medical decisions, perhaps the most surprising (and worrying) area of philosophical work that is yet to be done is the provision of a clear interpretation and defence of EBM's hierarchy of evidence.

This is not to suggest that *aspects* of EBM have not been debated. Many aspects of trial methodology have been extensively discussed within clinical epidemiology, statistics and philosophy. The role of randomisation pro-

vides one prominent example.[1] There is also an abundance of literature in which EBM is advocated or taught—as opposed to philosophically justified—including, for example, the well known EBM 'guidebooks' (Straus et al. 2005; Guyatt and Rennie 2002). What is missing is a systematic justification of EBM's methodological hierarchy that survives critical analysis.[2]

EBM puts forward the methodological hierarchy as a tool for making *good* medical decisions. Ideally, any justification of EBM needs to, first, describe how the hierarchy should be applied, and, second, justify how this application of the hierarchy improves medical decision making. This chapter focuses of the first part of this task. Proponents have made bold claims about what can be achieved by making decisions in accordance with the EBM hierarchy. Specifically, randomised interventional studies are seen to provide an especially secure form of evidence.

> Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the 'gold standard' for judging whether a treatment does more good than harm. (Sackett et al. 1996)

A number philosophical critiques of EBM have shown that not all the claims made by proponents of EBM on behalf of randomised interventional studies can be justified (Grossman and Mackenzie 2005; Worrall 2007a, 2002). I wish to extend these critiques in a particular way. Advocates of EBM propose that medical decisions—and even more specifically *therapeutic* decisions—are better informed by reference to the evidence hierarchy. I show that the interpretation of EBM's hierarchy that is most often put forward by proponents cannot be justified.

An unambiguous interpretation of the hierarchy has not been provided. Early papers, and the EBM 'guidebooks', provide the clearest account. On this account, the hierarchy is interpreted categorically. The categorical interpretation of the hierarchy holds that evidence from higher up the hierarchy trumps evidence from lower down. I describe this interpretation in Section 2.2. The philosophical treatments of EBM are examined in Section 2.3.

[1] See, for instance, Armitage (1982); Lindley (1982); Suppes (1982); Urbach (1985); and Worrall (2007a,b)

[2] A good recent attempt to collate some of the key arguments at the heart of EBM's claims is provided by Rothwell (2007c).

These accounts respond to the view of EBM that has been provided, and expose its problems. Ambiguity about how the hierarchy should be interpreted, however, gives proponents of EBM some 'wriggle room'. Restricting the claims of EBM, by explicitly narrowing the domain of application, and accepting that randomised trials are fallible, avoids *some* of the criticisms that have been raised. In the final section, I show that even if these moves are made, the categorical interpretation cannot be justified. And moreover, that imposing any further limits impedes the application of the hierarchy to therapeutic decisions. Hence, the chapter is predominately negative. If EBM is to inform therapeutic decisions, the hierarchy cannot be interpreted as proposed by advocates.

2.2 EBM according to the advocates

EBM's history is recent, and localised. As discussed, it developed as a distinct approach to medical practice and education at McMaster University, Canada, during the 1980s and 1990s. The McMaster faculty involved in disseminating the central ideas of EBM, including David Sackett, Gordon Guyatt, Brian Haynes, and Deborah Cook, continue to be prominent among EBM's leading proponents. The driving idea of EBM is that the skills of clinical epidemiology should play a more prominent role in clinical decisions made at a patient's bedside.

The first paper to outline EBM in detail best illustrates how proponents conceive EBM.

> A new paradigm for medical practice is emerging. Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature. (Evidence-Based Medicine Working Group 1992)

EBM is seen as a move from basing medical decisions on the 'unsystematic' judgement of an individual clinician, based on experience or the findings of the bench or basic sciences, to the more 'systematic' and 'relevant' outcomes

of patient-related clinical research. The 'basic' or 'bench sciences' are physiology, pharmacology, and related disciplines such as pathophysiology. These sciences are primarily focussed on developing a theoretical basis for how the body works (physiology), how drugs interact with physiological processes in the body (pharmacology) and the physiological abnormalities involved in disease (pathophysiology). Theory-based sciences such as these provide a contrast to the empiricism of EBM.

Proponents continue to label EBM a Kuhnian paradigm shift in medicine.[3] While they are using idea of 'paradigm' more informally than Kuhn,[4] the continued insistence that EBM is a paradigm shift simply illustrates the conviction of proponents that there is a marked distinction between EBM and the pre-EBM process of medical decision making. Medical authority, personal clinical experience, instinct, pathophysiologic rationale and external evidence gained from the systematic observations of experiments are some of the facets involved in clinical decision making. Any particular decision is likely to rely on a number of these facets. This is as true within the model for medical decisions proposed by EBM as it was in pre-EBM decision-making. The position put forward by EBM proponents is that focus should be given to experimental evidence. More especially, according to EBM's hierarchy, the focus should be given to evidence derived from experiments of a particular design, namely, randomised interventional studies.

EBM claims that good medical decisions involve the appropriate interpretation of evidence:

> Understanding certain rules of evidence is necessary to correctly interpret literature on causation, prognosis, diagnostic tests, and treatment strategy. (Evidence-Based Medicine Working Group 1992)

These 'rules of evidence' are provided by EBM's methodological hierarchy.

[3]The original evocation of Kuhn is provided in Evidence-Based Medicine Working Group (1992, p. 2420); the continued insistence is provided in Guyatt and Rennie (2002, p. 8).

[4]As noted previously, EBM is most certainly not a paradigm shift in the Kuhnian sense; there is no incommensurability between the new and old theories of medical decision making. Further, the shift to the EBM model of medical decision making has been (and continues to be) piecemeal—this would not be possible if EBM really was a Kuhnian paradigm shift.

EBM puts forward different hierarchies for different types of medical decisions. Hierarchies have been provided for decisions relating to therapeutic decisions, prognosis, diagnosis, symptom prevalence and economic and decision analyses; each relying on similar methodological distinctions (Guyatt and Rennie 2002; Phillips et al. 2001). As promised, I focus on the hierarchy provided for treatment and harm. EBM's largest influence has been on therapeutic decision making.

Being specific about what therapeutic decisions entail is important to this analysis. Recall, by 'therapeutic decisions' I mean both population and individual therapeutic decisions. Population therapeutic decisions rely on answering the question of whether the benefits of a particular medical therapy outweigh its harms in a defined population of patients. Such a population typically being defined in terms of average age, condition being treated, and presence of co-morbidities. Individual therapeutic decisions, by contrast, focus on the question of whether the proposed benefits of a particular medical therapy outweigh the possible harms in an individual patient, given his or her unique characteristics.

A number of hierarchies have been proposed for therapeutic decisions, but the differences between them are primarily in the level of detail. Table 2.1 is the hierarchy provided by Guyatt and Rennie, a more detailed version is given by Phillips et al. (2001) (and provided in Table 1.1).

Guyatt and Rennie (2002) place N of 1 randomised trials at the top of their hierarchy of evidence. N of 1 trials are conducted with a single patient. In these studies, the patient is randomly allocated to a period of treatment with the intervention under investigation (the 'active' treatment) or control. After a period of time the patient receives the alternative treatment (either active, or control). The patient's outcomes are monitored in each period. Ideally, both the patient and clinician are blinded to whether the patient is receiving active treatment or control. The set up mimics the very common 'unsystematic' clinical practice of giving a patient treatment and monitoring their outcome. N of 1 trials are particularly useful in some specific contexts;[5] but the effects of the treatment need to be rapid, and concurrent with treatment. N of 1 trials have the benefit that the patient involved in the trial

[5]For instance, this design has been used to assist osteoarthritic patients decide whether they need an anti-inflammatory agent rather than regular paracetamol to control their pain (March et al. 1994). Straus et al. (2005, pp. 172–175) provides some general principles for when N of 1 trials may be useful.

A Hierarchy of Strength of Evidence for Treatment Decisions
N of 1 randomised controlled trial
Systematic reviews of randomised trials
Single randomised trial
Systematic review of observational studies addressing patient-important outcomes
Single observational study addressing patients-important outcomes
Physiologic studies (studies of blood pressure, cardiac output, exercise capacity, bone density, and so forth)
Unsystematic clinical observations

Table 2.1: Guyatt and Rennie's (2002, p. 12) hierarchy of evidence for therapeutic decisions.

is the patient to whom the results of the trial will be applied. In this way, N of 1 trials avoid the challenges of external validity, and presumably this is why Guyatt and Rennie place this design at the top of their hierarchy of evidence for therapeutic decisions. However, N of 1 trials have a number of draw-backs. Most significantly, N of 1 trials are not feasible for determining long-term effects of treatments. N of 1 trials play a limited role in medical research, policy decisions, and therapeutic decisions regarding individual patients; considerably more focus is given to randomised interventional studies. I will not consider N of 1 studies further.

The hierarchy provided by Guyatt and Rennie (2002) and Phillips et al. (2001) highlight the distinctions important to EBM. *Systematic* evidence—that is evidence from studies, whether interventional, observational or laboratory studies—is valued higher than unsystematic experience. Of the systematic evidence, patient-related *clinical* evidence—that is, direct experimental evidence of the effects of treatments on patients—is valued higher than experimental evidence from the basic sciences. And finally, experimental evidence from clinical studies is distinguished according to *methodology*: randomised interventional studies, and systematic reviews of randomised interventional studies, are seen to provide better evidence for therapeutic decisions than

observational studies.

According to proponents of EBM, *systematic* evidence is superior to *non*-systematic evidence, systematic *clinical* evidence is superior to systematic *non*-clinical evidence, and systematic clinical evidence from *randomised* interventional studies is superior to evidence from *non*-randomised *non*-interventional studies. But how is this superiority achieved? To answer this question it is first necessary to examine how EBM applies the methodological hierarchy. In the account provided by the EBM guidebooks the notion that randomised interventional studies trump evidence from lower down the hierarchy is central.

> If the study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search. (Note: We can begin to rapidly critically appraise articles by scanning the abstract to determine if the study is randomised; if it isn't we can bin it.) Only if you can't find any randomised trials should you go back to it. (Straus et al. 2005, p. 118)

> The hierarchy implies a clear course of action for physicians addressing patient problems: they should look for the highest available evidence from the hierarchy. The hierarchy makes clear that any statement to the effect that there is no evdience addressing the effect of a particular treatment is a non sequitur. The evidence may be extremely weak—it may be the unsystematic observation of a single clinician or a generalisation from physiologic studies that are related only indirectly—but there is always evidence. (Guyatt and Rennie 2002, pp. 14–15)

These quotes show that EBM has a broad concept of 'evidence'; results of randomised interventional studies do not constitute the *only* source of evidence. But, equally, EBM has a narrow conception of what provides the 'best evidence'. According to EBM, when it comes to therapeutic decisions the 'best evidence' is provided by the results of randomised interventional studies. And this 'best evidence' is superior to evidence from lower down EBM's hierarchy, seemingly, without qualification. The EBM guidebooks suggest a *categorical* interpretation of the hierarchy.

On the categorical interpretation, the randomised interventional study design is seen to provide an incontrovertible epistemic good. The results of

randomised interventional studies are epistemologically superior to the results of non-randomised (observational) studies, and the superiority is absolute. *All* the results of a randomised interventional study are *always* superior to the results of studies from lower down the hierarchy—at least, for all those studies that are conducted that meet the standards of publication. How else could it be appropriate to 'bin' all 'non-randomised' studies relating to the therapeutic question we are investigating?

2.3 The Critic's View of EBM

Attention now turns to the critiques of EBM that have been provided in the philosophical literature. These criticisms can be seen as a subset to the general concerns noted in the previous chapter. While the philosophical criticism's of EBM have focussed on different aspects of the approach, each respond to a similar view of the hierarchy (Bluhm 2005; Grossman and Mackenzie 2005; Worrall 2007a, 2002, 2007b). Not surprisingly, the shared view is the one most clearly articulated in the EBM guidebooks. That is, that EBM's hierarchy should be interpreted categorically. It is possible to summarise the critical response into a number of broad themes: the claims that are made on behalf of randomisation; EBM's focus on a single aspect of methodology; problems of interpreting EBM's claims broadly; and the challenges of external validity. How some, but not all, of these criticisms may be avoided by proponents of EBM is discussed in the sections that follow.

Worrall (2007b, p. 452), examines the notion that randomised interventional studies provide especially secure knowledge in medicine. In particular, Worrall shows that the benefits of randomisation fall short of making randomisation 'essential' in the sense EBM often takes them to be. Contrary to what is often claimed by proponents of EBM, Worrall shows that randomisation does *not* ensure that all confounding factors, known and unknown, are equally balanced in the experimental groups. While randomisation has some benefits, such as preventing some types of selection bias, it certainly does not ensure infallibility. Nor, Worrall argues, does randomisation justify the *very special* scientific weight proponents of EBM place in randomised interventional studies.[6]

[6]Worrall is especially interested in the benefits or otherwise of *randomisation*. Arguments for and against randomisation *per se* need to be distinguished from arguments in favour of randomised interventional studies over observational studies. You can have

Jason Grossman and Fiona Mackenzie (2005) also highlight the fallibility of randomisation in experimental studies. In addition they illustrate the problems of measuring the quality of evidence according to a single methodological criteria.

> [...][W]hen one attempts to follow the guidelines, one discovers that whether or not the intervention in question is amenable to RCTs, if no RCTs have been performed the evidence obtained can never be better than level III. That is, even the most well-designed, carefully implemented, appropriate observational trial will fall short of even the most badly designed, badly implemented, ill-suited RCT.

Clearly, when evaluating evidence, much more needs to be considered in addition to whether a trial was randomised. (Notably, this is one criticism that is increasingly recognised in the medical literature, see Glasziou et al. 2004; Guyatt et al. 2008a; The GRADE Working Group 2004.)

Robyn Bluhm (2005) also reacts to a categorical interpretation of EBM's hierarchy. But her focus is directed towards the question of how broadly the hierarchy should be applied. In particular, Bluhm is concerned that *epidemiology* relies on the basic sciences. If EBM's hierarchy is applied broadly (say to all of science), then the basic or bench sciences, are seen to be 'lower' forms of evidence. And yet the basic sciences are essential for discovering 'effective causal interventions in the course of a disease in individual patients'—most certainly a key aim of epidemiology (Bluhm 2005, p. 543). Grossman and Mackenzie are also concerned about how broadly EBM's hierarchy is thought to apply. In particular, Grossman and Mackenzie are concerned about the application of EBM's hierarchy to public health policy.

> In recent years this preference for RCTs has extended beyond medicine, with researchers swept up with the ideals and methods of EBM in the promise of scientific recognition and increased

non-randomised interventional studies but not randomised observation studies. Ambiguity surrounding these terms is one of the reasons I prefer 'randomised interventional studies' to 'randomised controlled trials'; the later puts more emphasis on randomisation and fails to indicate that the study design is interventional. This said, clearly there is overlap in the discussion of randomised versus non-randomised, and interventional versus observational. Worrall's remarks about randomisation are relevant to both (just in different ways). These arguments are examined in considerable detail in Chapter 3, *EBM must be ...*, and Chapter 4, *Why randomised interventional studies.*

> funding. One important area in which this has happened is the evaluation of public health interventions, where (to take one example) a food policy program, evaluated observationally, has little chance of being accepted as effective, no matter how effective it actually is, and consequently has no chance of securing the sort of government funding available to phase III drug trials, even though food policy is probably more important to population health than all of these drug trials put together. (Grossman and Mackenzie 2005, p. 517)

The categorical interpretation of EBM's hierarchy also creates problems for external validity.[7] Recall, 'external validity' refers to the extent that results of a clinical trial can be generalised to patients other than those involved in the study. The problem arises because of the importance of the basic sciences in *interpreting* (and thus generalising) the results of randomised interventional studies.

> Because RCTs tend to report only average results in the treatment and control groups, the extent and sources of within group variability are not known. Both extrapolation of the results of an RCT to other patient groups and an understanding of the reasons for differences in outcomes within the study group require a knowledge of biological factors that may influence the effectiveness of a drug. This type of information, however, cannot come from epidemiological studies alone. Rather, it is often first discovered in the context of physiological studies on humans or animals (the second lowest level of evidence in the hierarchy) and of unstructured clinical observation (the lowest level).(Bluhm 2005, p. 537)

Randomised interventional studies examine the effects of a therapy in a very small sample of the patients who will eventually receive the drug. Often, though not always, the sample of patients that are included in trials are highly selected; they are considerably younger and suffering less comorbid illness. Applying the results of these studies to individual patients raises questions of extrapolation and interpolation. If the trial was highly selective

[7]Both Bluhm (2005) and Upshur (2005) recognise the problem of external validity in some form.

in its sample population it can be difficult to know whether the results of the trial extends to patients in routine care. And, for less selective trials, it can be difficult to know whether an individual patient, who resembles the individuals in the trial, would have been among the proportion of patients who benefited from the therapy under investigation.

To the extent these questions can be answered, they rely on basic science. Extrapolating the findings of a randomised trial to a patient under routine care often relies on a judgement of whether the patient's physiological characteristics are similar in relevant respects to patients included in the trial sample. If the patient under routine care is judged to be similar to the sample population, then there is an argument that the results of the trial can be extended to this patient. A judgement that the results of the trial do not extend to an individual in the clinic is also often due to the physiological characteristics of the individual—for instance, the patient may suffer a comorbid illness that will reduce the effectiveness of the therapy (or increase the risk of adverse effects).[8] The challenge that external validity provides for EBM is discussed in the later sections of this thesis. I will expand on this discussion then, but from what has been said already the problem of external validity for a categorical interpretation of EBM's hierarchy can be made clear. Extrapolating the findings of a randomised interventional trial requires a comprehensive understanding of the basic sciences. Interpreting the hierarchy categorically—that is, if evidence from basic science is *trumped* by evidence higher up the hierarchy—then making judgements regarding external validity becomes intractable. If the evidence provided by the basic sciences is as poor as their place in EBM's hierarchy suggests, then there is no principled way to apply to the results of these trials to patients.[9]

Proponents of EBM have elected not to engage with criticism directly (Buetow et al. 2006). Instead the account of EBM provided by proponents

[8]Another factor important to external validity, but not related to the basic sciences, are the circumstances under which the trial was performed. If patients included in the trial are treated in ways that are importantly different to how they are treated in routine care, the then external validity of the trial will be low. Assuming the trial treated patients under realistic conditions, then the reliance on the basic sciences to inform judgements about external validity is increased.

[9]Of course, different parts of pathophysiology and pharmacology will have different levels of plausibility. EBM, by placing all of the basic sciences low on the hierarchy, fails to differentiate those parts of the basic sciences in which we have a high degree of confidence with those parts that are currently more speculative in nature.

has subtly shifted over time.[10] Because of the lack of direct debate, and the absence of a rigourous defence of EBM's epistemological claims, pinning down EBM's 'current' view is difficult. There is certainly enough 'wriggle room' within EBM to avoid some of the criticisms discussed in this section. The view of EBM that would result, however, is considerably more complex, and yet to be adequately explicated by proponents.

I now examine how proponents of EBM can legitimately avoid some criticisms by suitably restricting their claims (while maintaining the primary aim of EBM as informing therapeutic decisions). This can be done by refining how the hierarchy is interpreted. Importantly, while some criticisms can be avoided by suitably restricting EBM's claims, the resolution of other problems comes at a cost to EBM's central aim of informing therapeutic decisions.

2.4 EBM can't be: How the hierarchy *can't* be interpreted

Two criticisms of EBM can be addressed, at least in part, by recognising that the hierarchy does not provide general epistemological rules. The domain of application for the hierarchy should be limited to the context for which it was developed: therapeutic decisions. This addresses Bluhm's concerns, and, less directly, provides an avenue for proponents to respond to Worrall's concerns regarding the *very special* weight EBM places in randomised interventional studies. Restricting EBM's claims to therapeutic decisions, however, is not enough. As much as EBM proposes an interpretation of the hierarchy, it is a *categorical* interpretation. According to this view, when looking for evidence to inform a therapeutic decision if it doesn't come from a randomised study 'bin it'. This view fails to acknowledge the complexity of the results provided by randomised interventional studies.

2.4.1 The EBM hierarchy does not provide general epistemological rules

Much rhetoric about EBM gives the impression that the hierarchy provides some general rules for all of science. It is the implicit assumption that the

[10]Worrall (2007a, p. 983) recognises, and provides some discussion on this point.

hierarchy provides such rules that fuels claims that the highest levels of the hierarchy provide especially secure evidence, and gives the impression that the hierarchy can be broadly applied. If EBM's hierarchy provides general epistemological rules, it would be expected to hold independent of context (or at least hold in a range of contexts defined by some general principles). On this view, randomised interventional studies would provide superior evidence to that of the 'basic sciences' in all (or at least many) scientific disciplines, not just clinical science.

It only takes a moment's reflection to see that this is simply false. Many sciences progress, in whole or in part, without randomised studies. Much of physics, for instance, does just fine without randomised interventional studies. Rather, if it makes any sense, EBM's hierarchy makes sense in the context of *therapeutic decisions*. While I do think a philosophical account of the hierarchy can be provided, it is far from general. Any account of evidence in medicine will be highly dependent on the specific context of the clinical sciences. Importantly, EBM proponents, when pushed, accept this limitation on the range of application of the evidence hierarchy.

'Evidence Based Medicine: What it is and what it isn't', is a reply by proponents of EBM to criticisms of the approach (it is one of the few papers in which proponents engage with criticism, even if they do so rather indirectly) (Sackett et al. 1996). This paper responds to claims that EBM focuses exclusively on randomised interventional studies and meta-analyses. The reply is telling. It makes clear that many types of medical decisions do not require randomised interventional studies. Questions of prognosis and the accuracy of a diagnostic test, for instance, are answered by observational studies. It is *therapeutic* questions that require randomised interventional studies.

> It is when asking questions about therapy that we should try to avoid the non-experimental approaches, since these routinely lead to false positive conclusions about efficacy. Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the 'gold standard' for judging whether a treatment does more good than harm. However, some questions about therapy do not require randomised trials (successful interventions for otherwise fatal conditions) or cannot wait for the trials to be conducted. And if no randomised

> trial has been carried out for our patient's predicament, we must follow the trail to the next best external evidence and work from there. (Sackett et al. 1996)

This passage reinforces the categorical interpretation of the hierarchy, but makes clear that the focus of the hierarchy is therapeutic decisions. Some therapeutic decisions may occasionally have to be made on the basis of alternative evidence; however, if the results of a randomised interventional study are available, then the decision should be based on these results.

Given the rhetoric that is sometimes employed, it is not surprising that some have interpreted proponents of EBM to view the hierarchy as providing general epistemological rules, but it is an over-reach. EBM's central claim is that evidence from study designs featured higher up the hierarchy more reliably inform *therapeutic decisions*. If experimental results from the basic sciences or observational studies are inferior to evidence from randomised interventional studies, it is only in terms of therapeutic decision making.

This limits EBM's claims considerably. And, it provides a response to Bluhm's concern regarding EBM undermining the importance of the basic sciences to epidemiology. The hierarchy simply does not extend that far. It should be applied only when considering the question of whether a particular therapy benefits a patient, or group of patients, more than it is likely to harm. The task of documenting the incidence, and discovering the cause, of a disease need not refer to EBM's hierarchy of evidence. This, of course, is implied by the differing hierarchies provided by proponents of EBM (Phillips et al. 2001). But is not made explicit enough in many discussions of EBM.

Recognising that EBM's hierarchy does not provide general epistemological rules also opens some avenues for proponents of EBM to respond to Worrall's concerns. Worrall (2002, 2007b,a) shows that randomisation does not provide any *guarantee* of the results of a randomised interventional study. Randomisation does not ensure experimental groups are equally balanced for all confounding factors. Randomisation is *not* a general requirement for deriving scientific conclusions from data. Recognising that randomisation is not a general requirement for gaining evidence in clinical science, opens the way for a much more limited—and thus more plausible—defence of randomisation in the context of therapeutic trials. Indeed, limiting EBM's claims in this way underlines the *need* for a positive account of why randomised trials are needed for therapeutic questions. Worrall (2007a) has shown this is yet to be provided by advocates of EBM—at least for the more ambitious claims

that are made by proponents of EBM. In Chapter 4, I provide an argument for randomised interventional studies in testing the efficacy of treatments. In providing this argument, I also show that it does not substantiate many of the grand claims made on behalf of randomised interventional studies by proponents of EBM.

Limiting EBM's claims to the context of therapeutic decisions also provides a response to the emerging epidemic of 'evidence based' disciplines. If a clear, and justifiable, interpretation of EBM is yet to be provided in the very context it was designed for, then the plight of these second generation 'evidence based' disciplines is not promising.

The 'evidence based' label has been extended to other areas of practice, such as nursing and pharmacy, other areas of health decision making, such as public health interventions, as well as a quickly increasing number of disciplines outside healthcare, including evidence based policy making. Though some do, not all of these second generation 'evidence based' disciplines explicitly import EBM's hierarchy along with its label. Whether or not they import EBM's heirarchy, the 'evidence based' claims of these disciplines are either problematic, or at best, unclear.

When these new 'evidence based' disciplines import the EBM hierarchy, such as in evidence based nursing, pharmacy and public health, it is usually a case of 'if it is good for medical decision making, then it is good for us'. In these situations, the EBM hierarchy is being extended to the decisions of interest to the discipline. Chapter 3 and 4 provide an argument for preferring large randomised interventional studies for testing the efficacy of drugs. This argument in turn provides justification for (partially) informing therapeutic decisions on the basis of randomised interventional studies. The argument, however, is specific to therapeutic decisions. Any use of EBM's hierarchy of evidence outside of therapeutic decision making is going to need an independent justification for the scientific context to which it is to be applied. This is not to say it can't be done. Some areas of these other disciplines may be similar enough to therapeutic questions so as to justify use of the hierarchy. But, a justification is needed. Furthermore, for some questions within these new 'evidence based' disciplines the hierarchy is simply inappropriate. As already discussed, one example is the application of EBM's hierarchy to some public health interventions. Grossman and Mackenzie (2005) show that randomised trials are ill-suited to address some research questions within public health. But, due to the hegemony of EBM's hierarchy, methodologies that are well

suited to address the research question are being ignored, or automatically and inappropriately downgraded.

Conversely, when disciplines take on the evidence based label without importing EBM's hierarchy, such as the way 'evidence based policy' is often used, then it is difficult to see what work the label is doing (other than sounding vaguely reassuring).[11] EBM without its hierarchy is meaningless. So too are other uses of the label without some explicit expression of what 'evidence' is being referred to, and how it is being used.

2.4.2 The EBM hierarchy can't be interpreted categorically

Recognising the hierarchy does not provide general epistemological rules, and limiting application of the hierarchy to therapeutic decisions, provides an avenue of response to some criticisms of EBM. But not all. EBM's account of how the hierarchy should be put into action relies on a categorical interpretation. When searching for evidence to inform a therapeutic decision:

> If the study wasn't randomised, we'd suggest that you stop reading it and go on to the next article in your search. (Straus et al. 2005, p. 118)

This does not suggest an interpretation of the hierarchy where only certain well defined questions are best answered by randomised studies. The categorical interpretation suggests that when it comes to therapeutic decisions *all* of the results of a randomised interventional study *always* trump evidence from lower down the hierarchy. Evidence from observational studies may sometimes be needed to help inform therapeutic decisions, but only in the *absence* of a randomised interventional study, and only then, when the considerably 'weaker' strength of this evidence is emphasised.

> Without clear confirmatory evidence from large-scale randomised trials or their meta-analyses, reports of moderate treatment effects from observational studies should not be interpreted as providing good evidence of either adverse or protective effects of

[11] It might be argued that 'evidence-based' is doing some work in 'evidence based policy'. Specifically, demarcating policy decisions based on emotion, or tabloid press, from policy decisions based on some form of 'evidence'. But, this use of 'evidence' is much too vague. To do something more than sound vaguely reassuring 'evidence based policy' needs to be clearer about what this 'evidence' is, and how it is being used.

> these agents (and, contrary to other suggestions, the absence of evidence from randomised trials does not in itself provide sufficient justification for relying on observational data). (Collins and MacMahon 2007, p. 24)

While the categorical interpretation is relatively straightforward, it is unsustainable.

First, as has already been discussed, the categorical interpretation equates quality of evidence with a single aspect of methodology. Many aspects of clinical trials affect the quality of the evidence they produce, not simply whether or not they are randomised (Grossman and Mackenzie 2005). Again, this is a criticism that the medical literature is responding to. The recently developed GRADE system for evaluating the quality of evidence explicitly recognises that randomisation is only one measure of quality (Guyatt et al. 2008b,a).

The second problem for the categorical interpretation holds even for the best designed (and implemented) randomised interventional studies. The categorical interpretation fails to distinguish between the different types of results furnished by randomised interventional studies. Randomised interventional studies supply many results, however, the warrant for each of these results is far from equal—even by EBM's reckoning. Randomised interventional studies are designed (statistically and methodologically) with a particular question in mind. Most often (in the studies of interest in EBM) the question is whether a given therapy will have a beneficial effect on a particular outcome in a defined group of patients. Most trials are set up to test a benefit hypothesis. Recall, the question for which the trial has been designed is called the primary hypothesis, and the outcome of interest to this hypothesis, the primary outcome or endpoint. For example, a randomised study might examine whether aspirin reduces the rate of death in a patient who is admitted to hospital suffering from acute coronary syndrome. The statistical test on this primary hypothesis will be adequately powered. But, in addition to the primary hypothesis there is usually two to three secondary hypotheses and related endpoints. These secondary hypotheses often relate to other benefits the therapy may have, as well as harms the therapy may cause. For example, with regard to the aspirin trial, secondary hypotheses may relate to whether aspirin reduces angina pain, and whether it increases the risk of bleeding. Therapeutic decisions rely on (or at least need to incorporate) the results of secondary endpoints, and the statistical results provided on such

analyses can be misleading because they are underpowered.

The results of an intervention on subgroups within the trial are also important to therapeutic decisions. For instance, regarding the aspirin trial above, a clinician with an elderly female diabetic patient will be particularly interested in the results of the intervention in the relevant subgroups; the female patients, the elderly patients and the diabetics. Subgroup analyses raise a number of thorny issues for the appropriate analysis and interpretation of randomised trials, and there are a range of views on the matter (many of which I discuss in Chapter 7). However, whichever view is taken with regard to the appropriate analysis of subgroups, it is undeniable that they provide evidence of importance to therapeutic decisions. This results in an

> ...unavoidable conflict between the reliable subgroup-specific conclusions that doctors and their patients want, and the unreliable findings that subgroup analyses of clinical trials might offer. (Collins and MacMahon 2007, p. 13)

Subgroup analyses and analyses of secondary endpoints, together with the results from the primary hypothesis test make up what is called the 'results' of randomised interventional studies. Any interpretation of the hierarchy needs to acknowledge the different warrant frequentist statistical methods provide to these results.

Randomised interventional studies are analysed according to the frequentist methods introduced earlier. Within these methods, *power* plays a vital role in establishing the warrant of the statistical test. Recall, 'power' is the pre-test probability that the statistical test will 'reject' the null hypothesis, on the assumption that the null hypothesis is false. From a pre-trial perspective the role of power is not contentious. Much effort is taken to ensure that the primary hypothesis test is sufficiently powered. Trials that are not sufficiently powered to test the primary hypothesis are often refused funding, or not given ethical approval. This is because underpowered trials are less likely to provide 'definitive' results according to the dictates of frequentist statistics. Statistical tests on secondary hypotheses and subgroup analyses, however, are often underpowered.[12]

[12]It should be noted that 'power' as defined within hypothesis testing does not play a direct role in estimation theory. However, the conceptual framework for hypothesis testing and estimation are similar, and the influence of a concept similar to power could be outlined within estimation theory.

Once the results of a trial have been observed the role of power is considerably more contentious. However, it is well recognised that the observed results of a trial are less reliable when the size of the trial is small relative to the true size of effect under investigation. Underpowered tests can result in false negative results—that is, fail to reject a false null hypothesis. After all, low power predicts—from a pre-trial perspective on the assumption the null hypothesis is false and a given size of trial—that observing a statistically significant result is unlikely. Somewhat less well recognised, but just as important, a low powered test can also result in false positive results. If the true effect size is small, and the power of the test for this small effect is low, then any result that is statistically significant will over-estimate the effect size. Land (1980) provides a description of this phenomenon, and uses it to explain 100 fold discrepancies in estimation of cancer risks due to low-dose radiation (I provide a thorough discussion of this phenomenon in Chapter 8). In this sense—that is, the possibility of false negative, or false positive, results—the results of subgroup analyses and analyses of secondary endpoints are unreliable (when they are underpowered). The unreliability of the results of subgroups and secondary endpoints, coupled with the importance of these results to therapeutic decisions, undermines a categorical interpretation of the EBM hierarchy.

To be clear, I am not suggesting that proponents of EBM do not recognise that the results of primary hypothesis tests have a different warrant to the results of secondary hypotheses and subgroups analyses. On the contrary, they will be the first to point out the differences. It is not contentious that these different types of results have a different epistemic standing.[13] Moreover, EBM has a fairly standard reply to the problems of subgroups analyses and analyses of secondary endpoints: await the results of meta-analyses. In the ideal case when you have a number of high quality randomised interventional studies that include similar enough patients and test the same treatment, meta-analysis will improve the reliability of subgroup analyses and secondary endpoints. The results of meta-analyses, however, are not always available, and when they are the realities of clinical research can undermine the improved reliability achieved in ideal circumstances (Egger et al. 1997; Egger and Smith 1995). More importantly, for our purposes, none of this is recognised by proponents of EBM when they describe how the hierar-

[13]Although, precisely what should be done about this different epistemic standing is highly contentious. I consider this debate in Chapter 7.

chy should be applied. The categorical interpretation is the interpretation of EBM's hierarchy that is provided by proponents. But it fails to acknowledge that *some* results of randomised intervnetional studies are unreliable—and because subgroup analyses and secondary endpoints are of particular interest to therapeutic decision makers, the unreliability of these results presents a particular problem for EBM.

I am also not suggesting that the results of outcomes within trials that have low pre-trial power are unimportant, or irrelevant (on the contrary these results are very important). My point is simply that the warrant provided for these results according to frequentist statistics is different to the warrant provided for the results of a well-powered primary hypothesis test. And, that the categorical interpretation of EBM's hierarchy fails to adequately recognise this difference.

This suggests a second limitation is needed to further restrict application of EBM's hierarchy of evidence. Not only does the hierarchy need to be restricted to therapeutic questions, but within therapeutic questions, application of the hierarchy should (at best be) limited to the results of primary hypothesis tests and well-conducted meta-analyses, as it is only for these tests that the optimal warrant of frequentist statistics is provided. While this is a positive move for EBM, as it provides a more justifiable interpretation of the hierarchy, there is a cost.

If my analysis is correct, applying EBM's hierarchy is not sufficient for informing therapeutic decisions, which after all is EBM's primary aim. Recall, therapeutic decisions rely on assessing the benefits and harms of therapies for groups of patients and individuals. Limiting the hierarchy to the results of primary hypothesis tests impedes this interpretation of the hierarchy informing therapeutic decisions in two ways.

First, the primary hypothesis under test in the vast majority of clinical trials is a 'benefit' hypothesis—trials are set-up, and powered to test whether a therapy produces a proposed *benefit* in a defined group of patients. Outcomes regarding the safety of the therapy are almost always relegated to a secondary hypothesis. Whereas the possibility of benefits and harms are symmetrically important to therapeutic decisions, the quality of evidence provided within EBM for benefits and harms is asymmetrical; according frequentist methods the benefits of therapies are tested more rigorously than the harms. The categorical interpretation of EBM's hierarchy obscures this asymmetry by proposing that therapeutic decisions be informed by reference

to a hierarchy that fails to recognise the differing warrant provided by the results of primary and secondary analyses. Again, while in other sections of the literature proponents of EBM acknowledge that randomised trials are not the best method for establishing *unsuspected* adverse effects, and recognise that the results of secondary endpoints and subgroup analyses can be unreliable, there is no recognition of any of this in what proponents of EBM say about applying the hierarchy of evidence.

Second, the results of secondary endpoints and subgroup analyses play a role in informing therapeutic decisions in an *individual* (Horwitz et al. 1998; Rothwell 2005b). The results of the primary hypothesis test gives information on whether the therapy benefits a defined population of patients. As discussed earlier, while the appropriate analysis of secondary endpoints and subgroups is highly contentious, these results play a role in decisions regarding individual patients. By comparing the unique characteristics of the patient in the clinic with the appropriate subgroups within the trial, therapeutic decisions (in some circumstances) can be refined so as to be more relevant to the individual. The categorical interpretation of the hierarchy fails to acknowledge the reduced warrant for findings from subgroup analyses. Indeed, in some circumstances, subgroup analyses can misleading (see Section 7.2 for discussion and examples). Further limiting the hierarchy to the results of primary hypothesis tests and the results of meta-analyses rectifies this failure, but rules out using analyses of subgroups and secondary endpoints to refine therapeutic decisions. Much more needs to be said on these matters (and later in the thesis, much more is said). The needs of therapeutic decision makers, especially their reliance on secondary endpoints and subgroup analyses, rules out the categorical interpretation of EBM's hierarchy.

The categorical interpretation of the hierarchy provides a simple message for decision makers: Base your decisions on the results of randomised trials and meta-analyses. The message however is too simple; the results furnished by randomised interventional studies are considerably more complicated. The interpretation of EBM's hierarchy can be further restricted to avoid this problem, but this more restricted interpretation severs the direct link between EBM's hierarchy and therapeutic decisions.

2.5 Conclusion

Proponents of EBM do not provide an unambiguous interpretation of the hierarchy of evidence. But as much as an interpretation is provided, the categorical interpretation of EBM's hierarchy is the interpretation most often put forward by advocates (either explicitly, or implicitly). The categorical interpretation holds that the results of randomised interventional studies more reliably inform therapeutic decisions than the results of observational studies. This interpretation, however, can not be justified without considerable qualification. Any successful interpretation of EBM's hierarchy of evidence will have to limit the claims of EBM. Two such limits are proposed. First, the application of the hierarchy should be limited to therapeutic decisions. EBM proponents, in their more careful moments, admit that the evidence hierarchy under consideration does not apply to other medical decisions, for example, decisions relating to prognosis, or unsuspected side effects of drugs. But, the reasons for this have not been documented, and as a result are forgotten, or under-emphasised in much of the EBM literature. Further, even once the application of the hierarchy has been limited to therapeutic decisions the categorical interpretation still does not hold. The second limit further restricts application of EBM's hierarchy to the results of primary hypothesis tests and meta-analyses. The second limit is proposed because findings regarding secondary hypotheses, and subgroup analyses, are less reliable according to frequentist statistics. And yet, adhering to this limit impedes EBM's capacity to inform therapeutic decisions.

As promised, this chapter has been mostly negative. It has shown that the dominant (and most clear) interpretation of EBM's hierarchy that has been provided by proponents cannot be justified. And while amendments can be made to how the hierarchy is interpreted to avoid some of the criticisms, this cannot be done without also restricting EBM's claims to be able to inform therapeutic decisions. In as much as there is a positive payoff to the conclusions of this chapter, it will be found in clearing the way for a considerably more restricted, and context dependent interpretation of EBM's hierarchy of evidence. I begin this task in the next chapter.

Chapter 3

Evidence based medicine must be ...

3.1 Introduction

This chapter focuses on how proponents of EBM have justified their claim that therapeutic decisions are better informed by evidence from randomised interventional studies. As John Worrall (2007a, p. 982) has suggested, the needed justification is inherently philosophical.

> [T]here surely is, at the underlying general level, nothing special about the role of evidence in medicine. Real evidence-based medicine results from applying the universal general principles of the logic of evidence to the particular case of medicine, and especially (though not of course exclusively) to claims about which treatment are and which are not genuinely therapeutic.

Proponents of EBM make strong claims on behalf of the evidence hiearchy. As previously discussed, randomised interventional studies are seen to provide especially secure evidence for therapeutic decisions. See, for instance, the quotes from Straus et al. (2005, p. 118) and Collins and MacMahon (2007, p. 24) provided in the previous chapter, on page 41 and 51 respectively. A systematic justification of EBM requires (i) an interpretation of the hierarchy, which describes clearly how it is to be applied in therapeutic decision making, and (ii) a justification for that interpretation, which explains why applying the hierarchy as proposed more reliably informs therapeutic decisions.

While there are few arguments provided explicitly for EBM's *hierarchy*, there are a number of arguments that have been provided for informing therapeutic decisions on the basis of evidence from randomised interventional studies. I follow the critical literature, and in particular the work of John Worrall (2007a; 2002; 2007b), in finding that these arguments do not substantiate the very special scientific weight placed in randomised interventional studies by proponents of EBM. But in contrast to Worrall I argue that a defensible interpretation of EBM's hierarchy can be provided—albeit an interpretation that substantiates much less than what is often claimed by proponents of EBM.

I propose that EBM's hierarchy should be interpreted as a hierarchy of comparative internal validity. 'Internal validity' is the degree to which the results of a study are accurate for the sample of patients included in the study (Fletcher et al. 1996, p. 12). By 'comparative internal validity' I mean, all other things being equal, that studies utilising methods higher in EBM's hierarchy, have higher internal validity than studies designed according to the methods lower down the hierarchy. Comparative internal validity is an under-appreciated argument for the hierarchy of evidence. While the argument is present in the clinical literature, the claims that it substantiates are considerably more circumscribed than those made by advocates of EBM. This and the other arguments that have been provided for EBM's hierarchy are examined in Section 3.2. In section 3.3, I sketch how EBM's hierarchy, viewed as a hierarchy of comparative internal validity can be applied, and illustrate the considerably more limited claims that can be justified on this basis.

3.2 Arguments for EBM's hierarchy

A justification for how EBM's hierarchy better informs therapeutic decisions is not found in the popular guidebooks. Straus et al. (2005, p. xii) suggest

> Those who wish, and have time for, more detailed discussions of the theoretical and methodological bases for the tactics described here should consult one of the longer textbooks on clinical epidemiology.

One of the textbooks they refer to is their own, Haynes et al. (2006) (then forthcoming). The argument for randomised interventional studies is sum-

marised in the chapter written by David Sackett (2006), I quote in full.

> [M]ightn't a high-quality cohort study be as good as, or even better than, an RCT for determining treatment benefit? Some methodologists have vigourously adopted this view. I disagree with them, for two reasons. First, there are abundant examples of the harm done when clinicians treat patients on the basis of cohort studies. Two recent examples of cohort-based treatment recommendations that failed in RCTs are postmenopausal estrogen plus progestin for healthy women and vitamin E for coronary heart disease. (Note my argument here does not apply to determining treatment harm, where observational studies are often the only way to detect a treatment's rare but awful adverse effects.)
>
> My second justification is an unprovable act of faith. It professes that the gold standard for determining the effectiveness of any health intervention is a high-quality systematic review of all relevant, high-quality RCTs. When the other study architectures are measured against this gold standard, they have generated less reliable estimates of effectiveness. For example, Regina Kunz and her colleagues performed a Cochrane Review of randomisation as a protection against selection bias in health care trials. They frequently found a worse prognosis at entry among control patients in nonrandomised studies. Moreover, they documented the overestimation of treatment effects when the randomisation schedule was not concealed from the clinicians who were inviting patients to join RCTs, converting these 'RCTs' into cohort studies.

The candour is refreshing, but the argument is far from compelling. Sackett's first argument is empirical—experience, he suggests, has shown randomised trials to be more reliable. Sackett's 'second justification' is difficult to differentiate from the first. Indeed, it appears to be a repetition of the first argument with some acknowledgement that, as this is an empirical argument, it is not—by Sackett's reckoning—compelling, rather, it is an 'unprovable act of faith'.

I review this and additional arguments that have been provided for the hierarchy by proponents of EBM. Particular focus is given to the distinction made between randomised interventional studies and observational studies. In doing so I closely follow the arguments provided by Worrall (2007a; 2002;

2007b). Worrall discusses the empirical justification as well as a number of additional arguments for the necessity of randomisation in clinical trials, and then comprehensively critiques these arguments. He comes to the view that randomisation, while often benign, is not *essential*. By 'essential' Worrall means that randomised interventional studies are not 'essential for any truly scientific conclusion to be drawn from trial data' (Worrall 2007b, p. 452). I review three of the arguments provided for randomised interventional studies that Worrall critiques: (i) the empirical justification of EBM's hierarchy, (ii) the view that randomising controls for all confounding factors, known and unknown,[1] and (iii) that randomising uniquely prevents selection bias.

It is important to be clear regarding the distinction between arguments for *randomised interventional studies*, and arguments for *randomisation* in interventional studies. At times Worrall appears to be more focussed on arguments regarding randomisation *per se*, but most of what he has to say is also relevant to the distinction between randomised interventional studies and observational studies. For instance, any argument which established that random allocation ensured experimental groups were equally matched for all possible confounding factors would provide an argument for the superiority of randomised interventional studies over observational studies. In other areas, however, such as when we come to the argument regarding selection bias, more care is needed in distinguishing whether we are discussing randomisation in and of itself, or randomised interventional studies. This chapter and the following chapter provide an opportunity to outline the epistemological issues that hang on these distinctions.

There is much to agree with in Worrall's analysis, in particular, he shows that many of the ambitious claims made by proponents of EBM on behalf of randomisation cannot be justified. But, in finding no argument for randomisation to be 'essential' for science, Worrall concludes that no epistemological distinction can be drawn between randomised interventional studies and observational studies.

> The best we can do (as ever) is test our theories against rivals that seem plausible in the light of background knowledge. Once

[1]In statistics, a 'confounding factor' is a third variable, which correlates with two variables that are being investigated for a potential causal relationship. The confounder may either mask a 'true' causal relationship between the two variables under investigation, or make it appear as though a causal relation exists between the variables under investigation when in fact both variables are under the influence of the confounder.

> we have eliminated other explanations that we know are possible (by suitable, or *post hoc*, control) we have done as much as we can epistemologically. (Worrall 2007b, p. 486)

This misses the argument that EBM's hierarchy can be justified as a hierarchy of comparative internal validity. Despite falling short of showing randomised interventional studies are essential for drawing scientific conclusions from data, the argument of comparative internal validity substantiates an epistemological distinction being drawn between randomised interventional studies and observational studies in clinical science. In this chapter I focus on the interpretation of EBM's hierarchy as a hierarchy of comparative internal validity, in the next chapter, I argue that there is an important epistemological distinction between randomised interventional studies and observational studies in more detail.

Comparative internal validity as an argument for the distinctions made in EBM's hierarchy can be found in the epidemiological literature, but it has been under-emphasised in the philosophical discussions to date.[2] One reason for this is that proponents of EBM have focused on disseminating, advocating and teaching EBM, rather than providing a philosophical justification of the view. In the literature focussing on EBM, the justification of the hierarchy is left implicit. The argument for EBM's hierarchy is found elsewhere—notably, in the clinical epidemiological literature. Another reason internal validity has been under-emphasised in the philosophical literature is that the philosophical analyses take the ambitious claims that proponents of EBM have made as a starting point, and search for arguments that could substantiate these claims (Bluhm 2005; Grossman and Mackenzie 2005; Upshur 2005; Worrall 2007a). While this is appropriate, the arguments regarding comparative internal validity are given little attention because they do not substantiate EBM's more ambitious claims. By changing tack, and focussing on arguments that are available in the epidemiological literature, the importance of comparative internal validity to the justification of EBM's hierarchy is highlighted. This approach permits examining what claims can be substantiated on the basis of this justification of EBM's hierarchy.

Randomised interventional studies play an important role in testing certain well-defined therapeutic questions. Clarifying why randomised interven-

[2]For instance, though they employ different terminology, and they attempt to draw a stronger conclusion, the main argument provided for EBM's hierarchy by Collins and MacMahon (2007) is that of comparative internal validity.

tional studies play this role may help explain why clinicians are so enamoured by this study design, while at the same time avoid the mistake (so often made within the EBM literature) of claiming too much on behalf of this methodological distinction.

3.2.1 The empirical justification of EBM's hierarchy

The empirical justification for EBM's hierarchy, as provided by Sackett in the previous section, cites studies finding randomised interventional studies provide more conservative estimates of treatment effects than non-randomised (observational) studies.[3] Proponents of this argument contend that observational studies provide less conservative estimates of treatment effects because of biases inherent in comparing groups that have not been randomly allocated.[4] The point at issue, however, is whether—and, importantly, how—randomised interventional studies provide more reliable evidence for therapeutic decisions. Let's accept the data Sackett is citing. How does this support the conclusion that the estimates provided by the randomised interventional studies are more reliable? What stops the opposite conclusion: that observational studies are *correct*, or more *likely to be correct*, and randomised interventional studies *under-estimate* treatment effects? The data alone provides no justification for asserting that randomised interventional studies provide the correct estimates. The empirical justification of EBM's hierarchy requires the premiss that randomised interventional studies *are* more reliable in order to make the claim that observational studies over-estimate treatment effects.

Empirical arguments do not provide a justification *for* EBM's hierarchy. They are circular. If you already accept that randomised interventional studies are the 'gold-standard', then the data cited by EBM proponents are grist for your mill. But the data won't compel a sceptic. Both Worrall (2002, p. S326) and Grossman and Mackenzie (2005, p. 520) make this point.

While this objection alone is strong enough to sink the empirical argument as a *justification* for EBM's hierarchy, there are further problems. As Worrall (2007a, pp. 1009–13) notes, recent reports comparing the findings of

[3]See for example Chalmers et al. (1977) and Sacks et al. (1982)

[4]Presumably, one of these biases is also publication bias. If observational studies are more biased, and publication bias was *not* present, then both over-estimation and under-estimation of treatment effects would be expected; not just over-estimation.

randomised interventional studies and observational studies contradict the earlier reports. Benson and Hartz (2000) and Concato et al. (2000) found estimates of effect size from observational studies were consistent with those found in randomised interventional studies in a range of therapeutic areas. Concato et al. (2000) suggest that earlier comparisons of the study methodologies focused on less rigourous observational studies. Therefore, even on its own terms, the empirical data underpinning the argument for the evidence hierarchy is poor.

At the heart of EBM is an epistemological claim: evidence from higher up the hierarchy provides more reliable evidence for therapeutic decisions. As such, EBM's hierarchy requires a philosophical justification.

3.2.2 Randomisation controls for *all* confounding factors

Clinical studies are conducted in order to test the effects of a therapy on a defined group of patients; typically studies are set up in order to test whether the therapy causes the beneficial effects suggested by research in basic science, or previous experience. Studies are set up in such a way as to ensure, as much as is possible, that any observed differences between the treatment and control group are due to the effects of the intervention. To achieve this the groups must be as similar as possible. One way to ensure the comparability of the groups is to match the treatment and control groups for all known confounding factors. Prospective observational cohort studies, for instance, match the experimental groups in this way. Obviously, however, cohort studies are unable to guarantee that the experimental groups are also matched for *unknown* confounders. A common claim in the clinical literature is that randomisation provides this guarantee.

Collins and MacMahon (2007, p. 23), complain that

> [...] non-randomised methods do not provide assurance that all sources of known and unknown bias are adequately controlled, and so cannot exclude the possibility that moderate biases have obscured or inflated any moderate effects, or have falsely indicated a treatment effect when none existed.

According to Collins and MacMahon, randomised allocation provides the assurance that all known and unknown confounders are adequately controlled.[5] While there is a sense in which randomisation provides this assurance, it rests on an important ambiguity in the use of the term 'bias' (an ambiguity which will be clarified shortly). Importantly, it is because randomisation is thought to eliminate bias due to confounding factors that it is seen as essential in clinical trials. Being clear on what sense of bias is 'eliminated', shows this claim to be false.[6]

'Bias' is a term that is variously applied in the clinical literature. Often it is used informally to refer to any factor that could obscure the 'true' results of a trial. (Worrall also uses the term in this way). However, 'bias' is used in statistics in a number of more formal ways. In parametric statistics, *statistical* bias refers to the *expectation* of an estimator of a parameter. An estimator is unbiased if its expectation is equal to the true value of the parameter. This is often informally referred to as an estimator's *long run* average.[7] It is *only* in the statistical sense that randomisation eliminates bias due to confounding factors, known and unknown. This is because statistical bias entails consideration of the entire sample space. (Note. In the quote above, Collins and MacMahon can only be referring to statistical bias; otherwise the claim is false).

Consider a population of a trial being randomly allocated to treatment or control. On any given allocation it is possible that the treatment and control group are not equally balanced for a particular confounder. Indeed, as Worrall (2002, p. S324) has pointed out, given that there could be indefinitely many possible confounding factors, the probability that a confounder is not equally balanced between the groups on any *particular* allocation is high. However, this is not so in the indefinite sequence of trials. If the trial is repeated indefinitely, with a new allocation of the population performed each

[5]Collins and MacMahon are not the only ones to make this claim. For instance, Kendall et al. (1983) state the following: '[...] by the very nature of the randomisation process, the effects of factors outside the experiment can show no favour to the factors inside it, and our inferences are free from bias.' Worrall provides a number of additional examples (see Worrall 2002, pp. S321–S324).

[6]I am indebted to conversations with Jason Grossman on this point. The comments I make about 'bias' pick up on points made in Grossman and Mackenzie (2005, p. 518).

[7]But this is not strictly correct. Rather it is the estimator's average over the sample space; any actual long run (even an infinite long run) won't necessarily equal its average over the sample space

time, then, in this indefinite sequence of trials, the effect of any unbalanced confounder in a particular allocation will be counteracted by the distribution of that confounder in other allocations. The net effect of all possible confounders on an unknown parameter in the indefinite sequence of trials will be zero. Hence, randomisation eliminates statistical bias due to confounding factors.

Now that the term 'bias' has been disambiguated, the problem for EBM can be made clear. That randomisation eliminates bias due to all known and unknown confounding factors is one of the key arguments EBM employs to justify randomised interventional studies having a special epistemic place in the hierarchy of evidence. Once it is clear what this 'elimination of bias' actually amounts to, it is reasonable to question whether it is sufficient to justify the emphasis placed in randomised interventional studies by EBM. Recall the advice of one of the EBM guidebooks: if a study is not randomised, 'bin it'. Statistical bias due to confounders is only eliminated in the indefinite sequence of trials, but (of course) the trials that inform therapeutic decisions are not repeated (at all, let alone indefinitely). In any particular trial allocation a confounder may be distributed unevenly between the experimental groups. While the groups can be examined for the imbalance of any known confounders, this is obviously not possible for unknown factors. Lack of statistical bias provides no assurance with regard to the actual allocation of the trial. In order to substantiate its ambitious claims, EBM needs randomisation to eliminate the informal sense of bias. But, of course, randomisation does not provide this kind of assurance.

While Worrall does not disambiguate 'bias' in this way, he makes a similar point in relation to randomisation.

> The fact is that the subjects have been randomised between control and experimental group only once, and that division either is or is not balanced for the unknown factor at issue. Suppose it is unbalanced, and that this throws the conclusion about the efficacy of the treatment off, then it seems to me scant consolation to be told that—although you don't and can't know it,—you were 'unlucky', and if the randomisation had been repeated indefinitely you would, in the indefinite long run, have inevitably realised your mistake. (Worrall 2007b, p. 484)

This, Worrall argues, undermines the view that randomisation is 'essential', or *sine qua non*, for clinical trials. And on this point, we agree. Importantly, rejecting these more ambitious claims is consistent with accepting that randomisation plays an important role in study designs in the clinical sciences. This role shall be outlined after considering a third argument for randomised interventional studies—that randomisation prevents selection bias.

3.2.3 Randomisation prevents 'selection bias'

The prevention of 'selection bias' is the one 'cast-iron' argument for randomisation that Worrall (2007a, p. 1009) concedes. But while he accepts that the prevention of selection bias provides a reason to randomise, Worrall argues that selection bias (as he defines it) can be avoided through other measures. Worrall's definition of 'selection bias', however, is considerably narrower than the conception of 'selection bias' in the epidemiological literature. Under the broader conception, 'selection bias' provides a rationale for randomisation that can not be provided by alternative measures.[8]

'Selection bias', according to Worrall, is the bias that can occur when trial investigators allocate patients to treatment or control. I will call this 'investigator-selection bias'. Having the investigators allocate patients to the experimental groups can obscure the analysis in a number of ways.[9] Perhaps the investigators (subconsciously or otherwise) preferentially allocate patients who they judge more likely to respond favourably to the treatment arm. Or, perhaps those patients with recalcitrant illness, or a propensity for side effects, are allocated to the control arm. Allocations such as these have the potential to significantly confound the analysis. Further, if the same investigators that allocated patients to experimental groups are responsible for the subject's treatment and collection of results—that is, the study is at best single blind—then there is a substantial risk of further bias to enter the analysis. For instance, the investigator's knowledge of a participant's

[8]While Worrall (2007a, p. 1008) acknowledges his notion of selection bias is narrower than that used in the medical literature, he does not appear to see the consequences of this difference for his argument.

[9]Here the informal notion of 'bias' is sufficient. That is, bias is any factor that will obscure the results of the trial from reflecting the real effect of treatment. When statisticians refer to bias, as in selection bias, it can be difficult to discern whether they are referring to bias informally, or 'statistical' parameter bias. Here, at least, the ambiguity does not cause any problem.

treatment allocation may influence how they treat the patient (even a higher level of implicit encouragement may make important differences for some conditions). And, perhaps even more importantly, knowledge of treatment allocation may influence how investigators interpret the patient's response. Clearly, these possible sources of bias undermine our ability to draw the right inference regarding the treatment's efficacy.

Worrall accepts that randomisation prevents investigator-selection bias. But he notes that it does this via two mechanisms: by taking allocation out of the control of the investigators, and by permitting double-blinding.

> Notice however that randomisation as a way of controlling for selection bias is very much a means to an end, rather than an end in itself. The important methodological point is that control of which arm of trial a particular patient ends up in is taken away from the experimenters—randomisation (as normally performed) is simply one method of achieving this. (Worrall 2007a, p. 1009)

Worrall takes this as further support for his conclusion that randomisation is not *essential* for evidence in medicine, but there are problems with this analysis.

Alternative methods for removing investigator-selection bias are only possible for certain study designs. In particular, the design needs to be one in which the allocation of patients into experimental groups *can* be taken out of the hands of investigators—that is, the study needs to be interventional. While alternative methods can be used to avoid investigator-selection bias in interventional studies, such methods are *not* available in the non-randomised studies that EBM is referring to. In observational studies the choices, and the myriad of other factors, that have caused patients to fall into the 'treatment' (or 'case') and 'control' groups have already played their part.

As it is used in the clinical literature 'selection bias' occurs when 'comparisons are made between groups of patients that differ in ways, other than the main factors under study, that affect the outcome of the study' (Fletcher et al. 1996, pp. 7–8). This definition incorporates Worrall's investigator-selection bias, but also includes other forms of selection bias. This broader notion of selection bias includes 'patient-selection bias'. This is the bias that can occur when patients 'select' which experimental group they will be a member of. In observational studies this 'selection' is typically anything but explicit. Observational cohort and case-control studies observe patients as

they go about their lives, exposing themselves as they do to certain treatments and risk factors, sometimes for identifiable reasons, but often as much due to circumstance. It is this form of selection bias that can be much more difficult to identify and remove from observational studies. For observational studies, taking the allocation of experimental groups out of the hands of the investigators is not possible.

Hence, the avoidance of certain types of selection bias (investigator-, and patient-selection bias) is an important argument for randomised trials in the clinical sciences. Or, more correctly, an important argument for prospective *interventional* studies. It is not randomisation *per se* that permits the avoidance of selection bias, but that these studies are prospective and interventional. Clearly, there are corollaries to Worrall's argument: *randomisation* is not essential—other methods could be used to take the allocation of subjects out of the control of investigators. But the advantage of interventional studies—that allocation of participants *can* be taken out of the hands of investigators—is important, and it is missed on Worrall's analysis due to his narrow conception of selection bias.

A second complicating factor is the ambiguity which surrounds precisely which methodological characteristic is emphasised in discussions on 'RCTs'; randomisation itself, or randomised interventional studies as a study design. Indeed, the important distinction in EBM's methodological hierarchy is not randomised versus non-randomised, but randomised interventional versus non-interventional. Whether randomisation, or an alternative method for reducing selection bias, be used in prospective interventional studies is certainly worthy of debate. But this debate is not central to the questions raised by EBM and its methodological hierarchy once the interventional/non-interventional distinction is made clear.

Prospective interventional randomised studies have a benefit over observational studies in that, when properly done, they rule out a certain type of selection bias (patient-selection bias, as I have labelled it here). Observational studies cannot rule out this bias. Of course, this benefit does not make randomised interventional studies infallible, nor does it mean that these studies are 'essential for any truly scientific conclusion to be drawn from trial data'. Both randomised interventional studies and observational studies employ a vast range of methods to provide results that are as reliable as possible. There are many potential biases, including other forms of selection bias, that can occur in randomised interventional studies. And there is also a host of

methods to assist isolating, and reducing the influence of selection bias in observational studies—including assuring that the control and treatment group are well matched according to background knowledge.

Worrall is right to call proponents of EBM on their over-ambitious claims with regard to randomisation. But it is entirely consistent, while rejecting the over-ambitious claims of EBM, to support arguments for randomised interventional studies that appeal to somewhat more modest benefits. This opens the way for a justification of randomised interventional studies, and EBM's hierarchy, that is both present in the clinical literature, and defensible.

3.2.4 All other things being equal, randomised interventional studies have higher internal validity compared to alternative methods

The reduction of selection bias that is achieved through randomised interventional studies is the key argument found in the clinical literature for marking an epistemological distinction between randomised interventional studies and observational studies. Any difference between the groups under comparison, other than treatment, that can influence the outcome of the study is a potential source of selection bias. Compared with randomised interventional studies, observational studies are more prone to selection bias—even when the experimental groups of a cohort study are well matched according to background knowledge, and *post hoc* adjustment has been conducted. Collins and MacMahon (2007, p. 16) provide a summary of the argument.

> As discussed, randomisation minimises systematic errors (i.e. biases) in the estimates of treatment effects, allowing any moderate effects that exist to be detected unbiasedly in studies of appropriately large size. By contrast, observational studies—such as cohort studies and case-control studies—involve comparisons of outcome among patients who have been exposed to the treatment of interest, typically as part of their medical care, with outcome among others who were not exposed (or comparisons between those with different amounts of exposure). The reasons why certain patients received a particular treatment while others did not are often difficult to account for fully, and, largely as a consequence, observational studies are more prone to bias than are

randomised trials.

This justification for EBM's distinction between randomised interventional studies and observational studies is seen *repeatedly* in the clinical literature and epidemiological textbooks. It is important to be clear as to what this is an argument for. The argument is one of comparative internal validity: all other things being equal, compared to alternative methods, randomised interventional studies have higher internal validity. What kinds of claims does this argument substantiate for EBM?

Internal validity asks the question: How likely are the results of the study to be true for the participants involved? By contrast, external validity refers to how well the findings of a study can be generalised to hold in patients not directly involved in the study. In clinical medicine, there is a focus on informing therapeutic decisions, and therefore an important balance needs to be achieved between internal and external validity. Clearly, it is important for the results of any study to be an accurate reflection of what has occurred in the sample population, but internal validity is not sufficient for reliably informing therapeutic decisions. Therapeutic decisions need not only the results of clinical studies to be accurate for the sample population, but also for the patients who will be treated with the intervention. (At the very least, a principled way of justifying how the results of a trial apply to patients presenting at the clinic is required.)

The useful distinction of efficacy versus effectiveness is also relevant here. Recall, 'efficacy' refers to the effects of the drug under experimental conditions. 'Effectiveness', on the other hand, refers to the effects of the drug in typical patients under routine care. Prior to the drug reaching the market you want to ensure the drug is efficacious. Once on the market it is the effectiveness of the drug in the patients who will be treated that is paramount.

Improvements in internal validity are achieved through two mechanisms; (i) by placing the participants of the study under 'experimental' conditions, to reduce, as much as possible, some of the many things that could influence the participants' progress other than the treatment under investigation, and (ii) by excluding participants who will complicate the analysis, that is, by ensuring the experiment is conducted on a relatively homogenous group of patients. Both of these mechanisms for improving internal validity, however, reduce external validity. Every imposed experimental condition removes the participant from their normal environment, making it difficult to infer the effect of the treatment in 'routine practice'. And, because once the drug is

on the market, clinicians will want to treat the full range of patients who suffer the condition, the narrow inclusion criteria of many trials raises the difficult question of whether patients excluded from participating in the trial will respond in the same manner as those included. How to best ensure the external validity of clinical research is recognised, by proponents and critics alike, as one of the most important issues for EBM to address (Black 1996; Rothwell 2005a; Upshur 2005).

Concerns about internal validity are concerns about trial methodology. To say a trial has high internal validity is to say that the trial employed methods to prevent the kind of errors known to occur when observing data on the effect of a therapy in a population. Randomising experimental groups in a prospective interventional study is one such method. So too is conducting an interventional study rather than an observational study when wanting to establish the efficacy of a treatment. EBM's hierarchy organises the study designs commonly used in clinical research according to internal validity.

Many methodological techniques are important to the internal validity of a clinical trial. While many of these methods are not explicitly included in the presentation of the hierarchy, it is clear they are very important to proponents of EBM. Examples include: adequate concealment of patient allocation to experimental groups (that is, maintenance of double blinding); proper treatment and analysis of 'drop outs' (trial participants who leave the study); the proper use and interpretation of statistical tests; and many many more. A cursory glance at any clinical epidemiological textbook, including the one written by the leading proponents of EBM, Haynes et al. (2006), is enough to confirm this; they are full of such methodological concerns.

Comparative internal validity provides a justification for EBM's entire hierarchy, not just the distinction between randomised interventional studies and observational studies. For instance, observational cohort studies are placed higher than case-control studies on the hierarchy provided by Phillips et al. (2001), because cohort studies have higher internal validity. Case-control studies rely on investigators to define a control group that has not suffered the outcome under investigation. In order to ensure the control group is appropriately comparable to the 'cases' a number of assumptions about exposure to the risk factor under investigation are required. Prospective cohort studies, by contrast, can follow a more natural group of patients, some exposed to the risk factor under investigation, others not exposed. Investigators do not have to construct a control group as they do in case-

control studies.[10] This means there is more opportunity for error in case-control studies, and hence, these studies possess lower internal validity.

There are a number of important points to make regarding comparative internal validity as an argument for EBM's hierarchy of evidence. First, the judgements of internal validity incorporated into EBM's hierarchy are specific to the clinical sciences. The methods to improve internal validity are canonical. Mostly, they have been collected through the experience of observing and testing the effects of drugs. Each method attempts to rule out, or minimise, a particular kind of error. Randomisation, blinding, prospective trials, intention-to-treat analyses, and so on, are all methods for preventing specific erroneous inferences in therapeutic trials. While these methods have been built up through experience, this does not mean that a philosophical account referring to a logic of evidence can not be provided. But any such account will be specific to the clinical sciences.

While there is no one factor that makes the clinical sciences unique, the particular confluence of factors that make up the clinical sciences *is* unique. For starters any account needs to recognise the high degree of unexplained inter-patient variability in response to therapy; theory does not adequately predict response in real-life patients. (Arguably, this is what makes statistical approaches so important—and controversial—in the clinical sciences.) In addition, a range of practical considerations of particular importance to the clinical sciences need to be recognised. For instance, the ability to conduct randomised trials (something impossible in many contexts), and the importance of health, and an appropriate conservatism toward risk when dealing with matters of health. Each of these factors (and many more besides) impinge not only on the kinds of error that can occur, but also the kinds of error that matter, and thus on the kind of methods that have been developed to avoid these errors. Identifying EBM's hierarchy with the internal validity of therapeutic trials not only provides a justification for the hierarchy, but also emphasises the problems of extending the hierarchy beyond this context.

Once outside of the context of a large drug trial other sources of error may become more important. Or, the method developed to reduce the error in the drug trial may no longer work. Randomisation, as Grossman and Mackenzie (2005, pp. 527–528) show, is a case in point. Simply changing the context

[10]There are variants of the case-control design that overcome this problem, for instance: nested case-control studies. Nested case-control studies use cohort studies to define the control group.

of an interventional study from testing a drug to testing, say, a social intervention in schools, can be enough to change randomisation from increasing internal validity to decreasing it. Large drug trials have many participants, and hence many 'units of analysis' that can be randomised. In such a situation, randomisation is a convenient way to both take treatment allocation out of the hands of investigators, and ensure the experimental groups are roughly equally balanced for a small number of known confounding factors. A trial of a social intervention in ten schools, with five schools receiving the intervention and five receiving control, might have as many pupils involved as the drug trial has participants, but because the intervention is school-wide it has far fewer units of analysis. With such a small effective sample size randomisation is much less likely to ensure the experimental groups are roughly equally balanced compared to if the investigators deliberately balanced the groups according to which factors are considered important. Providing investigators can justify their allocation, and ensure selection bias has been minimised, deliberate matching will result in higher internal validity in the test of a social intervention in schools than randomisation. Similar accounts can be given for the other methods of improving internal validity in EBM's hierarchy of evidence; care needs to be taken to ensure the methods that are being employed are pertinent to the case at hand.

It is also essential to be clear what the 'all other things being equal' part of this argument means for EBM's hierarchy. The guidebooks propose a categorical interpretation of the hierarchy. Viewing the hierarchy as a hierarchy of internal validity, however, requires considerably more nuanced judgements. The hierarchy only provides increasing internal validity when *all other things are equal.* Study designs higher up the hierarchy rule out more possible sources of error. But at each level of the hierarchy there are many sources of error, and many methods that can, and should, be applied to ensure the results of the study are as reliable as possible. It is only when all the additional measures that improve internal validity have been taken that there is an assurance that a randomised interventional study possesses higher internal validity than a prospective cohort study. As should be obvious, there is no assurance that a randomised interventional study that has ruled out patient-selection bias, but left another source of error unchecked, will be any more reliable than a carefully conducted cohort study, which has employed every method possible to ensure its results are valid. The quality of evidence that a study provides always requires a judgement of whether

all reasonable sources of error have either been ruled out or accounted for. Study designs higher up EBM's hierarchy are able to employ more methods to reduce potential sources of error, but this by no means ensures the quality of evidence that is provided by any particular study using a design listed high in the hierarchy is better than a study employing a design listed lower in the hierarchy.

Further, the increased potential for high internal validity in randomised interventional studies does not ensure *all* the results of a well-conducted randomised trial are reliable. Indeed, as discussed in the previous chapter, the statistical techniques employed in analysing randomised interventional studies provide optimal warrant only to the primary hypothesis under test. Subgroup analyses and findings on secondary endpoints are less reliable—at least to the extent that these outcomes are underpowered. This is important to keep in mind when considering the primary hypotheses that most clinical trials are set up to test. As noted previously, primary hypothesis tests in clinical trials almost invariably relate to the *benefits* of a therapy; questions related to a drugs safety are typically relegated to secondary endpoints. While there is no theoretical reason why randomised trials can't be set up to test the safety of a drug as a primary hypothesis, the practical constraints are considerable. First, there is the ethical question of whether it is appropriate to conduct a trial when the primary purpose of the trial is to detect adverse effects. Second, even if such trials are considered ethical, the problem of whether patients would consent to be included in such a trial remains. And third, even if these issues can be overcome, there is a certain degree of inertia in the system. The medical fraternity (as much as it acts as a single unit), the regulatory authorities who are responsible for outlining what research needs to be completed before a therapy will be permitted onto the market, and the pharmaceutical industry which funds the vast majority of large drug trials, currently hold that trials that test primary 'benefit' hypotheses are sufficient to 'prove' the drug is ready for the market. The proportion of trials that incorporate safety endpoints into the primary hypothesis under test is unlikely to increase while this view is pervasive.

Of course, concern about the reliability of analyses of secondary endpoints and subgroups is not new. Meta-analysis is the standard reply to the problem of providing reliable estimates for secondary endpoints and subgroup analyses. When a number of relevantly similar randomised interventional studies have been conducted they can be combined in order to conduct a meta-

analysis. Such meta-analyses will provide more reliable results for these secondary outcomes. EBM recognises this by placing meta-analyses and other systematic reviews of a number of randomised interventional trials at the pinnacle of the hierarchy. But this option is only available when multiple trials have been conducted, and will only improve the reliability of the estimates if the trials are similar enough in the relevant respects; needless to say this is not always the case.[11]

These considerations substantially restrict the claims that can be made by EBM's hierarchy on the basis of considerations of internal validity. The importance of restricting the claims of EBM to therapeutic questions and the results of primary hypothesis tests and meta-analyses is emphasised. This avoids much of the rhetoric sometimes employed by proponents of EBM. On this view EBM's hierarchy of evidence does not provide general epistemological rules, nor are randomised interventional studies infallible, nor do randomised studies carry *very special* scientific weight when compared to observational studies. The comparative benefits of randomised interventional studies can be put simply: randomised studies have the capacity to employ more methods that improve internal validity than observational studies. The quality of evidence that any particular study provides is a separate judgement taking much more into consideration.

3.3 Evidence based medicine must be

I have argued that EBM's hierarchy of evidence is best interpreted as a hierarchy of comparative internal validity. Furthermore, I have shown that identifying EBM's hierarchy with comparative increases in internal validity emphasises that EBM's claims should be limited to the results of primary hypothesis tests and meta-analyses. In the introduction, I suggested that any justification of EBM's hierarchy of evidence needs to provide (i) a clear interpretation of the hierarchy (that is, how it should be applied), and (ii) a justification for why applying the hierarchy in this way more reliably informs therapeutic decisions. I now examine how well identifying EBM's hierarchy with comparative increases in internal validity fulfils these requirements.

Significantly, the need to limit EBM's claims to the results of primary

[11] For some discussion on when meta-analyses are not ideal, see Egger and Smith (1995); Egger et al. (1997) and Smith and Egger (1998).

hypothesis tests and meta-analyses is just as important to the interpretation of EBM's hierarchy as a hierarchy of comparative internal validity as it was in the categorical interpretation of the hierarchy. Indeed, since this problem results from a combination of the kind of clinical trials that are typically conducted, and the use of frequentist statistical analyses, this problem will arise on any interpretation of EBM's hierarchy—providing the aim of EBM remains informing therapeutic decisions. Much medical research regarding drug therapies is focussed on establishing the efficacy of the treatment, and it is on questions of efficacy that statistical methods are also focussed. Questions of effectiveness requires an extension of these methods; the later chapters of the thesis focus on the challenge of extending the frequentist statistical methods in this way.

Applying the hierarchy viewed as a hierarchy of internal validity is considerably more complicated than the simple advice provided by the EBM guidebooks. When interpreting EBM's hierarchy as a hierarchy of internal validity the type of question under consideration is vital. Say, *Drug X* has passed the early stages of clinical testing. The drug appears safe, and has a promising pharmacological profile. If the question is whether *Drug X* is *efficacious* in a defined population of patients, then EBM's hierarchy provides clear advice. The most accurate method to measure the efficacy of *Drug X*, all other things being equal, is via a well conducted randomised interventional study. While these methods are fallible, they have the capacity to rule out more potential sources of error than methods lower down EBM's hierarchy.

Things become more complicated when the question changes to whether *Drug X* possesses a particular side effect. Note, this question just asks whether *Drug X* possesses the side effect, not whether any particular individual will suffer that side effect; that is, what is being considered is the side effect equivalent of 'efficacy'. It is still possible to suggest the 'best' methodology to address this question is a well conducted randomised interventional study, but for this to be so, some hefty assumptions are required. First, the side effect has to be suspected in order to set up a trial to test the hypothesis that the side effect exists. (Clearly, this is not always so in practice). And second, you have to be able to conduct the randomised interventional study. While there is no theoretical impediment to this, the practical constraints discussed in the previous section mean that such a trial is less likely to be conducted.

Relying on a randomised interventional study set up to test a primary 'benefit' hypothesis to detect adverse effects creates a number of problems. Not only because safety outcomes are relegated to secondary endpoints, but also because a trial set up to test a benefit hypothesis will select patients who have less complex medical histories than the population of patients with will eventually receive the drug. While selecting patients with only the condition under investigation makes sense when testing a drug's efficacy, this distorts both the detection and estimation of side effects more likely to occur in patients with multiple pathology. Randomised interventional studies are the 'best' way of confirming the possible side effect of *Drug X* in only the most theoretical sense. The randomised interventional studies typically conducted in practice will *not* provide the most reliable evidence relating to the drug's side effects.[12] Judging whether the side effect of *Drug X* exists will instead rely on: the serendipitous findings of randomised interventional studies set up to test benefit hypotheses; the accumulation of evidence from randomised interventional studies in meta-analyses; or, evidence from lower down EBM's hierarchy—or, if available, on a combination of these sources of evidence.

When the standpoint of the question is changed to that of a clinician regarding a marketed drug, the same questions can be addressed. But here the limitations on the claims substantiated by EBM's hierarchy are even more important. The clinician's most pressing question is whether the drug's benefits will outweigh its potential for harm in the patient presenting to the clinic. Interpreting EBM's hierarchy as one of comparative internal validity makes it explicit that reference to the hierarchy provides only *partial* assistance in answering this question. All things being equal, estimations of the drug's efficacy are more likely to be accurate in well conducted randomised interventional studies. Questions regarding the drug's potential for harm are more difficult. Because safety endpoints are likely to be secondary endpoints (and under-powered), the information that typical randomised interventional studies provide can be less reliable. Evidence from observational studies may be more helpful in this case, but here a judgement on the potential for any selection bias needs to be made. Crucially, EBM's hierarchy (interpreted as a hierarchy of increasing internal validity) provides some assistance, but it

[12]This assumes the incidence of the side effect is lower than the degree of benefit afforded by the drug—as one would hope. If this is not so, and the ill effects of the treatment are as large as the benefits, then the randomised interventional studies conducted to confirm the benefits will be large enough to reliably detect the harms.

does not answer the *effectiveness* question asked by the clinician. The clinical question raises questions of external validity, and they are additional to the questions addressed by EBM's hierarchy.

3.4 Conclusion

EBM's hierarchy is best understood as a hierarchy of comparative internal validity. The constraints this justification places on the EBM's hierarchy are too often under-appreciated (or at least under-emphasised) in the clinical literature discussing EBM.

First, the increased validity is only substantiated in a small subset of highly specific questions; its strongest claims are made for questions relating to a drug's efficacy. Even extending the question to the same drug's side effect profile raises considerable challenges.

Second, the justification is comparative, not absolute. In the situations in which it is obtained the greater opportunity for internal validity does not ensure infallibility. And the incremental benefits of being able to employ additional methods for ruling out certain types of error will only be realised if other appropriate methods to rule out or reduce sources of error are utilised. Far from placing very special weight in randomised trials, this justification emphasises the need for careful judgement on all sources of error as well as on the methods utilised to reduce them.

Third, this justification makes clear that high internal validity is not sufficient for reliably informing therapeutic decisions. Interpreted as a hierarchy of comparative internal validity, EBM's hierarchy of evidence provides a framework for developing arguments about evidence and the application of evidence to therapeutic decisions. The details of the question and the case at hand always matter. Interpreting EBM's hierarchy as one of comparative internal validity makes the challenge of external validity explicit. Being clear as to what questions are well answered by EBM's hierarchy assist framing the additional questions that need to be addressed to inform the therapeutic decision.

Chapter 4

Why Randomised Interventional Studies

4.1 Introduction

The previous two chapters focussed on the interpretation and justification of EBM's hierarchy of evidence. Specifically, I argued for an interpretation of EBM's hierarchy as a hierarchy of comparative interval validity. In addition to being a defensible interpretation of EBM's hierarchy, this interpretation has the benefit of highlighting the *indefensibility* of some of the claims that have been made by proponents of EBM on behalf of the methodological distinctions made in the hierarchy. Rather than reject EBM on the basis of some of the more ambitious claims that have been made, interpreting the hierarchy as a hierarchy of comparative internal validity will help rein EBM in. Randomised interventional studies have the capacity to rule out a potential source of selection bias that can not be avoided in observational studies. This chapter examines the argument for randomised interventional studies in more detail. Specifically, I'll argue for randomised interventional studies over observational studies in testing the efficacy of drug treatments.

Worrall (2002, 2007a,b) argues that there is no *unique* justification for randomisation in experiments. In this argument he follows the work of Peter Urbach (1985; 1993) and Colin Howson and Urbach (2006) in suggesting that randomisation is not essential, or *sine qua non*, in science. On this point, Worrall and I are in complete agreement. Worrall, however, goes one step further. He suggests that experimental designs that randomly allocate

subjects to treatment and control provide *no* epistemological good in addition to that which can be achieved through alternative means. On Worrall's view, if experimental groups are comparable according to background knowledge, then whether the groups were formed by random allocation, or deliberate matching, is of no epistemological import. Randomised trials can be done away with. Clearly, this view directly opposes current medical orthodoxy, especially EBM.

If Worrall is correct, how medical research is conducted can be revolutionised. Currently, prior to permitting a new drug on the market, large scale randomised interventional studies are required to show the therapy is efficacious. Considerable time and expense could be saved if equally reliable evidence regarding the efficacy of therapies could be gained from observational or historically controlled studies.[1] Some study designs, especially case-control and historically controlled studies, can be conducted much more quickly than randomised interventional studies. And, while prospective observational cohort studies are conducted on similar timeline, recruitment is much easier in cohort studies compared to interventional studies.

But Worrall is not correct; the evidence provided by an observational study (or a historically controlled trial) regarding a treatment's efficacy is not equivalent to the evidence provided by a well-conducted randomised interventional study, even when each of the studies involve experimental groups that are comparable according to background knowledge. There are good reasons for wanting to establish claims of a treatment's efficacy in a randomised interventional study rather than an observational study. In Section 4.2, I outline two claims that Worrall makes. The first is that randomised interventional studies are not essential for science (a claim we agree on), and the second is that there is no epistemological distinction between randomised interventional studies and observational studies providing the experimental groups are comparable based on background knowledge (a claim I wish to contend). Then, in Section 4.3, I provide a positive argument for testing claims of efficacy in randomised interventional studies rather than observa-

[1]In this context historically controlled trials might be considered quasi-interventional. As discussed in section 1.1, page 18, the main reason historically controlled trials are infrequently conducted in contemporary medicine is because they are seen to be epistemologically inferior to randomised interventional studies. And due to this epistemological inferiority historically controlled trials are considered unethical for testing new, or 'experimental' treatments. If Worrall's epistemological are arguments successful, then the ethics of these trials needs to be reconsidered.

tional studies.

4.2 Randomised interventional studies are not essential for drawing scientific conclusions in medicine

The claim that *only* randomised studies can provide the right kind of evidence for medicine is easy to find in the literature. In contrast to randomised trials, observational studies are seen as inherently inferior. (For an example of such claims, recall the quote from Collins and MacMahon 2007, provided in section 2.4.2, page 51). In line with these claims, medical orthodoxy appears to believe randomised studies are 'essential', or *sine qua non*, for gaining evidence about therapies. As Worrall notes,

> It is widely believed that RCTs [randomised controlled trials] carry special scientific weight—often indeed that they are *essential* for any truly scientific conclusion to be drawn from trial data about the effectiveness or otherwise of proposed new therapies or treatments. This is especially true in the case of clinical trials: the medical profession is overwhelmingly convinced that RCTs represent the 'gold standard' by providing the only 'valid', unalloyed, genuinely scientific evidence about the effectiveness of any therapy. Clinical science may occasionally have to rest content (perhaps for ethical or practical reasons) with evidence from other types of trial ... but this is always very much (at best) a case of epistemic second best. (Worrall 2007b, p. 452, emphasis in the original)

This claim—that randomised interventional studies are essential for drawing scientific conclusions in medicine—is Worrall's primary target. Worrall (2002, 2007a,b) considers the prominent arguments that have been provided for this claim and finds them wanting.

What does it mean for randomised interventional studies to be 'essential' for science or medicine? Worrall's argument follows, and extends upon, arguments provided by Peter Urbach (see Urbach 1985, 1993 and Howson and Urbach 2006). Urbach is clear as to what kind of justification is needed for randomisation to count as an essential.

> The leading justifications [for randomisation and control] have been epistemic in character; that is, they argue that control and randomisation are required by a particular logic of inference. So these justifications, if successful, and if based on a correct logic of inference, would show that the two conditions are formally required by any clinical trial, not just as a matter of convenience, or under certain circumstances, but as *sine qua non*. (Urbach 1993, p. 1421)

If randomisation is to be essential, it is to be *formally* required, for *all* clinical trials, in *every* context. Urbach sets the epistemological bar pretty high for randomisation to be considered essential.

Worrall (2007a, p. 983) acknowledges that even the staunchest proponents of EBM, when pushed, back away from endorsing the claim that randomised interventional studies are essential for science. For instance, Sackett et al. (1996) claim that EBM embraces a broad notion of evidence in medicine.

> By best available evidence we mean clinically relevant research, often from the basic sciences of medicine, but especially from patient centred clinical research into the accuracy and precision of diagnostic tests (including the clinical examination), the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventative regimens. (Sackett et al. 1996)

Yet, despite comments such as these, there is no question that proponents of EBM, particularly when it comes to claims regarding therapies, give evidence from randomised interventional studies 'very special weight' (to use Worrall's 2007a, p. 983 phrase). For instance, in the same paper cited above, Sackett et al. (1996) clarify

> It is when asking questions about therapy that we should try to avoid the non-experimental approaches, since these routinely lead to false positive conclusions about efficacy. Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the 'gold standard' for judging whether a treatment does more good than harm.

Neither Worrall, nor (more pertinently) proponents of EBM, are clear as to what giving 'very special weight' to evidence from randomised interventional studies amounts to. Worrall, for his part, analyses a number arguments

for conducting randomised interventional studies and finds that none of these arguments support the stronger assertion—that randomised interventional studies are *essential* for drawing scientific conclusions from data. Presumably, Worrall takes this to establish that the 'very special weight' placed in randomised interventional studies by EBM can't amount to much. The most pertinent of Worrall's arguments against randomised interventional studies were discussed in Chapter 3, and as far as showing that randomised interventional studies are not essential for drawing scientific conclusions from data, Worrall's arguments are correct. Here, focus is given to a claim he makes while establishing this argument.

On Worrall's view, not only are randomised interventional studies not essential for drawing scientific conclusions from data, but they provide no epistemological benefit over observational studies, providing experimental groups are well-matched according to background knowledge.

> The best we can do (as ever) is test our theories against rivals that seem plausible in the light of background knowledge. Once we have eliminated other explanations that we know are possible (by suitable deliberate, or *post hoc*, control) we have done as much as we can epistemologically. (Worrall 2007b, p. 486)

According to Worrall, providing the groups under comparison are suitably controlled according to background knowledge, non-randomised non-interventional (that is, observational) studies are epistemologically equivalent to randomised studies. While the claim that randomised interventional studies are essential for science (or medicine) is so strong that it is no surprise that it is false, Worrall's counter-claim that there is *no* epistemological distinction between randomised interventional studies and observational studies (when groups appear matched according to background knowledge) is also quite strong, just in the opposing direction. I provide a positive argument for marking an epistemological distinction between randomised interventional studies and observational studies in medicine. Components of the argument can be found in different parts of the clinical, epidemiological and philosophical literature. To my knowledge it has not been provided in the one place with a focus on the claims of EBM.

It is important to be clear on the study designs under comparison. EBM, and medicine more generally, is focussed on the distinction between randomised interventional studies and observational studies, as opposed to the

distinction between randomised and non-randomised interventional studies. Recall, studies are 'observational' when it is the participant, in consultation with their regular health carers, who select which treatments they take—no intervention by study investigators is made on the participant's treatment. Observational study designs rely on deliberate matching, or *post hoc* adjustment, to compare the groups under investigation. Historically controlled studies, which compare patients taking a 'new' treatment with a historical cohort of patients who were treated with the conventional treatment, rely on similar matching and *post hoc* adjustment to analyse clinical data.[2] Interventional studies, by contrast, *allocate* patients to active treatment or control. In randomised interventional studies, allocation is achieved using a random process. Providing the study is large relative to the number of potential confounders, randomised allocation ensures that the experimental groups are roughly equally matched for characteristics that may influence a patient's response to treatment. To the extent this rough matching for confounders is achieved, deliberate matching and *post hoc* adjustment is not required.

In non-randomised interventional studies some other method is used to allocate patients to treatment or control. Perhaps experimental groups are deliberately matched by individuals blinded to which group will eventually receive the treatment—this may allay concerns about allocation bias in the groups at baseline. Allocation bias is another term for 'investigator-selection bias', and refers to investigators influencing the make-up of the experimental groups by preferentially selecting patients for one group over another due to patient characteristics (a process which may be subconscious). Observational studies do not suffer from allocation bias (because they do not 'allocate' patients), but they do suffer from selection biases more generally. As seen in the previous chapter, selection bias, which is any bias that comes about due to differences between the experimental groups other than the treatment under investigation, is a bigger problem for observational studies because it can be extremely difficult to identify, isolate and adjust for the multitude of factors that may have influenced participants to take, or not take, the

[2]I will focus on the distinction between randomised interventional studies versus *observational* studies, rather than historically controlled trials. Observational studies, such as cohort and case-control studies are the studies more frequently used in contemporary medical research. However, the argument I provide for an epistemological distinction between randomised interventional studies and prospective cohort studies is just as relevant if historically controlled cohort studies are used for comparison.

treatment under investigation.

In some discussions of the role of randomisation in the literature it can be unclear whether the distinction of interest is randomised interventional studies versus observational studies, or randomised interventional studies versus non-randomised interventional studies. The distinction is important for a number of reasons. First, proponents of EBM are predominately concerned with the first distinction (though, as I have noted a number of times, this is sometimes muddled by the terminology employed; proponents of EBM typically refer to randomised interventional studies as randomised controlled trials (RCTs) and to observational studies as 'non-randomised studies'). Second, the distinction between interventional study designs and observational study designs raises more questions of practical and ethical importance than the distinction between randomised and non-randomised interventional studies. Interventional studies are considerably harder to run, and cost more than observational studies. Further, the ethical questions raised by interventional studies are more troubling than the questions raised by observational studies. In interventional studies participants must consent to be *allocated* treatment or control, whereas in observational studies it is the participants (usually in conjunction with their health carer) who decide whether or not to take a treatment. Thus, unless there is an epistemological benefit in conducting randomised interventional studies, we might as well conduct observational studies.

By contrast, there is no practical imperative in favour of non-randomised interventional studies over randomised interventional studies. Indeed, as the Bayesian argument for randomisation provided in the next section makes clear, in the large drug trials of interest, non-randomised interventional studies are typically slightly *harder* than randomised interventional studies to conduct. In such trials, deliberately matching patients will require more work than random allocation. In addition to being easier, randomised interventional studies have a number of epistemological advantages over non-randomised interventional studies in testing drug treatments.

4.3 Randomised interventional study designs are epistemologically superior to observational study designs

Worrall is right to criticise the proponents of EBM for claiming too much on behalf of randomisation. Proponents are wrong to claim (or imply) that randomisation *guarantees* the results of a clinical trial. When judging evidence for therapeutic decisions, the design of the study providing the evidence is but one of many important considerations. EBM's claims, in as much as they overshadow these considerations, are far too broad. Rather than all of science, or all of medicine, well-conducted randomised interventional studies are epistemologically superior to well-conducted observational studies in a much narrower domain. My aim is to show that, *ceterus paribus*, randomised interventional studies provide superior evidence in testing the efficacy of drugs.

Recall, *efficacy* refers to whether the drug works as expected under experimental conditions. Most clinical trials are set up to test the question of whether the drug is efficacious. Questions of efficacy can be made precise. Does the drug benefit the selected patients in the primary clinical outcome? The quality of the evidence that the trial provides depends on whether the trial employed appropriate methods to rule out, or minimise, as many sources of systematic error (bias) as possible.

The argument for preferring randomised interventional studies in tests of a treatment's efficacy proceeds in two parts. Chapter 3 showed that interventional studies have the capacity to rule out a form of bias that observational studies are particularly prone to. Here, I'll focus on what it is about testing the efficacy of drugs that makes study designs that have the capacity to rule out this bias important. Knowledge of causal processes in clinical science is highly uncertain. It is because of this uncertainty that the deliberate matching and *post hoc* adjustment employed in observational studies can be problematic. In addition, the argument for randomised interventional studies as it is presented in the epidemiological literature will be illustrated. In the epidemiological literature the argument for randomised interventional studies equivocates on the sense of 'bias' that is eliminated or reduced by this study design. I'll attempt to disambiguate this argument. What results is an argument for *interventional* studies over observational studies in tests of efficacy. In the subsequent section, I'll provide a Bayesian rationale for

randomisation in interventional studies.

4.3.1 Why *interventional* studies

Causal knowledge in clinical medicine is provided by the basic medical sciences, such as physiology and pharmacology. Despite the expanding knowledge in physiology and pharmacology, and the empirical knowledge provided by the extraordinary number of randomised interventional studies conducted each year, when it comes to the clinical sciences—inferences regarding how a patient or group of patients will respond to a therapy—there is often more unknown, or uncertain, than known.

Basic and clinical sciences have a very different focus. Pharmacology and pathophysiology are extensions of biology; focus is given to the biological effects of drugs and diseases respectively. To gain an understanding of these biological effects, systems and mechanisms *within* the body are isolated. The individual, as a single complex whole, is the first thing that is abstracted away. This abstraction is necessary in order to develop, test, and improve the causal explanations provided by these basic sciences. Such causal explanations are vital for the clinical sciences, but focus is given to a entirely different question.

Clinical sciences work at the level of individuals (or groups of individuals with similar characteristics), and try to predict, or understand, what effect treatment will elicit. Two difficulties present themselves at this level. Sometimes there is a lack of sufficient biological knowledge to be able to predict the outcome in an individual. Other times there is a wealth of biological knowledge, but insufficient detail to differentiate how the systems and mechanisms at play will interrelate in a given individual; that is, despite a comprehensive understanding of the underlying causal processes, it is impossible to predict what the overall effect of treatment will be. It is at this level that the empirical information provided by randomised interventional studies is invaluable, and superior to that provided by observational studies.

The set-up and analysis of observational studies requires investigators to estimate the effects of causal processes that influence the effects of a treatment in a patient—more so than is necessary in randomised interventional studies. Observational studies follow patients undergoing routine care. It is the patients, and their regular health carers, who have made the choice whether or not to take the treatment under investigation, often sometime

prior to the study. There are many possible factors that may influence whether or not patients choose to take a particular treatment, and any one (or a combination) of these may also influence the patient's outcomes. Due to the uncertainty in causal knowledge in clinical science, it is usually impossible to isolate which of the many *possible* causal factors differentially distributed in the experimental groups *may* influence the observed effects of treatment. The investigators of observational studies, by deliberate matching or *post hoc* adjustment, attempt to minimise the effects of such possible confounding factors. This requires estimation of the effects of different patient characteristics on their response to treatment, which in turn requires an extensive understanding of the causal processes at play. The correct analysis of observational studies is dependent on the veracity of these assumptions.

Interventional studies that employ methods to avoid allocation bias don't have this problem. There is no need to identify causal factors that may influence the outcomes under investigation from the myriad of causes that may have led a patient to be on, or not on, the treatment under investigation. In interventional studies, participants are *allocated* treatment or control. The analysis of interventional studies does not rely on assumptions made about the possible influences of factors that have led participants to choose whether or not to take the therapy.

Of course, in interventional studies many other assumptions and methods are needed to ensure the observed results are reliable. Bias, including other forms of selection bias, can occur in interventional studies. There can be important differences in experimental groups at baseline in randomised interventional studies. While the random allocation can be checked to ensure that known confounders are equally distributed, no checks are possible for unknown confounders. More on this in a moment. And even if the allocation appears appropriate at baseline, the possibility of differences arising after the allocation still needs to be minimised. This is achieved by maintaining allocation concealment (from participants and investigators), and by ensuring that the analysis does not rely on any post-allocation characteristics. Considerations such as these emphasise the care needed in conducting and interpreting any type of clinical reasearch. But, importantly, none of this undermines the benefit that interventional studies possess over observational studies. One specific source of bias is ruled out; bias due to patients, and their regular health carer, *choosing* whether or not to take the treatment under investigation. And because this source of bias is eliminated, well-conducted

randomised interventional studies, when compared to well-conducted observational studies, do not rely to the same extent on assumptions about the causal processes at play.

The importance of the decisions made on the basis of tests of efficacy also plays a role in the argument for interventional studies. Tests of efficacy are used by regulatory agencies to decide whether a new drug should be put on the market. In this situation there is typically not a rich understanding of the causal processes at play—at least not enough of an understanding to be able to isolate and adjust for all of the causal factors that could influence the effects of a therapy (and be differentially distributed in groups of patients who would select to take the therapy versus those patients who would elect not to). Given that results of tests of efficacy are used to decide an issue of such importance to public health, it is important that the test is rigourous. There is good reason to want to rule out or minimise as many potential sources of error as possible, and conducting interventional studies rather than observational studies as tests of efficacy help to achieve this rigour.

This is the argument for interventional studies in tests of a treatment's efficacy. While a form of this argument is provided in the epidemiological literature, the argument is often less than clear due to ambiguity in use of the term 'bias' (the ambiguity noted in the previous chapter, see 64). A good illustration is provided by Collins and MacMahon (2007, p. 16).[3] This particular quote was provided in the previous chapter; by returning to it I hope to illustrate the ambiguity often present in the epidemiological argument for randomised interventional studies.

> As discussed, randomisation minimises systematic errors (i.e. biases) in the estimates of treatment effects, allowing any moderate effects that exist to be detected unbiasedly in studies of appropriately large size. By contrast, observational studies—such as cohort studies and case-control studies—involve comparisons of outcome among patients who have been exposed to the treatment of interest, typically as part of their medical care, with outcome among others who were not exposed (or comparisons between those with different amounts of exposure). The reasons why certain patients received a particular treatment while others did not are often difficult to account for fully, and, largely as a conse-

[3] A similar argument is also provided in (Yusuf et al. 1984)

> quence, observational studies are more prone to bias than are randomised trials.

Two arguments are muddled together here. The first, relying on the informal notion of bias, is the argument for interventional studies that we have been discussing. The second argument, which relies on statistical bias, is perhaps best seen as an argument for randomisation *in* interventional studies. While the first argument is persuasive, the second is not.

Fletcher et al. (1996, p. 7) adopt the definition of bias as 'a process at any stage of inference tending to produce results that depart systematically from the true values'. This is the sense of bias referred to by Collins and MacMahon in the statement that '[randomised interventional studies] minimise systematic error (i.e. biases) in the estimates of treatment effects'. More specifically, the bias that is minimised in randomised interventional studies, in contrast to observational studies, is the systematic error that can occur when experimental groups differ systematically in some factor, other than treatment, that influences patient outcomes and is differentially distributed in patients who choose to take the treatment and those who choose not to. Observational studies are prone to this particular source of bias because 'the reasons why certain patients received a particular treatment while others did not are often difficult to account for fully'. Note, it is the *possibility* of this form of selection bias that differentiates observational studies from interventional studies—a possibility that cannot be ruled out because of the fallibility of causal knowledge in clinical science.

The second sense of bias is the more technical sense used in parametric statistics. Recall, statistical bias refers to the expectation of a estimator of an unknown parameter. An estimator is unbiased if the mean of the estimator over the entire sample space equals the true value of the unknown parameter.[4] A test-statistic is an estimator of the unknown parameter under investigation in a clinical trial. If the test-statistic provides an unbiased estimator of the unknown parameter, then, were the trial repeated indefinitely, the mean value of the test-statistic would equal the true value of the unknown parameter. Of course, trials are never repeated indefinitely.[5] Nevertheless

[4]Stuart and Ord (1991, p. 609) provide this definition of bias. Stuart and Ord emphasise that this sense of bias 'should not be allowed to convey overtones of a non-technical nature'.

[5]Even if a trial (which was perfectly internally valid) was to be repeated an incredible number of times, the mean of the observed values of the test-statistic would not necessarily equal the true value of the unknown parameter. The assurance of unbiasedness refers to

statistical bias is put to use in the epidemiological literature to argue for randomisation in interventional studies. This argument is hinted at in Collins and MacMahon's statement that randomisation allows 'any moderate effects that exist to be detected unbiasedly in studies of appropriate size'. Note the shift from randomised interventional studies minimising bias in comparison to observational studies, to randomisation allowing the moderate effects of treatment to be detected *unbiasedly*.

This argument for randomisation in interventional studies (as opposed to randomised interventional studies) is often made in the clinical literature. It is one of the arguments Worrall (2007a; 2002; 2007b) shows to be flawed—at least as an argument for randomisation to be *sine qua non*. Devereaux and Yusuf (2003, p. 107) provide a further example of how the argument is stated in the epidemiological literature

> The placement of RCTs at the top of the therapeutic research hierarchy has occurred due to the realisation that RCTs are superior to observational studies in evaluating treatment because RCTs *eliminate* bias in the choice of treatment assignments and provide the only means to control for unknown prognostic factors. [emphasis added]

The argument is sound, but as I discussed in the previous chapter, the soundness of the argument relies on understanding 'bias' in the statistical sense. And once this made clear the assurances that the argument provides the clinician ring hollow.[6]

---

the estimation of the test-statistic over the *entire sample space*. In any long run of trials, even an infinite long run, the estimator will not necessarily equal the true value of the unknown parameter.

[6]Recall, randomisation in interventional studies eliminates statistical bias arising from the differential distribution of confounding factors in experimental groups at baseline. And randomisation achieves this whether or not the confounding factor is suspected. Randomisation in interventional studies achieves this seemingly magical outcome because statistical bias refers to the influence of confounding factors in *expectation*—that is, the expected influence of confounding factors over the entire sample space. On any *particular* randomised allocation of experimental groups there is a possibility that confounding factors will be unevenly distributed between the groups. While the distribution of known confounders can be checked after any single random allocation, this is clearly not possible for unknown confounders. Indeed, as noted by both Worrall (2002, p. S324) and Howson and Urbach (2006, pp. 195–6), the probability that any single confounder is unevenly distributed on any particular random allocation ranges from zero to one. However, were the trial to

Proponents of the argument that randomisation is essential in medicine—because it eliminates statistical bias from uneven distribution in confounding factors at baseline—all too often neglect to explicate what eliminating 'statistical bias' actually amounts to. Often any reference to expectation, or the indefinite repetition of trials, is obscured or absent (as is this case in the quote provided from Devereaux and Yusuf). The argument is made *as if* randomisation ensures the even distribution of known and unknown confounders in a *particular* random allocation. Were randomisation to achieve this, then it may be considered *sine qua non*. But randomisation achieves nothing of the sort.

Randomisation in interventional studies does not ensure the even distribution of all possible confounders at baseline. But, interventional studies do rule out bias (of the first, more informal sense) originating from the differential distribution of confounders linked to whether or not a patient chooses to take a particular therapy. Interventional studies have the capacity to avoid a source of bias that cannot be avoided in observational studies. And this establishes an important epistemological distinction between interventional and observational studies in the context of tests of efficacy. Observational studies can certainly provide important scientific information. And, when we have a rich understanding of the causal processes at play, the epistemological benefits of interventional studies will be minor. But none of this undermines the importance of the distinction between interventional studies and observational studies in the context of testing the efficacy of drugs.

Worrall (2007a, pp. 1017–1018) acknowledges that observational studies are more prone to certain types of bias than randomised interventional studies, but gives this very little emphasis. Instead, Worrall makes the point that providing the treatment effects are sufficiently large, the effects of selection bias in observational studies are likely to be too small to entirely swamp the beneficial effects of the treatment. This is true, but does not resolve the problem for decision makers. Estimating the magnitude of any treatment effect is vital for decisions (a point that Worrall 2007a, p. 1017 accepts). Selection bias may lead to either an over- or under-estimation of the effect

be repeated indefinitely, with a new random allocation of trial participants performed on each occasion, then, in the indefinite sequence of trials, the imbalance of any confounder on a particular random allocation will be counteracted by the distribution of that confounder on another allocation. The overall effect of known and unknown confounders in the indefinite sequence of trials will be zero.

of treatment witnessed in an observational study. Even if the potential bias is smaller than the size of the treatment effect, it will often be impossible to estimate the magnitude of the bias—or for that matter its direction—with any degree of confidence. The difficulty this poses for therapeutic decision makers emphasises the importance of the superiority of interventional studies when testing the efficacy of therapies.

4.3.2 Why *randomised* interventional studies

The final part of the argument for *randomised* interventional studies relies on Bayesian considerations to provide a rationale for randomisation as the preferred method of ruling out allocation bias in interventional studies. The Bayesian argument for randomisation in interventional studies has been provided by Lindley (1982) and Suppes (1982). An outline of the argument is provided below. Lindley's position is of particular interest because Worrall attributes to him a different view. Worrall suggests there is no straightforward Bayesian argument for randomisation, and enlists Lindley's support.

> As always with Bayesianism, there are a variety of positions on offer (the phrase '*the* Bayesian account' always makes me smile), but the most straightforward one articulated, for example, by Savage (who later however, for reasons it seems difficult to fully understand, decided it was 'naive') and Lindley, as we in effect noted earlier,[7] sees no role for randomisation here at all. (Worrall 2007b, p. 482)

Perhaps Worrall is best interpreted as suggesting there is no Bayesian argument for randomisation to be *sine qua non* (and on this we would agree), because Lindley certainly acknowledges a role for randomisation in certain contexts.

Lindley recognises the importance of the experimental groups being evenly matched for known confounding factors. And also that whether the experimental groups are considered well-matched is ultimately a subjective judgment. He refers to an allocation in which the investigator judges the experimental groups evenly matched as 'haphazard'. The need for the allocated

[7]This appears to be referring an earlier recognition of Lindley's suggestion that as the number of confounders increases, the probability is high that any one of these confounders ends up being unevenly distributed in the randomly allocated groups.

groups to be haphazard is more important than randomisation (Lindley 1982, p. 439). But, while it is possible for the Bayesian to deliberately allocate the experimental groups, and hence ensure a haphazard design, the complexity accrues quickly. The expected utility of each possible allocation needs to be calculated. And, in order to calculate this utility, the effect of each possible confounder needs to be estimated. This is no easy task when many of the confounding factors are merely plausible rather than well-known or understood, which is precisely the situation in the clinical sciences. In response to this problem, Lindley suggests a 'reasonable approximation to the optimum design' is to randomise and then check to ensure the groups are haphazard (Lindley 1982, p. 439).

> Consequently the two, apparently conflicting, views of the randomiser and the Bayesian have been brought into agreement. It is the haphazard nature of the allocations, not the random element, that is important; and the use of a haphazard design saves the Bayesian a lot of trouble, with small chance of any appreciable gain, by producing a situation relatively easy to analyse. A further point is that a detailed, Bayesian consideration of possible covariates would almost certainly not be robust in that the analysis might be sensitive to small changes in the judgements about covariates.

The final sentence of this quote recognises the importance of context. If the number, and overall effect, of possible confounders is uncertain, then the Bayesian calculation needed for deliberate matching becomes difficult. This is certainly the case for the large medical trials conducted to test the efficacy of a new drug treatment. In much smaller trials, with few possible confounders, deliberate matching of the experimental groups may be more convenient than randomising and checking whether the groups are well-matched (and re-randomising if necessary).

Suppes (1982) also provides Bayesian reasons to randomise. In situations such as in the clinical sciences, where knowledge of causal processes is available, but unable to accurately predict the response to treatment, randomisation both simplifies computation, and aids communication to (and acceptance by) a sceptical scientific audience. These two reasons for the Bayesian to randomise are linked. Randomising simplifies the Bayesian's computation for the reasons noticed by Lindley. There is no shortage of

plausible causal processes in the clinical sciences, rather, what the clinical sciences lack is knowledge of how these plausible causal processes will interrelate in any particular therapeutic situation. Thus, in addition to a small number of well known, and reasonably well understood confounders—which can be checked to ensure they are evenly distributed in the experimental groups—there is a potentially limitless combination of additional causal processes that may affect the outcome of the treatment. The Bayesian can take (a least some of) these additional plausible causal processes into consideration in forming a prior based on a particular allocation of experimental groups, but the resulting prior will have a high variance, and be dependent on personal judgements. Randomising (with a check to ensure the allocation is haphazard) and *assuming* that randomisation has either resulted in the even distribution of additional plausible (but uncertain, or yet to be fully elucidated) causal factors, or that such factors are likely have a minimal effect on the intervention, provides a rationale for adopting a much simpler prior distribution, as well as helping narrow the many ways that the experimental results could be incorporated into the likelihood function (see Suppes 1982, pp. 464–6).[8]

This simplification of the Bayesian computation also aids communicating the experimental results. Simplification directly assists communication, a point that should not be ignored. But perhaps more persuasively, randomising can provide some common ground for agreement on both the Bayesian prior and the analysis of the experimental results (that is, randomisation may aid agreement on the experimental distribution—whether a Bayesian or frequentist analysis is to be conducted). Of course, this requires the audience to grant the assumption that merely plausible confounding factors are either evenly distributed or unlikely to affect the results. (Remember, any well understood confounders will be checked to ensure they are evenly distributed). But if this assumption is granted, there is a much improved possibility of reaching agreement on how the experiment should be analysed.

The possibility of reaching agreement on the personal judgements needed by the Bayesian to justify a particular deliberate allocation is much less likely.

[8]Clearly, simplifying the prior distribution and the likelihood function are benefits if a Bayesian analysis of the clinical trial is to be conducted. It should be noted that randomising also simplifies the experimental distribution for frequentist statistical analysis. I discuss some of the ramifications of adopting a Bayesian justification for randomisation for the continued frequentist analysis of clinical trials below.

The possibility of consensus is an important advantage of randomisation. It blocks a certain type of sceptical reply—a scepticism towards the personal judgements of the investigator. The more assumptions needed for the analysis, the more difficult it is to persuade a sceptical scientific audience. This is directly analogous to the problem encountered in observational studies—interpretation of observational studies is more difficult because the analysis relies on the assumptions made by the investigators when they deliberately match the experimental groups or make *post hoc* adjustments.

A sceptical reply can, of course, also be made against a randomised interventional study, but it is of a different character. A sceptical audience may question the analysis of a randomised trial on the basis of an unevenly distributed plausible confounding factor. Indeed, a reply of this sort is always possible given that randomisation does not ensure the even distribution of confounders on any particular allocation. In contrast to non-randomised interventional studies, however, the burden of proof is on the sceptic rather than the investigator. The sceptic needs to argue that the differential distribution of the causal factor is sufficient to explain the experimental results. In the clinical sciences the debate often follows such a path. The alternative hypothesis proffered by the sceptic, if plausible, can then be tested in a new trial. Such toing-and-froing plays an important role in how knowledge accumulates in the clinical sciences.

Where does the Bayesian justification for randomising in interventional studies leave a dyed-in-the-wool frequentist? After all, contemporary clinical trials are almost exclusively analysed by frequentist methods. Can this continue if the justification of randomised interventional studies relies on Bayesian considerations? While I am sympathetic to a completely Bayesian approach to the design, analysis and interpretation of clinical research, I won't argue for that view here. Rather, I'll limit myself to a couple of brief comments that may convince a frequentist that the Bayesian rationale for randomising should not be seen as too much of a problem for continued frequentist analysis of clinical trials (if that is your view on how clinical trials should be analysed).

First, randomisation—done for whatever reason—provides the experimental distribution that underpins the frequentist hypothesis test. Surely, the important issue for the frequentist statistician is that the experimental

groups were randomised, not why they where randomised.[9] Second, a sticking point in the argument for randomisation for the frequentist is presumably its reliance on personal judgements. The frequentist position is often construed as rejecting *any* reliance on subjective judgement in inference. Such a position, however, is not tenable (and I doubt it is held by too many frequentist statisticians when explicitly discussed). As Howson and Urbach (2006) show (time and time again) the judgement of the investigator (or analyst) plays a central role in frequentist inferences. Instead, the frequentist might restrict their view to a rejection of subjective judgement playing a role in drawing inferences from observed data once the experimental analysis has been specified.[10] If this restriction is accepted, then perhaps the frequentist can accept the Bayesian justification for randomisation, or develop a justification along similar lines to the argument that has been provided, without being explicitly Bayesian.

4.4 Conclusion

Randomisation does not provide the guarantee that all possible confounders are evenly distributed in experimental groups, and therefore does not provide some irrefutable epistemic good. However, given fallible access to knowledge of causal processes in the clinical sciences, *some* epistemic good is provided by conducting randomised interventional studies rather than observational studies. Randomised interventional studies rule out a source of bias that can occur in observational studies. Worrall is right to argue for a more positive view of observational studies than that provided by EBM (Worrall 2002, p. S329). But randomised interventional studies are not epistemologically equivalent to observational studies when known confounders appear adequately balanced. There is good reason to conduct randomised interven-

[9]What has been shown, by both Worrall (2007a, pp. 996–1001) and Howson and Urbach (2006, pp. 188–194) is that randomisation underpinning the frequentist tests provides no justification for randomisation to be essential for science. This argument in isolation is not an argument against the frequentist hypothesis test. (Of course, Howson and Urbach (2006) provide a range of arguments against frequentist statistical tests, but, presumably, given the frequentist is still a frequentist, he or she does not find these arguments compelling).

[10]I think this is a more accurate representation of the view, but in any case, it is certainly more tenable. It is difficult to see how subjective judgement can be excluded from the specification of an experiment.

tional studies rather than observational studies when testing the efficacy of drugs.

Of course, efficacy is only part of the story when it comes to informing therapeutic decisions. Clinicians need to make judgements about the *effectiveness* of a therapy in individuals and groups. These judgements rely on the external validity of clinical research, and raise a number of considerable challenges.

Chapter 5

The Challenge of External Validity

5.1 Introduction

Clinical research aims to achieve two primary goals. The first is to establish the *efficacy* of a therapy; that is, establish that the therapy *works* (in some population). The second, more difficult goal, is to inform therapeutic decisions. This second goal attempts to define *who* the therapy works for (as well as when it works, and even perhaps, in the ideal situation, how it works when used under routine circumstances). These are questions of *effectiveness*. EBM focusses squarely on the second goal: informing therapeutic decisions, and assessing the effectiveness of therapies. To do this EBM proposes that decisions be made in accordance with its 'hierarchy of evidence'. Evidence gained from randomised trials and systematic reviews of randomised trials reside at the pinnacle of EBM's hierarchy.

While EBM aims to inform therapeutic decisions, the hierarchy it advances only substantiates claims regarding efficacy. As seen in Chapter 3, EBM's hierarchy is best viewed as a hierarchy of comparative internal validity. There is an important gap between having evidence of high internal validity, and having the kind of evidence that clinicians require to inform a decision regarding the patient facing them in the clinic. This is the difference between having good reason to expect the accuracy of the observed results of an experiment, and having good reason to expect these results appropriately generalise to an individual. The challenge of external validity is the challenge

of traversing this gap.

Recall, 'internal validity' refers to the degree to which the results of a study accurately reflect the effects of the intervention on the participants in the study. 'Bias' occurs when any process produces results that systematically depart from the true effects of the treatment. Hence, studies which are subject to more possible sources of bias—or biases of greater magnitude—have lower internal validity than studies with fewer sources of bias (or biases of smaller magnitude).[1] 'External validity' is the degree to which the results of a study can be extended to individuals not involved in the study.

EBM proponents recognise that the practice of EBM requires more than simply consulting the hierarchy; incorporating patient values, and ensuring the 'evidence' is provided in the right form to statistically savvy clinicians are also important (Straus 2004). But as for judging the *evidence* for a therapeutic decision, EBM's hierarchy is put forward as both necessary and sufficient. This is seen in each of EBM's guidebooks (see the quote from Straus et al. 2005 provided above, and the quote from Guyatt and Rennie 2002 provided on page 41). I argue that consulting EBM's hierarchy is not sufficient for informing therapeutic decisions from an evidential perspective. The hierarchy, by organising study designs by their capacity to minimise the magnitude of bias in large therapeutic trials, provides an important input into therapeutic decisions. Ensuring the input from clinical studies is as accurate as possible is necessary for appropriate therapeutic decisions. But, because applying the results of clinical studies to individual patients raises a set of *additional* questions, consulting EBM's hierarchy is not sufficient for judging the evidence for a therapeutic decision. Indeed, slavish attention to EBM's hierarchy will lead to poor decisions; evidence from observational studies (where available), and the theoretical resources of the basic medical sciences are vital tools for appropriate therapeutic decision making.

I outline the challenge of external validity for therapeutic decisions in Section 5.2. The difficulty of applying the results of well conducted randomised trails to individual patients is examined. In Section 5.3 I show how evidence from observational studies, considered in conjunction with randomised trial evidence, can improve external validity. In Section 5.4, I consider the question of why proponents of EBM might deny the important role of observational studies in therapeutic decision making.

[1]This is the more general notion of 'bias', as opposed to 'statistical bias' discussed in previous chapters.

5.2 The Challenge of External Validity

Clinical epidemiology has developed many methods to reduce the possibility of systematic error, or *bias*, in large drug trials. For instance, assessing the efficacy of a drug in a randomised interventional study, rather than an observational cohort study, reduces the risk of systematic differences in the experimental groups at baseline. There are many methods in addition to conducting randomised interventional studies that are used to increase internal validity, and thus improve the accuracy of assessing the efficacy of a therapy. Examples include selecting a relatively homogenous population in which to test the treatment and focussing on the overall average effect on a major endpoint in the trial rather than the effect observed in smaller subgroups. When used appropriately these methods increase internal validity. However, these same methods, may reduce external validity.

For instance, if the sample in a trial is relatively homogenous it can be difficult to infer whether an individual who would have been excluded from the trial would reap the benefit observed in the study. The difficulty arises whether the members of the trial sample are of similarly 'low risk' or similarly 'high risk'. Often studies early in drug development select patients that are younger and only have the condition being treated; this raises the question of whether the therapy will be effective in older, sicker patients. Conversely, later in drug development, trials will often enrol older 'high risk' patients who present to specialist units and suffer the treated condition more severely. This is done to ensure a sufficient number of 'events' will occur to adequately test the effects of the therapy on a major outcome such as mortality. In trials such as these, it can be difficult to infer the effectiveness of the therapy in patients that present in general practice.

Most of the methods that can be employed to improve internal validity have the potential to reduce external validity. Even randomisation may reduce the external validity of a trial. Systematic differences may arise between the patients who are willing to be randomly allocated treatment, and the entire population with the condition. Peter Rothwell (2007a, p. 64) provides an example of a trial of treatments for breast cancer. If women at different stages of the disease have a strong preference for particular treatment modalities, this preference may lead to a difference between the women who are willing to receive a particular treatment via random allocation, and those women who will decline entry into the study. If the effects of the therapy are

also linked to the stage of the disease, then the results of any randomised trial will not be applicable to those women who would refuse to participate in the trial.[2]

Another difficulty for the external validity of clinical trials is that many of the methods that improve internal validity are conservative. The primary aim of most randomised trials is to assess whether the drug is efficacious. Many large drug trials are conducted in order to pass regulatory requirements. The sponsor is seeking marketing approval for a new drug, or a new indication for an existing drug. Hence regulatory bodies require that these trials employ methods that reduce the risk of finding a drug efficacious when, in truth, it is not. This leads to a tendency to minimise the risk of false positive findings, even if they increase the risk of false negatives.

Intention-to-treat analysis provides a good example of this conservatism. Intention-to-treat analysis is often recommended in preference to 'per protocol' analysis (Altman et al. 2001, p. 681). 'Per protocol' analyses include only those participants of the study that take the treatment under investigation and have their outcomes assessed. The concern is that attrition from the trial may not be random—that is, there is a risk of a systematic difference between those patients who continue to take the therapy and remain available for follow-up, and those patients who choose not to take treatment, are removed from the study, or are otherwise lost to follow-up. For example, if an unsuspected adverse effect of the treatment causes a number of patients to leave the study prior to the planned assessment, the per protocol analysis may falsely suggest the drug is more beneficial than it is (only participants who did not suffer the adverse effect will be left). But, the decision to conduct an intention-to-treat analysis is made prior to the trial, and typically it will be reported without the per protocol analysis, even when there has not been any differential drop-out between the groups. If compliance with the trial therapy is low (and this low compliance is unrelated to any effects of the

[2]See Rothwell (2007a) for a dizzying list of factors that can reduce the external validity of randomised trials, and systematic reviews of randomised trials. Examples include: differences in the healthcare system between where the trial was conducted and where results will be applied; how centres and clinicians participating in the trial were selected (were they more enthusiastic about the treatment, or are the clinicians involved early 'adopters', or in some other way different); and how were 'recruit-able' patients selected, even prior to applying the selection criteria of the trial the process of considering which patients to invite into the trial may introduce differences between the trial population and those who will eventually be treated.

therapy), then the intention-to-treat analysis will under-estimate the effect of the drug in those willing to take the therapy. While it can be argued the intention-to-treat analysis provides the clinician with information on what to expect from the therapy in those patients she offers treatment, many may well prefer to know what to expect in those patients who actually take the drug.

Questions of external validity arise from two different directions. Call the 'target population' of a clinical trial the population of patients that will be considered for treatment in routine care. The first set of questions regard whether the sample of patients in the trial reflect the target population. Ideally, the sample population of the trial would be a random sample from the target population; then (assuming the study is free from bias) the average effect of the treatment observed in the trial would be the same as the expected average effect of treatment in the target population. However, samples in clinical trials are not random samples from the target population. Rather, the trial population is made up of patients the investigator has access to, and that meet the trial's inclusion and exclusion criteria.[3] Therefore extrapolating the results of clinical trials always requires judgement. If the sample population is considered representative of the target population in the relevant respects, then the observed results of the trial would be expected to reflect the effects that, on average, would occur in the target population. Typically the sample of the trial will differ from the target population in some respect. Perhaps a trial selects a 'high risk' sample of patients. In this situation, if the trial sample is considered to be representative of high risk patients within the target population, the findings of the study might, in the first instance, be restricted to this set of the target population. Further judgement is necessary as to whether the findings of the study could be extrapolated to patients at lower risk. This set of questions regarding external validity can be summed up in the following way: Can we expect the average result observed in the trial to accurately predict the average response in the target population?

A second set of questions regarding external validity arises by considering whether the observed results of a trial are accurate for an individual. The results of clinical trials reflect the average response to the therapy; while

[3]While it is possible to mentally construct a target population from knowing the type of patients the investigators had access to, and the inclusion and exclusion criteria, there is no straightforward way that the sample population is a random sample from this constructed target population.

the response of some patients reflect this average, some respond better and others worse. Often, patients with different characteristics seem to respond differently to the treatment. For instance, treatment effects might vary according to gender, severity of disease, or membership to particular centres in a multi-centre trial. It is difficult, however, to determine the extent to which the differential response is explained by the differing patient characteristics. Any observed variety in response can be explained by random error, or the patient characteristic. Even when no variability in response is apparent among the *reported* patient characteristics, there is the possibility that patient response differs according to some unsuspected, and undocumented, characteristic. This creates a dilemma. Clinicians can rely on the overall results of clinical trials when making decisions about their patients. In which case, assuming the trial has been set up appropriately, these results will have the warrant associated with them that is provided by frequentist methods; for instance, the primary endpoint of the trial will have been adequately powered to provide a reliable test of the primary hypothesis. Or, clinicians can endeavour to match their patient to subgroups of participants within the trial, who share what the clinician judges to be the relevant characteristics of their patient. In which case, the findings from the trial for the subgroup of participants of interest *may* provide data more relevant to the patient in the clinic, but, according to frequentist methods, will not carry the same warrant associated with the findings (due to the smaller sample of participants involved). This dilemma, in some form, has exercised the leading proponents of modern clinical trials,[4] and the appropriate response continues to be vigourously debated (details are discussed in Chapter 7). All participants to this debate agree that the question is central to the appropriate application of clinical research to individual patients.

The two sets of questions that arise regarding the external validity of clinical research are specific examples of the much more general *reference class problem*. The reference class problem arises when the probability of an event differs depending on how the event is classified. For instance, we may wonder what the probability is that an 86 year old woman with atrial fibrillation living in rural Queensland will benefit from taking the anticoagulant warfarin. Each of the following could be used to calculate this probability: the overall findings from a systematic review of the large clinical trials; the findings from

[4]Both Archibald Cochrane (1971) and Austin Bradford Hill (1966) have lamented the difficulty of this question.

the subgroup of elderly female patients included in the systematic review; or, data from a much smaller study looking at the effects of warfarin in patients in rural communities with limited access to monitoring facilities. Arguments can be made for each of these 'reference classes'—that is the problem.

As Alan Hájek (2007) shows, the reference class problem is not restricted to the frequentist interpretation of probability, but rather any interpretation that can serve as a 'guide to life'. The epistemological difficulty posed by the reference class problem may not be 'solvable'—in the sense that one particular reference class will become the indisputably *right* one—but, good and bad arguments can be given for adopting a particular probability based the available reference classes (at least in the medical context[5]). In this chapter I argue that observational studies play an important (and under-recognised) role in deciding the appropriate probabilities for therapeutic decisions.

Observational studies can help answer the question of whether the observed response in randomised interventional studies will accurately predict the response in patients receiving the therapy in routine care. That is, evidence from observational studies assists the assessment of the effectiveness of therapies in the target population. Proponents of EBM, by arguing that therapeutic decisions should be informed by evidence from randomised trials *rather* than evidence from observational studies, deny—or at the very least, devalue—this role for observational studies.

5.3 The importance of observational studies to therapeutic decisions

Proponents of EBM claim that applying EBM's hierarchy of evidence better informs therapeutic decisions. Patient care, they suggest, is better informed, and more likely to result in positive outcomes, when therapeutic decisions are based on the results of randomised trials rather than observational studies, or the findings of the basic sciences. This claim, however, fails to acknowledge the gap between high internal validity, and achieving therapeutic outcomes

[5]In the medical context selecting the 'appropriate' reference class will depend on which causal factors are thought to be most important in the situation at hand. Ignorance about the causal details allows a debate to take place. In formal instances of the reference class problem, such as Bertram's paradox, no such debate is possible; there are two or more equally legitimate solutions.

in patients under routine care. EBM's hierarchy is justified on the basis of increasing internal validity. While high internal validity is important, therapeutic decisions need more than just well established claims of efficacy. Evidence from observational studies can be as, or more important to therapeutic decisions as evidence from randomised interventional studies. Indeed, evidence from the two methodologies is often complementary. Observational studies should be recognised for their methodological strengths as well as their weaknesses.

The weakness of observational studies in assessing claims of efficacy can be a strength when it comes to assessing a therapy's effectiveness. As Nick Black (1996) notes, observational studies often preserve the context of care better than randomised interventional studies. Preserving context of care is a weakness for observational studies in tests of efficacy. Because observational studies follow patients going about their lives it can be hard, if not impossible, to remove or account for all the factors that may confound an apparent association between treatment and effect. Randomised interventional studies eliminate this specific source of bias. But this gain in internal validity is achieved by taking patients out of routine care.

Observational studies often utilise data collected during routine care. While ensuring the accuracy and completeness of this data is a challenge, the possibility of using population-level data provides a degree of inclusiveness that randomised interventional studies are unable to match. Observational studies can also be considerably quicker and cheaper. This is particularly so for case-control studies, which can examine existing data registries retrospectively, and thus provide timely information on a variety of topics important to therapeutic decisions. While prospective observational cohort studies take as long as randomised trials, the reduced effort needed to recruit and monitor patients in observational cohort studies can substantially reduce costs. The reduced cost and relative convenience of observational studies, permit these methods to focus on topics that receive inadequate attention in randomised interventional studies. Because randomised interventional studies take time, and the evidence from observational studies would be useful while we await the results, observational studies would be important even if all therapeutic questions were equally well researched by randomised interventional methods. Observational studies become considerably more important in a system in which only *some* therapeutic questions are optimally addressed by randomised interventional studies.

Drug safety is an excellent example of why observational studies are important for therapeutic decisions. Observational studies are often superior to randomised interventional studies for detecting, understanding, and estimating the adverse effects of therapies. This is partly due to methodology. External validity is vital for assessing drug safety; clinicians need information on the possible harms of a therapy as it is used in routine practice. Safety data needs to be relevant to both *who* uses the treatment, and *how* they use it. Often many patients who would be treated with the therapy under routine conditions are excluded from randomised trials. And, for those that are included, treatment is often administered under considerably different conditions. Other reasons why observational studies are important for safety information are more political, and due to the way clinical research is conducted.

The possibility of financial gain is a key incentive for conducting randomised trials. Recall that most randomised interventional studies of interest in therapeutic decision making are conducted by industry sponsors for the purpose of regulatory approval. For this reason, randomised interventional studies are much more likely to be set up to address whether a drug works rather than assess a drug's side effect profile. Of course, safety information is gained from large scale randomised trials that are primarily set up to test a benefit hypothesis, but such trials are not optimal for gaining information regarding the safety of the drug. Trials set up to test a benefit hypothesis will select particular patients, and provide the drug under restricted conditions.

I provide two brief examples to illustrate the problems of the external validity of randomised interventional studies in gaining evidence regarding the safety of therapies. First is the use of warfarin to prevent stroke in patients with atrial fibrillation. This case highlights the problems that some large randomised trials can have with external validity. The second case concerns the use of aspirin and clopidogrel in combination in patients with acute coronary syndrome. This example illustrates the contribution of observational studies to therapeutic decision-making.

Randomised clinical trials have consistently shown that warfarin is superior to both placebo and aspirin in preventing stroke in patients with atrial fibrillation (Lip and Lowe 1996). While warfarin presents a higher risk of adverse effects compared to aspirin, especially the risk of bleeding, the benefits of stroke prevention in the clinical trials outweighed the risks. However, generalising these results to patients under routine care is difficult. The tri-

als were conducted in a highly select population; over 90% of the patients screened for entry into the trials were excluded (Lip and Lowe 1996).[6] Further, the risk of bleeding while taking warfarin was potentially reduced by the close monitoring undertaken in the randomised interventional studies. Aspirin was found to be less efficacious, but is less likely to cause bleeding, and does not require the same level of monitoring. The important *therapeutic question* is whether, on balance of benefit and harm, warfarin is superior to aspirin in the target population. This question is only partially answered by the randomised studies that have been conducted. While some argue that a large randomised interventional study recruiting a sample population that reflects patients in general practice will fill the inferential gap (see Morgan 2004, p. 546), others recognise that observational studies can provide important evidence to assist making this therapeutic decision (Reynolds et al. 2004, p. 1944).

There are many therapeutic questions, like the use of warfarin in atrial fibrillation, that can not be addressed adequately by the available evidence from randomised trials. Contrary to the claims of EBM, observational studies can provide important evidence for therapeutic decisions in such situations. This is because observational studies are more inclusive, and examine the effects of therapies on patients undergoing routine care. It might be argued that in some (or perhaps many) of these instances there is—in theory at least—a randomised interventional study that could be conducted that would provide the required evidence. But this misses the point. EBM makes claims about how clinicians should make therapeutic decisions. (EBM makes other claims as well, but let's focus on this one for the moment). Specifically, proponents advise that therapeutic decisions should be based on the results of randomised interventional studies rather than observational studies. This

[6]Sophie Morgan (2004, pp. 544–5) gives further information on the highly selected population involved in the clinical trials examining warfarin in atrial fibrillation:

> Around 50% of patients with [atrial fibrillation] are over 75 years of age whereas only 20% of the trial patients were in this age bracket, 32% of patients are over 80 years and were not included in the trials. Exclusion criteria included old age (>75 years), serious illness (liver, kidney, brain, or malignancy), alcoholism, risk of falls, forgetfulness, use of non-steriodal anti-inflammatories and uncontrolled hypertension. Overall, the studies tended to select younger patients at a lower risk of harm from treatment than the population with [atrial fibrillation] encountered in clinical practice.

is bad advice. Randomised trials have high internal validity, but often suffer from poor external validity.[7] Observational studies by contrast have lower internal validity but can provide information on the effects of therapies under routine conditions. This information can assist judgements about the external validity of randomised trials. Therapeutic decisions need both accuracy and clinical context, and hence should rely on both randomised interventional studies and observational studies.

It is important to disambiguate two claims that EBM makes. One claim suggests randomised interventional studies are the *ideal* design to *conduct* to answer therapeutic questions. And, the second claim suggests that, when we are considering the totality of *available* evidence, randomised interventional studies provide the best evidence for *informing* a therapeutic decision. These two claims are distinct. It is possible to accept, for the sake of argument, the claim that randomised interventional studies could be conceived that would adequately address the 'inferential gaps' in clinical practice, while at the same time to reject the claim that therapeutic decisions should be based on randomised trials rather than observational studies (or alternative evidence). It is EBM's approach to informing therapeutic decisions that I am arguing against (that is, the second claim). Evidence from observational studies, where available, provides an important input into therapeutic decisions. In many ways it is an easy argument to make—an example where observational studies provide important information for therapeutic decisions is provided below.

Proponents of EBM clearly make the second claim—that therapeutic decisions should be made on the basis of randomised interventional studies rather than obervational studies. See the quotes from EBM's guidebooks: Straus et al. (2005, p. 118), provided above, and Guyatt and Rennie (2002, p. 14), provided on page 41.[8] It is, however, unclear whether they would, when pushed, retreat to the first (easier to defend) claim.

In any case, proponents of EBM rarely put the first claim forward in an unproblematic way. This claim suggests that randomised interventional studies are the ideal experimental design for answering existing therapeu-

[7]This is not to suggest that randomised interventional studies can't have relatively good external validity. In a similar way, recognising that randomised interventional studies are able to employ many methods that can improve internal validity, does not suggest that no randomised interventional studies have poor internal validity.

[8]See also the quote from Paul Glasziou et al. (2004, p. 40) provided below.

tic questions. There are two problems for this claim. First, this claim is almost never suitably qualified. Randomised interventional studies are not always the ideal design for therapeutic questions. Randomised studies are optimal for particular questions for particular reasons. For instance, randomised interventional studies are often the optimal method for testing the efficacy of a drug (some of the reasons for this have been outlined in the previous chapters). But even small changes in context can render randomised interventional studies unsuitable for some therapeutic questions. Rothwell (2007a, p. 64) (as discussed previously) provides an example of women suffering from breast cancer who, according to the severity of their condition, differ in their willingness to accept randomised allocation to treatment. In this case, a randomised interventional study of the available treatments is not the optimal design to answer the question of which breast cancer treatment is more effective in the population of women with breast cancer. Proponents of EBM often make—or, are easily interpreted as making—much more absolute, and context independent claims about the benefits of randomised trials. When considering the optimal design for a therapeutic question, the nature of the question being asked, and the context of the research, need careful consideration.

The second problem is practical. Being able to *conceive* of a randomised trial that would fill an inferential gap doesn't get that trial conducted. Randomised interventional studies take much time, and cost a lot of money. The existence of a particular gap in clinical knowledge is rarely sufficient incentive to motivate the expenditure of time and money that randomised trials require.[9] Again, most large drug trials are conducted by commercial sponsors and are designed to meet regulatory requirements for marketing approval. Thus, most drug trials are set up to test benefit hypotheses. Hypothesis regarding the drug's safety are secondary, and assuming any adverse effects are less likely than the benefits of the drug, the trial will not provide an optimal test of these safety hypotheses. Identifying a therapeutic question that *would* be assisted *were* a large randomised interventional study conducted on a particular population of patients provides scant assistance to clinicians if this trial is never conducted. This first claim of EBM, which regards what

[9]There are, of course, counter-examples. One is the Women's Health Initiative. This research program was funded by a number of governmental departments responsible for health and research in the United States to address the insufficient clinical evidence available for prevention of a number of diseases prevalent in post-menopausal women.

trials should *ideally* be conducted, becomes even more problematic if, as a result of EBM's influence, observational studies are not conducted.

Back to EBM's claim that clinicians should base therapeutic decisions on randomised interventional studies rather than observational studies (EBM's second claim). The following example illustrates how observational studies provide important evidence for therapeutic decisions. Randomised interventional studies and observational studies have different strengths and weaknesses for providing evidence for therapeutic decisions; both are important. Aspirin and clopidogrel are antiplatelet agents. They are used to prevent platelet aggregation and clot formation in conditions in which clots cause damage to the body. Most notably they are used in acute coronary syndrome where the formation of a clot causes chest pain, and possibly leads to the permanent damage of heart tissue, as in myocardial infarction. Because aspirin and clopidogrel prevent platelet aggregation by two different mechanisms, combining the agents should provide additive effects. This will, however, also increase the risk of the main side effect of these agents, bleeding. The CURE study, a randomised, double-blind, placebo-controlled trial, showed that the combination of aspirin and clopidogrel (compared to aspirin alone) reduced the risk of the combined primary endpoint—death caused by cardiovascular disease, non-fatal myocardial infarction or stroke (The CURE Investigators 2001). In CURE, treating approximately 50 patients with high risk acute coronary syndrome with aspirin and clopidogrel for 9 months prevented one additional instance of the combined endpoint compared to treating with aspirin alone (this expresses the results of CURE as a 'number needed to treat (NNT)'). In the same time period, for patients taking the combined antiplatelets, there was an absolute increase of 1% in the risk of major bleeding. This is equivalent to a number needed to harm (NNH) of 100.[10] On the basis of this evidence guidelines recommend the use of aspirin and clopidogrel in patients with high risk acute coronary syndromes (Aroney et al. 2006, S21–S22).

The results of CURE have been implemented in practice; many patients who suffer acute coronary syndromes receive concurrent treatment with aspirin and clopidogrel. And this, in large measure, is appropriate. However, there is some evidence that the combination is prescribed to a broader group

[10]'Major bleed' was defined as: 'substantially disabling bleeding, intraocular bleeding leading to the loss of vision, or bleeding necessitating the transfusion of at least 2 units of blood' (The CURE Investigators 2001, p. 495).

of patients than the population that has been shown to benefit in the randomised studies (Hallas et al. 2006; Simpson et al. 2008). And, like most randomised interventional studies, questions can be raised regarding the external validity of CURE. Is the slim margin of benefit over risk retained when combination therapy is given to patients under routine care? Will the combination of aspirin and clopidogrel benefit patients with low risk acute coronary syndrome, or patients with high risk acute coronary syndrome, but who would have been excluded from CURE?

Under the ideal conditions of a randomised trial, CURE suggests that if you treat 100 patients with the combination of aspirin and clopidogrel for 9 months (rather than aspirin alone) you will prevent two patients from suffering a non-fatal myocardial infarction, a stroke, or death due to cardiovascular disease; and you will cause one patient to have a major bleed. Even if it is accepted that the benefits of combination therapy outweigh the risks for the trial population (when treated under the conditions of the trial), clinicians need to judge for which of their patients these results are likely to hold in routine care. Any shift in the risk to benefit ratio will influence therapeutic decisions. In these situations observational studies play a vital role.

The number of patients assessed for eligibility for the CURE trial, but not randomised, is not reported (nor, is any descriptive information provided about excluded patients). This makes judging how well the trial sample reflects the target population more difficult.[11] However, the exclusion criteria for the trial are given. Patients considered at 'high risk of bleeding' were excluded from the study. Further, it is plausible to assume that under trial conditions patients were monitored closely for any signs of bleeding—this level of monitoring may not be available in some institutions for patients under routine care.

Hallas et al. (2006) report a case-control study, which used population-based data registries from a county in Finland to assess the use and risks of combined antiplatelet therapy. Evidence provided by observational studies such as this provide information for therapeutic decisions that randomised interventional studies like CURE can't provide. First, this study illustrates the dramatic increase in use of combined aspirin and clopidogrel following publication of CURE and similar studies—a more than four-fold increase in use of

[11] In a positive move, the CONSORT guidelines published about the same time as CURE, but, at the time of publication, yet to be implemented widely, recommend providing this information. If CURE was to be published now, this information would be reported.

the combination was observed between 2000 and 2004. Second, the adjusted odds ratio for association between use of the aspirin-clopidogrel combination and serious upper gastrointestinal bleeding was 7.4 (95% CI; 3.5–15). Thus, once adjusted for known confounding factors, use of combined aspirin and clopidogrel was 7.4 times more common in the 'cases' (patients who had suffered serious upper gastrointestinal bleeding) than in the 'controls' (patients similar to the cases in other respects, but who did not suffer serious upper gastrointestinal bleeding). This, in conjunction with other results from an observational study, Simpson et al. (2008), supports the concerns raised above; the combination appears to be used in a broader population of patients than that examined in the randomised trials, and the risk of bleeding in patients who receive the combination therapy under routine circumstance may be higher than that observed in CURE. Many questions remain. But, arguably, this observational data suggests combination therapy should be restricted to those patients very similar to those recruited to CURE, and all patients should be monitored closely for any signs of bleeding. Data from observational studies helps inform decisions about *who* should get therapy, and *how* it should be administered.

Before leaving this case, it is worth pausing to note the important role that theory is playing in the background. The benefits and the harms of combination antiplatelet therapy in patients with acute coronary syndrome are predicted by pharmacological and pathophysiological theory. It is because the two agents work by different mechanisms that it was hypothesised the combination may be beneficial in patients at high risk of clots (such as those suffering acute coronary syndrome). And similarly, theory also predicts an increase in the risk of bleeding. Hence, the 'evidence' of importance here is not just the empirical results of the clinical trials, *theory* provides independent support for the observed results. Results from clinical trials, and physiological-pharmacological theory are fallible. That these two different types of evidence are consistent in the case lends a measure of credence to the conclusion. (More on this in the next chapter).

5.4 Clarifying the role of observational studies in EBM

I am not suggesting observational studies can, or should, replace randomised interventional studies in assessing the efficacy of therapies (at least not when randomised interventional studies can be conducted). When randomised interventional studies are possible for tests of efficacy, the role observational studies can play is largely supplemental. But this does not negate the importance of observational studies to therapeutic decisions. Because observational studies provide information on the effectiveness of therapies, they are typically more important once the efficacy of the therapy has been shown. It is once the drug is on the market and being used that the evidence provided by observational studies becomes pertinent. While it runs counter to their official doctrine, I doubt many proponents of EBM would disagree that observational studies can aid therapeutic decisions in this way. So why do they claim otherwise? Considering this question will assist clarifying the position presented in this paper, as well as forestall possible objections.

Recall, the ambiguity that exists between two of EBM's key claims. Understanding why this ambiguity exists may partly explain why proponents of EBM appear to reject a role for observational studies in informing therapeutic decisions. Recall the two claims that EBM makes. First, that randomised interventional studies are the ideal design to conduct to answer therapeutic questions, and second, that when faced with a therapeutic decision clinicians should rely on evidence from randomised rather than observational studies. This distinction between these claims is more relevant (and more obvious) in the context of contemporary clinical research. No new drug currently reaches the market without evidence from randomised interventional studies. The prominence of randomised interventional studies in clinical research, however, is relatively recent. Salim Yusuf, Rory Collins and Richard Peto argued that large simple randomised trials were necessary for many therapeutic questions as recently as 1984 (Yusuf et al. 1984). At this time large clinical studies such as CURE and the trials involving warfarin were a rarity. While substantially smaller randomised trials have been conducted since the 1960s, none were of the size commonly seen in trials testing new treatments now.

Hence, twenty years ago it made much more sense to claim both that large randomised interventional studies are better designed to test claims of

efficacy, *and* that clinicians should use such trials, where available, to inform therapeutic decisions. If efficacy is yet to be established, and the therapy is one in which large randomised interventional studies will provide the best assessment of the therapy's efficacy, then a focus on randomised interventional studies is appropriate. Now that efficacy has been demonstrated prior to the drug reaching the market, emphasis appropriately shifts to effectiveness; it is here observational studies play an important role.

I suspect that some EBM literature conflates EBM's two key claims because it is arguing for a conclusion that has now largely been accepted: compared to alternatives, and provided the research question is amenable, randomised interventional studies provide better tests of the efficacy of a therapy. This construes EBM's claims somewhat charitably, but in doing so it exposes a problem that runs through much of what proponents of EBM claim. In much of the EBM literature there is inadequate recognition that the 'hierarchy of evidence' can best be justified as a hierarchy of internal validity. The hierarchy gives guidance on claims of efficacy, not effectiveness. 'Which research design provides the best test of the efficacy of a therapy?' is all too often conflated with 'Which research design provides the best evidence for a therapeutic decision?'. Therapeutic decisions require information of the benefits and harms of a therapy. While proponents recognise that observational research is often the only possibility for the 'rare' adverse effects of therapies (see, for instance, Sackett 2006, p. 177 and Guyatt and Rennie 2002, pp. 77–78), there is little recognition that even relatively common adverse effects are not optimally tested in the trials that are typically conducted, and even less recognition that the claims made by EBM need to be scaled back from 'informing therapeutic decisions', to claims of efficacy. Some have bucked this trend. For instance, Paul Glasziou et al. (2004, p. 40) explicitly state

> For interventions, the best available evidence for each outcome of potential importance to patients is needed. Often this will require systematic reviews of several different types of study.

But even these authors are less than clear. On the same page, in the same paper, they give the typical EBM advice.

> For example, to answer a therapeutic question, the hierarchy would suggest first looking for a systematic review of randomised controlled trials. However, only a fraction of the hundreds of thousands of reports of randomised trials have been considered

> for possible inclusion in systematic reviews. So when there is no existing review, a busy clinician might next try to identify the best of several randomised trials. If the search fails to identify any randomised trials, non-randomised cohort studies might be informative. (Glasziou et al. 2004, p. 40)

Once again, 'therapeutic questions' are conflated with questions of 'efficacy'.

Of course, my suspicions about why proponents of EBM claim that therapeutic decisions should be based on evidence from randomised trials *rather than* observational studies, may be off the mark. Perhaps the claim should be taken at face value. An avenue for defending devaluing, or ignoring, evidence from observational studies in forming therapeutic decisions is the risk of selection bias—more particularly, that it may be impossible to isolate all possible confounding factors on the outcome of interest from the baseline characteristics of the compared groups of observational studies. Indeed, this is how proponents of EBM argue for randomised interventional studies over observational studies for tests of efficacy. Their concern would be that, due to the risk of selection bias, the findings of observational studies should not be relied upon for investigating claims of effectiveness, just as they should not be relied upon for investigating claims of efficacy. Clearly, observational studies will not assist establishing the effectiveness of a therapy if the apparent effects of the therapy are actually due to an unsuspected, or 'under-adjusted', confounder.

But there are problems with this argument. First, it overstates the problems that selection bias poses. The avoidance of selection biases do not provide an absolute need for randomised interventional studies in all situations. Some advocates of randomised interventional studies concede that observational studies will provide reliable evidence in some situations.[12] The argument for randomised interventional studies is clearest when claims of efficacy need to be established, particularly in the regulatory context. It makes sense to be conservative when you want to eliminate, or minimise as much as possible, any potential sources of bias from experiments that are conducted in order to inform the decision of whether the drug should be marketed.

Second, this argument fails to recognise the limitations of randomised

[12]See for example, Collins and MacMahon (2007, p. 17) who accept the reliability of observational studies when the outcome of interest is rare among individuals not exposed to treatment; the effects of the treatment in individuals exposed is large; and when there are no obvious sources of bias likely to account for most, or all, of the observed effect.

interventional studies. As has been discussed, and the examples show, a well established claim of efficacy is not the only factor important to therapeutic decisions. Indeed, the methods utilised to gain reliable information on efficacy often limit the utility of this evidence for clinicians needing to judge the effect of a therapy in routine care. Once the efficacy of a therapy has been established, the results of observational studies can complement the knowledge gained from randomised interventional studies. Again, therapeutic decision-making will benefit if a move is made from 'either randomised interventional studies, or observational studies' to better recognition of the relative strengths and weaknesses of each approach.

And finally, all of this follows from first principles, and concurs with common sense. Randomised interventional studies, and observational studies provide two different methods for gaining information on the clinical use of therapies. Each method has strengths and limitations. It should not be a surprise that the challenge of making appropriate therapeutic decisions can benefit from the information provided by both of these methods.

5.5 Conclusion

Previous chapters focussed on which claims the methodological distinctions in EBM's hierarchy of evidence could substantiate. I argued that EBM's hierarchy is a hierarchy of comparative internal validity. Study designs higher up the hierarchy have the capacity to rule our more opportunity for systematic error. These benefits in internal validity, however, do not necessarily correspond to improvements in the external validity of medical research. And since EBM focusses on applying clinical research to patients or groups, external validity is paramount. The challenge of external validity has been outlined, and the importance of additional sources of evidence (other than randomised interventional studies) to informing judgement about external validity has been established with particular emphasis on evidence from observational studies.

EBM puts very low value on the evidence provided by observational studies. I have argued that when it comes to therapeutic decisions, this is inappropriate. Evidence from randomised interventional studies can suffer from low external validity. Because observational studies are more inclusive, and follow patients undergoing routine care, they can play an important role in bridging the inferential gap between efficacy and effectiveness. In many, if not

most, instances randomised interventional studies are the best design for testing a drug's efficacy. But well established claims of efficacy provide only part of the input required for therapeutic questions. Observational studies can provide important additional input. Proponents of EBM, by advising therapeutic decisions be informed by randomised interventional studies rather than observational studies, deny this claim.

Chapter 6

The Role of Basic Science in Evidence Based Medicine

6.1 Introduction

The previous chapter illustrated the importance of evidence from observational studies to therapeutic decisions. The task of this chapter is to do the same for basic science. A central aim of biological science is to discover, understand and refine *mechanisms*. When proponents of EBM refer to basic medical science they principally refer to the mechanistic explanation provided by these sciences.[1]

Evidence based medicine (EBM) is put forward as a 'paradigm shift' in medical decision making (Guyatt and Rennie 2002, p. 8). Proponents argue that decisions should be informed by applied clinical research—especially randomised interventional studies—rather than pathophysiologic principles and clinical experience (Evidence-Based Medicine Working Group 1992).[2] Proponents of EBM provide a 'hierarchy of evidence', and suggest that therapeutic decisions should be informed by evidence from as high up the hierarchy as possible (Guyatt and Rennie 2002, p. 13). As has been discussed,

[1] While the step from basic science to the mechanisms of basic science is rarely stated explicitly in the EBM literature, the move is continuously implied. For instance, mechanism is suggested in the use of the term 'pathophysiologic rationale' in the original statement of EBM (Evidence-Based Medicine Working Group 1992) .

[2] Despite the terminology, I suggest that this should not be seen as a paradigm shift in the Kuhnian sense. Rather, proponents of EBM are referring to a shift in emphasis in what is considered ideal as a basis for therapeutic decisions.

randomised interventional studies are placed at the top of the hierarchy, and basic science at the bottom (see Table 1.1, and Table 2.1). This implies that basic medical science, despite providing the theoretical basis of disease and the effects of treatments, plays a minor role in therapeutic decision making. This is a remarkable claim.

The shift in focus from theoretical science to applied clinical research is reasonable when understood in a certain way. A randomised interventional study showing a therapy is beneficial provides more compelling evidence that the therapy will be effective in patients under routine care, than pharmacological evidence that a drug works by a mechanism understood to treat the disease. This is because much is unknown in clinical science; many drugs have promising pharmacological properties that, for one reason or another, do not bring about the expected beneficial outcome in patients.[3] Sometimes an unsuspected adverse effect outweighs the expected benefit, sometimes other pathophysiological mechanisms dampen the effects of the drug, and sometimes, for reasons that can not be discerned, the plausible pharmacological effects of a drug are simply not realised. It also sometimes occurs that a drug works exceedingly well in patients, but the pharmacological mechanism by which it has this effect can not be elucidated. In clinical science, theory sometimes predicts patient outcomes, and sometimes it doesn't; in any particular instance, it is often unknown which will be the case until applied clinical studies have been conducted.[4]

But this is too one dimensional. True, theory is not always predictive in the clinical sciences. And, the intuition that applied clinical research provides more compelling evidence for therapeutic decisions than basic theoretical science may be justified when these sources of evidence are *considered in isolation*. But this fails to recognise how entwined basic science and applied clinical research are. Basic science plays a role at all stages of applied

[3] A widely-cited example is the use of the anti-arrhythmic flecainide in patients after they had suffered an acute myocardial infarction (see, for instance Guyatt et al. 2008b). There is an increased risk of arrhythmia once a lack of blood supply has damaged the heart. It was thought that use of flecainide would prevent such arrhythmia. However, patients who received flecainide in a randomised interventional study had *higher* risks of sudden death compared to patients who did not received flecainide.

[4] Proponents of EBM provide a number of cases in which treatment strategies based on the best available basic science were later shown to be harmful in large randomised interventional studies. Hormone replacement therapy is one prominent example (see, for instance, Sackett 2006, p. 177.)

clinical research. Medical theory appropriately plays an important role in: selecting which therapies undergo clinical testing (treatments are chosen on the basis of their pharmacological profile); the design of experiments (theoretical considerations are important for choosing who the treatment should be tested in); and, perhaps most importantly, the analysis and interpretation of applied clinical research.

Basic science plays a role in therapeutic decision making, but how should we understand this role? EBM has surprisingly little to say on the positive role basic science does, and *should*, play in therapeutic decisions. The problem of theory in EBM is an example of the more general problem of describing the relationship between theory and observation in science.

The relationship between theory and observation in EBM can be understood by recognising the series of intermediary theories (and models of these theories) which exists between raw observation and high level theory. Patrick Suppes' (1962) has called this a 'hierarchy of data models'. I follow aspects of Suppes' work to provide a framework for the role of theory in EBM.[5] The hierarchy of data models illustrates that basic science is central to the design, analysis and interpretation of applied clinical research. And that therefore basic science, in conjunction with the results of applied clinical research, has a prominent role to play in therapeutic decision making. Therapeutic decisions cannot, and should not, be based on the results of applied clinical research *rather than* pathophysiologic principles.

Suppes' account of the relation between theory and observation is one of a number of plausible accounts that could be used to clarify the role of basic science in therapeutic decision making. Suppes' hierarchy of data models has been taken up by a number of philosophers—for instance, Suppes' account is central to Deborah Mayo's philosophy of experiment. However, the central idea that I take from Suppes' account has been discussed by others, and it pre-dates Suppes. Indeed, the importance of intermediary theories between direct observation and general theory, can be seen in Francis Bacon's *'Novum Organon'* (Klein 2008).

[5]Patrick Suppes' account of what models and theories are has been influential (see Suppes 1960, 1962, and Frigg and Hartmann 2008 for discussion). Suppes has a semantic account of theories. Theories *are* a family of models, and models are set-theoretic structures. There is, however, a range of views on these matters. Holding a different view on what models are is not necessarily incompatible with the aspects of Suppes' hierarchy of data models put to use here.

The hierarchy of data models, and the account of the relation between theory and data provided, captures the arguments that are made in analysing data from clinical research. Indeed, the hierarchy of data models explicates aspects of the analysis that are too often left implicit or neglected. (Proponents of EBM, for instance, neglect the role of basic science in applying applied clinical research to therapeutic decisions). Explicitly detailing the hierarchy of data models also provides a sound basis for experimental observations in testing theories. This approach to the relation between theory and observation in science is part of what has since become known as a philosophy of experiment. Philosophers of experiment include Ian Hacking, Nancy Cartwright, Alan Franklin and Deborah Mayo. While there is a wide range of views within this group, each share the belief that some of the traditional problems of the relation between theory and observation in philosophy of science can be avoided by paying close attention to the kind of arguments scientists make in experiments (Mayo 1996, pp. 57–69). An additional benefit of the hierarchy of data models is that this approach is open to a range of views within the philosophy of statistics.[6]

The 'hierarchy of data models' is outlined in Section 6.2. The framework links the abstract theory of basic science with the statistical findings of clinical research. In Section 6.3, I outline the connection between this framework and the standard frequentist analyses conducted in clinical trials. I show that basic science plays a vital, and largely uncontroversial, role in the *specification* of the applied clinical research, and, further, that the (more contested) role basic science plays in the subsequent analysis and interpretation of clinical trials is equally important. In the final section, I discuss the challenges of applying the results of randomised interventional studies to therapeutic decisions involving individual patients. Contrary to EBM's hierarchy of evidence, basic science is vital for applying clinical research to therapeutic decisions.

[6]Suppes' approach has been incorporated into both frequentist and Bayesian accounts. For instance, Mayo (1996, pp. 128–132) utilises Suppes' framework to develop her frequentist approach to a philosophy of experiment, and Stephan Hartmann (2008) incorporates Suppes' approach into his account of Bayesian Networks. The hierarchy of data models provides an account of the role of basic science in therapeutic decisions independent of which account of statistical inference is adopted.

6.2 The hierarchy of data models

The hierarchy of data models makes the role that basic science plays in the design, analysis and interpretation of clinical trials, and thus therapeutic decision making, explicit. Suppes' account explicates the relationship between scientific theories and raw data. The concept of a model is central. Suppes, following Tarksi, defines a model of a theory as a 'possible realisation in which all the valid sentences of the theory are satisfied' (Suppes 1962, p. 252). As there are also theories of experiment and theories of data, 'models of experiment' and 'models of data' are defined in a similar way. The primary insight is that there is a hierarchy of intermediatory models between theory and raw data. Basic medical science, the primary scientific theory in our context, makes claims about how the pharmacological characteristics of drugs will interact with the physiological features of patients. The theory is described at a level of abstraction, and deals with just the predicted interactions between theoretical entities. In order to test the theory, models of the theory, models of experiment, and models of data are required.

This is best illustrated with an example. Consider the case of rofecoxib, the anti-inflammatory agent withdrawn following evidence that it increased the risk of heart attacks and strokes (Bresalier et al. 2005). (This case was discussed in the prologue, and a more detailed analysis is provided in Chapter 8.[7]) Recall, rofecoxib is used to treat inflammation and pain in patients with arthritis. It is part of a relatively new class of drugs, the cyclo-oxygenase-2 (COX-2) inhibitors. The development of COX-2 inhibitors was driven by theoretical considerations; by selectively inhibiting COX-2 it was hoped that COX-2 inhibitors would cause less gastrointestinal damage when compared to older anti-inflammatories, such as aspirin, indomethacin and the like. (Note. This use of basic science is not seen as contentious by proponents of EBM; it is the use of basic science in therapeutic decision making that is questioned.) Unfortunately, while rofecoxib did reduce the risk of gastrointestinal bleeding, the Adenomatous Polyp Prevention Trial (APPROVe) provided evidence that it also *increased* the risk of blood clots, specifically the risk of heart attacks and strokes. This is likely to be a consequence of selective inhibition of COX-2, that is, that same mechanism by which rofecoxib reduces the risk of serious gastrointestinal bleeds. This example illustrates the links between the primary scientific question of whether rofecoxib increases the risk

[7]For another example, see Deborah Mayo (1996, pp. 141–4).

of thrombosis, and the data observed in APPROVe.

The primary theoretical question is based on basic science. Is the mechanism by which rofecoxib works likely to increase the risk of blood clots? In order for a theoretical question such as this to be tested, a model of the theory is required. One possible realisation of this theory is a comparison of the incidence of thrombotic events in patients taking rofecoxib, call this μ_R, with the incidence of thrombotic events in patients not taking rofecoxib, call this μ_C. μ_R and μ_C can be compared in a number of ways. One, fairly standard way, is to consider the relative risk of thrombotic events in patients taking rofecoxib, $RR_\mu = \mu_R/\mu_C$. If the theory is correct, $RR_\mu > 1$. Linking this model of the theory with the experimental and data models permits testing the theory.

A range of methods are employed at the level of the experiment to test the theory with observable data. Indeed, too many methods are employed to list here. The experiment has its own theory (and a model of that theory) by which it provides a test of the basic science. Some of the choices that need to be made in setting up experiments include: which patients to include in the experiment, whether or not the groups to be compared are randomised, how this randomisation is achieved, and whether the control group receives placebo or active control. For instance, in order to appropriately compare μ_R and μ_C it is important that the patients that have taken rofecoxib are similar to those patients that have received control. One way of achieving this is to conduct an interventional study and randomise recruited patients to treatment with rofecoxib or control. This does not ensure the two experimental groups are exactly alike, but does ensure that a number of patient characteristics are roughly equally balanced providing the sample is large relative to the number of patient characteristics. Whether the two experimental groups are roughly equally balanced can be independently tested. If the experimental groups differ in important ways, then the observed results are undermined. If the data suggests that rofecoxib increases the risk of blood clots, but the experimental groups were not comparable, then this is an example of the failure of the experimental model, and doesn't necessarily provide any information about the primary scientific theory under test.

At the next level down are models of the data. One aspect of the data model has already been assumed in the specification of the theory. RR_μ has been suggested to be a measure of whether rofecoxib increases the risk of blood clots. But there are a range of other measures that could be used in

specifying the theory. The absolute risk of thrombotic events, $AR_\mu = \mu_R - \mu_C$, could be taken as the appropriate comparison rather than RR_μ. While the model of the theory refers to unknown infinite sequences (such as μ_R) the data observed in the experiment is finite. APPROVe has a given sample size. If RR_μ is taken to be the unknown parameter of interest, the test statistic observed in APPROVe may be specified as RR_X, where $RR_X = X_R/X_C$, and X_R is the observed rate of thrombotic events in the patients treated with rofecoxib in APPROVe, and X_C is the observed rate of thrombotic events in the patients that received placebo. RR_X, X_R and X_C are each random variables. It is this *test statistic*, RR_X, in conjunction with the model of the experiment and the observed data, which tests the theory.

With the test statistic specified, the distribution of the test statistic under different assumptions can be considered. As outlined in Section 1.2, in frequentist statistics the value of the test statistic observed in the experiment is considered in light of the distribution of the test statistic under the assumption that the null hypothesis is true. In the case of APPROVe, the null hypothesis of interest is that $RR_\mu = 1$, that is, rofecoxib does not increase the risk of blood clots. Let RR_x represent the observed relative risk of thrombotic events for patients in the APPROVe trial. Some values for RR_x are unlikely if the null hypothesis is true. For instance, you would rarely expect to observe a value for RR_x that is considerably greater than one if the null hypothesis, that $RR_\mu = 1$, is true. This is the basis on which frequentist statistical methods inform inferences based on data. If RR_x is greater than one to an extent that observing such a value for the test statistic (or a value greater) is sufficiently unlikely on the assumption that the null hypothesis is true, then, according to frequentist statistics, the null hypothesis can be rejected. In this situation the alternative hypothesis, that rofecoxib increases the risk of thrombotic events, is provisionally accepted.

The details of hypothesis testing within frequentist statistics was discussed earlier, and the finer details of the rofecoxib case are provided in Chapter 8. What I wish to emphasise here is the continuity between basic science and the statistical findings of applied clinical research. The two are inextricably linked, and this is represented by the way the model of the theory and the model of the raw data are specified. What is known or supposed at the level of basic science plays a vital role in specifying both the model of the experiment and the model of the data. And once specified, the statistical findings based on the observed data provide information on the primary

scientific theory only if the assumptions made in the experimental and data model hold. Informally, which experiments are conducted and what data are analysed as tests of the primary scientific theory, depend on the details of the theory. And the ability of the data and the experiment to provide important information about the theory, depends on the adequacy of the assumptions made in specifying the experiment and the data. Theory, experiment and data are all linked; the results of applied clinical research are not separable from the basic science that specified the research.

Suppes' (1962) reason for discussing the hierarchy of data models was to illustrate the problems of an overly simplified account of the relation between theory and observation.

> One of the besetting sins of philosophers of science is to overly simplify the structure of science. [...] What I have attempted to argue is that a whole hierarchy of models stands between the model of the basic theory and the complete experimental experience. Moreover, for each level of the hierarchy there is a theory in its own right. Theory at one level is given empirical meaning by making formal connections with theory at a lower level. (Suppes 1962, p. 260)

The hierarchy of data models describes how experimental inquiry progresses. Data that appears inconsistent with the primary scientific theory will only undermine the theory if the assumptions of the data and experimental models can be shown to hold. In this way Duhem-type under-determination problems can be circumvented.

Within any given inquiry, the heirarchy of data models clearly demarcates questions of the primary scientific theory, from questions of the data or experimental model. Mayo (1996, pp. 147–8) provides discussion on this point. The 'theory-ladeness' of observation causes problems for any view of science that relies on too simple a notion of 'observed data'—for instance, the view of theory change put forward by the logical empiricists. Suppes' account, and philosophies of experiment more generally, avoid the problem of 'theory-laden observation' while maintaining a rational basis for the assessment of theories. Experimentalists accept an observation may 'theory-laden', but argue that it is not 'theory-laden' in a sense that rules out the use of the observation in experimental arguments. The 'theory-ladeness' of observation is not problematic in testing a theory when the 'theory' involved in establish-

ing the observation is 'independent' (in some sense[8]) from the theory that is under test in the experiment. Hacking (1983) illustrates this argument with the example of having to decide whether the observation of 'dense bodies' in platelets should considered 'real' or an artefact of how the platelets are observed. Utilising different physical techniques to view the platelets, for example, flourescence micrographs and electron micrographs (which rely on different physical theories), provides strong evidence that the observed dense bodies are real.

Mayo (1996) develops the argument regarding a hierarchy of models into an epistemology of experiment—an epistemology that incorporates the frequentist statistical methods used to analyse clinical trials (such as the methods proposed by Neyman and Pearson). Central to Mayo's account is the idea that scientists use experiments to develop an argument from error. On Mayo's account, the models of data, experiment and theory are successful to the extent that they are able to rule out, or minimise, sources of error. The more the different levels of models cohere and minimise canonical sources of error, the more confidence we can have in the results of research.

Increased confidence in the results of research when the different models cohere is a version of the 'variety-of-evidence thesis'. On this widely-held view, 'varied' sources of evidence—such as that provided by the different levels of models—better confirms a hypothesis than an equivalent amount of 'similar' sources of evidence. Bovens and Hartmann (2002) have refined this somewhat intuitive notion within a Bayesian framework, and shown that the variety-of-evidence thesis does not hold with any generality; positive evidence from multiple less-than-fully-reliable sources does not necessarily confirm a hypothesis more than multiple positive evidence from one less-than-fully-reliable source. Perhaps unsurprisingly, when it comes to the variety-of-evidence thesis, the details matter—when the reliability of our sources of evidence is low, we may gain more confidence in the hypothesis under investigation from coherent results from a single source (see Bovens and Hartmann 2002 for details). Nevertheless, when the data, experimental and theoretical models are accurate (or likely to be accurate), Bayesian networks, such as those provided by Bovens and Hartmann, support the notion that the variety

[8]Howson and Urbach (2006, pp. 125–126) interpret 'independence' in terms of whether the items of evidence are 'similar' or 'dissimilar'. Two items of evidence, e_1 and e_2, are 'similar' if $P(e_2 \mid e_1) \approx 1$. Two items of evidence are 'dissimilar' if $P(e_2 \mid e_1)$ is considerably less than 1.

of evidence provides confirmation of the theory in question.

My reason for discussing the hierarchy of data models is to illustrate the importance of basic science to therapeutic decisions—something that is downplayed by EBM's hierarchy of evidence. First, there is the uncontroversial role that basic science plays in specifying applied clinical research. This is already clear in what has been said so far, but I will extend this discussion in the next section. The second, and more important reason for discussing the hierarchy of data models, is that it helps illustrate the vital role basic science plays in *applying* clinical research to the problems of individual patients. EBM's hierarchy of evidence is provided as a guide for forming therapeutic decisions about individual patients. In Section 6.4, I show that the results of clinical research cannot be applied to the care of individual patients without the theories of basic medical science.

6.3 The hierarchy of data models and frequentist analysis of clinical trials

The relationship between theory and data outlined by Suppes is consistent with contemporary frequentist statistical analysis of clinical trials. Basic science assists specifying the theoretical, experimental and data models used in applied clinical research, and also plays an important role in the interpretation of results. Frequentist statistical approaches typically use experimental data in one of two ways. In the first, the underlying theoretical and data models are assumed, and the data are used to to provide information on an unknown parameter within the theoretical model. In the second approach, experimental data assesses the adequacy of the theoretical, experimental and data models in a process sometimes called *model criticism.*[9] The former is more common in the large drug trials and other clinical studies that have the most influence on evidence based medical practice, but even in these studies there is often some degree of model criticism that is conducted in supplementary analyses.

As suggested, theoretical considerations of basic sciences, such as pathophysiology and pharmacology, play a prominent role in the set-up and subsequent analysis of clinical trials. Basic science helps define the question being asked as well as the type of analysis that will appropriately answer the

[9]See Chapter 1 of D. R. Cox (2006) for some discussion

question. For instance, understanding the pathophysiological processes that occur in acute coronary syndrome suggests a number of treatment strategies. Early in acute coronary syndrome a dynamic process of platelet aggregation and clot formation takes place in the coronary arteries; if the clot forms it will occlude the artery and damage will occur to the area of the heart serviced by the artery. Because there are a number of pathways by which platelet aggregation occurs, it would be expected that combining antiplatelet agents that work on different pathways may improve the outcomes of patients suffering from acute coronary syndrome. This, in abbreviated form, is the basic science that underpins the Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial (CURE) (The CURE Investigators 2001). CURE compares the outcomes of patients suffering high risk acute coronary syndrome when treated with the combination of aspirin and clopidogrel with patients treated with aspirin alone. Basic science predicts that combination treatment will be beneficial in these patients; earlier treatment will be better than later treatment; those suffering 'high risk' acute coronary syndrome may receive more benefit than those at a lower risk;[10] and the combination of two antiplatelet agents will also likely increase the patient's risk of suffering a serious bleed. This information is incorporated into the theoretical, experimental and data models in numerous ways. In CURE, patients are randomised and treatment initiated as soon as possible on recruitment to the trial, and patients with low risk acute coronary syndrome, or a high risk of suffering a serious bleed, are excluded from the trial.

Basic science's influence on the specification of the analyses conducted in clinical trials flows through to how the results of the trial can be interpreted. In CURE patients who received the combination of aspirin and clopidogrel suffered the combined endpoint of death from cardiovascular causes, nonfatal myocardial infarcts, or strokes less often than patients who received aspirin. Assuming the theoretical, data and experimental models are adequate, these results support a number of inferences. First, treating patients with high risk acute coronary syndrome with the combination of aspirin and clopidogrel appears beneficial overall. Second, when possible, the combination should be started early. And third, caution is needed if treating patients at high risk of suffering a serious bleed. Without further argument, the results of CURE can't be interpreted as providing evidence that the combination of aspirin

[10]'High risk' acute coronary syndrome is defined on a number of criteria; one of which is the level of damage done to the heart as observed on an electrocardiogram.

and clopidogrel will assist different patients, or patients in different circumstances. In the next section, I show that extending the results of CURE to situations outside of the trial—that is, to questions of external validity—depends crucially on basic science. Here, I wish to emphasise that due to the role basic science plays in specifying applied clinical research, it has a direct influence on how the results of the research can be appropriately interpreted for the participants of the study. Positive results of clinical research support the claims of basic science within the constraints of the trial's specification. In one sense this not contentious; basic science has always played this role in the interpretation and application of research. EBM, however, in supplying its hierarchy of evidence, and in advising that therapeutic decisions should be made in accordance with the hierarchy, denies, or at the very least, greatly diminishes the importance of this role for basic science.

Basic science can also play an important role in the interpreting clinical research when the science is not explicitly incorporated into the set-up of the study. A somewhat controversial example of this occurred with rofecoxib. APPROVe was seen by most as confirmation that rofecoxib can increase the risk of thrombotic events; APPROVe provides an estimate of the unknown parameter, RR_{μ}. However, prior to APPROVe another large clinical trial had also provided evidence that rofecoxib may increase the risk of thrombotic events. VIGOR (the Vioxx Gastrointestinal Outcomes Research Trial) was set up to compare the risk of adverse gastrointestinal events—especially the risk of ulcers and bleeding—by randomising patients with rheumatoid arthritis to either rofecoxib or naproxen. While the risk of thrombotic events was not specified among the main clinical endpoints assessed in VIGOR, a statistically significant difference was observed between the groups in the rate of myocardial infarcts. 0.4% of the rofecoxib group suffered a myocardial infarction, compared to only 0.1% of the group given comparator naproxen. The relative risk of having a myocardial infarction in the naproxen group compared to the rofecoxib group was reported as 0.2 (95% confidence interval, 0.1 to 0.7) (Bombardier et al. 2000, p. 1523).[11] Importantly, basic science—in particular the pharmacology of naproxen—was used to explain

[11]The different assumptions required for the data and experimental model in this example should be noted. In contrast to APPROVe, gathering data on thrombotic events was not pre-specified in VIGOR. The only data available is the overall incidence of myocardial infarcts in each group. It is possible that thrombotic events, other than myocardial infarction, were missed in VIGOR.

the observed difference in the incidence of myocardial infarctions between the groups.

The authors of VIGOR suggested that the superior platelet inhibition of naproxen caused the observed difference in myocardial infarctions. The contention was that VIGOR included a number of patients at high risk of coronary artery disease, and that naproxen prevented more myocardial infarctions in this group than did rofecoxib. That is, the observed difference in risk of myocardial infarction was attributed to a benefit of naproxen, rather than a harm of rofecoxib. While the pharmacological claim that naproxen inhibits platelet aggregation can be substantiated, the argument put forward in VIGOR is undermined both by the data later observed in APPROVe, and VIGOR data that was subsequently submitted to the Food and Drug Administration.[12] Even though the argument was later overturned, it was largely accepted at the time of publication, and played a role in how the data observed in VIGOR was interpreted. While the benefits of the argument provided by the VIGOR authors can be debated in this particular case, I emphasise the more general point. Basic science is a necessary component in the interpretation of all clinical research. Once set up, results of randomised interventional studies require interpretation and application, they do not issue as if from some 'black box'. This is so when interpreting the results of a clinical trial in terms of internal validity, and, as I argue below, becomes even more important when it comes to applying these results to therapeutic decisions.

[12] A retrospective subgroup analysis on updated safety data from VIGOR shows a statistically significant increase in adjudicated thrombotic serious adverse experiences in patients taking rofecoxib who were judged not to be at sufficient cardiovascular risk to warrant aspirin prophylaxis, relative risk 1.87 (95% confidence interval 0.29-0.97) (Food and Drug Administration Advisory Committee 2001). This group included 96% of the patients involved in the trial. Thus, the bulk of patients in VIGOR were not at high risk of coronary artery disease, not in need of an antiplatelet agent, and the increased risk of suffering a myocardial infarct in patients taking rofecoxib was still observed among these patients. If the claims made by the VIGOR authors were true, it would be expected that the bulk of myocardial infarcts would have occurred in patients at high risk of coronary artery disease.

6.4 Basic science and the application of clinical research to therapeutic decisions

EBM is about therapeutic decisions, especially therapeutic decisions involving individual patients. EBM's hierarchy of evidence is put forward with the aim of improving therapeutic decisions by ensuring they are based on the most reliable evidence possible. As discussed in Chapter 3, EBM's hierarchy is best viewed as a hierarchy of comparative internal validity. Studies higher up EBM's hierarchy are able to employ more methods to reduce or eliminate possible sources of systematic error. In this way the results of randomised interventional studies are more likely to be accurate for the patients involved in the study. This is all for the good, but none of it ensures that studies higher up EBM's hierarchy provide evidence that is applicable to patients not included in the clinical studies. Therapeutic decisions require judgements about external validity. Despite the lowly place of basic science in EBM's hierarchy, it plays an irreplaceable role in judging the extent to which the results of clinical research are applicable to an individual patient. Indeed, it is *basic science*—or at least the combination of basic science and clinical research—that is applied to therapeutic decisions, rather than the results of clinical research in isolation.

The challenge external validity poses EBM, and therapeutic decision making more generally, was outlined in the previous chapter. Two separate questions arise in judgements of external validity. The first is the degree to which the overall findings observed in the trial can be expected to reflect the average response to treatment in the target population (where the target population is the population of patients who will be treated with the drug in routine care). And the second is whether clinicians should rely on the overall result observed in the trial, or base their inferences on a subgroup of patients within the trial that most closely match the relevant characteristics of their patient. We have seen that observational studies play an important role in informing judgements regarding external validity. Here I argue that basic sciences' role is even more fundamental.

There are many possible differences between a patient in a clinical trial and a patient presenting to the clinic. The condition that is being treated may well be one of the few similarities that patients in the trial and patients in the clinic share (and even with regard to the condition being treated there can be important differences). Clinical studies are not conducted on random

samples from the target population.[13] If samples in randomised interventional studies *were* random samples from the target population, judging the expected average response in the target population would be much easier. Rather, as previously discussed, the sample of patients involved in clinical research, particularly randomised interventional studies, are typically convenience samples. The sample is created from the patients that present to the clinical unit in the geographical areas that the study is conducted in (typically, it is the catchment area for tertiary referral hospitals in a small number of western countries). But, even if we managed to overcome this problem, and the sample of patients involved in a trial were, or could be considered to be, a random sample of the target population, the second question of external validity would still arise. Does the overall result of the trial obscure subgroups within the trial that respond particularly well or particularly poorly to treatment? Judging whether the patient in the clinic is *similar enough* to the patients enrolled in a clinical trial, or subgroups within a trial, represents a tough challenge for therapeutic decision making.

But we shouldn't set the epistemic bar too high. It is not the *possible* differences between patients in a clinical trial and patients presenting to the clinic that matter, but the much smaller set of *relevant* differences. Certainty that the patient in the clinic will experience the precise effect witnessed in clinical research is too much to expect. An indefatiguable sceptic will always be able to conjure up a possible reason for the results of a clinical trial not to be applicable to a particular patient. What is needed is a principled way of separating the factors that are likely to influence a patient's response to treatment, from the many possible differences that may occur between patients in clinical studies and patients presenting to the clinic. It is basic science, backed up by Suppes' hierarchy of data models, that provides this principled method of applying clinical research to individual patients.

Basic science does not always predict patient outcomes. Knowing that selective inhibition of COX-2 should have beneficial effects in reducing pain and inflammation without increasing the risk of gastrointestinal bleeding is not enough to ensure that rofecoxib will relieve pain in patients with rheumatoid arthritis and reduce the risk of these patients suffering a gastric ulcer compared to when they take an alternative anti-inflammatory. The clinical relevance of the pharmacology of COX-2 inhibitors is confirmed by applied

[13]Of course, recruiting a true random sample from the target population in any rigourous way would be impossible in most situations.

clinical research. The hierarchy of data models helps explicate the link. It is important to emphasise that each tier of the hierarchy is supported by independent, and independently testable, information.

The pharmacology of COX-2 inhibitors is confirmed using animal models. The benefits of COX-2 inhibitors in relation to gastrointestinal safety are thought to be due to the selective inhibition of COX-2 over COX-1 (traditional anti-inflammatory agents inhibit both COX-1 and COX-2 more or less equally). Measures, such as the selective *in vivo* inhibition of COX-2 relative to COX-1 can be quantified within the animal models. An experiment testing whether rofecoxib treats pain and inflammation with a lower risk of adverse gastrointestinal effects in patients with rheumatoid arthritis sets out to test the clinical relevance of this pharmacological knowledge. It does this by employing an experimental model. Within this model, decisions are made about how many, and what type of patients are included in the trial. And assumptions are made about what the methods employed in the trial achieve. As discussed earlier, if the trial is randomised, it is assumed that, providing the trial is sufficiently large, the two groups will be roughly matched for characteristics that could influence the effects of treatment. Statistical tests can be employed to assess whether the groups are indeed roughly matched for characteristics suspected to influence the effects of treatment.[14] The same is true for the data model employed in the trial; the data model can be independently assessed. A test statistic, or multiple test statistics,[15] are selected, and the distribution of each test statistic is used to test the hypotheses under investigation. Even while using the assumed distribution of the test statistic to inform inferences about the data, whether the correct probability model for the test statistic has been selected can be assessed. For instance, proportional hazards is a common assumption underpinning statistical tests involving relative risk. The proportional hazards assumption holds that the relative risk remains constant over time; if this assumption is violated the statistical analysis may be invalid. For this reason, when statistical tests are used that rely on this assumption, the data will be tested to assess whether the assumption of proportional hazards is supported.[16]

[14]Of course, these statistical tests fall short of certainty; and some may be better for the job than others. But the point remains. The assumptions of the experimental model can be independently tested.

[15]Clearly more than one will be required for an experiment to test both that rofecoxib works to reduce pain and inflammation, and to lower the risk of gastric ulcers.

[16]See Lagakos (2006) for a discussion of the proportional hazards assumption within the

The findings of clinical research gain credibility because of the interrelated links of the theoretical, experimental and data models. Each tier is based on independent sources of information, and as such each can be independently tested. When each tier is considered reliable, and provides a consistent representation of the process under investigation, some degree of confidence in the findings is warranted. As discussed, this is a version of the variety-of-evidence thesis. Bovens and Hartmann (2002) have shown that the variety of evidence thesis is not completely general, but providing the models are reliable this process of inquiry provides important information for therapeutic decisions.

The integrated whole—the conjunction of basic science with the statistical findings of applied clinical research—needs to be considered when making therapeutic decisions. The integrated whole assists identifying, from the many possible differences between patients in the trial and patients in the clinic, the *relevant* differences—those characteristics that are likely to influence a patient's response to treatment. APPROVe provides an estimate of RR_{μ}, and this estimate is considered broadly reflective of the relative risk of thrombotic events in patients taking rofecoxib. But there are many differences between the sample of patients involved in APPROVe and the population of patients who would typically be considered for treatment with rofecoxib. One obvious difference is that APPROVe was conducted in patients who had suffered a colorectal adenoma, whereas the vast majority of patients treated with rofecoxib have rheumatoid arthritis, or some other chronic inflammatory condition. Indeed, patients with chronic inflammatory conditions in need of anti-inflammatory treatment were *excluded* from APPROVe. So why are the findings of APPROVe considered relevant to this part of the target population, when they were excluded from the study?

The reason comes from the conjunction of basic science with applied clinical research. While there may be differences in the baseline risk of cardiovascular disease between patients with rheumatoid arthritis and those without, there is no plausible, or demonstrated, reason that patients with rhuematoid arthritis (or another inflammatory condition) would respond differently to thrombotic events caused by rofecoxib than the patients recruited to APPROVe. And further, there is a plausible account for why these two groups of patients would respond in a similar way. The results of APPROVe in isolation do not permit generalising the inference to patients outside of APPROVe, but

context of APPROVe.

the results, when placed in the broader context of what is understood about the process under consideration, does permit the generalisation. It is always possible that subsequent information could prove the generalisation wrong; but based on our current understanding, the generalisation is warranted.

Applied clinical research plays an important part in gaining knowledge in the clinical sciences. Sometimes clinical research replaces or refutes basic science, but more often it refines and improves understanding how the theories described in basic science are realised in terms of patient care. Just as basic science alone fails to predict patient outcomes, the statistical findings of clinical research alone fail to give direction on how the results can be applied appropriately. Rather than view basic science and the statistical findings of applied clinical research separately, more progress can be made by recognising the connections between these sources of information.

6.5 Conclusion

Proponents of EBM provide little justification for placing basic science so low in EBM's hierarchy. Given the importance of basic science to therapeutic decisions—and indeed the necessity of basic science in applying clinical research—the only reasonable interpretation that can be given is that, as Haynes suggests, 'basic science *alone* does not provide valid and practical guidance' (my emphasis). On this interpretation, however, the distinction made within EBM's hierarchy between applied clinical research and basic science is facile. The hierarchy of data models illustrates that the statistical results of applied clinical research sit within a much richer framework of models, including theoretical models based on basic science. And it is this richer framework that must be considered when applying the results of clinical research to therapeutic decisions. It should not surprise that the strong rhetoric sometimes employed by proponents of EBM in arguing that therapeutic decisions should be informed by applied clinical research *rather than* basic science, has lead to some confusion on the appropriate role for basic science in therapeutic decision making. One example of this confusion is witnessed in the ongoing debate regarding the appropriate interpretation of subgroup data in clinical trials. This debate is a direct consequence of the challenge of external validity—in particular, the question of whether clinicians should base their decisions on the overall results of a clinical trial, or on the basis of a subgroup of patients within the trial that closely matches

the patient facing them in the clinic. I consider this debate in the following chapter.

Chapter 7

External Validity, Subgroup Analysis and Basic Science

7.1 Introduction

The last two chapters have focussed on external validity and judging the effectiveness of therapies. I have been interested in the extent to which the results of studies generalise to individuals in routine care (external validity), and whether and how therapies work in such patients (effectiveness). Chapter 5 and 6 illustrated the importance of evidence from observational studies and basic science to therapeutic decision making. This chapter extends this discussion to the 'problem of subgroup data'.

Medicine is applied to individuals. This is one of the things that makes medical science challenging, and unique. While basic medical sciences, such as pharmacology and pathophysiology, may adequately describe the fundamental process, they are not able to predict the outcome of giving a therapy to a particular patient. And unlike some other sciences like, say, parts of physics where the background conditions can be controlled, in medicine, the 'background conditions' are whatever the individual patient brings with them. This is one of the reasons proponents of EBM emphasise the need for applied clinical research, especially randomised interventional studies. Applied clinical research is so important because theory doesn't necessarily predict patient outcomes. However, the results of applied clinical research provide information on the average effect of a therapy on the sample of patients included in the trial. This creates a dilemma when it comes to applying the

results of clinical research to individual patients. Should clinicians base therapeutic decisions on the overall results of randomised interventional studies, or should they base decisions on the results of a subgroup of patients within the trial who are considered relevantly similar to the individual facing them in the clinic?

The question of how best to interpret subgroup findings is at the heart of an ongoing debate within evidence based medicine. One group, I'll call them the 'trialists', argue that only the overall results of a trial are reliable enough to provide a basis for decisions. Subgroup data, the trialists suggest, is too unreliable—even when the subgroup is formed on the basis of plausible basic science (that is, scientific theory supports delineating the 'group'). The trialist's view is consistent with EBM's hierarchy of evidence. Basic science on this view plays no, or very little, role in the analysis and interpretation of applied clinical research. A second group, I'll call the 'pathophysiologists', are a somewhat looser coalition, but share the view that theory should play a role in interpreting clinical trials. Whereas the overall results of randomised studies provide the average effect of a therapy in a population, the effects of a therapy in subgroups within the trial may provide more relevant information for the treatment of individual patients. Views within the pathophysiologists, however, differ as to the extent to which clinicians can base decisions for an individual patient on the observed effects of a therapy on subgroups within a trial.

The 'subgroup debate' highlights the importance of theory to therapeutic decision making. I discussed the contribution of Patrick Suppes' hierarchy of data models to this issue in Chapter 6. Here, I apply the hierarchy of data models to the subgroup debate. I do this in Section 7.5. Before this, however, I need to outline the problem posed by subgroup data (Section 7.2), and to examine the response to the problem made by the trialist and pathophysiologist respectively (Section 7.3 and 7.4). Much of the debate between trialists and pathophysiologists, I suggest, is the result of the absence of an appropriate framework for incorporating basic science into the inferences drawn from clinical trials. The hierarchy of data models provides a useful framework for approaching the challenge of interpreting subgroup data, and hence, a framework for approaching the challenge of external validity.

7.2 The problem of subgroup data

Questions regarding external validity arise from two directions (as noted, in Section 5.2 on page 103). The first type of question concerns how well the trial sample reflects the target population (where the target population is the population of patients that will receive the treatment under routine care). Can we expect the results observed in the clinical trial to predict the response in the target population? That many randomised interventional studies recruit a highly selected trial population, creates a problem for assessing the external validity of much applied clinical research. The second type of question concerns whether the overall results of a clinical study is relevant for an individual. The results of randomised trials are rarely quantitatively equivalent in all subgroups involved in the trial. Thus, in the hope of 'individualising' therapy, matching the characteristics of the individual to a relevant group of patients within the trial is tempting. But such individualisation is risky; differences in subgroups are to be expected simply on the basis of random error.[1]

The desire for reliable subgroup data comes about due to the understandable and practical concerns of clinicians. Many well-established therapeutic decisions rely on understanding how the pharmacological effects of treatments interact with the physiological features of patients. An asthmatic should not be given a betablocker because the action of beta-agonists at beta-receptors are important for keeping the asthmatic's lungs dilated. When it comes to new therapies, how the treatment will effect certain physiological systems is usually at least partly unknown, but it is reasonable to suspect that some physiological features will promote the beneficial effects of the therapy, and other physiological features make the patient prone to adverse effects.

It is usually possible to hypothesise about the effects of the therapy in patients with particular conditions prior to clinical testing, but it is impossible to predict these effects with any degree of certainty. Enter applied clinical studies. These studies provide an opportunity to test a range of hypotheses about the effects of any given therapy. Early trials usually test the therapy in

[1]The emerging field of pharmacogenomics provides an interesting context for the challenges of interpreting subgroup data. Pharmacogenomic studies need to interpret data from subgroups defined according to genetic characteristics; typically these subgroups will be too small to provide reliable data. La Caze (2005) considers some of the ethical challenges of pharmacogenomic information.

patients younger and healthier than those in primary care. It can be difficult to infer from such trials what effect the therapy will have in the patients who present to general practice. Later trials include a broader range of patients but the primary findings of these trials are typically the average overall result of the entire sample on a major outcome. Such trials permit the claim that a drug is efficacious in patients similar to those included in the trial (providing the circumstances of treatment are also relevantly similar). But, whatever the average result for the major outcome of the trial, there will often be subgroups within the trial that appear to have responded differently. Again, there is a strong motivation for clinicians to attempt to identify any subgroups that respond differently due to some identifiable characteristic. The correct identification of such subgroups improves care. The difficulty is knowing when the inference is justified.

ISIS-2 (The Second International Study of Infarct Survival) provides an often used illustration of why caution is needed when forming inferences about the differential effects of treatment in subgroups (ISIS-2 Collaborative Group 1988). ISIS-2 is a landmark 'mega-trial' that established the benefit of streptokinase and aspirin in patients suffering from acute myocardial infarction. It randomised 17,187 patients, from 16 countries, to treatment with either streptokinase, aspirin, streptokinase and aspirin, or placebo. Aspirin, streptokinase, and the combination of aspirin and streptokinase, when compared to placebo, reduced the absolute risk of vascular death within the first 5 weeks by 2.4%, 2.8%, and 5.3% respectively. To make such a large trial possible the investigators kept recruiting and follow-up procedures as simple as possible. With the broad range of patients recruited, the question arises as to whether any subgroup of patients can be identified who respond particularly well or particularly poorly to treatment.

The investigators reported the effects of aspirin in patients with the astrological sign of Gemini or Libra, compared to those born under the other birth signs to make a point about undue focus on subgroup analyses. In contrast to other patients, those born under Gemini or Libra appeared to do worse when given aspirin rather than placebo (there was a non-statistically significant *increase* in deaths in patients treated with aspirin). The investigators warn

> Even in a trial as large as ISIS-2, reliable identification of subgroups of patients among whom treatment is particularly advantageous (or among whom it is ineffective) is unlikely to be pos-

> sible. When in a trial with a clearly positive overall result many subgroup analyses are considered, *false* negative results in some particular subgroups must be expected. (ISIS-2 Collaborative Group 1988, p. 356)

Clinical trials are set up to to provide a statistically adequate test of one hypothesis, the 'primary hypothesis'. According to frequentist statistics a central criteria for conducting an adequate test is that the rate of false positive test results is minimised. A 'positive' result is when the p value for the observed test statistic is less than 0.05, when this occurs the test suggests the rejection of the null hypothesis. Such a test will suggest rejecting the null hypothesis, when the null hypothesis is assumed to be true, on less than 5 out of every 100 repetitions of the experiment. A good test will also minimise the possibility of false negatives—that is, minimise when the test recommends accepting the null hypothesis when the null hypothesis is assumed to be false.[2] But the assurances provided by these methods only hold for the primary hypothesis (and even then only when a host of additional assumptions hold). Conducting more analyses on groups within the trial sample, or analyses on endpoints other than the primary endpoint, increases the chance of observing a false positive or a false negative result on tests within the trial.[3]

Brookes et al. (2001) conducted a simulation study to quantify the risks of false-positive and false-negative results in subgroup analyses. Simulation studies are useful for estimating error rates such as these because they allow setting the distributions for (what would normally be) 'unknown' parameters. With the distributions of the parameters set, the performance of statistical tests can be assessed by monitoring the results of the test on repeated simulations of the trial. Brookes et al. (2001, pp. 35–39) found that despite setting the effects of treatment to be equivalent in two subgroups, when there was *no* overall effect from treatment, 7–26% of trials showed one subgroup analysis gave a statistically significant result; and when there *was* an overall effect from treatment, only one of the two subgroup analyses gave

[2]The primer on frequentist hypothesis testing in Section 1.2 provides further discussion on these methods.

[3]Consider a trial conducted comparing two placebo arms (no difference should arise on any measure). If there are 20 independent statistical tests conducted on treatment effects in different patient subgroups, and each test has the standard cut-off of $p < 0.05$, the probability of at least one false positive is over half.

statistically significant results in 41–66% of trials. Thus, when there is no effect from treatment, subgroup specific tests often falsely suggest there is (false-positive), and when there is an effect from treatment, subgroup specific tests often falsely suggest there isn't (false-negative). These results are conservative. Whereas only two subgroups were considered in the simulations, the typical clinical trial often reports analyses on more.

Due to the problems of subgroup-specific analyses, separate formal tests of subgroup interaction are recommended (Rothwell 2007b, p. 173). Rather than assess whether the difference in treatment effects are statistically significant (as happens in subgroup-specific analyses), formal tests of subgroup interactions assess a separate hypothesis; typically, that the true difference in treatment efficacy in patient subgroups is equal—Byar (1985) and Gail and Simon (1985) provide discussion. In their simulation study, Brookes et al. (2001, pp. 35–39) showed—as would be expected—that the frequency of tests of subgroup interactions falsely finding an interaction, whether there was an overall treatment effect or not, was 5%. This is much better than the subgroup-specific analyses (as seen in the figures given in the previous paragraph). The problem for formal tests of interaction, however, is that they are often under-powered. The power of a formal test of subgroup interaction is a function of the size of the interaction relative to the overall treatment effect, and the power of the statistical test conducted on the overall treatment effect. Brookes et al. (2001, p. 37) consider a situation in which the size of the interaction is the same as the overall treatment effect. In a trial with 80% power to detect the overall treatment effect, a formal test of subgroup interaction will detect the interaction in only 29% of cases. A four-fold increase in sample size would be required to detect the interaction with the same power as the overall treatment effect. Clinical trials are typically powered for overall effects (not subgroup interactions), and interactions between subgroups may be smaller in size than the overall treatment effect. Thus, in practice, formal tests of subgroup interaction are a blunt tool.

Alvan Feinstein (1998, p. 299) aptly refers to the problem of subgroup data as a 'clinicostatistical tragedy'. Clinicians have impeccable reasons for *wanting* reliable subgroup data; when available, it helps tailor the right treatments to the right patients. However, the statistical problems of subgroup analyses are compelling. Because the problem of subgroup data is a problem for clinicians attempting to *apply* clinical research to therapeutic decisions regarding individual patients, it is a preeminent problem for EBM.

Some have suggested a role for basic science in assessing the plausibility of subgroup data.

> How can we determine whether the observed effect in a special subgroup is real rather than due to chance? This is a difficult task, but we can place more confidence in an observation if certain conditions are met. Specifically, an effect is more likely to be real if a biologic explanation for it can be found, if there is a 'dose-effect' relationship between the baseline characteristic (upon which the subgrouping is based) and outcome, if the observation is supported by independent findings within the trial, and most importantly, if it is replicated in another independent study. (Furberg and Byington 1983, p. I-99)

Such criteria appear promising, but more detail is needed. I focus on the role of basic science in judging whether the data observed in subgroups are likely to be genuine. Within the participants of the debate there is far from consensus as to whether basic science should play a role at all. Whereas the trialists argue against a role for basic science in the interpretation of subgroup data, the pathophysiologists view basic science as playing an important role (but views differ within this group as to what kind of role basic science can play). Neither response is completely adequate. What is needed, I suggest, is a more explicit framework for theory in therapeutic decision making.

7.3 The trialist's response

The trialist's response to the problem of subgroup analysis is best illustrated by the views of members of the Clinical Trial Service Unit at The University of Oxford, such as Richard Peto, Rory Collins, and Salim Yusuf (who is no longer at the unit), and a number of eminent statisticians, such as Douglas Altman (see, for example, Altman 1998; Collins and MacMahon 2007; Peto et al. 1995; Yusuf et al. 1984). While the trialists acknowledge the importance of 'reliable' subgroup data for therapeutic decisions, they argue such data is not typically available from clinical trials.

> The treatment that is appropriate for one patient may be inappropriate for another. Ideally, therefore, what is wanted is not only an answer to the question 'Is this treatment helpful on average

> for a wide range of patients?', but also an answer to the question 'For which recognisable categories of patient is this treatment helpful?' This ideal is, however, difficult to attain, for the direct use of clinical trial results in particular subgroups of patients is surprisingly unreliable. (Peto et al. 1995, p. 35)

As a result, trialists advise clinicians to base their inferences on the overall findings of clinical studies, or on other endpoints for the subgroup (should such an endpoint be available that is more statistically stable).

> There are two main remedies for this unavoidable conflict between the reliable subgroup-specific conclusions that doctors want and the unreliable findings that direct subgroup analyses can usually offer. But, the extent to which these remedies are helpful in particular instances is one on which informed judgements differ. The first is to emphasise chiefly the overall results for particular outcomes as a guide (or at least a context for speculation) as to the qualitative results in various specific subgroups of patients, and to give proportionally less weight to the actual results in that subgroup than to extrapolation of the overall results.
>
> The second is to be influenced, in discussing the likely effects on mortality in specific subgroups, not only on the mortality in these subgroups, but also by the analyses of recurrence-free survival or some other 'surrogate' outcome. (Peto et al. 1995, p. 35)

In addition to the statistical problems discussed earlier, trialists are led to this position for a number of reasons. Prominent among these is the view trialists take on the role of basic science in assessing subgroup analyses. Consistent with EBM's hierarchy, the theories of basic science play no, or very little, role in forming inferences regarding subgroup analyses. Another important factor that leads trialists to adopt their view on subgroup data, is their position on the broader question of what kind of trials should be conducted. Before the phrase 'evidence based medicine' was coined, the trialists argued that the *efficacy* of treatments should be tested in large and simple randomised interventional studies. The trialists provide a strong argument for the need of large and simple trials to test most contemporary treatments. However, in addition to arguing that large and simple trials are needed for tests of efficacy, the trialists argue that such trials provide compelling evidence for

questions of effectiveness—an argument less convincingly made. For trials to be relevant to clinicians the trialists have to make a number of assumptions. Notably, trialist's assume that genuine unanticipated differences in the effects of treatments in subgroups is unlikely. This assumption restricts the range of replies open to the trialist to the problem of subgroup data.

7.3.1 The trialist's scepticism towards basic science playing a role in interpreting subgroups

The trialists view all subgroups with suspicion, especially *post hoc* analyses, whether or not they are undergirded by basic science. Senn and Harrell (1997, p. 749) encapsulate the trialists view of *post hoc* analyses with wit

> Hindsight is so much more precise than foresight and but for its unfortunate habit of arriving too late, it would surely be used for prediction all the time.

Some *post hoc* analyses of subgroups are suspicious. Data dredging is appropriately repudiated. The trial sample can be divided in many ways. If you wait for the data to come in, and then continue to conduct analyses on different ways the group can be divided, you are sure to find a subgroup that appears to have responded differently to the main group. *Post hoc* analyses that are 'data dependent' are especially problematic. Creating subgroups on the basis of whether or not a patient (or group) responded to therapy, without any independent rationale for creating the subgroup, will lead to spurious results.[4] The ISIS-2 analysis of the effects of birth sign on response to aspirin is a particularly striking example of *post hoc* data-dependent subgroup analysis; a second example is provided in the following section.

A range of errors are possible in *post hoc* analyses. However, an important purpose of applied clinical research is to gain new information on the effects of therapies in patients. Ruling out all *post hoc* analyses, as the trialists do, makes interpretation of unexpected results difficult. Not all subgroup-specific

[4]More generally, any subgroups defined by post-randomisation characteristics require caution. Collins and MacMahon (2007, pp. 9–10) provide the example of a trial for a cholesterol lowering agent. Forming appropriate inferences based on subgroups created by separating patients into groups that achieved large cholesterol reduction versus small cholesterol reductions is problematic because such groups are likely to differ in more ways than simply what reduction in cholesterol was achieved.

effects are suspected, and not all *post hoc* subgroup analyses are examples of data dredging, or rely on data-derived subgroups. Ideally, what is needed is an independent method for assessing the plausibility of particular subgroup findings, taking into account statistical considerations such as whether or not the subgroup analysis was pre-specified. Marshalling the theoretical resources of basic science to aid this assessment is perhaps the most obvious avenue to explore. Trialists, however, reject this possibility.

Trialists are sceptical of basic science being able to assess of the plausibility of subgroup analyses. Douglas Altman expresses this scepticism in his response to the conditions suggested by Furberg and Byington (quoted above).

> My view is that biological plausibility is the weakest reason [for thinking a difference observed in a subgroup is genuine], as doctors seem able to find a biologically plausible explanation for any finding. (Altman 1998, p. 301)

This scepticism regarding biological plausibility denies theory a role in assessing subgroup data. As a result trialists are left to focus on purely statistically considerations.

> It is no doubt that a plausible explanation can help to understand the subgroup finding. In reality, however, finding such an explanation can be a difficult task, if not impossible. This task involves subjective judgement and the process is not clear. For these reasons, the assessment of the quality of a subgroup analysis primarily relies on the assessment of the intrinsic statistical properties of the analysis. (Cui et al. 2002, p. 356)

Trialists are left in the position of acknowledging the importance of 'reliable' subgroup information for therapeutic decisions, but having no way to assess the reliability of a subgroup analysis, apart from their intrinsic statistical properties. And, on the basis of the trials typically conducted, the statistical properties of most subgroup analyses are poor. Subgroup-specific analyses are neither sensitive nor specific. And, while formal tests of subgroup interaction have improved positive predictive value, the rate of false negatives is high because of the low power of most of these tests. This is why trialists argue therapeutic decisions should be based, as much as possible, on the overall results of randomised studies.

Due, in part, to their view that basic science is unhelpful in interpreting the results of clinical studies, trialists are left to raise a white flag to the challenge of interpreting subgroup findings. The important question of course is whether this attitude to subgroup analyses is appropriate. *Wanting* reliable subgroup analyses does not make subgroup analyses reliable.[5] It is here that a formal framework, such as that provided by Suppes' hierarchy of data models, is invaluable. This will be discussed later in the paper, the point I emphasise here is the *generality* of the trialist's rejection of a role for basic science in assessing the findings of clinical trials. Trialist's have no role for basic science in drawing inferences from subgroups. Excluding theoretical science from interpreting clinical studies risks leaving the statistical problems intractable.

7.3.2 The argument for large and simple trials

In their highly influential paper, trialists Yusuf et al. (1984), argue that, given the kinds of therapies investigated in contemporary clinical medicine, randomised interventional trials should be large and simple. Understanding the trialist's argument for large and simple trials provides further insight into their views on subgroup analysis. Indeed, for their argument regarding large and simple trials to go through, trialists are committed to the view that results in subgroups within the trial are reflective of the overall result (regardless of how findings appear in subgroups). The possibility that subgroups of patients may respond differently to therapy in large simple trials undermines the trialists argument that the results of such trials are *relevant* to clinical practice. By 'relevant to clinical practice', the trialists mean that the large and simple trials they advocate—on the basis of their improved internal validity—also have high external validity. It is, after all, the external validity of well-conducted clinical research that is the primary concern of therapeutic decision makers.

Yusuf et al. (1984) argue for large and simple trials in the following manner. First, they recognise that the effects of most contemporary treatments

[5] Rothwell (2007b, p. 169) makes this point by quoting John W. Tukey.

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. (Tukey 1986)

under investigation will at best be modest. If the treatment possessed larger effects, the efficacy of the therapy could be established via alternative methods.[6] Given the treatment effects under investigation are more likely modest (rather than large) the trial needs to be designed to ensure any possible errors are smaller than the modest effects of treatment.

> Methodology that may introduce moderately large biases or moderately large standard errors is not appropriate for the assessment of the type of moderate treatment effects that are all that is usually plausible to hope for. It is chiefly because one needs be able to distinguish reliably between moderate and null effects that trials need to be *strictly* randomised, analysed and interpreted completely unbiasedly, and much, much larger than is currently usual. (Yusuf et al. 1984, p. 410)

Systematic errors (biases) are minimised by selecting the appropriate methods for the trial. For instance, conducting a randomised interventional study rather than an observational study, ensuring allocation concealment is maintained, and conducting intention-to-treat analyses each minimise or rule out the possibility of a specific bias. 'Random errors' are the effect of chance on the observed outcomes. These errors are minimised by ensuring enough patients are recruited such that a sufficient number of outcomes occur in the treatment and control group. Large numbers of outcomes increases the precision of results, and this in turn increases the reliability of the frequentist statistical methods utilised to interpret the results.[7]

[6]For instance, randomised interventional studies are not needed to establish the efficacy of administering subcutaneous insulin to patients with type I diabetes. The disastrous effects of withholding treatment quickly manifest. (By contrast, randomised interventional studies may well be useful for comparing the efficacy of different regimens of insulin treatment.) Randomised interventional studies are not required when there is an 'all-or-none' effect from treatment. A sky-diver does not require evidence from a randomised interventional study to convince her of the need of a parachute (Smith and Pell 2003). Glasziou et al. (2007) provides discussion on when randomised interventional studies may be unnecessary.

[7]'Reliably' here is defined by frequentist statistics. For instance, a large number of outcomes in each of the experimental groups narrows the confidence intervals around the findings for each group. This increase in precision allows frequentist methods to 'reliably' reject the null hypothesis if the groups appear different (if the 95% confidence interval for the test statistic in the treatment group excludes the value of the test statistic observed for the control group), and accept (or not reject) the null hypothesis if the groups appear

Hence, when a treatment's effect is likely to be modest, large and simple randomised interventional trials are the best way to establish the efficacy of the treatment. According to the trialist, large and simple trials with broad inclusion criteria are the best way to establish whether the treatment is helpful on average to a wide range of patients. This is the conclusion that the argument in (Yusuf et al. 1984) substantiates,[8] however, they go on to make a further claim. Not only do large and simple randomised intervention studies establish the efficacy of treatment, the findings they provide are also *relevant* to clinical practice. Here the trialists shift focus from whether the treatment is helpful on average to a wide range of patients, to whether the treatment is helpful to an individual or group of patients with shared characteristics, from internal to external validity, and from efficacy to effectiveness.

To make this move, trialists assume that subgroups within the trial are unlikely to respond differently to the therapy in question. Yusuf et al. (1984, pp. 413–416) differentiate between 'quantitative' and 'qualitative' interactions within subgroups. A 'quantitative' interaction is one in which the direction of treatment effect is the same among subgroups, but the magnitude of benefit, or harm, is different. In 'qualitative' interactions the effect of treatment in the two subgroups is in opposing directions; one group benefits from therapy and the other is harmed. The trialist's argument for the clinical relevance of large and simple randomised interventional studies relies on the assumption that differential effects in subgroups are unlikely. Yusuf et al. (1984) explicitly assume that unanticipated qualitative interactions are unlikely.

> ... *[U]nanticipated qualitative* interactions (whereby treatment is

similar (because the 95% confidence interval for the test statistic observed in the treatment group includes the value of the test statistic observed in the control group).

[8]It is noteworthy that Yusuf et al. accept that they have not provided an argument for randomised interventional studies to be 'essential for any truly scientific conclusion to be drawn from trial data' (which is a conclusion that Worrall (2007b) spends much time showing cannot be substantiated).

> The argument for randomisation is not that no truths can emerge without it—indeed, the history of medicine contains many examples where uncontrolled clinical observation has reliably established the value of certain treatments—but that without it *moderate* biases can easily emerge. (Yusuf et al. 1984, p. 416)

This point, however, is too infrequently recognised in the EBM literature.

> of substantial benefit among one recognisable category of patients in a trial and not among another) are probably extremely rare, even though in retrospective subgroup analyses they may seem extremely common. Of course, one can recognise *a priori* certain categories of patients for whom certain drugs are contra-indicated (e.g. for patients with severe heart failure or advanced heart block, a beta-blocker is so clearly contra-indicated that such patients would probably have been formally ineligible for a beta-blocker trial). Our expectation is not that *all* qualitative interactions are unlikely, but merely that *unanticipated* qualitative interactions are unlikely, especially if attention is restricted to one mode of death. (Yusuf et al. 1984, p. 413)

The trialists claim that if this assumption is granted, the overall results of a trial are generaliseable, first to subgroups of patients within the trial (whatever treatment effect was observed within the subgroup), and second to patients who share the characteristics of the subgroup who were not involved in the trial. On the trialist's view, if the assumption is granted, the findings of large and simple randomised interventional studies are relevant to clinical practice.

But there are problems, both with the assumption that unanticipated qualitative interactions are unlikely, and with the claim that if this assumption is granted then large and simple trials are relevant to practice. First, it is hard to see why the assumption that qualitative interactions are unlikely should be granted. Qualitative interactions are common in clinical science. For pretty much any therapy, groups can be identified who benefit, and groups can be identified who either don't derive the benefit of the therapy, or who are harmed. The trialists seem to accept this by focussing not on qualitative interactions but on *unanticipated* qualitative interactions. But many of the qualitative interactions that are now considered confirmed were once unanticipated. As clinical science progresses groups of patients who are particularly benefited, or harmed by a therapy are identified. (Ironically, this process is illustrated by the very example the trialists choose to discuss. At the time Yusuf and colleagues were writing, beta-blockers were contra-indicated in patients with heart failure. Since then, the benefits of beta-blockers in patients with severe heart failure have been established (Whorlow and Krum 2000).) Given many large and simple trials examine relatively new therapies, unanticipated qualitative interactions would be

expected—and, it would be hoped that applied clinical research would be able to uncover them.[9]

Second, knowledge of *both* qualitative and quantitative interactions are important to therapeutic decisions. Trialists accept that quantitative interactions are common, but do not appear concerned that their method leaves such interactions undetected. Indeed, the trialists fail to recognise the importance of quantitative interactions to therapeutic decision makers. Therapeutic decisions are a matter of weighing the benefits of a therapy with any potential for harm. If the risks of a therapy are constant, but an identifiable subgroup of patients derives less benefit than the average, then this reduction in benefit might be enough to tip the balance against using the therapy in that subgroup. This is not merely a theoretical concern. Rothwell (2007d) provides a number of examples in which the absolute treatment effects in different subgroups of patients are markedly different based on the patients' risk without treatment. While the relative risk reduction is similar in different subgroups, those patients with lower risk without treatment achieve a smaller absolute reduction of risk on treatment. This smaller absolute risk reduction may be outweighed by either the risks or the costs of the therapy.

Third, Yusuf et al. (1984) appear to equate the anticipation of a difference in the effect of the treatment between two subgroups, and the incorporation of that anticipation into the specification of the trial. This is seen in the example they provide. Because beta-blockers were thought to be harmful to patients with heart failure, they assume that any trials designed to test the effects of beta-blockers would exclude these patients. In this way the anticipated effects of beta-blockers are included in the specification of the trial. This is the only combination of an anticipated difference in effect and incorporation of that anticipation into the trial specification that Yusuf and colleagues consider; but there are a range of alternatives. In addition to excluding patients based on the anticipation of differential effects in a subgroup, subgroup analyses may be pre-specified for groups in which variation is anticipated, or despite there being an anticipated (or at least foreseeable) variation in a particular subgroup, it is possible that this anticipated variation is not included in the

[9]It seems appropriate that unanticipated findings from a large randomised trials are treated with caution. Perhaps replication of the findings should be required before they are considered confirmed—that is, treat the findings as 'hypothesis generating' rather than 'hypothesis confirming'. More needs to be said about such an approach, but it is far better than interpreting studies on the assumption that no unanticipated interactions will occur.

specification of the trial.

If a quantitative interaction is anticipated, then it can be pre-specified. In this situation, it is expected that the effects of treatment might not be as large in one subgroup as in another. Again, any evidence of such a quantitative interaction is important for therapeutic decisions, and thus important that applied clinical research have a method for identifying these subgroup interactions.[10]

Finally, subgroup interactions (both qualitative and quantitative) can be tentatively anticipated on the basis of basic science, but not included in the specification of the trial. There is often many open questions about the effects of therapies in certain subgroups of patients. Any one clinical trial can only hope to answer a small subset of these questions. Rothwell (2007d) provides a list of situations in which clinically important differences should be anticipated in subgroup analyses. Each of these rely on theoretical considerations to some extent (Rothwell 2007d, p. 141). To return to one example, treatment effects are often relative to a patient's underlying risk without treatment. Thus, the absolute benefits of a therapy may differ between groups of patients that have a different risk of an event without therapy.[11] Most clinical trials are not set up to answer *all* questions that can be anticipated on the basis of basic science, thus, it is often the case that some observed effect of a therapy could have been predicted based on existing knowledge (and thus anticipated), but not included in specification of the trial. Despite these results being important to therapeutic decisions,

[10]Including anticipated subgroup interactions into the specification of the trial through pre-specifying subgroup analyses is less of an option for qualitative interactions. If it is anticipated that the drug may *harm* an identifiable subgroup, then you would expect this subgroup to be excluded from the trial.

[11]Rothwell (2007d, p. 141) uses the example of trials in patients with hypertension. Younger and older patients with hypertension receive similar relative risk reductions in cardiovascular disease when they are treated with anti-hypertensives. But younger patients have a substantially lower absolute risk of cardiovascular disease without treatment compared to older patients. This results in a lower absolute benefit from anti-hypertensives in younger patients. Due to this lower absolute benefit from treatment, the blood pressure thresholds for treating younger (or otherwise low risk) hypertensive patients are considerably higher than older (or otherwise high risk patients). Rothwell (2007d) provides further examples where quantitative interactions should be anticipated. For instance, when there are different pathologies underlying the disease under treatment, and when treatment benefit differs with differing severity of disease.

the trialists provide no avenue for their interpretation.[12]

The difference between questions of efficacy and questions of effectiveness is important in this context. The efficacy of a treatment under investigation can be expressed as a focussed question: Does the treatment produce the expected benefits in a defined population of patients? Clinical trials can be set up to answer such questions. Indeed, many of the methods developed within clinical epidemiology have been developed to improve the reliability of answering questions of efficacy. These methods are largely conservative, and are well-placed to assess the efficacy of a treatment. In a trial set up to test a treatment's efficacy, if the outcomes of patients given the treatment are on average superior to patients given control, then the treatment is judged to be efficacious. Assessing the effectiveness of a therapy opens up a different set of questions. Does the drug work in patients under routine care? Are there any groups of patients that clinicians should expect to respond particularly well, or particularly poorly to treatment? These are questions of external validity and treatment effectiveness. While the conservative approach of clinical epidemiology is appropriate for questions of efficacy—you want to ensure beyond any reasonable doubt that a treatment is efficacious before you put it on the market—some methodological latitude is needed for questions of effectiveness. At minimum, the conservatism of some clinical epidemiological methods need greater recognition. (I return to this discussion in Section 7.5, and take it up again in the following chapter).

The trialist's reply to the problem of subgroup data in large clinical trials is essentially a re-statement of their assumption that qualitative interactions are unlikely. Yusuf, Collins, and Peto (1984) argue that treatments should be tested in large and simple randomised interventional studies. For these studies to provide clinically relevant results, they assume that qualitative subgroup interactions are unlikely. Based on this assumption, the trialists recommend that overall study findings are generalised in favour of findings in specific subgroups. A decade later, and large-scale randomised interventional studies are the norm rather than the exception—at least when it comes to new drug therapies. Now Peto, Collins, and Gray (1995) argue that due to the intrinsic statistical properties of subgroup analyses, clinicians should base inferences on overall study findings rather than observations in particular

[12]The rofecoxib case, which will be discussed in the next chapter, provides a good example of some of the problems that arise when ancillary results of a trial are the most important results for therapeutic decisions.

subgroups.

Yusuf et al. (1984) addresses two methodological questions—how best to establish a therapy is efficacious in a wide range of patients, and how best to identify which patients respond well to therapy (that is, the effectiveness of a therapy). The argument provided by the trialists for questions of efficacy is considerably stronger than the argument provided for questions of effectiveness. The identification of which categories of patients are most likely to benefit from therapy is treated as a subsidiary to the question of how best to establish efficacy. In addition to their views on the role of basic science, the trialist's reply to the problem of subgroup data is forced by the assumptions they make to establish the need for large and simple randomised studies.

7.4 The pathophysiologist's response

Many are unsatisfied with the trialist's response to the problem of subgroup data. Alvan Feinstein (1984, p. 421) made the following comment in discussion of (Yusuf et al. 1984)

> The main problem, it seems to me, is again the question of whether we are evaluating two treatments or are we evaluating treatments for the care of patients? The different kinds of patient that are being lumped together into these heterogenous pastiches under the name of the same disease or under the name of the same therapeutic agents may produce results with excellent statistical ability to compare two treatments, but will be relatively worthless when people try to use the consequences in practice.

Applied clinical research continues to have these two somewhat mismatched aims. First, to establish, with as much rigour as possible, the efficacy of a given treatment. And second, to use that same data to inform therapeutic decisions involving individual patients. I'll group together, as the 'pathophysiologists', those with views that attempt to answer this second question with an appeal to the theoretical considerations of basic science. There is a range of positions within this group about the extent to which subgroup data can be interpreted in light of pathophysiological, or other theories of basic science. Importantly, not all of the pathophysiologists appropriately address the statistical problems that arise in subgroup analysis.

Ralph Horwitz and colleagues (1996) caused considerable controversy with their re-analysis of the β-Blocker Heart Attack Trial (BHAT) (β-Blocker Heart Attack Trial Reseach Group 1982). BHAT was a multicentre randomised trial that tested the effects of the beta-blocker propranolol in patients who had suffered an acute myocardial infarction. Overall, treatment with propranolol rather than placebo reduced mortality in the randomised patients. To the surprise of Horwitz et al. the beneficial effects of propranolol were not observed in one third of the centres involved in BHAT. With the aim of providing improved data for 'the clinician focusing on a single patient', Horwitz and colleagues compared data from the centres in which propranolol benefited patients, with data from the centres in which propranolol did not benefit patients. Horwitz et al. (1996) labelled those centres in which patients responded to propranolol, 'dominant', and those centres in which patients did not respond, 'divergent'. Using a formal test for qualitative interactions, they compared patients from dominant and divergent centres and found that the test was statistically significant, indicating—according to Horwitz et al. (1996, p. 397)—that 'this qualitative treatment interaction was unlikely to have occurred by chance alone'. On the basis of this result they then went on to find a number of baseline features that varied between patients in the dominant and divergent centres. Horwitz et al. (1996, p. 399) conclude

> Our recommendations for managing the analysis of multicentre trials are simple. When divergent centres are noted in mulitcentre trials, tests for qualitative interaction should be conducted. When these tests exclude chance as a cause for the divergence, and when clinical or biological features identify groups whose response to treatment differs from the overall result, treatment recommendations should be modified to reflect these findings.

The move to incorporate basic science into the analysis of subgroup data provides an avenue of response to the question of which categories of patients might respond particularly well to treatment. However, the methods adopted by Horwitz et al. are statistically problematic.

A range of commentary followed (Horwitz et al. 1996), much of it critical of the methods employed (Altman 1998; Senn and Harrell 1997; Smith and Egger 1998). Senn and Harrell (1997, p. 749) highlight that Horwitz and colleagues' use the formal test for interaction inappropriately.

> [The formal test for subgroup interaction] requires that, 'the sub-

> sets should be disjoint *and specified in advance*'. Choosing subsets based on observed rate differences is an extremely serious violation of the standard assumptions for sampling-based statistical inference.

The problem is not only that the analysis was *post hoc*, but that the two subgroups compared were derived from the data. It is no surprise that there is a statistically significant difference between two groups that were created on the basis of whether or not a response was observed. A more appropriate, though still *post hoc*, use of the formal test of interaction, would be a comparison of all 31 centres of the multicentre trial—especially if there were independent reasons for suspecting important differences in treatment effect between the centres. A formal test of interaction on the null hypothesis that no difference existed among the 31 centres was not conducted, but is unlikely to return a statistically significant result. Smith and Egger (1998, p. 292) use simulation to show that up to one third of centres showing no response, despite an overall and constant risk reduction in favour of propranolol, is entirely consistent with chance variation.

In two responses to this criticism, Horwitz et al. (1998, 1997) have failed to appropriately acknowledge this statistical flaw.[13] Instead, Horwitz and colleagues, have focussed on clinician's *need* for appropriate subgroup analyses.

> It is also clear from their commentary that while making their criticisms, they have completely missed the central point—stated in the first sentence of our paper—concerning the character of the information that is needed by clinicians to guide the management of individual patients.

But, rather than legitimise the statistical methods of Horwitz et al. (1996), the clinical need for reliable subgroup data emphasises the need for careful and appropriate statistical analysis. Part of the problem is the view that basic science and the statistical results of applied clinical research are separable—a view that is implied in EBM's hierarchy of evidence. The motivation of Horwitz and colleagues is correct, clinical and biological features

[13]Horwitz et al. (1997, p. 754) accepts the charge of using the formal test for subgroup interactions *post hoc*, but fails to recognise that the bigger problem is that their subgroups were derived from the data.

should be considered when interpreting subgroup analyses, but their methods are inappropriate.

Feinstein (1998) and Smith and Egger (1998), are just as clear on the clinical importance of data based on appropriate pathophysiologic subgroups, but refrain from endorsing the statistical methods employed by Horwitz et al. (1996). Neither, however, are clear as to what alternative methods would be appropriate. Feinstein (1998, p. 299) focusses on ensuring *some* subgroup analyses are permitted.

> My main concern in these comments is not to take sides in the controversy about the appropriateness of the subgroups formed in the 'inquiry' conducted by Horwitz et al. Instead, I want to rescue the scientific importance of valid pathophysiologic subgroups from being forgotten or destroyed by excessive vehemence in suggestions that all subgroups are evil.

And, the only 'solution' provided by Smith and Egger (1998) is that of acceptance.

> What is required in a degree of humility in the face of an issue for which there is no statistical or clinical solution. [...] The development of randomised clinical trials since Mackenzie's time has provided a much sounder basis for making decisions about abstract patients and—if representative samples of patients are included in the trials—for deciding if the overall effect on population health of a treatment is beneficial or harmful. Randomised trials have not, however, answered the question of which individuals actually benefit from medical interventions. This, surely, is the key issue in clinical research in for the next millennium. (Smith and Egger 1998)

Rothwell (2007b), by contrast, proposes a set of guidelines for the analysis of subgroup data. These guidelines represent an important step forward. Rather than assume subgroup interactions are unlikely, Rothwell's (2007b) general approach is to incorporate the theoretical concerns of basic science into the specification of the trial. Rothwell (2007d) illustrates the kind of subgroup interactions we should anticipate based on basic science, and recommends these subgroup analyses should be incorporated into the design, analysis and interpretation of the trial. A small number of subgroup analyses

should be specified prior to starting the trial, and assessed using the formal test for subgroup interactions—for which the trial should be appropriately powered to test. In addition, Rothwell suggests that statistical results of treatment effects in specific subgroups should not be reported, and that, if they are, the high error rate for such tests kept in mind. Further, Rothwell suggests that genuine unanticipated subgroup interactions discovered *post hoc* should be treated with caution, and that reproduction, in further clinical research, is necessary before any subgroup-specific findings should be considered confirmed.

Rothwell's wish-list for subgroup analysis provides a marked improvement on the advise provided by trialists. Qualitative and quantitative subgroup interactions are recognised as possible and important, and where such interactions are anticipated, Rothwell recommends including them in the specification of the trial. This is done while continuing to acknowledge the statistical challenge of correctly interpreting such subgroups.

Despite the advances made by Rothwell's guidelines for subgroup analysis, a number of problems remain. The first is practical. Rothwell's advice is optimistic. Many trials will have to be considerably larger than they already are if they are to reliably test for important subgroup interactions. Recall, Brookes et al. (2001, p. 37) showed that a four-fold increase in sample size was required to perform a test for interaction with the same power as the statistical test of the overall treatment effect. Such a large commitment from regulatory agencies and the funders of trials would be required to improve the reliability of subgroup analyses.[14] In cases where such large trials are not possible, some flexibility in interpreting these studies will be required.

Also, Rothwell's recommendations are limited to anticipated subgroup interactions that are included in the specification of the trial. As discussed in the previous section, and Rothwell appears to accept, there will often be more anticipated subgroup interactions than can realistically be specified in most contemporary trials. There is a limited number of hypotheses any clinical trial can reliably aim to test.

The benefits of pre-specification need acknowledgement. Pre-specifying subgroup analyses guards against undisclosed data-dredging. (The problem of multiple testing still arises if many subgroup analyses are pre-specified, but at least it is explicit.) Pre-specification also provides the opportunity for

[14]This assumes the continued use of frequentist statistics. Other statistical methods will define 'reliable' in different ways.

incorporating the pre-specified subgroup analyses into the trial design. For instance, stratified randomisation may be considered. In stratified randomisation patients are randomised *within* the specified subgroup. This ensures roughly equal numbers are randomised into treatment and control within each stratified subgroup. Pre-specification also allows investigators to ensure data collection is appropriate for the subgroup under consideration. *Post hoc* creation of subgroups may rely on unsuitable data if the data utilised was not collected with this purpose in mind.

But none of this rules out potentially useful *post hoc* analysis of subgroups being conducted. And such analyses, even if they are informal, can play a vital role in judging whether the results of a given study are applicable to an individual patient. Rothwell's guidelines for subgroup analyses continues to be silent on the problem of how to interpret anticipated (or at least foreseeable) subgroup effects that were not specified within the trial protocol.

In the final section, I consider Patrick Suppes' hierarchy of data models. My aim in doing so is to provide an explicit framework for incorporating basic science into the interpretation of subgroup analyses. Such a framework is implicit in Rothwell's guidelines (though, as I have shown, the role for basic science is limited to the specification of the analysis undertaken in the trial). The hierarchy of data models doesn't solve the problem of subgroup data, but it does illustrate the links between basic science and the statistical findings of applied clinical research. By shedding light on the links between basic science and the statistical analysis of clinical trials, this framework also opens the way for a more flexible approach to the interpretation of clinical research. This is especially important for informing therapeutic decisions (as opposed to testing the efficacy of a treatment).

7.5 The hierarchy of data models and the analysis of subgroup data

All participants in the subgroup debate agree that clinicians need reliable subgroup analyses. There is, however, considerable scepticism regarding the possibility of conducting such analyses reliably. This scepticism is not limited to the trialists, most of the pathophysiologists also share this concern. Indeed, given the standard statistical approach to analysing clinical trials, and the intrinsic statistical properties of subgroup analyses in these trials,

scepticism is perhaps the appropriate attitude—at least if we adhere to the methods employed in testing the efficacy of treatments. However, the methods employed to test a treatment's efficacy are general and conservative. Any method that *may* lead to error when applied as a general rule is repudiated—even, if there are particular circumstances in which the method may not lead to error. Limiting consideration of basic science to the specification a trial, and ruling out *post hoc* analyses are good examples of this conservative attitude.

Applying clinical research to individual patients raises a different set of questions. While conservative methods may be appropriate when we want to ensure that a treatment is efficacious—once again, tests of efficacy play an important role in regulatory decisions about whether a drug should be marketed—adhering to these methods when wanting to inform therapeutic decisions involving individual patients leaves clinicians hamstrung. However, if we are to relax some of these normative methodological rules, it is important that it is not done in a way that disproportionately increases the risk of error; it is, of course, *reliable* subgroup analyses that clinicians need. Suppes' formal framework is one way of providing a principled account for determining the circumstances under which particular subgroup analysis may be reliable.

Suppes' hierarchy of data models provides a plausible account of the links between the theory of basic science with the statistical findings of applied clinical research (Suppes 1962). Recall, at least three models are at work in any clinical trial, a model of the theory, a model of the experiment, and a model of the data. The hierarchy of data models can be put work on two important tasks within therapeutic decision making. First, the hierarchy of data models rejects the separation of basic science from the statistical results of applied clinical research; a separation that is implied by EBM's hierarchy of evidence. Basic science is integral to the specification, analysis and interpretation of applied clinical research (as seen in the previous chapter). Second, this account of the relation between theory and data provides a method for distinguishing between subgroup specific results that are the result of random error, and those results that may be due to an underlying process. Subgroup specific results for which the necessary assumptions of the theoretic, experimental, and data models hold, have more warrant than subgroup specific results for which the assumptions of the models do not hold.

Trialists deny basic science a role in analysing and interpreting applied clinical research. The only time basic science is appealed to is in the specification of the trial. Trialists view basic science as unable to discern between true subgroup-specific variation in treatment effects, and random variation. (Or, as Altman expresses the view, 'doctors seem to be able to find a biologically plausible explanation for any finding'.) In a number of circumstances the trialist's scepticism towards basic science playing a role in interpreting clinical studies is appropriate—especially when testing whether a treatment is efficacious.

The basic science theories that are incorporated into the specification of the clinical trial are tested by the data. If the assumptions of the theoretical, experimental and data models hold, and the data support the basic science, then the basic science can be considered to have received a degree of support. This is considered appropriate by the trialist, but any appeal to basic science that was not included in the specification of the trial is rejected. This rejection of the use of basic science in interpreting the results of clinical trials is appropriate when there is no strong independent support for the basic science being appealed to. As already discussed, subgroup specific variation in the effects of treatment are to be expected on the basis of random error. Basic science without strong independent plausibility neither explains, nor is confirmed by an observed variation in a particular subgroup.

Ruling out a role for basic science that has not been incorporated into the specification of the trial may be justifiable in tests of efficacy. Tests of efficacy more neatly fit into the standard statistical analysis of clinical trials. Efficacy can be formulated as a single question, and a trial can be powered to adequately test this question. Also, the basic science that underpins tests of efficacy is likely to be fairly well established. The basic pharmacological properties will have been confirmed in extensive animal testing, and the basic safety profile will have been tested in pre-clinical studies. And because the test of efficacy is focussed on a single question, all the basic science that is anticipated to be relevant to this question can be included in the specification of the trial. In this situation, basic science that was not involved in the specification of the trial, *appropriately* plays a limited role in the interpretation of whether the therapy is considered efficacious. Notably, subgroup data plays a limited role in tests of efficacy—again, the question is simply whether the treatment is beneficial overall for a defined sample of patients. Subgroup data is more relevant for questions of effectiveness; more on this

in a moment.

Post hoc analyses should also be approached with caution in tests of efficacy. First, as discussed, there are a number of ways *post hoc* analyses can be done poorly. Data dredging and data-dependent analyses increase the risk of erroneous inference. Second, with a focussed question like that of testing the efficacy of a treatment, *post hoc* analyses should not be required. The proposed efficacy of the treatment can be defined prior to the trial by an outcome on a single endpoint (or perhaps, a small number of endpoints combined). If the data from the trial are inconclusive, or do not find the therapy efficacious according to this measure, focus on alternative endpoints, or other *post hoc* analyses are rightly eyed with suspicion.

If the treatment's claims to efficacy from data in a clinical trial rely on anything other than the pre-specified analyses, then alarm bells should ring. Questions of effectiveness, relying on the external validity of clinical trials, are different. They are not the central focus of most clinical trials, and the question will vary considerably depending on the patient for whom the results of the trial are to be applied. In this context, a more flexible approach to the interpretation of data from clinical trials is required.

A different criteria should be applied to questions of effectiveness as opposed to questions of efficacy. The contemporary methods in clinical epidemiology emphasised by proponents of EBM are particularly focussed on rigourously testing whether treatments are efficacious. Randomised allocation, prospective interventional studies, intention-to-treat analysis and the like are employed to improve the internal validity of applied clinical research. And, providing what these methods achieve is suitably understood, they are successful in this task. Therapeutic questions however are questions of how well, and in whom, the treatment works in routine care; therapeutic questions are questions of effectiveness. And as such, these questions require judgement of the external validity of clinical research. Employing the same normative methodological rules to these questions—for instance, rejecting *post hoc* analyses, and a role for basic science in the interpretation of data—risks rendering such questions unanswerable. Trialists find themselves in this position, they are caged in by the poor intrinsic statistical properties of subgroup analyses and their denial of basic science playing a role in interpreting clinical research.

Not all *post hoc* appeals to basic science are necessarily fallacious. The hierarchy of data models provides a framework for considering subgroup anal-

yses. Basic science can play an important role in interpreting clinical trial data—provided that the basic science appealed to is *independently* plausible. Basic science may explain observed variation in subgroups only if good independent reason for that basic science can be provided. The extent to which such appeals to basic science are credible relies heavily on the strength of this independent evidence. The same goes for *post hoc* analyses more generally. Data-dredging, and data-dependent analysis should always be avoided, and the risk of false positives for multiple pre-specified tests explicitly taken into consideration. But *post hoc* analyses that have independent justification, whether that be via independently plausible basic science or reliable data from other clinical trials, can be appropriate. The strength of the interpretation they provide depends directly on the strength of the evidence independent of the trial under discussion.

The hierarchy of data models illustrates the argument that has to be sustained to support interpreting an observed variation in a subgroup as genuine. A consistent representation of the underlying process is required from the three tiers of models, the model of the theory, the model of the experiment and the model of the data. ISIS-2 provides an example. Recall, ISIS-2, among other things, assessed whether streptokinase reduced death due to vascular disease in patients presenting with acute myocardial infarction. Overall, streptokinase was found to be beneficial in a wide range of such patients. One subgroup in which benefit from streptokinase was not observed were patients with ST segment depression on their pre-randomisation electrocardiogram. The question that arises is whether this observed variation is genuine, or best attributed to chance.

There are independent pathophysiological reasons for expecting patients with ST depression on their electrocardiogram to derive less benefit from fibrin-lysing drugs such as streptokinase. Patients suffering from acute coronary syndrome typically present with either ST depression or ST elevation on their electrocardiogram. In patients suffering an acute myocardial infarction, their electrocardiogram often evolves from ST depression, which signifies that the heart is receiving less oxygen, to ST elevation, which signifies that damage to heart tissue from lack of oxygen is beginning. ST depression suggests an early platelet-rich clot, which is only partly occluding the coronary artery. ST elevation suggests the clot has progressed, that it is rich in fibrin, and that it has completely occluded the artery. Thus, pathophysiology predicts that streptokinase—which dissolves the fibrin in the clot that is occluding the

coronary artery—will particularly benefit patients with ST elevation on their electrocardiogram. In addition to this plausible pathophysiologic rationale, many of the assumptions of the experimental and data models that hold for the overall analysis will hold for this subgroup. Because patients were randomised after their electrocardiogram, whether or not they present with ST elevation is not due to some aspect of treatment. And, because patients are randomised, it can be assumed that patients in the subgroup who received streptokinase are roughly similar to patients who received placebo. This can, and ideally should, be checked using standard statistical analyses (though, the demographic data of this subgroup was not reported in the original publication). Because the endpoint under discussion is the primary endpoint of the trial, the data model, test statistic, and related assumptions are the same for this subgroup as in the overall analysis. In addition, further support for the variation observed in this subgroup being genuine, is provided by data available prior to ISIS-2 (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) 1986).

The strength of the argument supplied above reduces the possibility that the reduced benefit observed from streptokinase in patients with ST depression is due to chance. Compare this with other subgroup specific results observed in ISIS-2. The observed variation of effect from aspirin seen in patients born under different star signs provides a particularly stark contrast. Overall, patients in ISIS-2 that received aspirin had better outcomes than those who did not receive aspirin. Patients born under Libra or Gemini, however, did not receive the benefit.

The argument provided for interpreting the variation in results witnessed in patients with ST depression as genuine can not be provided in the star sign subgroups. No basic science rationale for the variation observed in outcomes in patients born to different star signs can be supplied. And the method by which these results were found can be questioned. What justification can be given for combining results from patients born under Libra and Gemini and comparing them to patients born under other star signs? The only apparent rationale for combining patients in this way is that patients born under Libra and Gemini were observed to respond less to aspirin—a clear case of data-dependent analysis. Further, it is unknown how many other analyses were conducted before this variation was observed. What other patient characteristics were checked? Not only was the analysis provided by the star sign subgroups data dependent, but it likely arose after a process

of data dredging. By contrast, while assessing the effects of streptokinase in patients with ST depression versus ST elevation was not part of the primary statistical analysis in ISIS-2, at least the subgroup was pre-specified, and an independent rationale for creating the group can be provided.

The trialists do not have the framework to differentiate between the *post hoc* subgroups created by ST-elevation or ST-depression on the electrocardiogram, and the *post hoc, data-dependent and data-dredged* subgroups created according to birth sign. Rather, the trialists use the later to argue against the former on general grounds (recall the quote provided on page 143). By contrast the framework provided by the hierarchy of data models clearly distinguishes the two examples.

It is important not to overstate what can be achieved by explicitly incorporating Suppes' framework into the interpretation of subgroup data. Interpreting subgroup analyses is difficult, and there is a substantial risk of error when the emphasis of the analysis shifts once the data are available. Subgroups should be formed on the basis of independent biologic or pathophysiologic information, rather than an observed variation in the effect of a treatment. Whether the independent information from basic science is a plausible explanation of the observed variation is a matter of judgement. And while judgement is also required when the analyses are pre-specified, additional caution is needed because the focus on the subgroup has come about after the data has been observed. It is important to acknowledge the different warrant provided by pre-specified and *post hoc* analyses.[15]

In most cases, interpretation of subgroups will be more difficult than the example provided by ISIS-2. But this does not diminish the valuable contribution the hierarchy of data models makes to the interpretation of subgroup analyses. To make therapeutic decisions, clinicians need to make judgements about how data observed in clinical research applies to their patients. Adhering, without reflection, to the methodological rules put in place to reliably test whether a treatment is efficacious hinders making such judgements. The hierarchy of data models provides a framework for judging whether an ob-

[15]It is also important to acknowledge that while basic science plays an important role in understanding and explaining the results of clinical research, it is neither sufficient nor necessary. There is good evidence for the effectiveness of many contemporary treatments that nevertheless lack a clear pharmacological explanation of how they work. In these situations reproducibility, and an absence of alternative explanations for the results, is sufficient.

served variation in a subgroup may be genuine—as well as a framework for assessing the judgements of others. Using Suppes' framework in this way provides an improvement on the methodological conservatism of the trialists, and explicates a process by which the pathophsiologists can incorporate basic science into the appropriate interpretation of applied clinical research.

7.6 Conclusion

Analysing subgroup results in light of the hierarchy of data models redresses an imbalance present in the clinical literature. EBM proposes that therapeutic decisions be based on the evidence supplied by applied clinical research, especially randomised interventional studies. To achieve this, clinicians need to be able to apply the results of clinical research to the patients under their care. This requires the clinician to make judgements about the external validity of clinical research, and the effectiveness of the treatments that have been assessed. Proponents of EBM provide little helpful advice in how these judgements should be made.

Statistical results can not be interpreted in isolation—even if the basic science is uncertain, this in itself is relevant to the appropriate interpretation of the statistical results of clinical trials. On the basis of the conservative methods utilised to test the efficacy of treatments, trialists warn against the use of any subgroup analyses from randomised trials. This presents a difficulty for decision makers. Not all types of patient within the trial will respond equally to treatment, but EBM, as described by proponents, provides no method for ascertaining which groups of patients are more likely to benefit or be harmed from treatment. A clearer separation of questions of effectiveness from questions of efficacy, a recognition of the importance of subgroup analyses and external validity to clinicians, and a framework for considering the question of when an observed variation of treatment effect in a subgroup is likely to be genuine, provide an important, if partial, response to the challenge of interpreting subgroup analyses.

Chapter 8

Power and Inference: The Rofecoxib Case

8.1 Introduction

Conservative methodological and statistical rules are employed to test the efficacy of therapies. And, for the most part, there are good reasons for employing these rules—providing, of course, the strengths and weaknesses of these rules are recognised. Assessing the *effectiveness* of therapies, however, raises substantially different questions. And to answer these different questions we need different resources. This chapter has two aims. First, highlight, through an important contemporary case, that therapeutic decisions are often made on statistical tests that are not optimal according to Neyman-Pearson methods. And second, following from the reliance of therapeutic decision makers on less than optimal Neyman-Pearson tests, I use the rofecoxib case to examine whether there are situations in which we can appropriately relax the statistical strictures in place for tests of efficacy.

Since randomised interventional studies are analysed according to frequentist statistics, EBM explicitly ties frequentist methods to therapeutic decision-making. While a number of general philosophical criticisms have been levelled at frequentist statistics (for instance, Howson and Urbach 2006), the dominance of these methods in the analysis of clinical trials goes undiminished. Granting the frequentist methods on their own terms, I argue there are problems in how these methods are *applied* in the analysis of clinical trials. Specifically, while the normative rules supplied by frequentist methods

can be understood for Neyman-Pearson tests set up optimally, clinical trials typically conduct a number of additional tests that are not set up in this way. These additional tests are important to therapeutic decisions, and should *not* be interpreted in the same way as optimal Neyman-Pearson tests

The controversy surrounding rofecoxib well illustrates the use of frequentist methods to inform therapeutic decisions. Recall, rofecoxib reduces pain and inflammation in conditions such as rheumatoid arthritis. A lower risk of gastrointestinal adverse effects has been observed with rofecoxib (and similar agents) compared to traditional antiinflammatories (such as aspirin, ibuprofen and the like) (Bombardier et al. 2000). Introduced with a heavily marketed promise of superior safety, rofecoxib was widely prescribed in many countries.[1] However, after only five years on the market, rofecoxib was withdrawn following the Adenomatous Polyp Prevention Trial (APPROVe). APPROVe found patients taking rofecoxib suffered a *statistically significant* increase in heart attacks and strokes compared to patients taking placebo (Bresalier et al. 2005).

APPROVe was not explicitly set up to test whether rofecoxib increases the risk of thrombotic events (such as heart attacks and strokes). Indeed, as will be shown, evidence external to the trial suggests APPROVe had low *power* to test this outcome. The power of a statistical test is the pre-experimental probability that the statistical test will yield a statistically significant result for the measure of a particular outcome on the assumption the null hypothesis is false (see Section 1.2 for discussion). Power is calculated either on the basis of prior evidence, or practical considerations of what is considered a clinically important effect. Well designed clinical trials will have a power of 80–90% to test the the primary clinical endpoint. Rather than to provide a powered test of whether rofecoxib increases thrombotic risk, APPROVe was set up (and powered) to test whether rofecoxib would prevent the relapse of adenomatous polyps. Since the most important inferences made on the basis of APPROVe relate to the risk of thrombotic events (rather than adenomatous polyps), APPROVe raises the question of what role statistical power should play in the inferences we draw?

Power is of interest to this analysis because it is the way in which Neyman-Pearson methods take into account evidence regarding the unknown parameter under investigation. APPROVe's statistical test of rofecoxib's thrombotic

[1]Cutts et al. (2002) show that rofecoxib was prescribed to a broad range of patients, many of whom were older and suffering multiple pathologies.

risk was set up without consideration of the available evidence regarding this risk—*this* is my primary concern. The question is whether this lack of explicit consideration of the available evidence regarding the thrombotic risk of rofecoxib is relevant to the interpretation of the statistical test provided by APPROVe. The focus on power will seem somewhat convoluted to a Bayesian. On the Bayesian account the available evidence regarding the unknown parameter will be included in the prior probability distribution. As seen in Section 1.2, for the frequentist conducting a Neyman-Pearson test, the available evidence is incorporated into the statistical test via the pre-experimental α and β error rates. β and power, $(1 - \beta)$, is especially important because β is a function of the test statistic under the assumption of the alternative hypothesis. This, however, is only done for the primary endpoint in a clinical trial. Rofecoxib raises the question of how best to interpret secondary endpoints when evidence external to the trial suggests the statistical test is under-powered.

Within medical statistics, power is considered important *prior* to conducting the experiment. It is widely accepted that trials should only be conducted if they have a reasonable chance of providing a result that is conclusive within the terms of frequentist statistics—that is, provide a statistically significant result that rejects the null hypothesis. Ensuring the statistical test is well powered for the primary clinical endpoint provides this assurance. Power is also seen as relevant to the interpretation of a statistical result when the result of the test is 'non-significant' and the power of the test is low. In such a scenario the statistical test is considered inconclusive *because* the test is underpowered.

Once a *statistically significant* result has been *observed*, however, the standard view holds power irrelevant to the *interpretation* of the result. If the result of the test is statistically significant, then the null hypothesis is rejected (according to the terms of frequentist statistics), and the alternative hypothesis accepted. I will label the view that the power of a test is irrelevant once a statistically significant result has been observed the 'standard view'. There are many examples of this view in the clinical literature. Here is one from a well known textbook on clinical epidemiology:

> Calculation of statistical power based on the hypothesis testing approach is done by the researchers before a study is undertaken to ensure that enough patients will be entered to have a good chance of detecting a clinically meaningful effect if it is present.

> However, after the study is completed this approach is no longer as relevant. There is no need to estimate effect size, outcome event rates, and variability among patients; they are now known. (Fletcher et al. 1996, p. 200)

Part of the aim of this paper is to illustrate when this view is appropriate and when it is not. Specifically, I will show that this view can be justified for adequately powered endpoints (such as is the case for many tests of efficacy). By contrast, however, this view cannot be justified when evidence external to the trial suggests the statistical test is underpowered (which often occurs for secondary endpoints).

In recent years more emphasis has been given to the frequentist approach to estimation. This is because estimation theory provides more information for therapeutic decisions. The frequentist approach to estimation uses the observed data to *estimate* the true value of the unknown parameter. As noted previously, despite the different focus, both hypothesis testing and estimation can be outlined within the same framework. And, both hypothesis testing and estimation are employed (often simultaneously) in drawing inferences from clinical trials.[2] It is difficult to speculate what would have happened if an increased, though not statistically significant, risk was observed in APPROVe. But the immediate withdrawal of the drug, and the international controversy that ensued, seems much less likely. It is generally the case in medical research that estimation becomes important once a hypothesis has passed a statistical test, and this is specifically so for APPROVe. For this reason, and because power is most naturally discussed in terms of hypothesis testing, I will focus on the frequentist approach to hypothesis tests. I stress, however, that the comments I make in relation to power are easily transferable to the theory of estimation.

EBM places a greater focus on the results of randomised interventional studies, and this focus extends beyond the results of the primary hypothesis test. In practice, in each trial many statistical tests are conducted, and inferences drawn, on endpoints in addition to the primary endpoint. Whereas frequentist statistical methods ensure the optimal warrant for the tests set up according to ideal Neyman-Pearson methods, this warrant does not extend to

[2]See Gardner and Altman (1989) and Ware et al. (1992) for discussion regarding the use of estimation rather than hypothesis testing. Both p values and confidence intervals are reported in clinical trials (as seen in APPROVe). Most who argue for confidence intervals argue for them as a valuable *addition* to the reporting of p values.

the additional statistical tests. I suggest the *application* of the standard view to these additional statistical tests in clinical trials is problematic. While the standard view provides a clear interpretation of statistical tests set up according frequentist methods, such as is typically the case for primary hypothesis tests, this interpretation can break down in the case of secondary endpoints and subgroup analyses.

Drawing appropriate inferences in relation to secondary endpoints and subgroup analyses is all the more important due to the significance of these analyses to therapeutic decisions. As the 'subgroup debate' discussed in the previous chapter makes clear, analyses of secondary endpoints and subgroups play a role in informing decisions regarding *individual* patients. Judging whether results of a trial are relevant to a particular patient takes into consideration the results of the trial in the subgroups most relevant to the patient. Further, most primary hypotheses test whether a therapy *benefits* a group of patients, this means safety endpoints are relegated to secondary hypotheses. As therapeutic decisions require clinicians to weigh up both the benefits and harms of a therapy, analyses of secondary endpoints are crucial. This creates a well recognised problem for therapeutic decision makers. As Collins and MacMahon note, there is an

> [...] unavoidable conflict between the reliable subgroup-specific conclusions that doctors and their patients want, and the unreliable findings that subgroup analyses of clinical trials might offer. (Collins and MacMahon 2007, p. 13)

Alvan Feinstein's (1998) description of the tension between what frequentist methods provide, and clinicians need, is apt (if somewhat dramatic): the clinician's need for a reliability the statisticians can't provide is a 'clinico-statistical tragedy'. The standard approach to this problem is to rely on systematic reviews and meta-analyses. Collating the subgroup analyses or secondary endpoints from a number of similar clinical studies will improve the reliability of these analyses if a number of sufficiently similar clinical studies are conducted. But this, of course, relies on similar studies being conducted. And even if they are, the time it takes these studies to be conducted creates a substantial problem when therapeutic decisions need to be made in the interim. I grant that subsequent meta-analyses and systematic reviews may eventually provide the necessary evidence for therapeutic decisions, but to see this as a *reply* to the problems of subgroup analyses and secondary

endpoints is wrong-headed. Surely, better understanding precisely why, and when, subgroup analyses and secondary endpoints are unreliable is vital for any adequate reply to the problem.

It is only appropriate to consider power after a statistical result has been observed in certain, well defined, circumstances. Some of the attempts that have been made to interpret statistical findings in relation to the power of the test are patently incorrect. For instance, 'post-hoc power', where the power of the test is calculated based on the estimate provided by the trial, is a clear case of contorted reasoning. By definition, any result that is non-significant will have a low 'post-hoc power', and any result that is significant will have high 'post-hoc power' (Goodman and Berlin 1994 provide discussion). The power of a test can be appropriately calculated in one of two ways: on the basis of available evidence regarding the parameter under investigation, or in the absence of any such evidence, on the basis of the smallest clinically important effect.

In clinical trials statistical tests take place within one of the following three contexts (i) the test can be adequately powered based on the available evidence, (ii) the test can be adequately powered, in the absence of evidence, on the 'smallest clinically important effect', or (iii) the test can take place when evidence external to the trial suggests the test is underpowered. I will call this the 'epistemic context of the statistical test'. In general, the standard view ignores the epistemic context of the statistical test. For instance, while evidence external to the trial suggests APPROVe was underpowered to test whether rofecoxib increases the risk of thrombotic events, this evidence is considered irrelevant (or inadmissible) when interpreting the observed statistically significant result. I argue that in such situations evidence external to the trial undermines the warrant of the statistically significant result. The insensitivity of the standard view to evidence external to the trial means that the reduced warrant of the statistical test goes unrecognised. Clearly, this is a serious problem when frequentist methods are being used to inform therapeutic decisions.

8.2 The Rofecoxib Case

The main features of the rofecoxib case are outlined. Two related aspects of the frequentist statistical approach are important: hypothesis testing and estimation. To aid exposition I shall simply report the standard statisti-

cal findings relevant to the case. The statistical concepts employed were discussed in Section 1.2, and will be reviewed in the following section.

APPROVe was set up to test whether rofecoxib would reduce recurrent gastrointestinal adenomas. APPROVe recruited male and female patients who were over 40 and had at least one large-bowel adenoma removed within 12 weeks of entry to the study. Participants were randomised to treatment with rofecoxib or placebo. It was planned that the study would continue for three years. Due to the plausible pharmacological rationale for rofecoxib increasing the risk of thrombotic events, and some suggestive (though inconclusive) clinical trial evidence, an independent committee was installed to monitor cardiovascular events. The rate of thrombic events was a secondary clinical endpoint in the APPROVe trial. This endpoint was used to test the hypothesis that rofecoxib increases the risk of thrombotic events. Call the hypothesis that rofecoxib increases the risk of thrombotic events, H_A. Remember, despite H_A being the hypothesis everyone is interested in, it was not the primary hypothesis APPROVe was designed to test.

H_A can be specified in a number of ways. To be consistent with the analyses conducted in APPROVe, specify H_A as the following: let μ_R be the unknown incidence of thrombotic events in patients taking rofecoxib, and μ_C be the incidence of thrombotic events in patients taking control. H_A holds that the relative risk of thrombotic events is greater than one: $RR_\mu = \mu_R/\mu_C > 1$. The *null hypothesis*, H_0, holds that the relative risk of thrombotic events equals one: $RR_\mu = \mu_R/\mu_C = 1$. There is also an estimation problem at hand. If H_A is supported the magnitude of this increased risk becomes important. Therapeutic decisions are an act of weighing benefits with possible harms. Estimating effect size is vital for this deliberation. If the thrombotic risk posed by rofecoxib is real, but sufficiently rare, the marginal gastrointestinal safety may justify its use in some patients. The magnitude of the increase in risk of thrombotic events due to rofecoxib is estimated by the absolute risk increase, $ARI_\mu = \mu_R - \mu_C$.

Although no formal stopping rule was in place, the study was halted 2 months early due to a statistically significant increase in the rate of confirmed thrombotic events (this combined endpoint included cardiac, cerebrovascular and peripheral vascular events). The relative risk for having a confirmed thrombotic event in the rofecoxib group was 1.92 (p value = 0.008; 95% confidence interval, 1.19–3.11) (Bresalier et al. 2005). This increase in risk was predominately made up of patients suffering heart attacks and strokes. AP-

PROVe estimates an ARI_{μ} of 0.72 thrombotic events per 100 patient years.[3] The VIGOR study estimated treatment with rofecoxib rather than naproxen resulted in a absolute reduction of 0.8 gastrointestinal events per 100 patient years (Bombardier et al. 2000). Hence, APPROVe suggests taking rofecoxib increases the risk of suffering a heart attack or stroke roughly as much as it reduces the risk of suffering a gastrointestinal event. If this estimate is accurate then it would be difficult to argue for a widespread role for rofecoxib.

The decision of Merck, the manufacturer of rofecoxib, to withdraw the drug would have been due to a range of factors (including evidential, ethical, legal and financial). There is little doubt, however, the findings of APPROVe were crucial. (Rofecoxib was marketed under the name Vioxx).

> Merck has always believed that prospective, randomized, controlled clinical trials are the best way to evaluate the safety of medicines. APPROVe is precisely this type of study—and it has provided us with new data on the cardiovascular profile of VIOXX. While the cause of these results is uncertain at this time, they suggest an increased risk of confirmed [cardiovascular] events beginning after eighteen months of continuous therapy. While we recognize that VIOXX benefited many patients, we believe this action [withdrawal] is appropriate.[4] (Merck 2004)

The statistically significant increase in thrombotic events observed in APPROVe played a vital role in the withdrawal of rofecoxib.

But evidence regarding the thrombotic risk of rofecoxib was not incorporated into the specification of the test conducted in APPROVe. This raises the question of whether the power of APPROVe to test H_A is relevant to

[3] In addition to point estimates, the approach to estimation used in clinical trials provides an interval. The 95% confidence interval for the absolute risk of thrombotic events observed in APPROVe was 0.15–1.62 per 100 patient years. Interpretation of this interval (in relation to RR_{μ}) is discussed in the following section.

[4] It is questionable whether randomised interventional studies actually *are* the best way to evaluate medicines. Indeed, the rofecoxib case provides a good counter-example to this claim—specifically, the delay in identifying the adverse effects of rofecoxib despite the many randomised interventional studies conducted prior to APPROVe. As I have argued throughout the thesis, randomised interventional studies and the statistical methods we employ to analyse these studies are often not the best source of evidence regarding safety outcomes (see Section 5.3 for discussion). Further, it should be noted the suggestion that confirmed cardiovascular events begin *after* eighteen months of therapy is based on an error. For discussion, see Lagakos (2006).

the interpretation of the observed statistically significant result. There are two senses in which APPROVe was not powered to test H_A. First, a powered test of H_A was not explicitly part of the set-up of APPROVe. This is equivalent to pointing out that H_A is not the primary hypothesis under test. The second sense is more important to this analysis. The second sense refers to evidence on rofecoxib's thrombotic risk, and *on the basis of this evidence* suggests APPROVe was under-powered to test H_A.

A recent meta-analysis by Kearney et al. (2006) provides an estimate of the thrombotic risk of rofecoxib. This estimate can be used to calculate the power APPROVe had to test H_A. Thirty-seven trials comparing rofecoxib and placebo are included in the meta-analysis. The relative risk for thrombotic events in these trials appears similar to the overall estimated relative risk for thrombotic events for the class of COX-2 inhibitors (the actual figure for rofecoxib is not reported). Taking θ_A to be 1.42 (the reported relative risk) it is possible to calculate the power APPROVe had to test H_A: approximately 14%.[5] Recall that a large clinical trial is expected to have a power of 80–90% to test the primary hypothesis. Clearly APPROVe's power to test H_A is considerably lower than recommended.

Specifying θ_A in this way is contentious for a number of reasons. First, there are substantial problems on relying on the Kearney meta-analysis to estimate θ_A. Perhaps most notable, the meta-anlysis was published after, and thus includes, APPROVe. While the influence of APPROVe on the estimate for the combined placebo controlled trials is not reported, it plausibly contributes markedly.[6] Hence, if APPROVe is brought into question so too is the estimate provided by Kearney's meta-analysis. There is, however, no easy way to overcome this problem. Because Kearney includes all important placebo controlled trials involving rofecoxib, the meta-analysis arguably

[5]Power calculators using the binomial distribution are available online (for example, `http://www.swogstat.org/stat/public/Binomial/binomial.htm`). A power of 22% for APPROVe's test of H_A is based on the following: total sample size of APPROVe, $N = 2586$; the proportion of events in the placebo group as per APPROVe, $P_1 = 0.020$; the proportion of events in the rofecoxib group (calculated on the basis of $\theta_A = 1.42$), $P_2 = 0.026$; and a one-sided $\alpha = 0.05$.

[6]Most of the 37 placebo controlled trials included in the meta-analysis were of short duration (4–12 weeks). The only long term studies (longer than 12 months), other than APPROVe, which compared rofecoxib and placebo, were three trials in Alzheimers disease—none of which showed a significant trend for rofecoxib to increase thrombotic events (Aisen et al. 2003; Reines et al. 2004; Thal et al. 2005).

provides our current *best guess* for θ_A. It is not ideal, but it is the best we have. Also, the Kearney meta-analysis provides a conservative estimate for the claim that APPROVe was underpowered to test H_A. It is, of course, reasonable to ask what inferences should have been made based on evidence available at the time APPROVe was conducted. At this time the results of systematic reviews, and other data, gave conflicting accounts of rofecoxib's thrombotic risk. An estimate drawn on the basis of this data would would be lower than that provided by Kearney; and, thus, suggest that APPROVe had even lower power to test H_A.

Second, and perhaps more importantly, specifying θ_A on the basis of *any* evidence is highly contentious. Recall, the standard view holds that once a statistically significant result has been observed, the power of the test is irrelevant. On this view, APPROVe *confirms* H_A, and the magnitude of the observed risk is used to inform therapeutic decisions. This however is precisely the view I wish to question.

The healthcare community's response to APPROVe was consistent with the standard view. The manufacturer withdrew rofecoxib from the international market. And, the overwhelming response of the medical fraternity was to question how it took so long to confirm such an important adverse effect in a drug that had passed through all the regulatory processes of drug development, and been so widely prescribed.[7]

It is important to acknowledge that APPROVe does not provide the only evidence supporting H_A. Prior to APPROVe, there was both a plausible pharmacological rationale for how COX-2 inhibitors may increase the risk of thrombotic events, and evidence from, VIGOR, a randomised interventional study (Bombardier et al. 2000). VIGOR examined whether patients taking rofecoxib had less 'adverse gastrointestinal events' than those taking naproxen—such 'events' include gastric or duodenal ulcers, and related complications. Patients taking rofecoxib in this study did have less adverse gastrointestinal events but they also had statistically significantly more thrombotic events (observed as an increase in heart attacks). Since the withdrawal of rofecoxib, the findings of some randomised interventional studies, meta-analyses and observational studies also provide support for H_A (Kearney et al. 2006; Graham 2006; Juni et al. 2004; Kerr et al. 2007).

A couple of points are important. While there existed a range of evidence

[7]See, for example, the following editorials from leading medical journals: Abbasi (2004); Okie (2005); Topol (2004).

that supported H_A prior to APPROVe, this evidence was not considered sufficient for severe prescribing or regulatory restrictions. H_A was not considered sufficiently tested, perhaps, because alternative hypotheses were available. The pharmacological rationale for H_A, for instance, was not considered confirmed as it had not been demonstrated in a randomised interventional study. And while VIGOR provides some evidence from a randomised interventional study, the study investigators put forward an alternative to H_A as an explanation of the observed risk, namely, that naproxen prevented thrombotic events (Bombardier et al. 2000).

In contrast to some commentary,[8] I think a level of uncertainty regarding whether rofecoxib increased the risk of thrombotic events was appropriate. While some evidence supported a small increase in thrombotic risk associated with rofecoxib (and some of the more dismissive claims made in VIGOR are dubious[9]), the available data were far from conclusive. An industry sponsored meta-analysis, which combines much of the evidence prior to APPROVe, failed to show an increased incidence of thrombotic events in patients taking rofecoxib (Weir et al. 2003). In addition, a very large study involving lumiracoxib, a member of the same class of drugs as rofecoxib and a considerably more selective inhibitor of COX-2, failed to detect a statistically significant increase in risk of thrombotic events despite enrolling nearly 18,000 patients (Farkouh et al. 2004).

The results of APPROVe played an important (though not singular) role

[8]Some argue there was sufficient support for H_A for a several years prior to APPROVe (Juni et al. 2004). To be clear, I think, what most of the commentary suggests is very reasonable: we should not have waited so long to test H_A. Even if there was not enough evidence to support H_A there was enough to show it needed investigation. Recall that APPROVe did not explicitly test H_A; what has been taken as confirmation of H_A was accidental.

[9]VIGOR excluded patients who were taking aspirin. Part of the VIGOR authors' explanation for the increased thrombotic risk observed in VIGOR relies on the following claims: (i) the study enrolled patients at high risk of thrombotic events, patients who should have been taking aspirin to prevent blood clots and (ii) that it is likely that naproxen was better at preventing thrombotic events in these patients. While pharmacologically plausible, these claims do not entirely account for the observed increase in thrombotic risk. An FDA report, with updated data from VIGOR, provides results contrary to what this explanation would predict. Statistically significantly more adjudicated thrombotic events were observed in the subgroup of patients who did *not* have an indication for aspirin thromboprophylaxis. See Food and Drug Administration Advisory Committee (2001), and Section 6.3 on page 131.

in supporting H_A. Those faced with the decision of whether rofecoxib should be used in a patient, or population, need to draw an inference based on available evidence. How should decision makers respond to the data observed in APPROVe? More particularly, how should the statistically significant result be interpreted when evidence external to the trial suggests the statistical test was underpowered?[10] I wish to question the standard view. To do so it is necessary to review aspects of the frequentist statistical set-up.

8.3 Frequentist analysis of APPROVe

Clinical trials are primarily analysed according to the methods proposed by Neyman and Pearson introduced in Section 1.2. Here I outline the analysis of APPROVe focusing on the two warrants provided by frequentist methods for rejecting a null hypothesis. Importantly, the second warrant, which relies on Neyman and Pearson's Fundamental Lemma, does not hold in this case.

The test statistic of interest is RR_X, which is equal to X_R/X_C, where X_R is the rate of thrombotic events in rofecoxib treated patients, and X_C is the rate of thrombotic events in patients receiving placebo. RR_X is an *estimator* of θ, the unknown 'true' relative risk for thrombotic events in patients taking rofecoxib. RR_X, X_R and X_C are random variables. H_0 holds that θ_0, the value of θ assuming the null hypothesis is true, equals one. Since there is considerable evidence regarding the use of rofecoxib, I propose using this evidence to select θ_A, the value of θ proposed by H_A (specifically, the estimate provided by the Kearney meta-analysis discussed in the previous section).

Because H_A is not the primary hypothesis under test, APPROVe was not powered to test this hypothesis. The point of contention is whether the H_A should be specified, and power—or more directly, the available evidence regarding the thrombotic risk of rofecoxib—considered retrospectively in interpreting the test. More on this in a moment.

[10]The focus in this paper is what *inferences* are warranted on the basis of APPROVe. This should be distinguished from what decisions should follow. For instance, it is perfectly reasonable within a decision theoretic framework to deny that APPROVe establishes that rofecoxib increases the risk of thrombotic events, but hold that nevertheless (due to considerations of the utilities involved) rofecoxib should be withdrawn from the market. By focussing on the first component (the warrant provided by the data in APPROVe) I hope to aid the second component (decisions involving whether rofecoxib should be prescribed).

A test statistic falling into the rejection region ensures that $P(RR_x \mid H_0)$ is low. For adequately powered tests, Neyman and Pearson's Fundamental Lemma provides the additional assurance that $P(RR_x \mid \theta_A)/P(RR_x \mid \theta_0)$ is maximised. The likelihood ratio is important to Neyman-Pearson methods. But it is of secondary importance. Frequentist considerations regarding the entire sample space predominate (see page 28 for further discussion). Typical clinical trials report many tests in which the power of the test is never considered. The standard view, which holds power irrelevant once a statistically significant result has been observed, sees no problem in this.

Estimation theory considers which values of θ are supported by the data. Given RR_X is an estimator of θ, the lower and upper bounds of a confidence interval can be calculated such that the pre-experimental probability of θ being within the interval can be specified, that is $P(T_\alpha < \theta < T^\alpha \mid \theta) = 1-\alpha$ for all $\theta \in \Omega$, where T_α and T^α are the lower and upper bounds respectively, and Ω is the parameter space. In APPROVe, the reported 95% confidence interval for the relative risk of thrombotic events is 1.19–3.11. The observed data is used to calculate t_α and t^α (1.19 for the lower bound and 3.11 for the upper bound). Recall, θ is an unknown constant, so the probability of θ falling into any interval is zero or one. The probability statement refers to T_α and T^α and *not* t_α and t^α. Put into physical terms, if the trial was repeated indefinitely, and if the assumptions of the specification are correct, then the means of the respective distributions of values for the observed t_α and t^α, would capture θ 95% of the time.

If RR_x is not a good representation of RR_X, or the assumptions made in the specification fail, then the calculated confidence interval may be misleading. An example, relevant to APPROVe, is when evidence from outside the trial suggests RR_x is not a good representation of RR_X; that is, when external evidence suggests RR_x is extreme. 'Good representation' here refers to the observed RR_x falling in the central range of the random variable RR_X.

8.4 Why the Epistemic Context of the Test Matters

On the standard view, once a significant result has been observed, H_0 is rejected, and H_A, along with the calculated confidence interval provisionally accepted. Importantly, the null and alternative hypotheses have a different

logical standing. Inferences are based solely on the sampling distribution of RR_X under the null hypothesis. Here the only time the alternative hypothesis is considered is in the set up of the test; and this is done only for the primary hypothesis under test in the trial. This ensures the test has a good chance of rejecting the null hypothesis when false. Cox and Hinkley make this point repeatedly.

> So far as the significance test is concerned, however, the null and alternative hypotheses are not on equal footing; H_0 is clearly specified and of intrinsic interest, whereas the alternatives serve only to indicate the direction of interesting departures. (Cox and Hinkley 1974, pp. 88–89.)

The standard view's interpretation of power is consistent with this asymmetrical attitude to the influence of θ_0 and θ_A. Power is seen to be completely separate from the analysis of data.

> A final general comment on the idea of power is that it is never used directly in the analysis of data. When we are interested in the relation between data y and values of the parameter other than the null value, an interval or other form of estimate will be needed; the power function is a general property of a test and does not give us information specific to particular sets of data. (Cox and Hinkley 1974, p. 105)

What these quotes assume, however, is that statistical tests are always strictly set up according Neyman-Pearson methods—that is, if evidence regarding the unknown parameter is available, it is part of the specification (context (i)), or, in the absence of such evidence, the test is powered based on the smallest clinically important effect (context (ii)). Of course, in many areas outside of clinical trials this assumption may well be reasonable. General accounts of frequentist statistical inference do not discuss how statistically significant results should be interpreted following a test that is underpowered according to the available evidence—Neyman-Pearson methods direct investigators to set up *powered* tests. If statistical tests are set up according to the method, the question of how to interpret an underpowered test does not arise. As discussed, in the clinical sciences the primary hypothesis test is set up according to Neyman-Pearson methods, but many additional tests

are conducted outside of these contexts. This results in a great deal of contention, and indeed confusion, about how such tests should be interpreted in the clinical sciences where subgroup analyses and secondary endpoints are important to decision makers. This is why situations arise in clinical trials in which there is evidence external to the trial that suggests the test is underpowered (that is, context (iii)). What I wish to highlight, and the standard view appears to ignore, is that this changes the warrant of a statistically significant result from such a test. Whereas the standard view can be justified in context (i) and (ii), it can not be justified in (iii). I now consider the justification that can be provided by frequentist statistics for the standard view. I start by considering context (ii): when there is no evidence available on which to power the statistical test.

8.4.1 Statistical tests in the absence of evidence relating to the alternative hypothesis

The standard view's distinction between the relevance of pre- and post-trial power *is* justified when there is no empirical basis to specify θ_A. Say H_0: $\theta = 0$, and the null hypothesis is fully specified. The trial is designed to test this hypothesis against the alternative hypothesis, H_A: $\theta > 0$, but no evidence supports picking any particular value for θ_A. Here the trial is typically powered to test the null hypothesis against the smallest clinically important effect, using the power curve for values of θ under H_A.

Suppose a statistically significant test statistic has been observed. Let's also accept that the investigators designed the study to provide an adequately powered test of the null hypothesis based on what they propose to be the smallest clinically important effect. While the value of the smallest clinically important effect may be contested, there is no legitimate basis on which to argue the test underpowered. Any such argument requires evidence supporting a particular value for θ_A. In the absence of this evidence it is appropriate to accept what the *trial data* tell us about H_0. That is, the trial data, assuming the specification of the trial, suggests a discrepancy from the null hypothesis.

This is reported in shortened form: reject the null hypothesis. Of course *reject* has a particular interpretation within frequentist statistics. It means that the *observed* test statistic, call it t, falls into the rejection region of the sampling distribution for the random variable T assuming the null hypoth-

esis. If α is set at 0.05 this means the observed value of the test statistic, or a value more extreme, would be expected less than 5 out of every 100 repetitions of the experiment.[11] Hence, assuming the specification of the test is correct, the first warrant of a statistically significant result is provided. And this is the best warrant Neyman-Pearson methods can provide in this situation. Given there is no evidence on which to specify θ_A, the second warrant does not apply.

8.4.2 Statistical tests when evidence relating to the alternative hypothesis is included in the specification

The standard view can also be defended when empirical evidence exists about H_A. But this defence hangs on setting up the trial according to Neyman-Pearson methods. Thus, the evidence supporting a particular value of the alternative hypothesis is included in the specification of the trial, and therefore the statistical test is powered on the basis of this evidence.

In this situation Neyman and Pearson's Fundamental Lemma ensures, for a test statistic that falls in the rejection region, the likelihood ratio, $P(RR_x \mid \theta_A)/P(RR_x \mid \theta_0)$, is maximised. Hence, the nature of the test tells us something about how the statistically significant data relates to the null and alternative hypotheses. Where evidence exists about the alternative hypothesis, and it is included in the specification, the Fundamental Lemma provides some assurance about the 'fit' of the data. If a statistically significant result is observed, the data support the alternative hypothesis (and the evidence used to specify the alternative hypothesis). Both of the warrants for a statistically significant result within the Neyman-Pearson approach are provided.

[11]The proposed magnitude of the smallest clinically important effect may well be contested. The statistically significant result, it may be argued, is not clinically significant. But this does not argue the discrepancy from the null hypothesis does not exist. Only that the discrepancy is of insufficient size to affect therapeutic decision making.

8.4.3 Statistical tests when evidence relating to the alternative hypothesis is *not* included in the specification

APPROVe raises difficult questions because the evidence relating to thrombotic risk was not considered in the set up of the trial. Based on available evidence a considerably larger sample of patients would have been necessary for APPROVe to provide a sufficiently powered test of H_0. If this larger trial was conducted, and a significant test statistic observed, then the Neyman-Pearson test would provide information on how the data relates to both H_0 and H_A. This result, and its associated confidence interval, would possess both of the warrants associated with being set up according to Neyman-Pearson methods. However, the hypothesis of interest in APPROVe was not set up according to Neyman-Pearson methods; based on the available evidence the test was underpowered. As a result the observation of a statistically significant test statistic for this hypothesis creates a tension between the two warrants supplied by Neyman-Pearson methods.

The first warrant of Neyman-Pearson methods goes through: the observed thrombotic risk data from APPROVe support a discrepancy from H_0. This is implied by the observed test statistic falling into the critical region of the sampling distribution under H_0. But evidence external to the trial also suggests that the observed data are unlikely under the assumption that H_A is the true hypothesis; that is, $P(RR_x \mid H_A)$ is also low. This undermines the second warrant. The low power of the test means the likelihood ratio, $P(RR_x \mid \theta_A)/P(RR_x \mid \theta_0)$, is not maximised for all values of the test statistic that fall in the critical region. This opens an interpretative dilemma. Either RR_x is an extreme sample point within the distribution of RR_X based on available evidence (that is, the evidence used to specify θ_A), or the evidence used to estimate θ_A underestimates the true RR_X (and hence, the H_A). The standard view obscures this interpretative dilemma by ruling external evidence inadmissible.

What does this mean for the hypothesis test and estimation problem in APPROVe? First, let's accept the external evidence suggesting the test of H_0 in APPROVe is underpowered. The result of the hypothesis test and estimation problem is as for the standard view: reject H_0; and the 95% confidence interval for RR_X (calculated on the basis of RR_x) is 1.19–3.11. This is what information the *observed data* provide, assuming the specification of the trial.

But these results need to be *interpreted* in light of the evidence external to the trial. The observed RR_x fell in the rejection region of RR_X under H_0. The question becomes *how much* of a discrepancy the observed RR_x should be taken to support. Accepting the external evidence raises doubts that the true value of θ departs from H_0 as much as that suggested by the observed data.

This is illustrated by considering the confidence interval provided by APPROVe. The lower bound, t_α, is 1.19 and the higher bound, t^α, is 3.11. The probability this interval captures θ refers not to t_α and t^α, but to the random variables T_α and T^α. Evidence external to the trial suggests the observed data may be extreme, and should this be true, this would mean the calculated confidence interval does not represent the means of T_α and T^α. The only way to test this is to either conduct a study significantly larger than APPROVe, or combine a number of studies of similar size and set up to APPROVe, and use meta-analysis techniques to calculate a new confidence interval based on this data. In the absence of a larger study, or meta-analysis—or while awaiting such results—care needs to be taken while interpreting the confidence interval observed in APPROVe.

8.5 Conclusion

Epistemic context plays an important role in providing warrant for the inferences of Neyman-Pearson statistics. More specifically, the power of a statistical test—when calculated on the basis of available evidence—is important to the test's interpretation. And it remains important after a statistically significant result has been observed. The rofecoxib case, and APPROVe in particular, illustrate the importance of power to frequentist inferences, and hence to therapeutic decisions within EBM. The inference warranted by a statistically significant observation differs in the three contexts considered: an adequately powered test on the basis of available evidence; a test powered, in the absence of evidence, on the smallest clinically important effect; and a test, that evidence external to the trial suggests is underpowered. While the standard view recommends ignoring the power of a test when interpreting the observed results of a trial, this view can not be justified when evidence external to the trial suggest the test is underpowered.

Frequentist statistical methods put forward a particular approach to data. These methods are justified by how they are expected to perform in the long

run. The limitations of this approach, in interpreting an observed result, are very well recognised.[12] Indeed, even in proposing their approach to hypothesis testing, Neyman and Pearson accept limitations in interpreting a single observed result.

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. (Neyman and Pearson 1933, p. 291)

Yet, judging by the response to APPROVe, the implications of this approach for those needing to base decisions on a particular result, appears to go unnoticed. This can be understood if Neyman-Pearson results are mistakenly taken to hold the same warrant independent of the epistemic context of the test.

Bayesians and other critics of frequentist statistics may find the argument provided in this chapter somewhat laboured. Context (iii), when evidence regarding the unknown parameter is available but not included in the specification of the trial, is a problem particular to the frequentist analysis of clinical trials. Bayesian statistics can accommodate available evidence into the Bayesian *prior*. Sellke et al. (2001) have shown that Bayesian statisticians will reach conclusions that are different to their frequentist counterparts, even when they are both looking at the same data—these differences will only be exacerbated when the two approaches rule different data admissible as they do in cases such as APPROVe. Estimating the cardiovascular risks of rofecoxib is indeed just another case in which those with Bayesian intuitions and those with frequentist intuitions will lock horns. Despite such cases being well documented in the literature, the point bears repeating. I have argued that *even on the frequentist view* the influence of the epistemic context of statistical tests needs to be recognised if frequentist methods are to be used to help inform therapeutic decisions.

I accept for the sake of argument, that taking a *long run* approach to data diminishes the relevance of epistemic context to the warrant of Neyman-Pearson tests. For instance, should a number of trials similar to APPROVe

[12]Pretty much every text on theoretical statistics, and most papers in philosophy of statistics, recognise this point. This includes the two texts cited in this paper (Barnett 1999; Cox and Hinkley 1974).

be conducted then a meta-analysis of these trials will provide an appropriate estimate for rofecoxib's risk of thrombotic events. Similarly, if a considerably larger trial is conducted. It is the *practical* difficulties of gaining this data that raises the problem. First, it takes considerable time to conduct and analyse further trials. Even if what is considered needed is more trials; decisions often need to be made in the interim. Second, randomised interventional studies are not repeated. Subsequent trials may cover similar areas with the same drug, but will invariably be designed to test different hypotheses, involve patients in different situations and examine different outcomes. These differences raise considerable challenges for interpretation of meta-analyses. Some of the challenges presented by meta-analysis are new, for example, questions of homogeneity across the included trials. But some are familiar; the challenge presented by APPROVe is re-iterated. How should we interpret *this* meta-analysis as opposed to what is assured if we follow the method in the long run? Kearney's meta-analysis provides an example. Assuming APPROVe contributes considerably to the estimate provided by the combined placebo controlled trials, should we rely on this estimate, or await further results? Therapeutic decision makers need to interpret the observations at hand. All the evidence needs to be considered, not just the results of the Neyman-Pearson test. (A trivially true statement, but in need of emphasis given the focus given to the statistical results of randomised interventional studies in EBM.)

The problem of underpowered tests overestimating effect size, though less often recognised, has also been documented. In particular, Charles Land (1980) argued for precisely this phenomenon as an explanation of the wide discrepancies observed in data estimating cancer risk from ionising radiation.

> A negative estimate is not unlikely if power is low, and such a result can be interpreted, however improperly, as evidence that there is no excess risk associated with exposure to low-dose radiation, or even that such exposure may be beneficial. A more serious problem, because it is less well understood, follows from the fact that even when power is low, the chances of rejecting the null hypothesis are not negligible (Land 1980, p. 1198).

There are also conceptual similarities in the discussion of 'asymmetrical funnel plots' (Berlin et al. 1989; Egger and Smith 1995; Egger et al. 1997). Funnel plots graph observed effect size against sample size. It is not uncommon

for such plots to be asymmetrical; the smaller studies at the base of the plot provide estimates which are considerably more extreme than larger studies. Low power, in conjunction with publication and reporting bias, provide an explanation for the asymmetry.

While the underlying tension in the application of frequentist statistical methods, and the influence of power, have been recognised in sections of the literature—sometimes very well recognised—they are often not explicitly considered when using frequentist statistical methods to inform therapeutic decisions. APPROVe illustrates that the most important therapeutic decisions don't always come from the result a well powered primary hypothesis test. EBM applies frequentist statistical methods to a particular practical context. The demands of therapeutic decision making need to be kept in mind when applying the frequentist statistical approach.

The rofecoxib case recommends many lessons. One lesson is the need to recognise the gap between the resources of frequentist statistical methods and the needs of therapeutic decision makers.

Chapter 9

Conclusion

This thesis has examined the foundations of EBM. My principle argument can be summarised quite simply. Therapeutic decisions are complex. Clinicians need to incorporate evidence from a range of sources with patient preferences, the clinician's previous experience, and the political set-up of the health system. I have focussed solely on the evidential question, and specifically the claims made by proponents of EBM. EBM provides a hierarchy of evidence and a set of fairly simple instructions on how to apply this hierarchy to therapeutic decisions. EBM's central claim is that informing therapeutic decisions on the basis of evidence from higher up EBM's hierarchy, results in better patient outcomes compared to informing therapeutic decisions on the basis of evidence from lower down. Little explicit justification for this claim can be found in the EBM literature—and what is given is not sufficient to defend it. However, justification can be given for a more limited claim; study designs higher up EBM's hierarchy possess the capacity to rule out more sources of error, and thus, have higher *internal validity*. But, the distinction between internal and external validity—and the related distinction between efficacy and effectiveness—is consistently under-appreciated in the claims that proponents of EBM make on behalf of the hierarchy. Of course, proponents of EBM recognise these terminological distinctions, the problem is that they don't consistently *observe* these distinctions in the claims they make.

EBM's hierarchy, and the statistical methods employed in study designs in the top tiers of the hierarchy, can be justified in assessing claims of *efficacy*. Therapeutic decisions, however, rely on more than efficacy. Therapeutic decisions require a much richer framework for evidence.

The claim that evidence from higher up EBM's hierarchy better informs *therapeutic decisions* is false, at least without qualifications and clarifying caveats. Despite the claim being ubiquitously made in the clinical literature, it is rarely accompanied with the necessary clarifying remarks. One set of problems for the claim is that the appropriate interpretation of the hierarchy is less than clear. Chapter 2 canvassed the ways EBM's hierarchy can *not* be interpreted. EBM's hierarchy does not provide general epistemological rules; if a defence of EBM's hierarchy is to be provided, it needs to be provided within the context of therapeutic decisions (as opposed to science more broadly). And further, EBM's hierarchy can't be interpreted categorically; if EBM's focus is to remain on therapeutic decisions, then evidence from sources higher up EBM's hierarchy can't *trump* (in any straightforward way) evidence from lower down. A more complicated story needs to be told.

The more pressing problem for EBM's central claim is the narrow applicability of the hierarchy on the interpretation that *can* be justified. The foundations of EBM (to the extent that they are articulated) are found in clinical epidemiology, especially in clinical epidemiology's development of methods that have high internal validity. It is here the justification for EBM's hierarchy of evidence can be found. As outlined in Chapter 3, EBM's hierarchy is justified when viewed as a hierarchy of comparative internal validity. Study designs higher up EBM's hierarchy have the *capacity* for higher internal validity than study designs lower down. Providing all the additional methods that minimise error are utilised, studies utilising the methods listed high on EBM's hierarchy will possess high internal validity. This justification demarcates the kinds of questions the hierarchy can be applied to. Chapter 4 showed that good reasons can be given for testing the efficacy of drugs in well conducted randomised interventional studies—the study design favoured by proponents of EBM. But, again, efficacy, and the evidence provided by randomised interventional studies, are not the full story. Therapeutic decisions require judgements about the effectiveness of therapies, and these judgements require consideration of *all* the available evidence.

Chapter 5 outlined the challenge that external validity presents to therapeutic decision makers. Randomised interventional studies employ excellent methods for comparing the outcomes of two therapeutic interventions (usually an experimental treatment against standard care) but these methods are less able to identify *which* patients are likely to respond well to therapy, and which are likely to respond poorly. The challenge of external validity,

by its very nature, can't be met by the restrictive and conservative methods employed to test the efficacy of therapies. Judging the external validity of clinical research, and by extension the effectiveness of therapies, requires consideration of a broad range of evidence. EBM's central claim, as it is so often made, is false because, contrary to the simple advice provided by EBM's hierarchy of evidence, both observational studies and basic science play critical roles in therapeutic decisions. The role these sources of evidence play is insufficiently acknowledged in most presentations of EBM's position, including the definitive presentation provided in EBM's guidebooks.[1]

The distinction between efficacy and effectiveness is paramount. In many instances, limiting EBM's claims to tests of efficacy, makes the claims reasonable. EBM's hierarchy can be justified once it is understood as a hierarchy of comparative internal validity. And the statistical approaches employed in randomised interventional studies, which are both conservative and focussed on a single primary endpoint, derive more support when the context is limited in such a way (as discussed in Section 7.5).

It is once we move from assessing whether a therapy produces the expected benefits in a defined sample of patients, to assessing whether, and in whom, the therapy produces the expected benefits in routine care, that the hierarchy becomes problematic. Chapters 5 and 6 illustrated the importance of evidence from observational studies and basic science for therapeutic questions. These sources of evidence supplement the knowledge gained from randomised interventional studies, and highlight the need for a richer framework for evidence for approaching therapeutic questions. As I have shown, the necessity of this richer framework undermines any simple application of the hierarchy to therapeutic questions.

At times, proponents of EBM appear to accept these limitations on EBM's claims. Sackett et al. (1996) recommends 'integrating individual clinical expertise with the best available external clinical evidence' while recognising that basic medical science is among the sources of the 'best available external clinical evidence'. But such statements are left unclarified, and how these different sources of evidence should be *integrated* is left unclear. Further, when it comes to providing something more concrete about how EBM should be practised, Sackett and other proponents of EBM, contradict these statements (again, observe the advice given in EBM's guidebooks). The Grading of Recommendations Assessment, Development and Evalua-

[1]See, for instance, the quotes taken from the guidebooks on pages 41 and 41.

tion (GRADE) system, which has been around for a number of years but seems to be gaining momentum, is somewhat clearer on the limitations of EBM's hierarchy for informing therapeutic decisions (Montori and Guyatt 2008; Guyatt et al. 2008b). I'll return to GRADE in a moment.

An important practical consequence of EBM's failure to limit its claims regarding the benefits of randomised interventional studies also deserved emphasis. Randomised interventional studies, in conjunction with frequentist statistical approaches, are ill-suited to identify and confirm treatment *harms*. This is especially so for the trials that are actually conducted, as opposed to hypothetical trials that *could* be conducted to answer specific questions. This was most explicitly discussed in Section 5.3, but the threads of the argument have been provided throughout the thesis. EBM promotes the idea that randomised interventional studies provide superior evidence regarding therapies, *including* the assessment of a therapy's safety (Sackett et al. 1996, see quote provided on page 82). This idea has received broad acceptance;[2] but it is both wrong and dangerous.

As discussed, most randomised interventional studies of direct relevance to therapeutic decisions are set up to test a 'benefit' hypothesis. The selection of participants, the time-scale of the intervention, and the statistical methods are all geared towards establishing an expected benefit of the therapy. Of course, important safety information comes from these randomised interventional studies, and this information supplements knowledge gained from pre-clinical testing and other sources. But the randomised interventional studies that are conducted throughout a product's development and subsequent marketing are not ideal for identifying the harms of treatments. Notice, we would employ remarkably different methods if obtaining reliable safety information was our primary goal.

While proponents have, at times, remarked that randomised interventional studies can be ill-suited to elicit the 'rare and awful' adverse effects of therapies, for example Sackett (2006, p. 177) and Guyatt and Rennie (2002, pp. 78–79), the extent of the problem for even relatively frequent

[2]Witness Merck's statement that 'prospective randomised, controlled clinical trials are the best way to evaluate the safety of medicines' (Merck 2004, full quote on page 176), and Collins and MacMahon's claim that *nothing* other than 'large-scale randomised trials or their meta-analyses' are able to provide 'clear confirmatory' evidence of either the adverse or protective effects of therapies (Collins and MacMahon 2007, p. 24, and provided on page 51).

adverse effects is under-appreciated. The entire system leans towards detecting benefits. And EBM's hierarchy of evidence explicitly and automatically downgrades the evidence provided by methods that are often better suited to detecting treatment harms: cohort studies, case-control studies and basic science. Often, finding important side effects in randomised interventional studies is due to serendipity rather than good planning. Rofecoxib, as discussed in Chapter 8, is a case in point. The randomised trials providing evidence of the thrombotic risks of rofecoxib, VIGOR and APPROVe like the many randomised trials before them that did not detect the risk, were set-up to test a benefit hypothesis—the safety related findings were a fortunate by-product.

Perhaps cases like rofecoxib provide the best impetus for a renewed interest in methodological diversity. Even if it is possible to *imagine* a randomised interventional study designed to test a specific safety hypothesis, such perfectly designed 'safety trials' are not conducted. And yet, everyone—including, of course, the pharmaceutical industry—wants to identify the adverse effects of therapies as reliably and as swiftly as possible. This societal imperative, combined with the clear argument that the trials that are typically conducted on marketed therapies are not ideal for eliciting safety information (not to mention the existence of cases such as rofecoxib), should be enough to awaken clinicians, the pharmaceutical industry and policy makers from the dogmatic slumber induced by EBM's emphasis on randomised interventional studies.

This is not to suggest observational studies don't have their flaws for obtaining safety information on treatments; they do. But until the hegemony of randomised interventional studies is replaced by a more inclusive view of evidence in medicine, especially with regard to assessing safety, we won't be doing our best to prevent cases like rofecoxib recurring.

The role of statistical approaches within EBM has also been considered. While there is no necessary link between EBM and the frequentist statistical methods used to analyse clinical studies, the two are so deeply entwined it is hard to imagine EBM relying on alternative statistical methods. My aim in discussing frequentist methods throughout the thesis has been to clarify, and understand, the warrant provided by these methods. Making a case for a rival statistical methodology, such as a Bayesian approach to statistical inference, would require considerable further argument. It is, however, possible to draw a couple of important conclusions regarding the applicability of frequentist

statistics to therapeutic decisions.

There is a case for frequentist methods in tests of efficacy. Discussion on these points was provided in Section 7.5. Assessments of efficacy focus on the primary clinical endpoint in the randomised interventional study, and thus, providing the trial is adequately powered for the primary endpoint, the findings related to this endpoint receive the full warrant provided by Neyman-Pearson methods. Further, the conservative methods employed in the analysis of clinical studies can be justified. There is an appropriate focus on ensuring treatments are not judged efficacious when in fact they are not. This is especially important because tests of efficacy play a key role in regulatory decisions. In this context, intention-to-treat analysis and the inadmissibility of *post hoc* analyses make sense. Of course, there is still the more general arguments against frequentist methods, provided by Howson and Urbach (2006) and others. I am sympathetic to these arguments. However, providing frequentist methods, and the inferences they warrant, are clearly understood, these methods are well placed to provide rigid and conservative tests of efficacy. (Of course, it is often not the case that frequentist methods are clearly understood. But this is a different problem, and one which, at least on the face of it, can be rectified through education.) At minimum, we can make some sense of why frequentist methods are employed in tests of efficacy.

But, we can not be so sanguine about the applicability of frequentist methods to questions of effectiveness. Frequentist statistical methods are considerably less useful for identifying who is likely to benefit from therapy and who may be harmed. Adhering to the conservative analyses employed in tests of efficacy leaves practitioners in a quandary. As seen in Chapter 7, the inadmissibility of *post hoc* analyses, and a scepticism towards the role of basic science in interpreting data, limits the resources frequentist methods provide therapeutic decision makers. Patients presenting to clinics differ, in small or large measure, from patients enrolled in applied clinical studies. To adequately treat patients, clinicians must judge whether and *which* results observed in clinical studies are relevant to the situation at hand. Targeting treatments to individuals based on the data from relevant subgroups in clinical trials would would greatly assist therapeutic decisions. But, the reliability of subgroup analyses in clinical studies is very much in question.

This is partly unavoidable, the complexity of clinical practice will raise

more questions than the available data are able to answer. But the situation is made worse by the theoretical commitments of trialists (and other proponents of EBM). Scepticism about the role of basic science in interpreting subgroups, and the application of the same conservative methods used for questions of efficacy to questions of effectiveness, leave trialists focussed solely on the overall results of clinical trials. The extent of this problem for EBM—or, for that matter, *any* account of therapeutic decision making—is under-appreciated (mainly, I suspect, because the lack of easy answers promotes a begrudged acceptance of the problem). In Chapter 7, I argued that some methodological flexibility was needed to assist judgements of external validity. Specifically, a more permissive attitude to *post hoc* subgroup data can be supported, providing there is independent support for the inference from the three tiers of models involved in the experimental inquiry, the model of the theory, the model of the experiment and the model of the data. The hierarchy of data models re-emphasises the links between basic science, experiment, and data. (*Re-emphasises* because the link has always been there. The commitments of proponents of EBM, however, suppress this link.) The framework provided by the hierarchy of data models can assist clinicians to judge whether an observed difference in a subgroup is due to an underlying process or likely due to chance. It does not resolve all the difficulties of interpreting subgroup data, but being explicit about the hierarchy of data models highlights one type of argument that can be made for subgroup data to be considered genuine—a useful (if somewhat modest) addition to the armamentarium of therapeutic decision makers.

While some of the positive attributes of the frequentist approach to tests of efficacy have been noted, the deficiencies of our current methods for analysing clinical trials in a way that meets the needs of therapeutic decision makers have been clearly articulated. This leaves the door open for a more forceful argument regarding statistical methodology. Participants in the debate over statistical methodology—whether frequentist or Bayesian—agree on at least two points. First, clinicians need reliable information on subgroups of patients involved in randomised interventional studies. And second, in the trials that are typically conducted, frequentist analyses are unable to provide reliable analyses of subgroup data. While their reliability will be hotly contested, Bayesian approaches (due to their flexibility) are able to provide direct analysis of subgroup data. In this context the call for Bayesian approaches to the analysis of clinical studies has increased from an

occasional plea to a rising chorus (see, for example, Spiegelhalter et al. 1994, Berry 2006, and most recently, Rawlins 2008). Clinicians and statisticians of a frequentist persuasion will argue that the problems of Bayesian analyses outweigh these benefits. In addition to more clearly articulating the pros and the cons of frequentist statistical methods for therapeutic decisions, the contribution of this thesis has been to further underscore the practical importance of this debate. (It is noteworthy that the issues in this debate are essentially philosophical. For instance, one of the central questions regards what sort of probabilities should be involved in scientific inquiry.)

Bayesian approaches may be better equipped to analyse subgroup data in clinical trials. To the extent that Bayesian analyses have this benefit, EBM could avoid some of the problems raised in Chapter 7 and 8 by changing statistical methods. Presumably, even if such a change was to be made, proponents of EBM would retain their methodological commitment to the hierarchy of evidence and the superiority of randomised trials.[3] The arguments provided in Chapters 2–6 focus on the hierarchy of evidence, and the distinction between randomised and non-randomised studies, and therefore they hold irrespective of the choice of statistical analysis of clinical trails (providing EBM retains its focus on the hierarchy of evidence).

I expressed a view in the introduction that EBM *may* be viewed as an evolution in therapeutic decision making. It is now possible to be more precise. The main tool that EBM provides is its hierarchy of evidence. And, on the basis of this hierarchy, proponents of EBM make some strong, and indeed some indefensible, claims. It is entirely consistent with the arguments provided to reject 'EBM' on the basis that its central claim—as typically stated—is false. But this risks missing an opportunity.

EBM's rationalist pretensions are admirable. In this sense I agree with Adrian Smith (1996), who was quoted in the opening of Chapter 1. Prior to EBM it was not the case that medical decisions were *not* based on evidence, nor that only the loudest, or most eminent clinical voices were heard. But it is true that there was little focus on the project of providing a rational framework for therapeutic decisions; perhaps this project was thought either impossible or undesirable. I think EBM, providing it is properly conceived, legitimises and makes some headway towards providing a rational account of

[3] As noted in §4.3.2, a Bayesian justification for randomisation can be provided. Thanks to Jeremy Howick for pushing me to emphasize the lack of necessary connection between EBM and frequentist statistics.

therapeutic decision making.

Evidence from randomised interventional studies has substantially improved care in many cases. Paradigm cases include hormone replacement therapy and the use of thrombolytic agents in patients who have suffered a myocardial infarction (see Sackett 2006, p. 177 and Peto et al. 1995 respectively). Better evidence regarding the efficacy of therapies improves therapeutic decisions. The problem with EBM's central claim is not so much what is said, but what too often goes unsaid. EBM, and its claims, need clarification—not reduction to an overly simple hierarchy of evidence. This thesis provides clarification on the foundations of EBM. On my account, EBM's hierarchy is a hierarchy of internal validity. If appropriate methods are applied we can be more confident of our tests of efficacy. Tests of efficacy provide some information on the effects of a therapy in a sample of the population. This provides a good starting point for therapeutic decisions. But, since many therapeutic decisions involve extrapolation from and interpolation into this sample, a more sophisticated view of evidence is required. I don't pretend to have provided a full account of this more sophisticated view of evidence, rather the bulk of my arguments have been aimed at highlighting the deficiency of the account provided in the EBM literature. I do think that the hierarchy of data models provides one avenue for a more positive account. Either way, better understanding what clinical epidemiological methods achieve, and what they do not, is a first important step in improving therapeutic decisions.

All of this implies a clear course for improving EBM. *More* attention is needed on the foundations of decisions. We need to be more explicit about what the current methods fail to achieve, or achieve poorly. The methodological dogmatism needs to be put to bed, and a greater focus is required on developing methods for assessing external validity, and the effectiveness and safety of therapies. The GRADE system seems to put forward a different view for improving EBM.

GRADE is primarily focussed on guideline development rather than the direct judgements of clinicians. It recognises that external validity is an additional marker of research quality, and explicitly separates 'quality of evidence' (judged primarily according to study design) and 'strength of recommendation' (which takes issues such as patient preferences, and the balance of positive and negative outcomes into account). This separation of quality of evidence from strength of recommendation permits guidelines developed

according to the GRADE system to take more into account than simply the strength of evidence based on EBM's hierarchy. If patient preferences regarding the outcomes of a therapy differ greatly, high quality evidence may lead to a 'weak' recommendation for treatment (reflecting the fact that despite good evidence from randomised interventional studies demonstrating a benefit from treatment, some patients may put little value in that benefit). Alternatively, 'low quality' evidence from, say an observational study, may support a 'strong' recommendation due to the overwhelming (and uncontroversial) benefits of a therapy that lacks significant risk for harm.

Hence, the GRADE system makes some improvements; it adequately responds to some of the failings of an overly simplistic view of EBM's hierarchy. However, GRADE also risks perpetuating some of EBM's more fundamental flaws. It is good that GRADE recognises the importance of external validity, and the distinction between 'high quality evidence' and 'good evidence for use'. But after recognising these additional factors for therapeutic decision making (of which there are many forms) it reduces consideration of these factors to a dichotomous 'strength of recommendation'. Far more important than the eventual rating, are the considerations that go into it. And even more fundamental is how we go about the complex process of weighing evidence from a range of sources, while taking patient preferences and a range of additional factors into consideration. As I have argued throughout, it is the foundational issues which need greater focus. Adding a scale for the 'strength of a recommendation' risks obscuring important foundational issues in the same way the hierarchy of evidence obscures evidence 'quality'.[4]

EBM started as a simple (and good) idea in need of some clarification. Unfortunately as EBM gained momentum, as it was disseminated and operationalised, it became simpler. There is a cost to holding a more complex view of evidence in therapeutic decision making. On a more complicated view of evidence, there is no universally accepted source of evidence that is optimal to answer therapeutic questions. But the benefits of this view outweigh this cost. A more complex view of medical evidence, a view that legitimises methodological diversity, is much better placed to reply to the full spectrum of therapeutic questions that arise.

[4]Michael Rawlins (2008, p. 35) has argued along these lines recently.

Bibliography

Abbasi, Kamran. 2004. Is drug regulation failing? *British Medical Journal*, 329:0.

Aisen, Paul S., Kimberly A. Schafter, Michael Grundman, Eric Pfeiffer, Mary Sano, Kenneth L. Davis, Martin R. Farlow, Sheila Jin, Ronald G. Thomas, and Leon J. Thal. 2003. Effects of rofecoxib or naproxen vs placebo on alzheimers disease progression: A randomised controlled trial. *The Journal of the American Medical Association*, 289(21):2819–2826.

Altman, Douglas G. 1998. Within trial variation—a false trail? *Journal of Clinical Epidemiology*, 51(4):301–303.

Altman, Douglas G., Kenneth F. Schulz, David Moher, Matthias Egger, Frank Davidoff, Diana R. Elbourne, Peter C. Gotzsche, Thomas Lang, and the CONSORT group. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine*, 123:663–694.

Anonymous. 1995. Evidence-based medicine, in its place (Editorial). *The Lancet*, 346(8978):785.

Armitage, Peter. 1982. The role of randomization in clinical trials. *Statistics in Medicine*, 1:345–352.

Aroney, Constantine N., Philip Aylward, Anne-Maree Kelly, Derek P. B. Chew, and Eleanor Clune. 2006. Guidelines for the management of acute coronary syndromes 2006. *Medical Journal of Australia*, 184(8):S1–S30.

Barnett, Vic. 1999. *Comparative Statistical Inference*. Wiley Series in Probability and Statistics. West Sussex: John Wiley and Sons, 3rd edition.

Benson, Kjell and Arthur J. Hartz. 2000. A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine*, 342(25):1878–1886.

Berger, James O. and Donald A. Berry. 1988. The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics IV*. New York: Springer Verlag.

Berlin, Jesse A., Colin B. Begg, and Thomas A. Louis. 1989. An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association*, 84(406):381–392.

Berry, Donald A. 1987. Interim analysis in clinical trials: The role of the likelihood principle. *The American Statistician*, 41(2):117–122.

———. 2006. Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.

β-Blocker Heart Attack Trial Reseach Group. 1982. A randomised trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *Journal of American Medical Association*, 247(12):1707–1714.

Black, Nick. 1996. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312(7040):1215–1218.

Bluhm, Robyn. 2005. From heirarchy to network: A richer view of evidence for evidence-based medicine. *Perspectives in Biology and Medicine*, 48(4):535–547.

Bombardier, Claire, Loren Laine, Alise Reicin, Deborah Shapiro, Ruben Burgos-Vargos, Barry Davis, and et. al. 2000. Comparision of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis: VIGOR study group. *New England Journal of Medicine*, 343:1520–1528.

Bovens, Luc and Stephan Hartmann. 2002. Bayesian networks and the problem of unreliable instruments. *Philosophy of Science*, 69(1):29–72.

Bresalier, Robert S., Robert S. Sandler, Hui Quan, James A. Bolognese, Bettina Oxenius, Kevin Horgan, Christopher Lines, Robert Riddell, Dion Morton, Angel Lanas, Marvin A. Konstam, and John A. Baron. 2005.

Cardiovascular events associated with rofecoxib in a colorectal andenoma chemoprevention trial. *New England Journal of Medicine*, 352(11):1092–1102.

Brookes, S.T., E. Whitely, T.J. Peters, P. A. Mulheran, Matthias Egger, and George Davey Smith. 2001. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment*, 5(33):1–58.

Buetow, Stephen, Ross Upshur, Andrew Miles, and Michael Loughlin. 2006. Taking stock of evidence-based medicine: Opportunities for its continuing evolution. *Journal of Evaluation in Clinical Practice*, 12(4):399–404.

Byar, David P. 1985. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4:255–263.

Chalmers, Thomas C., R. J. Matta, Harry Smith, Jr., and A. M. Kunzler. 1977. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine*, 297:1091–1097.

Cochrane, Archie L. 1971. *Effectiveness and Efficiency: Random Reflections on Health Services.* London: Nuffield Provincial Hospitals Trust.

Collins, Rory and Stephen MacMahon. 2007. Reliable assesment of the effects of treatments on mortality and major morbidity. In Rothwell (2007c).

Concato, John, Nirav Shah, and Ralph I. Horwitz. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine*, 342(25):1887–1892.

Cox, D. R. 2006. *Principles of Statistical Inference.* Cambridge: Cambridge Univeristy Press.

Cox, D. R. and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman and Hall.

Cui, Lu, H. M. James Hung, Sue Jane Wang, and Yi Tsong. 2002. Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics*, 12(3):347–358.

Cutts, Christopher, Adam La Caze, and Susan Tett. 2002. A clinical audit of the prescribing of celecoxib and rofecoxib in Australian rural general practice. *British Journal of Clinical Pharmacology*, 54:522–527.

Devereaux, P. J. and Salim Yusuf. 2003. The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of Internal Medicine*, 254:105–113.

Egger, Matthias and George Davey Smith. 1995. Misleading meta-analysis. *British Medical Journal*, 310(6982):752–754.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association*, 268(17):2420–2425.

Farkouh, Michael E., Howard Kirshner, Robert A. Harrington, Sean Ruland, Freek W. A. Verheugt, Thomas J. Schnitzer, Gerd R. Burmester, Eduardo Mysler, Marc C. Hochberg, Michael Doherty, Elena Ehrsam, Xavier Gitton, Gerhard Krammer, Bernhard Mellein, Alberto Gimona, Patrice Matchaba, Christopher J. Hawkey, and James H. Chesebro. 2004. Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), cardiovascular outcomes: Randomised controlled trial. *Lancet*, 364(9435):675–684.

Feinstein, Alvan R. 1984. Why do we need some large, simple randomized trials? Discussion. In Yusuf et al. (1984), pages 421–422.

———. 1998. The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology*, 51(4):297–299.

Fletcher, Robert H., Suzanne W. Fletcher, and Edward H. Wagner. 1996. *Clinical Epidemiology: The Essentials*. Baltimore: Lippincott Williams and Wilkins, 3rd edition.

Food and Drug Administration Advisory Committee. 2001. Cardiovascular safety review of rofecoxib. Technical report, Food and Drug Administration/Centre for Drug Evaluation and Research. `http://www.fda.gov/ohrms/dockets/ac/01/briefing/3677b2_06_cardio.pdf` Accessed 28/11/08.

Frigg, Roman and Stephan Hartmann. 2008. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition. `http://plato.stanford.edu/archives/fall2008/entries/models-science/` Accessed 28/11/08.

Furberg, Curt D. and Robert P. Byington. 1983. What do subgroup analyses reveal about differential response to beta-blocker therapy? *Circulation*, 67 (Suppl I):I98–I101.

Gail, M. and R. Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372.

Gardner, Martin J. and Douglas G. Altman. 1989. Estimation rather than hypothesis testing: Confidence intervals rather than p values. In Martin J. Gardner and Douglas G. Altman, editors, *Statistics with Confidence*. London: British Medical Journal.

Glasziou, Paul, Iain Chalmers, Michael Rawlins, and Peter McCulloch. 2007. When are randomised trials unnecessary? picking signal from noise. *British Medical Journal*, 334(7589):349–351.

Glasziou, Paul, Jan Vandenbroucke, and Iain Chalmers. 2004. Assessing the quality of research. *British Medical Journal*, 328(7430):39–41.

Goodman, Steven N. 1993. p values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137(5):485–496.

Goodman, Steven N. and Jesse A. Berlin. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121(3):200–206.

Graham, David. 2006. COX-2 Inhibitors, other NSAIDs, and cardiovascular risk: The seduction of common sense. *The Journal of the American Medical Association*, 296(13):1653–1656.

Greenland, Sander and Keith O'Rourke. 2008. Meta-analysis. In Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash, editors, *Modern Epidemiology*. Philadelphia: Lippincott Williams and Wilkins, 3rd edition.

Grossman, Jason. ——. Statistical inference. Unpublished Manuscript.

Grossman, Jason and Fiona J. Mackenzie. 2005. The randomised controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine*, 48(4):516–534.

Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). 1986. Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet*, 1:397–402.

Guyatt, Gordon H. 1991. Evidence-based medicine. *ACP Journal Club*, 114:A–16.

Guyatt, Gordon H, Andrew D Oxman, Regina Kunz, Gunn E Vist, Yngve Falck-Ytter, Holger J Schunemann, and the GRADE Working Group. 2008a. What is"quality of evidence" and why is it important to clinicians? *British Medical Journal*, 336(7651):995–998.

Guyatt, Gordon H, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, Holger J Schunemann, and the GRADE Working Group. 2008b. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336(7650):924–926.

Guyatt, Gordon H. and Drummond Rennie, editors. 2002. *Users' guide to the medical literature: Essentials of evidence-based clinical practice*. Chigaco: American Medical Association Press.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Hájek, Alan. 2007. The reference class problem is your problem too. *Synthese*, 156:563–585.

Hallas, Jesper, Michael Dall, Alin Andries, Birthe Sogaard Andersen, Claus Aalykke, Jane Moller Hansen, Morten Andersen, and Annmarie Touborg Lassen. 2006. Use of single and combined antithrombotic therapy and risk

of serious upper gastrointestinal bleeding: population based case-control study. *British Medical Journal*, 333(7571):726–731.

Hartmann, Stephan. 2008. Modeling in philosophy of science. In M Frauchiger and W. K. Essler, editors, *Representation, Evidence, and Justification: Themes from Suppes*, Launer Library of Analytical Philosophy, Vol. 1. Frankfurt: Ontos Verlag.

Haynes, R. Brian. 2002. What kind of evidence is it that evidence-based medicine advocates want health care providers and consumers to pay attention to? *BMC Health Services Research*, 2. `http://www.biomedcentral.com/1472-6963/2/3` Accessed 28/11/08.

Haynes, R. Brian, David L. Sackett, Gordon H. Guyatt, and Peter Tugwell. 2006. *Clinical Epidemiology: How to Do Clinical Practice Research.* Philadelphia: Lippincott Williams and Wilkins, 3rd edition.

Henry, Steven G. 2006. Recognizing tacit knowledge in medical epistemology. *Theoretical Medicine and Bioethics*, 27(3):187–213.

Hill, Austin Bradford. 1966. Heberden oration (1965): Reflections on controlled trials. *Annals of the Rheumatic Diseases*, 25:107–113.

Horwitz, Ralph I., Burton H. Singer, Robert W. Makuch, and Catherine M. Viscoli. 1996. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs to clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4):395–400.

———. 1997. On reaching the tunnel at the end of the light. *Journal of Clinical Epidemiology*, 50(7):753–755.

———. 1998. Clinical versus statistical considerations in the design and analysis of clinical research. *Journal of Clinical Epidemiology*, 51(4):305–307.

Howson, Colin and Peter Urbach. 2006. *Scientific Reasoning: The Bayesian Approach.* Chicago: Open Court, 3rd edition.

ISIS-2 Collaborative Group. 1988. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *The Lancet*, 332(8607):349–360.

Juni, Peter, Linda Nartey, Stephan Reichenbach, Rebekka Sterchi, Paul A. Dieppe, and Matthias Egger. 2004. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*, 364:2021–2029.

Kearney, Patricia M., Colin Baigent, Jon Godwin, Heather Halls, Jonathon R. Emberson, and Carlo Patrono. 2006. Do selective cyclooxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? meta-analysis of randomised trials. *British Medical Journal*, 332:1302–1308.

Kendall, Maurice, Alan Stuart, and J. Keith Ord. 1983. *The Advanced Theory of Statistics*, volume 3. London: Charles Griffin and Company Limited, 4th edition.

Kerr, David J., Janet A. Dunn, M. D. Langman, Justine L. Smith, Rachel S. J. Midgley, Andrew Stanley, Joanne C. Stokes, Patrick Julier, Claire Iveson, Ravi Duvvuri, and Christopher C. McConkey. 2007. Rofecoxib and cardiovascular adverse events in adjuvant treatment of colorectal cancer. *New England Journal of Medicine*, 357(4):360–369.

Klein, Juergen. 2008. Francis Bacon. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition. `http://plato.stanford.edu/archives/fall2008/entries/francis-bacon/` Accessed 28/11/08.

Krumholz, Harlan M., Joseph S. Ross, Amos H. Presler, and David S. Egilman. 2007. What have we learnt from Vioxx? *British Medical Journal*, 334:120–123.

La Caze, Adam. 2005. Does pharmacogenomics provide an ethical challenge to the utilisation of cost-effectiveness analysis by public health systems? *Pharmacoeconomics*, 23(5):445–447.

———. 2008a. Evidence-based medicine can't be .... *Social Epistemology*, 22(4):353–370.

———. 2008b. A problem for achieving informed consent. *Theoretical Medicine and Bioethics*, 29:255–265.

Lagakos, Stephen W. 2006. Time-to-event analyses for long-term treatments—the APPROVe trial. *New England Journal of Medicine*, 355(2):113–117.

Land, Charles E. 1980. Estimating cancer risks from low doses of ionizing radiation. *Science*, 209(4462):1197–1203.

Lindley, Dennis V. 1982. The role of randomization in inference. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982:431–446.

Lip, Gregory Y H and Gordon D O Lowe. 1996. ABC of atrial fibrillation: Antithrombotic treatment for atrial fibrillation. *British Medical Journal*, 312(7022):45–49.

March, L., L. Irwig, J. Schwarz, J. Simpson, C. Chock, and P. Brooks. 1994. N of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. *British Medical Journal*, 309(6961):1041–1044.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations. Chigaco: University of Chicago Press.

Merck. 2004. News release: Merck announces voluntary worldwide withdrawal of Vioxx. `http://www.vioxx.com/vioxx/documents/english/vioxx_press_release.pdf` Accessed 28/11/08.

Montori, Victor M. and Gordon H. Guyatt. 2008. Progress in evidence-based medicine. *The Journal of the American Medical Association*, 300(15):1814–1816.

Morgan, Sophie Victoria. 2004. Between the devil and the deep blue sea—balancing the risks and potential benefits of warfarin for older people with atrial fibrillation. *Age and Ageing*, 33(6):544–547.

Neyman, Jerzy. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380.

Neyman, Jerzy and Egon S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A.*, 231:289–337.

Okie, S. 2005. What ails the FDA? *New England Journal of Medicine*, 352(11):1063–1066.

Peto, Richard, Rory Collins, and Richard Gray. 1995. Large-scale randomized evidence: Large, simple trials and overviews of trials. *Journal of Clinical Epidemiology*, 48(1):23–40.

Phillips, Bob, Chris Ball, David L. Sackett, Doug Badenoch, Sharon E. Straus, R. Brian Haynes, and Martin Dawes. 2001. Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001). `http://www.cebm.net/?o=1023` Accessed 18/2/08.

Rawlins, Michael David. 2008. De Testimonio: On the use of evidence about the use of therapeutic interventions. The Harveian Oration. `http://www.rcplondon.ac.uk/pubs/brochure.aspx?e=262`. Accessed 28/11/08.

Reilly, Brendan M. 2004. The essence of EBM. *British Medical Journal*, 329(7473):991–992.

Reines, S. A., G. A. Block, J. C. Morris, G. Liu, M. L. Nessly, C. R. Lines, B. A. Norman, and C. C. Baranak. 2004. No effect on alzheimer's disease progression in a 1-year, randomized, blinded, controlled study. *Neurology*, 62:66–71.

Reynolds, Matthew V., Kyle Fahrbach, Ole Hauch, Gail Wygant, Rhonda Estok, Catherine Cella, and Luba Nalysnk. 2004. Warfarin anticoagulation and outcomes in patients with atrial fibrillation: A systematic review and meta-analysis. *Chest*, 126:1938–1945.

Ross, Joseph S., Kevin P. Hill, David S. Egilman, and Harlan M. Krumholz. 2008. Guest authorship and ghostwriting in publications related to rofecoxib: A case study of industry documents from rofecoxib litigation. *The Journal of the American Medical Association*, 299(15):1800–1812.

Rothwell, Peter M. 2005a. External validity of randomised controlled trials: "To whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93.

———. 2005b. Treating individuals 2: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

———. 2007a. Assessment of the external validity of randomised controlled trials. In Rothwell (2007c).

———. 2007b. Reliable estimation and interpreation of the effects of treatment in subgroups. In Rothwell (2007c).

Rothwell, Peter M., editor. 2007c. *Treating Individuals: From randomised trials to personalised medicine.* Philadelphia: Elsevier.

Rothwell, Peter M. 2007d. When should we expect clinically important differences in response to treatment? In Rothwell (2007c).

Sackett, David L. 2006. The principles behind the tactics of performing therapeutic trials. In Haynes et al. (2006), pages 173–243.

Sackett, David L., William Rosenberg, J A Muir Gray, Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: What is it and what it isn't. *British Medical Journal*, 312(7023):71–72.

Sacks, Henry, Thomas C. Chalmers, and Harry Smith, Jr. 1982. Randomized versus historical controls for clinical trials. *The American Journal of Medicine*, 72(2):233–240.

Sehon, Scott R. and Donald E. Stanley. 2003. A philosophical analysis of the evdience-based medicine debate. *BMC Health Services Research*, 3(14):1–10. `http://www.biomedcentral.com/1472-6963/3/14` Accessed 28/11/08.

Sellke, Thomas, M. J. Bayarri, and James O. Berger. 2001. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.

Senn, Stephen and Frank Harrell. 1997. On wisdom after the event. *Journal of Clinical Epidemiology*, 50(7):749–751.

Shahar, Eyal. 1997. A Poperian perspective of the term 'evidence-based medicine'. *Journal of Evaluation in Clinical Practice*, 3(2):109–116.

Simpson, Colin R., Karen Alford, and David Williams. 2008. Evidence for prolonged prescribing of aspirin-clopidogrel combination in Scottish primary care. *Pharmacoepidemiology and Drug Safety*, 17:397–400.

Smith, Adrian F. M. 1996. Mad cows and ecstasy: Chance and choice in an evidence-based society. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):367–383.

Smith, George D. and Matthias Egger. 1998. Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. *Journal of Clinical Epidemiology*, 51(4):289–295.

Smith, Gordon C. S. and Jill P. Pell. 2003. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *British Medical Journal*, 327(7429):1459–1461.

Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K. B. Parmar. 1994. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):357–416.

Straus, Sharon E. 2004. What's the E for in EBM? *British Medical Journal*, 328:535–536.

Straus, Sharon E., W. Scott Richardson, Paul Glasziou, and R. Brian Haynes. 2005. *Evidence-Based Medicine: How to Practice and Teach.* London: Elsevier Churchill Livingstone, 3rd edition.

Stuart, Alan and J. Keith Ord. 1991. *Kendall's Advanced Theory of Statistics*, volume 2. London: Edward Arnold, 5th edition.

Suppes, Patrick. 1960. A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese*, 12(2):287–301.

———. 1962. Models of data. In E Nagel, P Suppes, and A Tarski, editors, *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, pages 252–61. Stanford: Stanford University Press.

———. 1982. Arguments for randomizing. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982:464–475.

Tanenbaum, Sandra J. 1993. What physicians know. *New England Journal of Medicine*, 329:1268–1271.

Thal, Leon J, Steven H. Ferris, Louis Kirby, Gilbert A. Block, Chistopher R. Lines, Eric Yuen, Christopher Assaid, Michael L. Nessly, Barbara A. Norman, Christine C. Baranak, and Scott A. Reines. 2005. A randomized, double-blind study of rofecoxib in patients with mild cognitive impairment. *Neuropsychopharmacology*, 30:1204–1215.

The CURE Investigators. 2001. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *New England Journal of Medicine*, 345(7):494–502.

The GRADE Working Group. 2004. Grading quality of evidence and strength of recommendations. *British Medical Journal*, 328:1490–1498.

Topol, Eric J. 2004. Failing the public health—rofecoxib, Merck, and the FDA. *New England Journal of Medicine*, 351(17):1707–1709.

Tukey, John W. 1986. Sunset salvo. *The American Statistician*, 40(1):72–76.

Upshur, Ross E.G. 2005. Looking for rules in a world of exceptions: Reflections on evidence-based practice. *Perspectives in Biology and Medicine*, 48(4):477–489.

Urbach, Peter. 1985. Randomization and the design of experiments. *Philosophy of Science*, 52:256–273.

———. 1993. The value of randomization and control in clinical trials. *Statistics in Medicine*, 12:1421–1431.

Ware, James H., Frederick Mosteller, Fernando Delgado, Christl Donnelly, and Joseph A Ingelfinger. 1992. P values. In John C. III Bailar and Frederick Mosteller, editors, *Medical Uses of Statistics*, pages 181–200. Boston: NEJM Books.

Weir, Matthew R., Rhonda S. Sperling, Alise Reicin, and Barry J. Gertz. 2003. Selective COX-2 inhibition and cardiovascular effects: A review of the rofecoxib development program. *American Heart Journal*, 146:591–604.

Whorlow, Sarah L. and Henry Krum. 2000. Meta-analysis of effect of beta-blocker therapy on mortality in patient with New York Heart Association class IV chronic congestive heart failure. *American Journal of Cardiology*, 86:886–889.

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

———. 2007a. Evidence in medicine and evidence-based medicine. *Philosophy Compass*, 2(6):981–1022.

———. 2007b. Why there's no cause to randomize. *British Journal for the Philosophy of Science*, 58(3):451–488.

———. 2008. Evidence and ethics in medicine. *Perspectives in Biology and Medicine*, 51(3):418–431.

Young, Jane M, Paul Glasziou, and Jeanette E Ward. 2002. General practitioners' self ratings of skills in evidence based medicine: validation study. *British Medical Journal*, 324(7343):950–951.

Yusuf, Salim, Rory Collins, and Richard Peto. 1984. Why do we need some large, simple randomized trials? *Statistics in Medicine*, 3:409–420.