The Role of Basic Science in Evidence-Based Medicine

Adam La Caze
The University of Queensland
a.lacaze@uq.edu.au*

Abstract

Proponents of Evidence-based medicine (EBM) do not provide a clear role for basic science in therapeutic decision making. Of what they do say about basic science, most of it is negative. Basic science resides on the lower tiers of EBM's hierarchy of evidence. Therapeutic decisions, according to proponents of EBM, should be informed by evidence from randomised studies (and systematic reviews of randomised studies) *rather than* basic science. A framework of models explicates the links between the mechanisms of basic science, experimental inquiry, and observed data. Relying on the framework of models I show that basic science often plays a role not only in specifying experiments, but also analysing and interpreting the data that is provided. Further, and contradicting what is implied in EBM's hierarchy of evidence, appeals to basic science are often required to apply clinical research to therapeutic questions.

1 Introduction

Clinical medicine is a teleological science. What matters is the outcome: improving the health of a patient. In this way clinical medicine shares challenges with other teleological sciences. Decisions need to be made, often quickly, and invariably on limited information. Like many applied sciences, the evidential inputs into decisions come from a range of fields. Clinical medicine relies on the basic medical sciences of physiology, pathophysiology and immunology, as well as applied clinical research in the form of observational studies and randomised trials—not to mention the many additional fields that complement these sources of information. While basic medical science provides a rich theoretical basis for medical decisions, applying the best supported mechanisms of basic science to patient care is not straightforward.

There is much unexplained inter-patient variability in clinical medicine. The pathophysiological mechanism of the disease, and the pharmacological action of

*This is a pre-print version of a paper published in *Biology and Philosophy*, 26(1):81–98; http://dx.doi.org/10.1007/s10539-010-9231-5. Please reference published version.

the medicine employed to treat the disease may be well known, and yet it remain impossible to predict whether patients will respond to the therapy (this is true whether the prediction refers to an individual, or the average response of a group of patients). The unexplained inter-patient variability is part epistemic, and part stochastic. Learning more about the pharmacological/pathophysiological mechanisms can help to discern which patients are more likely to benefit from therapy. But medical knowledge remains incomplete, and relies on observing the average response of a therapy in a defined group of patients. The complexity of the human body and the seemingly endless variability of individual physiological response means that a considerable proportion of inter-patient variability is attributed to chance. While basic medical science becomes finer grained, and some variability that was attributed to chance is re-evaluated, clinical questions also become finer grained, emphasising the importance of observing and measuring how different patients respond to a treatment. This is the task of applied clinical research.

Clinical studies may be randomised or observational. In randomised studies, patients are randomly allocated to the treatment under investigation or control and then assessed on pre-defined outcome measures. Observational studies, by contrast, compare patients undergoing routine care. The patients under observation have chosen their therapy, often with the assistance of their health professional. A complicated relationship exists between applied clinical research and basic medical science. In situations in which the basic science is well worked out, applied clinical research can be seen as a test of the *clinical applicability* of basic science. But applied clinical research is not always a test of a well understood biological theory—sometimes little is known at the level of basic science. In areas in which the basic science is less well developed, the relationship can operate in the opposite direction with the results of clinical research suggesting hypotheses for future basic research.

How the mechanisms (or theories) of basic science and the statistical results of clinical research should be used to inform medical decisions is a perennial topic for debate.[1] While the links between basic science and applied clinical research are readily apparent, a comprehensive account of how these two sources of knowledge should be incorporated into therapeutic decisions is lacking. The evidential challenge is exacerbated by the extraordinary quantity of basic science and applied clinical research published each year, a high degree of professional specialisation, and the critical importance of timely decisions.

Evidence-based Medicine (EBM), an approach to therapeutic decision making developed by clinicians and epidemiologists working at McMaster University during the 1980s and 1990s, has come to play a prominent role in medicine over the last two decades. Founders include past and present faculty at McMaster, including David Sackett, Gordon Guyatt, David Haynes and Deborah Cook. These clinicians and academics, with numerous co-authors, have written two 'guidebooks' on how to practise and teach evidence-based medicine and many

[1] See, for example, the 1835 paper on the "application of the numerical method" to therapeutic decisions reprinted in Poisson et al (2001), and Feinstein (1996)

additional papers. While aspects of EBM continue to be debated—see, for instance, one of the several thematic issues of *Journal of Evaluation in Clinical Practice*—, the influence of EBM on medical practice has been profound. One needs to look no further than the numerous academic centres, the format of any clinical guideline (for almost any condition), or the increasing number of 'evidence-based' disciplines that have materialised. The wide-ranging impact of EBM led Brendan Reilly (2004, p. 992) to remark, '. . . anyone in medicine today who does not believe [in EBM] is in the wrong business'.

Proponents of EBM urge clinicians to base decisions on the outcomes of large randomised studies *rather than* the mechanistic understanding of pharmacology and physiology provided by basic science. I argue against this account. I use Deborah Mayo's account of experimental inquiry to illustrate how the mechanistic understanding provided by basic science is central to the design, analysis and application of clinical research. I introduce EBM and the attitude of EBM proponents to basic science in Section 2. In Section 3, I clarify what is meant by 'basic science' and 'mechanism' in this context. I outline the framework of models in experimental inquiry in Section 4. In Section 5, I use this framework of models to illustrate the important role that the mechanisms of basic science play in designing and interpreting clinical studies. This argument is extended in Section 6 to show how the mechanisms of basic science play an important role in *applying* the results of randomised studies to individual patients.

2 Evidence-based medicine and basic science

EBM was introduced to remedy a number of traditional and emerging problems with medical decision making. In the first paper to articulate EBM, proponents argued that medical decisions relied too heavily on clinical experience and basic science.

> 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 Working Group, 1992)

EBM's response to this problem is the 'hierarchy of evidence'. EBM suggests that therapeutic decisions should be based of the best available evidence. On EBM's account, methods that reside higher up the hierarchy of evidence provide better, and more reliable, evidence for therapeutic decisions. Randomised studies (and systematic reviews of randomised studies) are placed higher in the hierarchy than observational studies, which in turn are placed higher than basic science and unsystematic clinical experience.[2]

Operationally, proponents of EBM provide clear advice for therapeutic decision making. The popular EBM guidebooks recommend applying the hierarchy

[2]For examples of EBM's hierarchy of evidence, see Phillips et al (1998); Guyatt and Rennie (2002).

categorically. Evidence from randomised studies is taken to *trump* evidence from lower down the hierarchy, including evidence from basic medical science:

> The hierarchy implies a clear course of action for physicians adressing 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 evidence 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 physiological studies that are related only indirectly—but there is always evidence. (Guyatt and Rennie, 2002, pp. 14–15)

> If a 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 categorical interpretation of EBM's hierarchy of evidence implies that the mechanistic understanding provided by basic science should play little, if any, role in therapeutic decisions.

An advertised benefit of the hierarchy of evidence—and another problem that EBM was introduced to combat—is that it makes therapeutic decisions more expedient. There is far more research published each month than there is time to read it (Sackett et al, 1996, p. 72). The hierarchy of evidence allows busy clinicians to narrow the range of the medical literature they need to search. This is reflected in the advice provided by Straus and colleagues (quoted above)—if the study is not randomised, it can be set aside.

It is important to recognise the benefits that EBM has brought about. Most significantly, EBM has bolstered the view that medical decisions should be based on *evidence*, rather than eminence, or some version of the view that medical decisions are irreducibly tied to the tacit knowledge of an individual clinician. Leaving aside whether the specific account of evidence adopted by EBM is the best account, EBM has played a role in changing what is considered an appropriate justification of medical decisions. There are also a host of more tangible benefits. Those working in EBM have improved the accessibility of clinical research by developing abstracting and summarising services. They have aided the development of health professionals by emphasising the importance of fundamental literature searching and critical appraisal skills. And, those working in EBM are largely responsible for the large improvement in the reporting of clinical research through the development of guidelines such as the CONSORT Statement (Schulz et al, 2010), and publications auditing the quality of clinical research reports. But none of these benefits should impede reflection on the foundations of EBM.

EBM's central epistemic claim is that evidence listed higher in the hierarchy of evidence provides more reliable evidence for therapeutic decisions. EBM's claim to improve the expedience of therapeutic decisions relies on this epistemological claim and on the categorical interpretation of the hierarchy—clinicians need only worry about randomised studies because evidence from these studies trumps evidence from lower down the hierarchy.[3] There are problems for EBM's central epistemic claim and for the categorical interpretation of the hierarchy of evidence (see La Caze (2009) and (2008) for an examination of the justification of the hierarchy of evidence, and the categorical interpretation of the hierarchy respectively). The main problem for EBM's epistemic claim is that while the claim relates to *therapeutic decisions*, the justification of the hierarchy relates to the *comparative internal validity* of the methods listed. Internal validity is the degree to which the results of a study are accurate for the sample of participants involved in the study. All other things being equal, a study design listed higher in the hierarchy of evidence has the capacity to rule out (or minimise) more sources of error than studies listed lower down the hierarchy. But there is more to therapeutic decisions than studies with high internal validity. For a start, therapeutic decisions rely on *apply* the results of clinical studies to specific individuals—individuals who may, or may not, share the social or physiological characteristics of the sample of participants in the study. Multiple sources of evidence play an important role in deciding whether and how the results of clinical research apply to the individual in the clinic. So, contrary to EBM's hierarchy of evidence, multiple sources of evidence are important to therapeutic decisions. While there are other problems with using the hierarchy of evidence to directly inform therapeutic decisions, many of these problems have their root in EBM's extension of the hierarchy of evidence beyond the domain in which it is justified.[4] The focus of this paper is to examine the importance of basic science to therapeutic decisions, and in particular, to illustrate the importance of basic science to the design, interpretation and application of applied clinical research.

Proponents of EBM use a number of prominent case studies to argue for the use of randomised studies rather than basic science in therapeutic decisions. Of particular note are the Women's Health Initiative study providing compelling evidence that hormone replacement therapy with oestrogen and progestogen in postmenopausal women is not cardioprotective (Women's Health Initiative Writing Group, 2002) and the Cardiac Arrhythmia Suppression Trial providing compelling evidence that encainide and flecainide worsen rather than improve the risk of cardiac death in patients who have suffered a myocardial infarction

[3] A different 'expedience' argument could be mounted without interpreting the hierarchy of evidence categorically. If, on occasion, evidence from lower down the hierarchy could be more important to a decision about a specific patient (despite the availability of evidence obtained from a method listed higher), it might be argued that the cost of finding this evidence outweighs the benefits, either in terms of the individual patient, or over the course of the many decisions clinicians need to make. This argument, however, is not provided in the key EBM texts or articles.

[4] Again, La Caze (2009) provides further discussion.

(Echt et al, 1991).[5] Importantly, evidence from the Women's Health Initiative Study and the Cardiac Arrhythmia Suppression Trial overturned evidence from observational studies and basic medical science which suggested these interventions were beneficial. In both cases, the intervention was used widely prior to the large randomised study providing evidence of harm.

These sobering examples illustrate that evidence from randomised trials can improve decisions, and that sometimes randomised trial evidence will overturn findings from observational studies and basic medical science. But the examples don't support the categorical interpretation of the hierarchy of evidence. Specifically, the examples don't show that evidence from randomised trials should always trump the mechanisms provided by basic medical science when making decisions about patient care. An important role for basic science in therapeutic decisions is entirely compatible with these examples. This positive role for basic science is difficult to compartmentalise in the way that is implied by EBM's hierarchy of evidence. Framing the debate in terms of which is more important is unhelpful; basic science and clinical research are enmeshed. Even when evidence from a well conducted randomised trial is available, some therapeutic decisions might rely more on mechanistic evidence than the statistical evidence provided by the clinical trial. Often both types of evidence will be appealed to and the relative influence of mechanistic and statistical evidence on a particular therapeutic decision will be difficult to discern.

Proponents of EBM have occasionally (though prominently) endorsed the view that basic science plays a greater role in therapeutic decisions than suggested by the hierarchy of evidence. Sackett et al (1996) defend EBM as a means of integrating clinical expertise with the best available external clinical evidence. Sackett et al. clarify:

> By best available external clinical 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 clinical examination), the power of prognostic markers, and the efficiency and safety of therapeutic, rehabilitative, and preventative regimens. [Emphasis added.]

This definition of 'best available clinically relevant research' is hard to reconcile with EBM's hierarchy of evidence and *raison d'être*. Maybe Sackett and colleagues intended to emphasise patient-centred clinical research more than the suggestion that best evidence 'often comes from from the basic medical sciences'. In any case, everything else Sackett et al. say in this paper re-affirms EBM's hierarchy of evidence, especially in the context of therapeutic decisions.

In addition, the Evidence-Based Medicine Working Group (1992, p. 2421) acknowledge that basic science is necessary to apply the results of randomised studies.

[5] Sackett (2006, p. 177) discusses hormone replacement therapy (and the Women's Health Study indirectly), and Devereaux and Yusuf (2003, p. 108) cite the Cardiac Arrhythmia Suppression Trial.

> A sound understanding of pathophysiology is necessary to interpret and apply the results of clinical research. ... Understanding the underlying pathophysiology allows the clinician to better judge whether the results are applicable to the patient at hand ...

The EBM Working Group is right to emphasise the importance of basic science in applying evidence from randomised studies, but again, this advice conflicts with what is said about applying the hierarchy of evidence.

EBM makes inconsistent claims about the role basic science in therapeutic decisions. On one hand EBM places basic science on the lowest level of the hierarchy and advises clinicians to inform therapeutic decisions from evidence from as high up the hierarchy as possible. On the other hand, EBM acknowledges a role for basic science in applying the results of clinical research, but nowhere do they square this view with the hierarchy of evidence. The EBM literature focusses much more on the former claim than the latter. Before taking up the task of arguing that basic science plays an important role in therapeutic decisions, some clarificatory remarks are needed on what is meant by 'basic medical science'.

3 Basic science and mechanisms

The 'basic' or 'bench' sciences of medicine are biological sciences, especially physiology, pharmacology and related disciplines such as pathophysiology. A central aim of biological science is to discover, understand and refine *mechanisms*. When proponents of EBM refer to basic medical science they refer to the mechanistic explanation provided by these sciences.[6] There has been much recent work on mechanisms and mechanistic models in the philosophy of science. A number of philosopher's have provided mechanistic accounts, including Stuart Glennan (1996), Peter Machamer, Lindley Darden and Carl Craver (2000) and William Bechtel and Robert C. Richardson (1993). There are important differences between these accounts, but their central aim is similar; each highlight, and look to explain, the importance of mechanisms in biology and associated sciences. Any one of these accounts could be used to flesh out the points I wish to make in relation to medical decisions. To assist discussion I provide some details of the Machamer-Darden-Carver account.

Machamer et al (2000, p. 3) define 'mechanisms' as the following:

[6] 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. Mechanism is suggested in the use of the term 'pathophysiologic rationale' in the original statement of EBM quoted above. Mechanism is also implied in the notions of 'biological plausibility' suggested in the following quote from statistician Douglas Altman:

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

Quotes such as these can be multiplied.

> Mechanisms are entities and activities organised such that they are productive of regular changes from start or set-up to finish or termination conditions.

Mechanisms are analysed in terms of entities and activities. Entities are the things postulated by the mechanism, and activities are what the entities do. A 'mechanism schema' is an abstract description of a mechanism. In a schema, known aspects of the mechanism are abstracted away. Schemata can be used to (i) provide a mechanistic explanation (when the schema is instantiated); (ii) yield predictions and (iii) provide a 'blueprints' for designing research (that is, schemata can be used to design tests to refine and, if necessary, refute mechanistic hypotheses)(Machamer et al, 2000, pp. 15–18). Biologists, as Machamer et al (2000, p. 16) note, often use the term *theory* to refer to mechanism schemata.

Basic medical science provides a plethora of mechanisms; physiological mechanisms of how biological processes work, pathophysiological mechanisms of how physiological processes go awry, and pharmacological mechanisms of how drugs combat illness. In sections 5 and 6 I will discuss how mechanisms are important for analysing and applying the results of clinical research. In many cases the mechanisms of basic medical science have strong evidential support—for instance, the background conditions required for the mechanism to operate can be well known, the mechanism can be demonstrated *in vitro*, say, in paradigmatic animal models, and the mechanism can be intervened upon in a way that produces predictable and regular changes. These are how-actually mechanisms (following the Machamer-Darden-Craver terminology). However, less complete mechanistic explanations are also important to therapeutic decisions. Mechanism sketches and how-possibly mechanistic models can play a role in interpreting the findings of randomised trials, providing the provisional nature of these sketches and models are kept in mind.[7]

Biological mechanisms, and the biological evidence which support these mechanisms need to be distinguished from the *application* of these mechanisms in patient care. Many drugs have promising pharmacological properties that, for one reason or another, do not bring about the expected beneficial outcome in patients. This can occur even when the drug works by a well-supported biological mechanism that should assist patients suffering a particular condition.[8] Much is unknown in clinical science. Pharmacological/pathophsyiological mechanisms sometimes predict patient outcomes, and sometimes they don't; in any particular instance, it is often unknown which will be the case until applied clinical studies have been conducted. It is also possible to have reliable evidence on the effects of an intervention without knowing the mechanism by which these effects are brought about. For instance, the effects of paracetamol are very well

[7] Conversely, the findings of randomised trials can provide leads for biological research, and possibly help flesh out the mechanism sketch.

[8] Hormone replacement therapy and flecainide, as discussed earlier, provide striking examples of instances in which a reasonably well supported putative mechanisms do not necessarily lead to benefits in patients. Note, that examples such as these don't necessarily mean that the mechanism is not instantiated; there could, for instance, be unexpected harms which swamp the postulated benefits.

understood, but only a sketch of paracetamol's analgesic and antipyretic actions can be provided.

Proponents of EBM are right to suggest that mechanisms, and reasoning about mechanisms, are not (solely) sufficient for therapeutic decisions—especially when other kinds of evidence are available. A randomised study showing that a therapy is effective in patients under routine care can provide more compelling evidence for therapeutic decisions than pharmacological evidence that a drug works by a mechanism understood to treat a disease. But, EBM's account of medical evidence fails to recognise how interrelated the mechanisms of basic science and applied clinical research are. EBM is wrong to treat the different kinds of medical evidence as discrete. The problems of EBM's account of medical evidence become especially clear when judging the relevance of clinical studies for individual patients. This is the problem of 'external validity'. External validity is the extent to which the results of a study can be generalised to patients outside of the study, it is usefully contrasted with internal validity. Despite being well recognised, there is precious little in the EBM literature on how the problem of external validity might be overcome. This is because any reply to the problem of external validity relies on interpreting clinical research in light of basic science. EBM is left short because it lacks an account of the relation between basic science and clinical research.

4 Models: Primary models, models of experiments and models of data

Randomised studies are analysed according to frequentist statistics. Specifically, both hypothesis testing and estimation are central to trial analysis. Trial analysis proceeds by linking the generated raw data to the hypothesis under test and the parameter being estimated. To test the data against the hypothesis, and to provide rigourous estimates, assumptions are made at a number of levels. For instance, assumptions are made about the general theory and the primary hypothesis under test; about how the methods employed in the study provide a test of the primary hypothesis; and about how the raw data can be reduced to provide a convenient measure of the parameter under examination. The best way to keep track of these assumptions is to recognise the framework of models employed to gain information from randomised studies.

Mayo (1996, pp. 128–173) provides a helpful outline of the framework of models employed in the analysis of experiments.[9] There is nothing particularly controversial about the framework of models, it is after all an explication of the standard approach to statistical analysis.[10] My claim is that recognising the framework of models makes explicit the role that basic science plays in the

[9] Patrick Suppes (1962) is central to Mayo's (1996, pp. 132–144) account.

[10] There is, of course, much debate surrounding different approaches to statistical inference. But the framework of models, and their role in statistical analysis, is orthogonal to debates over preferred methods of statistical inference.

design, analysis and application of clinical trials. The mechanisms of basic science, even mere sketches, often play an important role in interpreting evidence from randomised studies, and thus are important for clinical decisions (I'll provide examples to make this claim more concrete in the following sections). To be clear, I don't think this claim is particularly controversial either, but it is obscured by EBM's hierarchy of evidence, and by what proponents of EBM say about putting the hierarchy of evidence into practice.

While delineation of an experimental inquiry into the three types of models is by no means exact, recognition of the models employed in trial analysis highlights the kinds of questions and concerns common to assessing the validity of most clinical trials. The 'primary model' puts the general theory or inquiry into a testable form. For instance, the first step in hypothesis testing is to specify the null hypothesis. The null hypothesis specifies a particular value for the unknown parameter, often that the intervention under examination will have no effect on a particular variable. What kind of variation from the null hypothesis would be considered important is also specified and labelled the alternative hypothesis. The null and alternative hypotheses, which are representations of the general theory, are considerably more specific than the general theory.

'Experimental models' link primary models with data models. To continue with the example of hypothesis testing, the experimental model includes the sampling distribution for the variable under investigation on the assumption that the null hypothesis is true. The sampling distribution considers how the random variable under investigation will vary on repetitions of the trial assuming the null hypothesis is true. How the process is understood to operate may suggest an appropriate distribution from the family of probability models typically used in statistical analysis, or the methods employed in the experiment may generate data consistent with a particular distribution. The experimental model is linked to the primary model. The experimental model is a more specific representation of the primary theory, a representation that takes into account the experimental conditions.

The experimental model is also linked to the data model. As Mayo (1996, p. 134) notes, the experimental model turns the primary theory into a statistical hypothesis. The sampling distribution under the assumption of the null hypothesis (and the data model) allows comparison of observed data with data predicted by the primary model. This permits statistical testing. For instance, a significance test specifies which values of the variable under investigation represent a large enough discrepancy from the null hypothesis such that the null hypothesis can be rejected (within the dictates of frequentist statistics).

The link between raw data and the experimental model is provided by the 'data model'. The data collected in any randomised study is likely to be vast, and include much that does not pertain to the hypothesis under investigation. The information drawn from the data needs to be reduced in a way that retains everything that is important to the hypothesis. Even once focus is reduced to those aspects of the data that are related to the primary hypothesis, there will be more than one way to represent that data. Consider an intervention that is thought to reduce mortality. Mortality in the intervention and control group can

be compared according to the fraction of patients alive at the end of the trial; or the mortality rate per year of the study; or the average time of survival since diagnosis, or one of a number of alternative measures of mortality. The data model provides a 'test statistic'. The test statistic provides a representation of the data, and this representation is used to compare predicted and observed data via the sampling distribution and the observed value of the test statistic. An example of the kinds of decisions that need to be made to specify the data model is provided in the next section.

Supplementing these three models are considerations of trial design as well as a host of ceteris paribus conditions. Many aspects of trial design relate directly to assumptions of the primary, experimental and data models. Some of the assumptions of trial design are formally tested. Randomisation and the assumption of proportional hazards are examples (and are discussed below). D. R. Cox (2006, p. 3) labels this process 'model criticism'. Other aspects of trial design, and many additional ceteris paribus conditions, are assumed to hold for the models employed in the experimental inquiry, but are not formally tested. This includes many intuitive considerations, such as the irrelevance of eye colour to the response to a therapy, and the irrelevance of the day of the week that patients are recruited on study outcomes (and many others).[11]

Mayo (1996) develops the framework 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 (and the models of experimental inquiry) to develop an argument from error. Data provides information on the hypothesis in as much as the hypothesis has been 'severely tested', where a severe test is a test that successfully rules out (or minimises) as many sources of error as possible.

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests taken together) that has a very high probability of detecting an error if (and only if) it exists nevertheless detects no error. (Mayo, 1996, p. 184)

Mayo's notion of a 'severe test' is embedded within the models of the experiment. The models of data, experiment and primary theory provide the framework for testing hypotheses. The models of experimental inquiry explicate and provide the means to assess the auxiliary assumptions that are employed to learn from data.

My reason for discussing the framework of models is to illustrate the importance of basic science to therapeutic decisions—something that is downplayed by EBM's hierarchy of evidence. First, basic science plays an important role in specifying and interpreting applied clinical research. I illustrate this role with an example in the next section. I provide additional details in this example

[11] As Mayo (1996, pp 139–141) observes, should a ceteris paribus condition come under question, a decision can be made to include it in the formal analysis of the trial.

in order to flesh out each level of the framework of models. The second, and more important reason for discussing the framework of models, is that it helps illustrate the vital role the mechanisms of basic science play in *applying* clinical research to the problems of individual patients. This is discussed in Section 6.

5 Basic science and internal validity

Rofecoxib is an anti-inflammatory agent that was withdrawn following evidence that it increased the risk of heart attacks and strokes (Bresalier et al, 2005). 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, in part, by mechanistic considerations. Developments in the mechanistic understanding of how non-steriodial anti-inflammatory drugs produce their benefits and harms led to the hypothesis that if a drug could selectively inhibit COX-2 it would cause less gastrointestinal damage compared to older anti-inflammatories, such as aspirin, indomethacin and the like.[12] 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, the same mechanism by which rofecoxib reduces the risk of serious gastrointestinal bleeds. I will use this example to illustrate the framework of models between the primary scientific question of whether rofecoxib increases the risk of thrombosis, and the data observed in APPROVe. There are many statistical considerations in specifying the models, I will focus on the role that mechanisms of basic science play in model specification and assessment. It is because of this role that basic science is important for assessing the internal validity of clinical trials.

The primary theoretical question is formulated on the mechanism schema. Is the mechanism by which rofecoxib works likely to increase the risk of blood clots? The primary model turns this question into a something that can be tested. The primary question may be represented by 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, $RR_\mu = \mu_R/\mu_C$. The null hypothesis can be specified as, $RR_\mu = 1$, and the alternative hypothesis can be specified as, $RR_\mu > 1$.

The experimental model requires specifying the sampling distribution to test the null hypothesis. If RR_μ is taken to be the unknown parameter of interest, the test statistic observed in APPROVe can 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

[12] Notably, 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.

are each random variables. Notice, while the primary model refers to unknown population values (such as μ_R) the test statistic is observable and finite. By making some assumptions about the process under investigation the sampling distribution of the test statistic assuming the null hypothesis is true can be specified. The sampling distribution under the null hypothesis can then be divided into accept and rejection regions, and the standard statistical tests conducted.

Mechanistic knowledge of the underlying process is important to the specification and assessment of the experimental model—it is by no means the *only* consideration, but it is an important consideration. For instance, the sampling distribution and the statistical tests rely on the assumption that random allocation has provided two comparable experimental groups. 'Comparable', here, means that patient factors that may plausibly influence the outcome measure are roughly evenly distributed in the intervention and control groups. As the sample size increases the chance that a small number of specified patient factors are unevenly distributed is reduced, but in any actual allocation there will always be differences between experimental groups. Whether the experimental groups are sufficiently comparable requires judgement and this judgement is informed, at least in part, by basic science.

Of all the possible patient differences, basic science helps to judge which patient factors need to be roughly evenly distributed in the experimental groups. Co-morbid illnesses such as diabetes or rheumatoid arthritis may influence the rate of thrombotic events, and thus confound the trial results if unevenly distributed; other factors, such as, say, eye colour are unlikely to be important. The separation of possible and unlikely confounders is made on the basis of basic science. There are pathophysiological mechanisms which explain how diabetes may promote thrombosis; there are no such mechanisms for the effects of eye colour. If factors that are known to influence the mechanism under investigation are distributed unevenly in the randomised study, then the experimental model—and the statistical tests based on the experimental model—will be invalid. Hence, one aspect of assessing the internal validity of a trial involves assessing the consistency of the experimental model with basic science.

Purely statistical considerations play a large part in selecting the data model, but basic science is also important. Again, assessing the validity of the model, in this case the choice of test statistic, will involve assessing the consistency of the model with basic science. Proportional hazards is one assumption that needs to hold for RR_X to be a suitable overall measure of risk in the trial. Proportional hazards assumes that the relative risk remains constant over time. If the assumption of proportional hazards is violated it can undermine (or at least change the interpretation) of the statistical tests conducted in the trial. For instance, the value observed for RR_X in APPROVe is consistent with a constant increase in risk of thrombotic events; an initial reduction in the risk of thrombotic events and then a slow longer-term increase in risk, and a large initial increase in thrombotic risk and a slow longer-term reduction in the risk. The main way of assessing whether proportional hazards holds in a trial is to test the assumption empirically, nevertheless if the assumption is at odds with

what is known about the process under investigation—the mechanism—then an alternative data model should be specified in the trial set-up.

Assessing whether proportional hazards is a valid assumption in APPROVe provides a striking example of the role of basic science in specifying and assessing data models. Time-to-event analyses were used to measure the occurrence of thrombosis in the experimental groups over time. The time-to-event analysis provides a cumulative incidence function for thrombotic events in the rofecoxib and placebo groups. The cumulative incidence function is a function of both the number of first thrombotic events, and the sum of the amount of time each patient remains in the population (a patient is considered a member of the population until they suffer a thrombotic event). Specifying the cumulative incidence function requires defining what will count as a thrombotic event. Among other things, a decision needs to be taken as to whether a thrombotic event which occurs during the study but after a patient has stopped taking the treatment will be included in the analysis (and, hence, attributed to the drug). A patient suffering a thrombotic event the day after stopping rofecoxib should be included in the analysis (after all rofecoxib will still be present in the patient's system), but it is less clear that a patient suffering a thrombotic event 3 years after stopping rofecoxib should also be included in the analysis. If a line is to be drawn, where should it be drawn? This is a debate over the choice of the data model, and it was a controversial matter in APPROVe (see Lagakos, 2006 and Nissen et al, 2006 for discussion).

In the original analysis a patient suffering a thrombotic event was included in the analysis if the event occurred during treatment or up to 14 days after treatment had ceased, however in data submitted to the Food and Drug Administration 12 months later, patients suffering a thrombotic event were included in the analysis if they suffered an event at any time of the study, regardless of whether or not they were still taking the experimental treatment (that is, the data were analysed according to 'intention-to-treat'). The latter analysis suggests that the risk of a thrombotic event for patients randomised to rofecoxib presents earlier and is larger than that suggested by the analysis presented in the original report. There is a tendency in Lagakos (2006) and Nissen et al (2006) to discuss this issue on entirely statistical grounds (focussing on the benefits or otherwise of intention-to-treat analyses of safety endpoints, and whether or not such analyses are considered conventional). Statistical issues aside, basic science makes an important contribution to this discussion. While what is understood about the mechanism of rofecoxib causing thrombosis might not settle the issue, it can help discern between data models, or at least help to interpret whichever data model is selected. As it stands nothing in the biological evidence for rofecoxib's effects suggests that the drug would increase the risk of thrombosis long after it has cleared the body.

Basic science is important to the interpretation of clinical research. Once set up, results of randomised interventional studies require interpretation and application, they do not issue as if from some 'black box'. I don't wish to overplay the role of basic science in interpreting the findings of clinical research—there are other important considerations—, but I do want to counter EBM's hierarchy

of evidence and the categorical interpretation of the hierarchy provided in some prominent accounts of EBM. The rofecoxib case illustrates that the mechanisms of basic science and the statistical findings of clinical research are not so easily separated. This can be seen in the way the primary model, experimental model and data model are specified. What is known or supposed at the mechanistic level plays a vital role in specifying both the models of experiment and the models of data. And once specified, the statistical findings based on the observed data provide information on the clinical applicability of the mechanism provided by basic science only if the assumptions made in the experimental and data model hold. Informally, which experiments are conducted and what data are analysed as tests of the clinical applicability of the mechanism, depend on the details of the mechanism. And the ability of the data and the experiment to provide important information about whether the activities specified by the mechanism work as expected in the clinical context, depends on the adequacy of the assumptions made in specifying the experiment and the data. Theory, experiment and data are all linked; the basic science that specified, and helps to assess, the models of the experiment are *part* of the results of applied clinical research.

6 Basic science and external validity

The framework of models shows that the result of a trial is not simply a statement of the results of the statistical tests conducted within the trial. The statistical tests sit within a broader epistemology of experiment, and this epistemology is fleshed out through the framework of models. The mechanisms of basic science play a role in specifying, assessing and interpreting the models employed in randomised studies. This richer account of experimental learning helps to approach the challenge of external validity (that is, the challenge of applying the results of clinical studies to patients outside of the trial).

Two 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 are likely to 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.

The mechanisms of basic science are important for formulating a response to the questions of external validity. Even if the first problem could be overcome, 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 remain. Does the overall result of the trial obscure subgroups within the trial that respond particularly well or particularly poorly to treatment? The interpretation of statistical analyses on subgroups within randomised trials is notoriously difficult (for discussion see Feinstein, 1998 and Peto et al, 1995,

especially, p. 35). Brookes et al (2001), in a simulation study, showed that subgroup analyses often fail to indicate a true difference between subgroups (7–26% of tests under the conditions they simulated), and frequently indicate a difference between subgroups when no difference exists (41–66% of tests under the conditions they simulated). The problem is one of selecting the appropriate reference class. For most clinical trials, it can be assumed that the trial sample is not a homogenous reference class (in the Salmon, 1971, pp. 40–47 sense). The second problem of external validity raises the question of whether the trial sample should be considered (until further notice) as a homogenous reference class, or whether the sample can be split into two or more homogenous references classes based on some property or properties. Once again, the statistical results of the trial alone do not answer this question.

Mechanistic reasoning helps to judge whether the reference class is homogenous (or, homogenous enough for practicable purposes). The very large ISIS-2 trial provides an example. ISIS-2 examined the effects of streptokinase and aspirin in patients suspected of suffering a myocardial infarction (Group, 1988). In addition to the planned overall study results, the authors provided a small number of *post hoc* analyses. In one analysis they found that patients who were born under the star signs Gemini or Libra and received aspirin appeared to do worse than patients who received aspirin and were born under the other star signs. In another *post hoc* subgroup analysis they found that patients with ST depression on their pre-randomisation electrocardiogram did not receive the same benefit from streptokinase as patients with ST elevation on their electrocardiogram. The statistical features of these two analyses are similar—both focus on *post hoc* subgroups, and both can be assumed to be of questionable reliability. There is an important difference however.

Independent pathophysiological mechanisms can predict the results of the second analysis, but not the first. Patients suffering a myocardial infarction often progress from ST depression to ST elevation as a complete fibrin-linked clot develops. The action of streptokinase is targeted towards dissolving the fibrin in such clots. Pathophysiology predicts that steptokinase will be more beneficial to patients with ST elevation on their electrocardiogram than ST depression. No such account can be given for the influence of star sign on the effects of aspirin. The independent pathophysiological mechanism in the case of the streptokinase subgroup analysis is not enough by itself to overcome to statistical problems of the *post hoc* subgroups analyses, but it is enough to distinguish this analysis from the subgroup analysis based on star sign. The streptokinase subgroup analysis is worthy of further investigation, further investigating the star sign subgroup analysis is considerably less likely to be fruitful. Basic science suggests how the sample in a trial might be partitioned. Indeed, in the time that has elapsed since the ISIS-2 trial, ST elevation has been shown to be a strong indicator of a positive response to clot-lysing drugs such as streptokinase.

When the statistical evidence provided by a randomised trial conflicts with the mechanistic evidence provided by basic medical science, it will sometimes be appropriate for the clinician to base their decision on the statistical evidence, and it will sometimes be appropriate for clinician to base their decision on the

mechanistic evidence. The details of the case will dictate as to which source of evidence is more appropriate—not some general rule provided on the basis of the hierarchy of evidence. But this formulation of the problem suggests that the results of applied clinical research are easily separated from the mechanisms of basic science. I have argued against this view. It is the integrated whole—the conjunction of basic science with the statistical findings of applied clinical research—that needs to be considered when making therapeutic decisions. It is the integrated whole that 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. Clinical research may refute basic science, but more often it refines and improves the understanding of how the mechanisms described in basic science are realised in clinical care. Just as basic science alone fails to predict patient outcomes, the statistical findings of clinical research alone fails 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, considerably more progress can be made by recognising the connections between these sources of evidence.

7 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, EBM can only be interpreted as making the rather facile claim that basic science *alone* does not provide sufficient evidence for therapeutic decisions. The statistical results of applied clinical research sit within a much richer framework of models, including theoretical models based on the mechanisms of basic science. It is this richer framework that must be considered when applying the results of clinical research to therapeutic decisions.

References

Altman DG (1998) Within trial variation—a false trail? Journal of Clinical Epidemiology 51(4):301–303

Bechtel W, Richardson RC (1993) Discovering Complexity: Decomposition and localization as strategies in scientific research. Cambridge Univeristy Press, Cambridge

Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargos R, Davis B, al e (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

Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, Lines C, Riddell R, Morton D, Lanas A, Konstam MA, Baron JA (2005) Cardio-

vascular events associated with rofecoxib in a colorectal andenoma chemoprevention trial. New England Journal of Medicine 352(11):1092–1102

Brookes S, Whitely E, Peters T, Mulheran PA, Egger M, Davey Smith G (2001) Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment 5(33):1–58

Cox DR (2006) Principles of Statistical Inference. Cambridge Univeristy Press, Cambridge

Devereaux PJ, Yusuf S (2003) The evolution of the randomized controlled trial and its role in evidence-based decision making. Journal of Internal Medicine 254:105–113

Echt D, Liebson P, Mitchell L, Peters R, Obias-Manno D, Barker A, Arensberg D, Baker A, Friedman L, Greene H, et al (1991) Mortality and morbidity in patients receiving encainide, flecainide, or placebo. the cardiac arrhythmia suppression trial. The New England Journal of Medicine 324(12):781–788

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

Feinstein A (1996) Two centuries of conflict-collaboration between medicine and mathematics. Journal of Clinical Epidemiology 49(12):1339–1343

Feinstein AR (1998) The problem of cogent subgroups: A clinicostatistical tragedy. Journal of Clinical Epidemiology 51(4):297–299

Food and Drug Administration Advisory Committee (2001) Cardiovascular safety review of rofecoxib. Tech. rep., Food and Drug Administration/Centre for Drug Evaluation and Research, URL `http://www.fda.gov/ohrms/dockets/ac/01/briefing/3677b2_06_cardio.pdf`

Glennan S (1996) Mechanisms and the nature of causation. Erkenntnis 44(1):49–71

Group IC (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

Guyatt GH, Rennie D (eds) (2002) Users' guide to the medical literature: Essentials of evidence-based clinical practice. American Medical Association Press, Chigaco

La Caze A (2008) Evidence-based medicine can't be .... Social Epistemology 22(4):353–370

La Caze A (2009) Evidence-based medicine must be .... Journal of Medicine and Philosophy 34:509–527

Lagakos SW (2006) Time-to-event analyses for long-term treatments—the APPROVe trial. New England Journal of Medicine 355(2):113–117

Lip GYH, Lowe GDO (1996) ABC of atrial fibrillation: Antithrombotic treatment for atrial fibrillation. British Medical Journal 312(7022):45–49

Machamer P, Darden L, Craver C (2000) Thinking about mechanisms. Philosophy of Science 67(1):1–25

Mayo DG (1996) Error and the Growth of Experimental Knowledge. Science and Its Conceptual Foundations, University of Chicago Press, Chigaco

Nissen SE, Furberg CD, Bresalier RS, Baron JA (2006) Adverse cardiovascular effects of rofecoxib. The New England Journal of Medicine 355(2):203–205

Peto R, Collins R, Gray R (1995) Large-scale randomized evidence: Large, simple trials and overviews of trials. Journal of Clinical Epidemiology 48(1):23–40

Phillips B, Ball C, Sackett DL, Badenoch D, Straus SE, Haynes RB, Dawes M (1998) Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001). Tech. rep., URL `\url{http://www.cebm.net/?o=1023}`

Poisson, Dulong, Larrey, Double (2001) Statistical research on conditions caused by calculi by Doctor Civiale. International Journal of Epidemiology 30(6):1246–1249

Reilly B (2004) The essence of EBM. British Medical Journal 329:991–2

Sackett DL (2006) The principles behind the tactics of performing therapeutic trials. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P (eds) Clinical Epidemiology: How to Do Clinical Practice Research, 3rd edn, Lippincott Williams and Wilkins, Philadelphia, pp 173–243

Sackett DL, Rosenberg W, Gray JAM, Haynes B, Richardson WS (1996) Evidence based medicine: What is it and what it isn't. British Medical Journal 312(7023):71–72

Salmon WC (1971) Statistical Explanation and Statistical Relevance. University of Pittsburgh Press

Schulz KF, Altman DG, Moher D (2010) Consort 2010 statement: Updated guidelines for reporting parallel group randomized trials. Annals of Internal Medicine 152(11):1–7

Straus SE, Richardson WS, Glasziou P, Haynes RB (2005) Evidence-Based Medicine: How to Practice and Teach, 3rd edn. Elsevier Churchill Livingstone, London

Suppes P (1962) Models of data. In: Nagel E, Suppes P, Tarski A (eds) Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress, Stanford University Press, Stanford, pp 252–61

Women's Health Initiative Writing Group (2002) Risks and benefits of estrogen plus progestin in health postmenopausal women. Journal of the American Medical Association 288:321–333.