10

Inductive Risk and the Role of Values in Clinical Trials

Robyn Bluhm

There is a clear consensus on the goal of clinical trials: it is to determine whether potential interventions are effective and safe, and thus to improve the health of patients by ensuring that they receive proven treatments. Yet a closer look at discussions in the clinical, bioethics, and philosophy literature shows that this broad consensus masks heated disagreement about how studies should be designed in order to best reach that goal. In this chapter, I consider three debates about how clinical trials should be conducted and show both that they can profitably be analyzed from the perspective of inductive risk and that they raise important issues relevant to the philosophical discussion of inductive risk. The three debates are: (1) whether randomization is the most important aspect of study design, as is suggested by the developers of evidence-based medicine (EBM); (2) whether clinical trials should be pragmatic or explanatory in design; and (3) when clinical trials should use placebo controls.

Although these three debates raise overlapping questions and issues, they have largely been conducted separately. I will show, however, they are all fundamentally disagreements about appropriate study design that can be understood as differing positions on how to handle inductive risk. Moreover, I will show that in all of the debates, methodological and ethical considerations are inextricably linked—and this linkage has implications for the philosophical question of the appropriate role for non-epistemic values in science. Specifically, I draw three lessons for the philosophical discussion of inductive risk. First, considerations of inductive risk need not take the form of a

trade-off between distinct consequences associated with false positive versus false negative results. Second, while discussions of inductive risk have tended to focus on the "quantitative" question of how much evidence is needed to support a hypothesis, the "qualitative" question of what kind of evidence should count also plays an important role. Finally, my analysis of the three debates in clinical research emphasizes the extent to which the data in support of a hypothesis depend on value-laden decisions about which methods to use; this complicates the issue of distinguishing between direct and indirect roles for values.

Clinical Trial Design and Inductive Risk

Clinical trials use methods derived from epidemiology to test whether new treatments are effective and safe. While these trials can have a variety of methodological characteristics, in this section, I will introduce the key characteristics of clinical trials and show how trials are relevant to the existing discussion on inductive risk, by describing a simple, two-arm randomized controlled trial (RCT).

RCTs are generally considered to be the "gold standard" test of potential new therapies. In an RCT, eligible study participants are divided into two groups, only one of which receives the experimental intervention. Outcomes of interest (e.g., occurrence of death or heart attack, or symptom severity as measured using a self-report scale or a physiological measurement) are measured throughout the duration of the trial and, where applicable, are compared with baseline measurements taken at the start of the trial. The outcomes are then compared in the treatment versus the control group. Random assignment of participants to one or the other of these groups is supposed to accomplish two major goals of the study: first, it facilitates allocation concealment, or "blinding," which ensures that study personnel and the participants themselves are unaware who has been assigned to the treatment or the control groups. This is important because knowledge of group allocation can bias assessments: if a study participant knows she is receiving the experimental therapy and believes that the therapy will be effective, this will (perhaps unconsciously) influence how she reports her experiences while on the medication, and possibly also, because of the placebo effect, how she responds on some "objective" measures. Similarly, a study clinician who knows that a patient is receiving active medication, or that she is receiving placebo, may be biased when assessing study outcomes for that patient.

The second thing that random allocation is supposed to achieve is to ensure that the treatment and the control groups in the study are similar with regard to the demographic and clinical characteristics of the study participants. This is important because it balances the potential effects of any factors (other than the experimental intervention) that can influence the effects of the treatment being tested. If, for example, the study drug is less effective in older patients than in younger ones, and one study group includes more older participants than the other, it will not be clear whether differences in the outcome being tested should be attributed to the intervention itself, or to physiological differences associated with age. The idea is that the effects of such confounding factors are "cancelled out" when the factors are balanced across the study groups, and so can be ignored when interpreting the study's results.

Because of these characteristics, RCTs face the kinds of issues raised in philosophical discussions of inductive risk. RCTs are, of course, supposed to provide evidence regarding whether the drug should be used in clinical practice. Yet scientists and clinicians may be mistaken in accepting the results of a clinical trial, whether because the trial shows that a treatment is effective and safe when it actually is not (a false positive result) or because it fails to demonstrate that a treatment is effective and safe when it really is (a false negative result). In the philosophical literature, the focus of inductive risk has traditionally been (following Rudner 1953) on the choice of a threshold for statistical significance. Applied to RCTs, this means that the more stringent the criteria for statistical significance, the harder it is for a study to demonstrate that a drug is effective, and therefore the more likely to produce a false negative result. This means, however, that the treatment will not be used in clinical practice and patients will be deprived of the opportunity to benefit from an effective therapy. By contrast, setting the standard for statistical significance too low increases the risks of false positive errors, in which an ineffective treatment is wrongly concluded to be effective. This means that, when the drug is adopted in clinical practice, patients will be exposed to the risks of side effects of the drug without a reasonable expectation of benefit. Moreover, they will not have access to an alternative, beneficial therapy that they might otherwise have taken instead of the experimental drug.

Yet the traditional discussion of inductive risk, with its focus on statistical thresholds, only addresses a single point at which inductive risk is relevant to scientific research. In a paper that reignited philosophical interest in inductive risk, Heather Douglas (2000) demonstrated that, in addition to the methodological question of the appropriate threshold for statistical significance, inductive risk occurs at other points in the scientific process. One

major goal of this chapter is to build on Douglas's insights by showing that a number of the other decisions that must be made in the course of designing a clinical trial are also relevant to inductive risk. In order to do this, the following sections introduce three debates about the appropriate methods for study design, and show that they all involve consideration of inductive risk.

The Hierarchy of Evidence and the Role of Nonrandomized Trials

I noted that the randomized controlled trial is considered to be the gold standard of evidence in clinical research. This fact is central to the approach to clinical research developed by proponents of evidence-based medicine (EBM). EBM was developed in the early 1990s by a group of physicians who aimed to ensure that clinicians had the skills necessary to find and to critically assess the quality of clinical research studies relevant to the care of their patients. Although a number of characteristics or features are relevant to study quality, the primary indicator of a high-quality study was held to be whether it used random allocation to assign study participants to the experimental or control groups.

The central importance of randomization to EBM is reflected in the hierarchy of evidence, which ranks study designs on the basis of how likely they are to provide high-quality evidence. The hierarchy originally proposed by members of the Evidence-Based Medicine Working Group is as follows:

- N of 1 randomized controlled trial[1]
- Systematic reviews of randomized trials
- Single randomized trial
- Systematic review of observational studies
- Single observational study
- Physiologic studies
- Unsystematic clinical observations (Guyatt and Rennie 2002, 7)

Although there have been different versions of the hierarchy proposed by different groups, they all have in common the placement of controlled trials above physiological research and clinical experience, and, key for this chapter,

1. This is a kind of randomized trial in which the effectiveness of a drug is tested for a single patient, by having that patient alternate between taking an experimental and a control therapy. The results of the trial inform the care of that patient, but are not intended to be generalized to other patients.

the placement of randomized studies above nonrandomized, "observational" studies. Randomization is so important that (on this hierarchy, at least) a single randomized study "trumps" any number of nonrandomized ones (since a systematic review or meta-analysis of observational studies falls below a single RCT on the hierarchy). Later refinements of the hierarchy of evidence build in other aspects of study design; most notably the GRADE system is flexible enough that well-designed nonrandomized trials can be rated higher than less well-designed randomized trials (GRADE working group). But even on this revised system, all else being equal, randomized trials outrank nonrandomized studies. The reasons for this are precisely the ones I outlined earlier: randomization is held to be the best way to balance potential confounders across the treatment and the control groups, and also to aid in concealment of which intervention (treatment or control) study participants are receiving.

Yet critics of the hierarchy of evidence have noted that randomization is not the only or even necessarily the best way to achieve these goals. For example, John Worrall has pointed out that randomization does not guarantee that potential confounders are balanced across study groups. This is why published trial results actually report the clinical and demographic characteristics of study groups—and conduct statistical tests to determine whether these characteristics are (roughly) the same in each group. In fact, Worrall (2002) argues, deliberately balancing potential confounders is a more effective means of achieving this goal.

The point of criticizing the hierarchy is not to say that randomization is not useful; rather it is to challenge the idea that it is the most important feature in determining the quality of a study. Critics worry that focusing so closely on whether or not a study is randomized causes all other kinds of study to be ignored, even in cases, such as in qualitative research, where randomization is not applicable (Grossman and Mackenzie 2005). Nor is this fear entirely unfounded. One EBM textbook advises clinicians who are examining the literature on a topic as follows: "If the study wasn't randomized, we'd suggest that you stop reading it and go on to the next article in your search. . . . Only if you can't find any randomized trials should you go back to it" (Straus et al. 2005, 118).

How is this debate about randomization relevant to questions of inductive risk? Recall that what is at issue with inductive risk is the worry that a hypothesis will be falsely accepted or rejected. The hierarchy of evidence is essentially a statement that randomized trials are much less prone to inductive risk than nonrandomized studies, that RCTs are the study design that is most likely to deliver the truth about whether a study is effective. This is because,

the argument goes, randomized trials are less likely to lead to biased results than are nonrandomized studies, where "bias" here is understood in the statistical sense, as any systematic deviation from the truth. A common theme in the literature explaining EBM and the hierarchy of evidence is to point to examples of therapies that had been believed, on the basis of nonrandomized studies, to be safe and effective, but that were eventually shown conclusively, via an RCT, to be unsafe or ineffective (see, e.g., Guyatt and Rennie 2002, esp. ch. 2B1). This line of argument emphasizes false positive results from nonrandomized trials, but Regina Kunz and Andrew Oxman (1998) have claimed that nonrandomized trials are also more prone than randomized trials to false negative results. They compared a number of randomized trials with nonrandomized trials of the same intervention and found that, compared to the randomized studies, nonrandomized trials might either significantly overestimate outcomes (i.e., give false positive results) or significantly underestimate outcomes (i.e., give false negative results), a phenomenon they dubbed the "unpredictability paradox."

Another possible interpretation of their results, however, is that random allocation does not necessarily have the benefits its proponents claim for it. Moreover, Kunz and Oxman appear to be begging the question in favor of randomized trials by using them as a benchmark to which nonrandomized studies must conform (Bluhm 2009). Finally, those who argue that nonrandomized studies have an important role to play in assessing therapies point to evidence that suggests that, other aspects of study design being equal, nonrandomized and randomized studies give similar results (e.g., Benson and Hartz 2000).

In summary, the debate over the necessity of random allocation, unlike the issues of setting statistical thresholds, does not involve a straightforward trade-off between the risks of false positive and false negative results. Rather, proponents of randomization claim, and critics of the evidence hierarchy deny, that random allocation minimizes both dangers.

Explanatory versus Pragmatic Trials

This section describes a second debate regarding the appropriate methods for clinical trials, which focuses on the influence of other methodological decisions made in designing a study. In describing the arguments given for random allocation of study participants to the arms of a study, I emphasized the importance of balancing potential confounding factors in the treatment and the control groups. These factors include demographic characteristics, such as age

and sex, as well as clinical characteristics, such as the severity of illness and the presence of additional health problems, other than the one being studied (i.e., of comorbid conditions). Again, random assignment of participants tends to result in these characteristics being roughly equally distributed in the treatment and the control groups.

But if these characteristics really do have an important effect on the disease or on the outcomes being investigated, they will have this effect within, as well as between, the treatment and the control groups. That is, if older patients are less likely to respond to a study medication, this is true even in cases where there are roughly equal numbers in the treatment and the control groups. This raises the question of whether the results obtained in a clinical trial can accurately predict the results that will be observed in the clinic. If there is a significantly higher proportion of older people in the study than will be in the population treated in clinical practice if the drug is shown to be effective, then the results obtained in the experimental group will be less dramatic than in the population as a whole. If (as is more likely) there are proportionally fewer older adults in the study than in the general population that will be treated with the study drug, the drug will, on average, be less effective in practice than it seemed to be in the original RCT.

What this example shows is that in addition to considering whether the treatment and the control groups in a study are clinically and demographically similar to each other, it is also important to consider whether the study groups are similar to the population of patients who will be treated on the basis of the results of the trial. This is the question of the external validity of the trial. If a trial has low external validity, then the study participants do not resemble the clinical population, so it is not clear that the results of the trial are applicable to this larger group. Generally, trials with low external validity tend to exclude patients with comorbid conditions, those taking additional medications, and older patients. By contrast, trials with high external validity are ones in which the participants are similar to the patients who will be treated in clinical practice. Another way of describing trials with high external validity is to say that they tend to be "pragmatic" in their design; in general, pragmatic trials "seek to answer the question 'Does this intervention work under usual conditions?' " (Thorpe et al. 2009, 465). They are therefore designed to be similar to the clinical settings in which the intervention will be used.

In addition to the similarity of the study participants to the larger population of patients who will be treated using the new intervention, there are a number of other ways that a study might be pragmatic in its design. Kevin

Thorpe et al. (2009) have identified ten features of clinical trial design that characterize pragmatic trials. One such characteristic has to do with the flexibility of the intervention being tested; for example, whether the dose of a medication can be modified based on patients' responses to the original regimen. Another feature of pragmatic trials may involve the characteristics of the control intervention; instead of a single, specific control intervention, investigators have considerable leeway in deciding what intervention(s) participants in the control group will receive, depending on the range of "usual practice" at the study site. A third characteristic is the lack of formal follow-up (i.e., the use of predetermined outcome measures at predetermined times); pragmatic trials may instead follow patients by examining their electronic health records. As Thorpe et al. are careful to point out, trials may have only some pragmatic characteristics and may have them to different degrees.

But the similarity to clinical practice that characterizes pragmatic trials comes at a cost. Because there is so much variability within the treatment and the control groups, it can be difficult to ascertain that the outcome differences between the groups are really caused by the drug being studied. This problem is analogous to the one discussed earlier with regard to the necessity of similarity between the treatment and the control groups in an RCT. To put the point somewhat differently, differences within the study groups with regard to the characteristics of the participants, of the interventions, or of the timing of outcome measurements may confound the assessment of the effects of the drug. Thus, while pragmatic trials do a good job of showing outcomes in clinical practice, they are not as good at isolating the treatment of interest as a significant cause of those outcomes.

Instead, isolating the causal efficacy of a potential therapy is best done in a trial that has an explanatory design. In these trials, variability is minimized as much as possible. This means that the outcomes to be measured must be specified precisely and measured at specific intervals, that the treatment regimen cannot be adjusted for individual study participants, and that the population eligible to participate in the study is fairly homogeneous (clinically and demographically) and does not have any comorbid conditions or take medications other than the study drug.

If pragmatic trials aim to determine whether an intervention will work in clinical practice, explanatory trials have the aim of showing that it actually causes the outcome(s) of interest. In one sense, these two study types (understood as representing the extreme ends of a spectrum of methods) are asking different questions or testing different hypotheses: one about what would be observed clinically and one about the drug's biological effect (Schwartz

and Lellouch 1967). Because of this, explanatory trials are often described as establishing efficacy, rather than effectiveness. Yet ultimately, both of them are concerned with the same question—whether treating patients using the new therapy being tested will improve their health. Moreover, as Kirstin Borgerson (2013) has discussed, a large majority of trials being conducted are explanatory in design, so that regardless of the way their purpose is described in the clinical literature, much of the evidence available to inform practice has come from explanatory trials.

As with the debate about randomization, arguments about the relative importance of explanatory and pragmatic trials can be understood in terms of inductive risk.[2] Proponents of explanatory trials argue that because pragmatic trials cannot give a precise, or "clean," estimate of the causal efficacy of a treatment, they cannot give us sufficient confidence in the claim that the treatment really has the desired effects. Implicit in this claim about precision is the view that pragmatic trials are more prone to both false positive and false negative results. By contrast, those who favor pragmatic trials point out that the variability that explanatory trials minimize is very important in a clinical context—in fact, it is minimized precisely because it is caused by factors that affect the drug's ability to bring about desired outcomes. Therefore, showing that a drug works under the idealized conditions of an explanatory trial does not justify concluding that it will work in clinical practice; only a pragmatic trial, designed to resemble clinical practice, can do so. Explanatory trials tend to enroll a relatively homogeneous group of participants who are not too old, not too sick, and not taking other medications. While, strictly speaking, they do not tend to give false positive results—if the target population can be assumed to have similar characteristics to the study participants—because this assumption is unlikely to be justified, taking the results of an explanatory trial to be generalizable beyond the study will tend to overestimate the effectiveness of a treatment in clinical practice.

Placebo Controls

There is also a long-standing debate in the clinical and bioethics literature regarding what kind of intervention is most appropriate to give the control

2. These debates are also related in that, while it is possible to do a randomized pragmatic trial, pragmatic trials that track long-term outcomes in clinical practice are unlikely to be randomized, while explanatory trials will almost certainly use random allocation.

group in a study, a placebo or another treatment for the condition being studied (i.e., an active control). To a greater extent than the other two debates, the ethical implications of the choice of control have been emphasized; however, both sides of the debate also claim that their position is supported by epistemological, as well as ethical, arguments.

Recall from the beginning of this chapter why a control group is necessary in a clinical trial: first, studying only one group that receives the experimental intervention does not allow investigators to determine whether changes in the outcomes measured (whether improvements or declines in health) are due to the intervention being tested or simply to changes in the natural history of the condition being studied. Second, it is well-known that our beliefs about an intervention can influence how effective it is: this is the basis of the placebo effect—if we believe that an intervention is likely to help, or to harm, us the probability that it will actually do so is increased. Because of this second point, clinical trials do not tend to use a "no treatment" control group. Instead, they control for the effects of patients' expectations by using a placebo or another control therapy.[3]

As should be clear, these are arguments for including a control group in a clinical study, but not arguments for using a specific kind of control. Critics of placebo-controlled trials have argued that assigning half of the participants in a study to a placebo group is (almost always) unethical, because they are thereby being deprived of not just the possible benefit from the experimental therapy but also the benefit from any standard therapy that they could have received if they had not chosen to participate in the trial. The major argument against using placebo controls was first presented by Benjamin Freedman (1987) and has been further developed by Charles Weijer (1999). Freedman introduced the concept of clinical equipoise as a way of determining whether the control arm chosen in a trial is ethical. Clinical equipoise exists when the relevant community of expert clinicians is not in agreement about a preferred course of therapy: applied to clinical trials, the principle can best be understood as requiring that both the experimental and the control interventions in a trial are ones that, in the judgment of this community, might be as effective

3. Note, too, that the use of a placebo that resembles the experimental therapy also helps with allocation concealment: if all of the study participants receive a daily yellow tablet, though only the tablets given to one group contain an active ingredient, then neither the participants themselves nor the clinicians who assess them can readily determine who is taking the active medication. In fact, because of this, studies that use an active control that does not resemble the experimental therapy may use a "double dummy" design: one group gets the experimental drug and a placebo that looks like the control drug, while the other gets a placebo resembling the experimental drug and the active control drug.

as other available interventions. Thus, an experimental intervention would not be tested in a trial unless it showed promise as a therapy comparable to already available treatments. And, key to the issue discussed here, a trial using a placebo control could only meet the requirements of clinical equipoise if the community did not believe that there were already existing therapies more effective than a placebo.

So far, I have emphasized the ethical rationale underlying the principle of clinical equipoise, but both Freedman and Weijer emphasize that it is also an epistemological requirement. This is because a trial should provide knowledge that is useful to those clinicians who would be using the results of a trial. What clinicians—and for that matter, patients—want to know about a promising new medication is not whether it is better than a placebo, but whether it is a better therapy (or at least as good a therapy) as the one(s) already available and used in clinical practice. The only way to answer this question is to actually test the new drug against a current therapy.

Although the concept of clinical equipoise has been very influential, there are still bioethicists who support the use of placebo controls. For example, Franklin Miller and Howard Brody (2003) describe an RCT that compared the antidepressant sertraline to both St. John's Wort and a placebo. They point out that the trial does not meet the requirements of clinical equipoise, not only because sertraline had been shown to be more effective than a placebo in previous trials but also because no psychiatrist would actually use St. John's Wort to treat patients with severe depression. Yet, they argue that the trial is ethical, in part because patients with severe depression sometimes want to take St. John's Wort instead of taking an antidepressant. The trial was intended to show definitively that the "natural" remedy was not as effective as sertraline.

In fact, in the trial, neither sertraline nor St. John's Wort was found to be more effective than the placebo. Miller and Brody argue that these results demonstrate why a placebo control is always needed: following Robert Temple and Susan Ellenberg (2000), they say that when a clinical trial does not show a statistically significant difference between two active drugs (whether this is due to a failure to demonstrate significance, as in the sertraline trial, or in a trial that is designed to show the equivalence of two active treatments), a third, placebo arm is needed to allow researchers to interpret the results. Without a placebo, the results "could mean that the treatments were both effective in the study, but it could also mean that both treatments were ineffective in the study" (Temple and Ellenberg, 456). The phrase "in the study" is key here: Temple and Ellenberg point out that it is

quite common for clinical trials to fail to identify an effective drug as effective. (They speculate that this failure could be due to, basically, quirks of the study sample or design.) A placebo control tests the ability of the trial to detect an effective drug, a property that Temple and Ellenberg call "assay sensitivity." In effect, the placebo functions as a sort of internal control that assesses the effectiveness, not of the intervention, but of the trial as a test of the intervention.

What this means, though, is that the choice of a control arm has implications for the amount and kind of evidence required before the results of a study should be accepted, which means that Temple and Ellenberg are concerned with inductive risk. They claim that, in a study that uses only an active control, when there is no statistically significant difference between the two treatment arms, it is necessary to look at evidence from outside of the trial, primarily evidence from other clinical studies, to determine whether both drugs were effective or ineffective (in the context of the trial). Like those who argue for the use of explanatory controls, Temple and Ellenberg are concerned with precision.

By contrast, the proponents of clinical equipoise argue that placebo controlled trials are (usually) neither ethical nor necessary. With regard to the latter, they argue that active control equivalence studies can establish whether a new treatment is as effective as an older therapy (Weijer 1999) and that placebo-controlled trials, just as much as active controlled trials, must be interpreted using information drawn from outside of the study (Anderson 2006). Moreover, the information gained from using active controls is directly applicable to clinical practice, in that it addresses the questions that physicians and patients really want to know by providing information about the relative merits of potential therapies—information that placebo-controlled trials cannot provide. In addition, it is easier to show a statistically significant difference between an experimental drug and a placebo than it is to show that a new drug is as good as, or better than, an already-available therapy because an active-controlled trial needs to detect a smaller difference, compared with a placebo-controlled trial, between the experimental and the control interventions. It is therefore possible that a drug tested against a placebo may be adopted in clinical practice, but be less effective than older therapies. Its use would offer patients less benefit than they would have received before the new drug was adopted. Like those who argue for the use of pragmatic trials, Freedman, Weijer, and Anderson are concerned with the applicability of trial results to clinical practice.

Clinical Research Methods and Inductive Risk

In order to understand why these debates should be of interest to philosophers writing about inductive risk, it is first important to recognize that in none of the debates is one side accusing the other of simply doing "bad science." Even the staunchest proponents of randomization accept that there is an important role for nonrandomized studies in clinical research, especially when it comes to detecting harmful side effects that are either rare or associated with long-term use of the therapy; conversely, nobody denies that randomization can be a useful methodological tool. Similarly, there is general agreement that both explanatory and pragmatic trials have a place in clinical research; the disagreement is about which kind of trial provides the most important kind of information, or which kind of trial should be performed more often (see, e.g., Borgerson 2013). Finally, those who argue that trials must meet the requirement of clinical equipoise acknowledge that there are cases in which placebo-controlled trials do so, while those who advocate for placebo controls acknowledge that they are not necessary in trials that show an experimental treatment is (statistically significantly) superior to an active control. In all cases, the disagreement is about which methods are best able to establish the effectiveness and safety of a treatment; that is, about which kinds of trial design best ground epistemological claims about the treatment.

But it is also important to note that the choice of method is made with both ethical and epistemological goals in mind. All of the debates are concerned with the consequences of using the results of clinical research to inform patient care. An error in accepting the results of a trial will mean that patients are exposed to a treatment that is ineffective, unsafe, or both. Erroneously rejecting the results of a trial will prevent patients from accessing a safe and effective treatment. In this, the three debates I discuss here echo the traditional example of inductive risk (i.e., the issue of setting a level for statistical significance).

At the same time, however, examining issues of inductive risk in clinical research expands the philosophical discussion of inductive risk and the related question of the appropriate role for (non-epistemic) values in science. In making this case, I am building on the work of Heather Douglas. One of Douglas's major contributions has been to revive interest in inductive risk by showing how thoroughly it permeates the scientific process. In her 2000 paper, she draws on research in toxicology to show that "significant inductive

risk is present at each of the three 'internal' stages of science: choice of methodology, gathering and characterization of the data, and interpretation of the data" (2000, 256).

Douglas uses the standard case of setting a threshold for statistical significance to show that methodological choices carry significant inductive risk, but her discussion of methodology is also closely tied to that of a second "internal" part of science where considerations of inductive risk may legitimately influence scientists' choices. This is the choice of a model for interpreting the data obtained in a study. In a threshold model of the relationship between exposure to a potentially carcinogenic substance and the occurrence of cancer, it is assumed that there is no biological effect of a substance below a threshold of exposure. By contrast, a linear extrapolation model is based on the idea that the substance will instead produce lower rates of an effect at lower doses. Because these models will (even when the same threshold for statistical significance is used) lead to different claims about the dose-response relationship, the choice of model also has implications for inductive risk and for the regulatory policies that would be based on the study. Therefore, scientists must weigh the relative consequences of false positive and false negative results in interpreting their data according to one of the models.

This brings us to one way in which my analysis of clinical research expands the philosophical discussion of inductive risk. Both the choice of a significance level and the choice of an interpretive model involve a trade-off between a higher risk of false positive results and a higher risk of false negative results. By contrast, the three debates I have reviewed here have a more complicated relationship with inductive risk. In the case of randomization, the proponents of randomization say that nonrandomized trials are more prone to both false positive and false negative results, while those who do not view randomization as essential to good clinical research deny this claim. In the other two debates, supporters of explanatory trials and of the use of placebo controls do not tend to explicitly couch their arguments in terms of inductive risk, but they do argue that their methods are more likely to give a true estimate of the effects of the intervention. This is because of the potential for confounding factors to influence the results of a trial (in nonrandomized trials and in pragmatic studies), or because of the lack of an internal baseline measure (in trials that do not include a placebo arm). Because estimating the "true" effects is a matter of the precision of the results (as reflected in a low *p* value, or a narrow confidence interval) this view implicitly also claims that explanatory trials and placebo controlled trials are best at avoiding both false positive and false negative results.

On the other side of the debates, critics of the hierarchy of evidence, proponents of pragmatic trials, and opponents of placebo controls all take the view that the best way to ensure that the results of clinical research can be extrapolated to clinical practice is to ensure that the research is designed to be clinically relevant. The less a study reflects clinical practice, the greater the risk of erroneously accepting the hypothesis that a treatment will provide therapeutic benefit (a false positive error). In the case of pragmatic trials, in particular,[4] the claim might also be made that such trials are less prone than explanatory trials to false negative errors in cases where, for example, the explanatory trial excluded a group of patients who do benefit from the treatment, or prohibited the use of a concomitant medication that would, in actual practice, be prescribed together with the drug being tested.[5] In summary, the debates I have discussed show that considerations of inductive risk can help illuminate the roots of scientific disagreement even when a straightforward trade-off between the two kinds of error is not necessarily involved.

A second way in which the case of clinical research expands the philosophical discussion is that it adds an irreducible qualitative dimension to the assessment of inductive risk. For both the choice of a threshold for statistical significance and the choice of a model, Douglas notes that increasing the sample size of the study would decrease inductive risk by lowering the uncertainty of the results (though she also recognizes that this is not always practically possible because of the cost of doing a larger study). More generally, Douglas tends to view the problems posed by inductive risk in terms of the amount of evidence available regarding a question of interest. For example, she says that in deciding whether to accept a hypothesis, "[a] scientist will need to consider both the quantity of evidence or degree of confirmation to estimate the magnitude of inductive risk and the valuation of the consequences that would result from error" (Douglas 2000, 565).

It is not clear, however, that having more evidence will settle issues of inductive risk in clinical research. This is because clinical scientists' assessments of whether to accept the results of a clinical trial involve not just how much evidence is required before accepting a claim but also what kind of evidence is required; they all involve deciding which study designs supply the

---

4. Recall, however, that pragmatic trials are much more likely than explanatory trials to be nonrandomized, so there is a connection between the two debates.

5. Although a detailed assessment of these claims is beyond the scope of this chapter, I have argued elsewhere that, for both ethical and epistemological reasons, clinical research should resemble practice (Bluhm 2009, 2010).

strongest evidence for a clinically relevant hypothesis. Nor is it clear that these debates about the quality of evidence can be reduced to debates about quantity, perhaps through some sort of weighting scheme by which each side of the debate can give "partial credit" to studies that use the methods they deem inferior. Recall from earlier in the chapter that some of the (admittedly more extreme) proponents of randomization think that it is not simply that RCTs provide better evidence than nonrandomized trials, but that if RCTs regarding a particular therapy exist, they are the only evidence that should be considered; they always trump evidence from nonrandomized studies. Somewhat more plausibly, advocates of pragmatic trials might insist that no matter how many studies have been done, evidence for the effectiveness of a therapy in a relatively young, relatively healthy population can never establish that a drug will work in a geriatric population with multiple health problems and the potential for drug interactions. These examples show that adding a qualitative dimension to the judgment of the sufficiency of the evidence means that more evidence is not guaranteed to solve the problem.

Finally, considering these methodological debates in clinical research has implications for the relationship between values and evidence, which also raises questions for Douglas's account of the roles that values can legitimately play in science. Douglas's quantitative assessment of inductive risk allows her to uphold the view, traditional in philosophy of science, that "whether or not a piece of evidence is confirmatory of a hypothesis . . . is a relationship in which value judgments have no role" (Douglas 2000, 565). On her view, the relative contributions of evidence and value judgments to decisions about inductive risk can be separated; with more evidence available, values will play less of a role in deciding whether to accept or reject a hypothesis (Douglas 2009, 96). This separation of evidential and value considerations prevents values from playing an illegitimate direct role in the assessment of evidence by preventing cases of wishful thinking, in which poor evidence is accepted in support of a hypothesis that supports one's ethical or political commitments.

The debates I have discussed here show that this sharp separation between evidence and values does not work.[6] This is because ethical (as well as epistemological) values influence the methods chosen by clinical researchers. In turn, methodological choices shape the data collected and thus the

6. At least, it does not work in clinical research, though I doubt that this area of science is unique.

evidence available to confirm (or to fail to confirm) the hypotheses being considered. To see this point more clearly, let us return to the example of setting an appropriate statistical significance threshold. While in practice, this decision is often made based on discipline-specific conventions, it can also be a purely value-based decision about whether it is more important to avoid false positive or false negative conclusions. But regardless of how high or how low the threshold is set, the data themselves are unaffected. All that changes is whether we accept them as significant. By contrast, the data obtained in a clinical trial clearly depend on such factors as which patients are eligible for the trial, and whether the control group is given a placebo or an active drug; both sides of the debates acknowledge this, though they draw different conclusions about what these methods imply for the quality of the result. The case of randomization is slightly different, as what is up for debate is whether use of this methodological feature affects the data obtained (by minimizing confounding), but this is still very different than the question of whether to consider data statistically significant. My discussion of clinical research shows that the relationship between evidence and hypothesis is influenced by values because the data themselves depend on methodological decisions that are defended on both ethical and epistemological grounds.[7]

How best to characterize this influence of values is unclear. Douglas has distinguished between direct and indirect roles for values and has sketched out legitimate instances of each. Although this distinction can be understood in several distinct ways (Elliott 2011, 2013), one interpretation of this distinction that is central to Douglas's arguments about where in the scientific process values play a legitimate role is her claim that values ought not to play a direct role by "act[ing] as reasons in themselves to accept a claim" (Douglas 2009, 96). As Elliott explains, Douglas "insisted that values should not play a direct role when scientists are evaluating what empirical claims to accept, because it would amount to something like wishful thinking—scientists would be treating their ethical, political, or religious values as if they were evidence in support of their claims" (Elliott 2013, 376). Ethical values can, however, legitimately play a direct role in the selection of methods, specifically by ruling out methods that are morally

7. Although I will not discuss this point further, I believe that this is also the case for Douglas's third example of a decision involving inductive risk, which examines the standards used for characterizing tissue samples as cancerous or non-cancerous (Douglas 2000).

unacceptable (e.g., ones that pose significant harm to human study participants). In these cases, ethics trump epistemology: "despite the cognitive value of such a test, the conflicting ethical and social values would overrule that value" (2009, 100).

Douglas does not, however, consider the role that values play in deciding among methods that are ethically permissible, though it seems that this would also be a direct role. And while this means that the data collected in the study are also shaped by these value choices, and are used as evidence for the hypothesis being tested, this is not the same thing as having the values "act as evidence" in the wishful thinking case. Overall, it is not clear that this role for values counts as a direct role, for Douglas.

Yet neither do they fit with Douglas's characterization of the indirect role that values can play. This role is the one that fits the "traditional" discussion of inductive risk, in which values "act to weigh the importance of uncertainty about the claim, helping to decide what should count as *sufficient* evidence for the claim." In this indirect role, "the values do not compete with or supplant evidence, but rather determine the importance of the inductive gaps left by the evidence" (Douglas 2009, 96). But in the examples I have discussed, while values do not compete with or supplant evidence, they do (directly!) help to determine what the evidence is. Thus, the third way that my analysis advances the discussion of inductive risk is by showing the extent to which methodological choices incorporate both epistemological and ethical questions, and the challenge this entanglement raises for understanding the role of values in science.

In summary, clinical trials provide a paradigm case of scientific research in which consideration of inductive risk is important, but they also draw our attention to new issues relevant to inductive risk and to the broader issue of value-laden science. The debates I have considered in this chapter show that ethical and methodological considerations are not separable in the design of clinical research.

References

Anderson, James A. 2006. "The Ethics and Science of Placebo-Controlled Trials: Assay Sensitivity and the Duhem-Quine Thesis." *Journal of Medicine and Philosophy* 31(1): 65–81.

Benson, Kjell, and Arthur J. Hartz. 2000. "A Comparison of Observational Studies and Randomized, Controlled Trials." *New England Journal of Medicine* 342: 1878–86.

Bluhm, Robyn. 2009. "Some Observations on 'Observational' Research." *Perspectives in Biology and Medicine* 52(2): 252–63.

Bluhm, Robyn. 2010. "The Epistemology and Ethics of Chronic Disease Research: Further Lessons from ECMO." *Theoretical Medicine and Bioethics* 31(2): 107–22.

Borgerson, Kirstin. 2013. "Are Explanatory Trials Ethical? Shifting the Burden of Justification in Clinical Trial Design." *Theoretical Medicine and Bioethics* 34(4): 293–308.

Douglas, Heather E. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67(4): 559–79.

Douglas, Heather E. 2009. *Science, Policy and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Elliott, Kevin C. 2011. "Direct and Indirect Roles for Values in Science." *Philosophy of Science* 78(2): 303–24.

Elliott, Kevin C. 2013. "Douglas on Values: From Indirect Roles to Multiple Goals." *Studies in History and Philosophy of Science Part A* 44(3): 375–83.

Freedman, Benjamin. 1987. "Equipoise and the Ethics of Clinical Research." *New England Journal of Medicine* 317:141–5.

GRADE Working Group. http://www.gradeworkinggroup.org.

Grossman, Jason, and Fiona J. MacKenzie. 2005. "The Randomized Controlled Trial: Gold Standard, or Merely Standard?" *Perspectives in Biology and Medicine* 48(4): 516–34.

Guyatt, Gordon, and Drummond Rennie, eds. 2001. *Users' Guide to the Medical Literature: Essentials of Evidence-Based Clinical Practice*. Chicago: AMA Press.

Kunz, Regina, and Andrew D. Oxman. 1998. "The Unpredictability Paradox: Review of Empirical Comparisons of Randomised and Nonrandomised Clinical Trials." *BMJ* 317:1185–90.

Miller, Franklin G., and Howard Brody. 2003. "A Critique of Clinical Equipoise: Therapeutic Misconception in the Ethics of Clinical Trials." *Hastings Center Report* 33(3): 19–28.

Rudner, Richard. 1953. "The Scientist qua Scientist Makes Value Judgments." *Philosophy of Science* 20(1): 1–6.

Schwartz, Daniel, and Joseph Lellouch. 1967. "Explanatory and Pragmatic Attitudes in Therapeutic Trials." *Journal of Chronic Diseases* 20(8): 637–48.

Straus, Sharon E., W. Scott Richardson, Paul Glasziou, and R. Brian Haynes. 2005. *Evidence-Based Medicine: How to Practice and Teach It*. Toronto: Elsevier.

Temple, Robert, and Susan S. Ellenberg. 2000. "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues." *Annals of Internal Medicine* 133(6): 455–63.

Thorpe, Kevin E., Merrick Zwarenstein, Andrew D. Oxman, Shaun Treweek, Curt D. Furberg, Douglas G. Altman, Sean Tunis, et al. 2009. "A Pragmatic-Explanatory Continuum Indicator Summary (PRECIS): A Tool to Help Designers." *Journal of Clinical Epidemiology* 62(5): 464–75.

Weijer, Charles. 1999. "Placebo-Controlled Trials in Schizophrenia: Are They Ethical? Are They Necessary?" *Schizophrenia Research* 35(3): 211–18.

Worrall, John. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69(S3): S316–30.