Part III
RESEARCH METHODS
(a) Evidence in Medicine
Introduction

When discussing the validity of clinical trials, we need to consider both internal validity and external validity. Internal versus external validity is a distinction made in clinical epidemiology to discuss how well a clinical study answers certain questions. In broad terms, a study has high internal validity if the results can be considered accurate for the sample included in the study. It is an assessment of whether the study’s design and conduct is sufficient to reliably answer the research question posed in the study. In clinical drug development, a study with high internal validity is considered to provide a reliable answer to the question: what is the effect of giving drug A to the sample studied in the trial? External validity, by contrast, refers to how well the outcomes observed in the study apply to people outside the sample included in the trial. When most people discuss internal and external validity, they do so in terms of randomized controlled trials, also known as RCTs or simply randomized trials. A common claim is that well-conducted randomized trials have better internal validity than other types of medical studies such as observational studies.

The first objective of this chapter is to examine arguments for the claim that randomized studies are methodologically superior to alternative study designs used in medicine. The section “Internal validity and hormone replacement therapy” introduces internal validity and bias in relation to the challenge of assessing the cardiovascular effects of hormone replacement therapy. More formal arguments for the benefits of random allocation in terms of internal validity are considered in the section “Arguments for randomization.” Philosophical criticism of the tendency to claim too much on behalf of randomization is then considered in “Criticisms of randomized trials.” The second objective of this chapter is to explore “The challenge of external validity.” This section outlines some of the limits of randomized trials and briefly introduces a philosophical approach to improve the assessment of whether the results of a randomized trial will apply to patients in the clinic.
Internal Validity and Hormone Replacement Therapy

Internal validity is defined in similar terms in most clinical epidemiology texts. Fletcher et al. (1996, 12) provide the following:

"Internal validity is the degree to which the results of a study are correct for the sample of patients being studied. It is “internal” because it applies to the conditions of the particular group of patients being observed and not necessarily to others. The internal validity of clinical research is determined by how well the design, data collection, and analyses are carried out and is threatened by all the biases and random variation discussed above. For a clinical observation to be useful, internal validity is a necessary but not sufficient condition."

Clinical research—medical studies focusing on patients and the effects of treatments—is conducted using studies of a variety of designs. For each of these study designs, methods are employed to reduce error and improve internal validity. Randomized trials play a prominent role in clinical research because it is argued that they have the capacity to reduce or eliminate more sources of error than the alternative study designs employed in clinical research.

The goal of much clinical research is to estimate the effect of an experimental intervention as opposed to a control. An intervention is considered to “work” if the overall benefits of the intervention are greater than some prespecified clinically important amount (e.g., a new drug for preventing heart attacks might be considered worthwhile if it reduces the absolute risk of heart attack by an additional 5% in an at-risk population who are treated for 5 years). Errors that occur when discerning the effect of an intervention can be subdivided into random errors and systematic errors. Random errors occur due to chance. To take a simple example, consider 10 tosses of a fair coin. The coin is fair, so the probability of “heads” is 0.5, but in any series of 10 tosses we expect some variation in the number of “heads” observed—this variation is random error. Random error is managed by ensuring the study is large enough to estimate the outcome with sufficient precision.

Systematic errors occur when bias leads to incorrect measures of effect. In epidemiology, the term “bias” refers to any process that systematically promotes erroneous inference. Many sources of bias can affect clinical studies. Rothman et al. (2008) classify biases into three general categories: confounding, selection bias, and information bias. These concepts are illustrated using a real-life example. Estimating the effect of hormone replacement therapy on cardiovascular disease in postmenopausal women provides a classic case study of the effects of different sources of bias. The influence of hormone replacement therapy on cardiovascular disease has been assessed over the past 30–40 years in studies employing a range of different methods. Prior to menopause, women tend to suffer less than men from cardiovascular disease. After menopause the rates of cardiovascular disease in women approaches the rates seen in men. On the basis of this observation, it was thought that long-term use of hormone replacement therapy might protect women from cardiovascular disease. This question was first assessed in the 1970s and 1980s with methods that are collectively labeled “observational studies”—so called because these studies observe rather than intervene upon participants. Though there were conflicting results, the majority of findings from the observational studies supported the idea that postmenopausal women who take hormone replacement therapy suffer less cardiovascular disease. In a notable review of the observational studies, Stampfer and Colditz (1991, 61) concluded:

"The preponderance of evidence from epidemiologic studies strongly supports the view that postmenopausal estrogen therapy can substantially reduce the risk of coronary
Prior to 2002, many women took hormone replacement wholly or partly for the purpose of reducing cardiovascular risk.

Clinical practice regarding the use of long-term hormone replacement therapy to reduce cardiovascular disease changed with the publication of the Women's Health Initiative Study in 2002 (Women's Health Initiative Investigators 2002). This large randomized trial reversed the findings of the earlier observational studies and provides compelling evidence that hormone replacement therapy does not reduce cardiovascular disease in older postmenopausal women—indeed, the Women's Health Initiative Study was stopped early due to excess risks in women taking hormone replacement therapy, including excess risks of cardiovascular disease.

What explains the disparity between the observational studies and the Women's Health Initiative Study? How does a single large randomized trial undermine the evidence provided in several large observational studies?

The short answer is that the Women's Health Initiative Study employed methods that rule out or minimize more sources of error than the methods used by the observational studies. Since the key observational studies and the Women's Health Initiative Study were all well conducted, analyzed and reported, the principal difference comes down to the capacity of the methods to rule out specific types of error. The key observational studies conducted to assess the association between hormone replacement therapy and cardiovascular disease were prospective cohort studies. These studies recruited a sample of postmenopausal women, some of whom were taking hormone replacement therapy and some of whom were not. The participants were observed over time and could start or stop hormone replacement therapy at their discretion. At the end of the study, the investigators compared the rates of cardiovascular disease in the women who were exposed to hormone replacement therapy with those who were not. If women who were exposed to hormone replacement therapy experience less cardiovascular disease, then it may be that hormone replacement therapy reduces the risk of cardiovascular disease.

The inferences licensed by observational studies such as these are valid only to the extent that a rather long list of additional assumptions are met. Sources of bias are legion. If socioeconomic status is positively correlated with exposure to hormone replacement therapy and negatively correlated with cardiovascular disease, then unadjusted estimates provided by a cohort study will exaggerate the putative benefits of hormone replacement therapy. If the impact of this confounder is not taken into account in an observational study, then participants exposed to hormone replacement therapy will have higher socioeconomic status and less cardiovascular disease compared to participants not exposed to hormone replacement therapy—and this will be the case without there being a causal relation between hormone replacement therapy use and cardiovascular disease. In this example, socioeconomic status is a confounder: (i) it is associated with exposure to the factor under investigation; (ii) it is correlated with the outcome; and (iii) the confounder is not an intermediate cause between exposure and outcome (Rothman et al. 2008).

Another type of bias can occur if factors that influence participation in the study also influence outcomes. For instance, the research question of interest is whether long-term use of hormone replacement therapy in postmenopausal women will reduce cardiovascular disease. The challenge here is that hormone replacement therapy is used for a range of indications. Perhaps the most common of these is short-term use in women experiencing symptoms around the time of menopause. If younger women are more likely to participate and their use of hormone replacement therapy poses a different risk of cardiovascular disease, then any estimate of the

heart disease. . . . This effect is unlikely to be explained by confounding factors or selection.
influence of treatment on cardiovascular disease will differ systematically from the estimate in older women using longer-term hormone replacement therapy. Bias due to the selection of individuals for participation is called selection bias. An observational study assessing the cardiovascular effects of hormone replacement therapy will be subject to selection bias to the extent that (i) women using hormone replacement therapy for perimenopausal symptoms are overrepresented in the study and (ii) any difference in effect on cardiovascular events in short-term use among perimenopausal women goes unrecognized. Selection bias seems to have played a part in the observational studies assessing the link between hormone replacement therapy and cardiovascular disease.

The third common type of bias is information bias. Information bias occurs from measurement error or misclassification. Studies assessing the effects of hormone replacement therapy need to accurately classify a participant’s exposure to the treatment. There are a number of ways this might go awry. Perhaps exposure to hormone replacement therapy is assessed in a database that collects health insurance claims for medicines. If not all types of hormone replacement therapy are covered by the policy, then exposure will be underestimated. Similarly, it may be difficult in the claims database to differentiate women who are taking estrogen for a gynecological indication as opposed to hormone replacement therapy. Any misclassification that occurs may influence the estimate of the effect of hormone replacement therapy on cardiovascular disease. Information bias does not seem to have played an important role in the assessment of the link between hormone replacement therapy and cardiovascular disease.

Well-conducted observational studies are designed to mitigate the influence of known confounders—this is achieved in part through careful selection and adequate control. Further, well-conducted observational studies will employ data-based methods to systematically identify and address additional potential confounders. The overall consistency of findings from observational studies and randomized trials regarding the effects of hormone replacement therapy on a wide range of outcomes other than cardiovascular disease is testament to the success of these methods (Lawlor et al. 2004a). Unfortunately, a risk of residual confounding through unsuspected confounding factors or through under- or over-control of known or suspected confounders remains.

The discrepant results observed regarding the cardiovascular risks of hormone replacement therapy from observational studies and the Women’s Health Initiative Study have been attributed to residual confounding. Lawlor et al. (2004a) make a case that the specific confounder that was not adequately controlled for in the observational studies was life-course socioeconomic status. Current socioeconomic status was controlled for in the observational studies, but socioeconomic status over the participant’s life was not. Women with higher life-course socioeconomic status are both more likely to take hormone replacement therapy and have less physiological risk factors for heart disease—and these relationships are independent of current socioeconomic status (Lawlor et al. 2004b). This hitherto unrecognized confounder plausibly accounts for the apparent benefits of hormone replacement therapy on heart disease in the observational studies. Significantly, the discrepancy between the observational studies and the Women’s Health Initiative Study prompted further examination of the association between life-course socioeconomic status, exposure to hormone replacement therapy, and risk factors for heart disease; without the surprising results of the Women’s Health Initiative Study, these investigators would not have searched for a possible residual confounding factor.

Observational studies play an important role in epidemiology and clinical medicine, and sophisticated methods can be employed for mitigating the influence of many sources of bias. But, like any methods, the methods employed in observational studies have their limits. One of these limits is that if a confounder is unsuspected and goes unidentified in the variety of checks performed on the data to identify potential confounders, then there is a risk that confounder
and its influence on the observed data goes undetected. Random allocation reduces the likelihood of this kind of error. The case of hormone replacement therapy and cardiovascular risk emphasizes the benefits of randomized trials in terms of internal validity. This, and two additional arguments for random allocation, are considered below.

**Arguments for Randomization**

R. A. Fisher (1966) was influential in promoting randomized allocation in experimental design and for developing frequentist statistical approaches for analyzing randomized studies (see especially, *The Design of Experiments*, which was first published in 1935). Fisher’s work remains instructive regarding the key benefits of random allocation in terms of internal validity. Fisher provides three interlinked arguments in favor of random allocation. Random allocation: (i) provides an objective basis for statistical analysis; (ii) allows adequate masking and avoids biases associated with non-random allocation; and (iii) reduces the influence of confounders (especially unsuspected confounders). Each of these arguments is considered below. Although these arguments will be discussed in frequentist terms, it is worth noting that many of the frequentist reasons to randomize are also good reasons to randomize on alternative approaches to statistical inference. Suppes (1982) and Kadane and Seidenfeld (1990), for instance, provide arguments for random allocation in Bayesian terms.

Fisher’s (1966, 11–26) tea-drinker experiment provides a helpful illustration of the first two arguments for randomization. Fisher also uses this case to outline and defend the significance test—the statistical test he is most famous for developing. The tea-drinker experiment sets out to assess a person’s claim to be able to discriminate whether or not milk is added to a cup of tea before the tea infusion. Fisher proposes the following experiment to test the tea-drinker’s claim. Present the tea-drinker with eight cups of tea, four of which have been prepared “milk-first” and four “tea-first.” The order in which the cups are presented to the tea-drinker is randomized, as opposed to, say, ordered “systematically” where the investigator allocates cups according to some system—e.g., alternate cups or blocks of two “milk-first” cups then two “tea-first” cups, or ordered “haphazardly” where the investigator places the cups in what appears to the investigator to be a “sufficiently random” order (in the sense that there is no apparent order to the sequence). The tea-drinker is considered to have passed the test if she or he correctly identifies all four “milk-first” cups. The tea-drinker is informed of all the details of the experiment, including that the order of the cups will be randomized. Random allocation provides a probability model against which the tea-drinker’s claim can be assessed. Fisher uses this probability model as the basis for the significance test. He outlines his approach in the following way:

In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them. Further, we must know by what argument this interpretation is to be sustained.

(Fisher 1966, 12)

The possible outcomes of Fisher’s experiment is that the tea-drinker correctly identifies zero, one, two, three, or four “milk-first” cups. Given what matters for the experiment is how many “milk-first” cups the tea-drinker correctly identifies (as opposed to which “milk-first” cups the tea-drinker correctly identifies), there are 70 possible permutations of the outcomes of the experiment. In only one of these permutations are all four “milk-first” cups correctly identified (of the remaining 69 permutations, there is one permutation in which no “milk-first” cups are
Fisher then assigns a pre-experiment probability to each of the experimental outcomes. He does this by hypothesizing that the tea-drinker has no special ability to discriminate “milk-first” cups—this is labeled the null hypothesis. If the tea-drinker has no special ability to discriminate “milk-first” cups, then the probability of each outcome of the experiment equals the expected number of successes if the four cups were selected by chance. Fisher thinks of this in terms of the expected outcomes of the experiment if we were to repeat the experiment indefinitely, while assuming the truth of the null hypothesis. In this scenario, the probability of each possible outcome of the experiment is equal to the number of permutations in which the tea-drinker correctly identifies the specified number of cups divided by the total number of permutations. Fisher is particularly interested in the probability of the tea-drinker passing the test if in fact the null hypothesis is true. Despite having no skill in discriminating “milk-first” cups of tea, you would expect the tea-drinker to strike it lucky and correctly identify four “milk-first” cups once in every 70 repetitions of the experiment. Fisher reasons that 1 in 70 is sufficiently unlikely that he is willing to grant the tea-drinker’s claim to be able to accurately discriminate “milk-first” cups of tea if she or he passes the test. Notice that the random allocation of the cups provides the objective basis for Fisher’s test.

The second argument for random allocation refers to masking. “Masking” in clinical studies (also frequently referred to in the medical literature as “blinding”) is the attempt to conceal which treatment a participant is allocated to (e.g., the experimental treatment or control; “milk-first” or “tea-first” cups of tea). Allocation concealment can take place at multiple levels; most importantly, allocation may be masked from the participant only (“single-blind”), or allocation may be masked from the participant and investigators (“double-blind”). Employing adequate methods to mask participants and investigators is an important part of trial design because it minimizes a wide range of biases that can occur when participants and/or investigators are aware of treatment allocation. Senn (1994) argues in relation to the tea-drinker experiment that only random allocation adequately masks the participant to the order of cups. This is because use of a systematic or haphazard method of allocation leaves the possibility open that the tea-drinker is able to make some educated guesses regarding likely orderings. Even if it is considered unlikely for the tea-drinker to guess the exact order of the cups, if she or he is able to rule some orderings more or less likely, then her or his chances of striking it lucky and correctly identifying all four “milk-first” cups will be considerably greater than 1 in 70, thereby undermining the rigor of the test.

Complete masking throughout a clinical trial is often difficult to achieve. Most experimental treatments cause effects that may unmask treatment allocation. The argument in favor of random allocation is that it is prone to fewer sources of bias than alternative methods of allocation. For instance, haphazard allocation is subject to subtle (perhaps subconscious) biases the investigator might have for allocating participants to experimental treatment or control. And, systematic allocation is subject to biases that may arise from unexpected influences on the order of recruitment. (Perhaps a clinical study sets out to allocate the first four participants enrolled every day to the experimental treatment, then the next four to control, and so on. This system of allocation is likely to result in biased results if, for instance, the first six patients seen in the clinic every day are the sickest.)

Finally, random allocation reduces the influence of extraneous risk factors. “Extraneous risk factors” are confounders: they are associated with both the exposure under investigation and the outcome under investigation without being an intermediate cause between the exposure and the outcome. Life-course socioeconomic status is an extraneous risk factor between

identified; 16 in which one “milk-first” cup is identified; 36 in which two “milk-first” cups are identified; and 16 in which three “milk-first” cups are identified).
hormone replacement therapy (exposure) and cardiovascular disease (outcome). Random allocation distributes extraneous risk factors in a statistically predictable way—and it does this independently of whether the extraneous risk factor is recognized. Exactly what random allocation achieves by distributing extraneous risk factors in a “statistically predictable way” is a key point of contention in the literature debating the merits of random allocation. Appreciating the statistical details of how random allocation influences the distribution of extraneous risk factors is important for understanding what random allocation does (and doesn’t) achieve. In what follows, I provide the frequentist account. Consider a single extraneous risk factor, $A_1$, that is present in 50% of a population. Randomly allocating members of this population to intervention or control in a 1:1 ratio is analogous to tossing a fair coin. The larger the trial, the more likely $A_1$ is similarly distributed in the treatment and control groups. If the population of prospective participants is infinite, and recruitment is continued indefinitely, then we expect the proportion of participants with $A_1$ in each arm of the trial to approach 50% as the trial size approaches infinity.

We get the same result with a slightly modified scenario. In the modified scenario, the population of prospective participants is finite and the size of each trial is fixed. Consider a trial that recruits, randomizes, and observes 1,000 participants. The size of any single trial is large enough that $A_1$ is expected to be found in similar proportions in the experimental groups (think of this as something like 1,000 tosses of a fair coin). Now consider replications of this trial. In the indefinite sequence of trials, we expect most trials to have similar proportions of participants with $A_1$ in each arm, and while we would expect some trials to have an uneven distribution of $A_1$ simply through the play of chance, this would balance out in the indefinite sequence of trials. The overall effect of $A_1$ on the trial outcome summed over the indefinite sequence of trials is 0. This simple case can be extended to take into consideration multiple extraneous risk factors, $A_i$ where $i = 1, \ldots, n$. Providing we consider the indefinite sequence of trials, the same can be said: the overall effect of $A_i$ extraneous risk factors on the trial outcome when summed across the indefinite sequence of trials is 0.

Random allocation provides a probabilistic model for how extraneous risk factors will be distributed in the indefinite sequence of trials. This formulation is disputed by Bayesians and other non-frequentists, who question the relevance of what happens in the indefinite sequence of trials. The important point to recognize is that many of the stronger statements made about what random allocation achieves refer to the indefinite sequence of trials, sometimes implicitly. For instance, Devereaux and Yusuf (2003, 107):

\[ \ldots \text{RCTs are superior to observational studies in evaluating treatment because RCTs eliminate bias in the choice of treatment assignments and provide the only means to control for unknown prognostic factors.} \]

Random allocation eliminates bias due to known and unknown extraneous risk factors, but only in the indefinite sequence of trials. In any single trial, the most that can be said is that, providing the trial is large, random allocation improves the probability that extraneous risk factors (known and unknown) are roughly balanced. This appears to have been enough for the Women's Health Initiative Study to provide a superior assessment of the effects of hormone replacement therapy on heart disease—assuming that there are residual confounders that explain the difference between the cardiovascular outcomes measured in the observational studies and the Women's Health Initiative Study, and that these confounders are sufficiently evenly distributed in the Women's Health Initiative Study. Importantly, this more fragile benefit of random allocation is less than what is implied by some proponents of randomized trials.
Criticisms of Randomized Trials

There are many examples in the general medical literature of strong claims made regarding the benefits of randomized trials:

Without clear confirmatory evidence from large-scale randomized trials or their meta-analyses, reports of moderate treatment effects from observational studies should not be interpreted as providing good evidence of either adverse or protective effects of these agents (and, contrary to other suggestions, the absence of evidence from randomized trials does not in itself provide sufficient justification for relying on observational data).

(Collins and MacMahon 2007, 24)

This is a strong claim in a carefully argued paper on the benefits of randomized trials in medicine. It is easy to find statements that echo this sentiment or make stronger claims. The overarching view of randomized trials provided in this strand of the literature is that they are categorically superior to alternative study designs. The expressed view is that randomized trials are necessary for drawing causal inferences about medical therapies. The justification for this view is the idea that random allocation provides some kind of guarantee of the observed results of a trial.

This view is targeted by critics, including Peter Urbach (1993) and John Worrall (2002, 2007). Worrall summarizes the problematic view of randomized trials in the following way:

It is widely believed that RCTs carry special scientific weight—often indeed that they are essential for any truly scientific conclusion to be drawn from the trial data about the effectiveness or otherwise of proposed new therapies or treatments.

(Worrall 2007, 452)

Worrall is successful in undermining this view of randomization, and his argument is an important corrective to a common tendency to overstate what randomized trials achieve. Worrall’s key argument hinges on what random allocation achieves with regard to the distribution of extraneous risk factors. Worrall suggests that those who hold the view that randomized trials are essential in medicine elide what randomization achieves on any particular allocation with what randomization achieves in the indefinite sequence of trials. If random allocation did ensure all possible confounders were equally distributed in the intervention and control groups, then randomized trials would have a distinct advantage in terms of internal validity. Other things being equal, any difference observed in such a trial could only be due to the effect of the treatment; this is the kind of guarantee randomized trials are assumed to provide by those who view randomized trials as essential. But, of course, random allocation does not achieve an equal distribution of all possible confounders in any specific trial (perhaps, for example, in a particular trial more women are allocated to the treatment group than men). Indeed, as Worrall (2002, S324) notes, and as Urbach (1993, 1426) did before him, the more possible confounders there are, the higher the probability that any one of the confounders will be mal-distributed in the experimental groups in any particular random allocation. Randomization provides no sure-fire guarantee for the results of the trial.

Most people who write about randomized trials in statistics and medicine know this, and most acknowledge it somewhere, if sometimes only in the small print. Nevertheless, the idea that randomized trials are necessary and sufficient to justify causal inferences about treatments is commonly expressed. And this simple, commonly expressed view is enormously influential.
in health care and a growing range of disciplines. Worrall takes the view that random allocation in clinical trials provides no additional benefits over those that could be gained from adequate control in an observational study, where “adequate control” for Worrall means comparable based on known confounding factors. But this swings the randomization debate too far in the opposite direction. As discussed above, there are good reasons to randomize even though random allocation does not guarantee the results of a trial. Random allocation provides an objective basis for statistical analysis, permits masking, and reduces the influence of unsuspected confounders. The benefits of randomized studies are, however, relative. Randomized trials employ methods that avoid sources of error that cannot be avoided in observational studies. The capacity to avoid more sources of error means that, other things being equal, the internal validity of a well-conducted randomized study will be higher than that of a well-conducted observational study.

The Challenge of External Validity

Being clear on the benefits of randomized trials—and the justifications for these benefits—helps to better identify the limits of randomized trials. Before releasing a new treatment into the market, we want to be sure that it works in a clearly identified group of patients. When thinking about a study testing whether a new treatment works, the focus is rightly on internal validity: can we trust the results? Well-conducted randomized trials provide more reliable results than do alternative methods. In this context the frequently offered advice to focus on randomized trials can be defended, but internal validity is only one aspect to consider. Once the treatment is on the market, clinicians will need to decide whether the treatment is likely to benefit the specific patients under their care; similarly, the patients will need to decide if it is in their best interests to take the treatment. Assessing whether a medicine will work for a specific person is often significantly harder than assessing whether it works, on average, in an experimental sample. These decisions rely on a broader range of evidence than that provided by well-conducted randomized trials. In this context, following the advice provided in many medical resources to focus on evidence from randomized trials leads to poor decisions.

External validity is the “degree to which the results of an observation hold true in other settings” (Fletcher et al. 1996, 12). Assuming an experimental treatment benefits the sample of patients included in a well-conducted randomized trial, the assessment of external validity focuses on the generalizability of the observed results to patients outside of the trial. A related term is effectiveness—an intervention is “effective” when the intervention benefits patients receiving the treatment undergoing routine care. Effectiveness contrasts with efficacy; an intervention is “efficacious” when it benefits patients in an experimental setting (e.g., a randomized trial). There is consensus in medicine as to what needs to be done to achieve internal validity. There is less consensus on the topic of external validity. Indeed, the main point of consensus within medicine is that external validity is difficult to judge (Rothwell 2007; Rawlins 2011).

Well-conducted randomized trials are a more specialized tool than is typically appreciated. The particular function in medicine to which they are well-suited is the rigorous assessment of whether a treatment is efficacious. All of the randomized trials conducted throughout drug development are designed to test a specific hypothesis: does the investigational treatment produce benefits on a specific health measure when compared to control in a particular group of patients? All of the key methodological decisions made in the trial are to ensure the primary hypothesis is assessed in a way that avoids systematic and random error. Many of these decisions limit external validity.

Variability is a threat to internal validity and trial efficiency. Trial participants can vary in many ways that influence their response to treatment; they may vary with respect to their
underlying health, the severity of their condition, their adherence to treatments, their ability to absorb, metabolize, and/or eliminate the pharmaceutical treatment, and so on. The more variability, the more likely some of this variability will influence trial outcomes, which means more participants will be needed to demonstrate the purported benefits of the investigational treatment. Variability in how patients are treated other than the investigational treatment is also tightly controlled in randomized trials. Many randomized trials conducted for the purposes of regulatory approval are multi-site, and a significant number are multi-national. Considerable effort is taken to ensure that trial participants are treated in similar ways at the various trial sites in order to reduce bias. For instance, non-experimental treatments, treatment progression and withdrawal, and patient monitoring are all standardized. These conditions are important to ensure that the trial results reflect the experimental treatment rather than some other aspect of care. But, at the same time, each one of these conditions shifts a participant’s treatment further away from what he or she would have received under routine care. This raises an inevitable question: will the benefit observed under strict trial conditions be observed in the clinic?

Typical patients recruited to clinical trials are often importantly different from typical patients with the same condition seen in the clinic. Patients enrolled in trials tend to be younger and suffer from fewer additional illnesses—this improves trial efficiency and reduces the risk of adverse effects from the investigational treatment. The duration of treatment for many interventions is also substantially shorter in a randomized trial compared to routine care. Furthermore, patients enrolled in trials often have a more severe presentation of the primary condition being treated. Patients with moderate-to-severe illness are more likely to experience clinically important outcomes, such as heart attacks or strokes. In this way, the trial will be able to demonstrate the benefits of a new drug in reducing heart attacks or strokes with fewer participants enrolled in the trial. There is no sense in which participants in randomized trials are a random sample of patients in the community with the primary condition being treated. Much progress has been made in improving the reporting of trials in recent years; specifically, inclusion and exclusion criteria, flow of patients through the trial, reasons for trial ineligibility, and dropouts are far better reported now than they were 10 years ago. Nevertheless, it remains difficult to judge the population of which the trial participants are a sample.

The upshot of all of this is that it can be difficult to generalize the results of randomized trials. One response to the challenge of external validity is to conduct more randomized trials, especially large pragmatic randomized trials. Large pragmatic trials (also known as “large simple trials”) attempt to conduct the trial in a way that is as close to clinical practice as possible (Yusuf et al. 1984). This is achieved by keeping the inclusion and exclusion criteria simple, testing a single intervention, and focusing on a single clinically important outcome (such as mortality). When available, large pragmatic trials provide clinically important evidence on the average outcomes of a treatment and avoid some of the problems of smaller randomized trials. The Women’s Health Initiative Study arguably falls into the category of a large pragmatic trial; another positive example is the ISIS-2 study (1988), which enrolled 17,187 patients and demonstrated the benefits of antithrombotic treatments and aspirin in patients suffering from an acute heart attack.

However, large pragmatic trials are not a comprehensive solution to the problem of external validity. The first problem is practical: large trials, no matter how simple, take time and are expensive to complete. If a large pragmatic trial has been conducted for the specific effectiveness question under consideration, it will likely be relevant, but for the vast majority of questions, it is unlikely a large pragmatic trial has been conducted. The second problem is more technical. The main selling point for large pragmatic trials is that by allowing considerably more variability in the patients recruited and in the non-experimental treatments that
they receive, the trial provides more insight into the likely effects of the treatment in routine clinical care. This is true to an extent. A well-conducted successful large pragmatic trial provides good evidence that the average effects of giving the treatment are positive. However, in extending the results of such a trial to a given specific population or individual, the critical assumption is that the positive average effects are consistent across the many subpopulations included in the trial. Sometimes this seems to be a reasonable assumption, but often it is an assumption that is difficult to justify. Subgroup analyses attempt to quantify treatment effects in participant groups within a clinical trial, but they are notoriously unreliable (Feinstein 1998; Brookes et al. 2001).

When the assumption of similar effects in different patient groups is brought into question, the results of large pragmatic trials are just as difficult to generalize as those of small trials. The problem in the medical literature is that there is too great a focus on randomized trials. It is often underappreciated how much evidence external to a randomized trial needs to be taken into account in order to assess whether the results can be generalized to another setting. Assessing internal validity is largely an assessment of the methods employed in the trial—too often, this focus on trial methods is exported to the assessment of external validity. The important role that evidence external to the randomized trial plays in the assessment of external validity is also often underappreciated. Well-conducted observational studies and the insights provided by basic medical science possess unique strengths in providing evidence that supports judging whether the results observed in an experimental setting are likely to translate to patient benefits in the clinical setting (Black 1996; Vandenbroucke and Psaty 2008; La Caze 2011).

Nancy Cartwright’s work highlights what needs to hold in addition to a successful randomized trial to be confident that a treatment will be effective. Cartwright and colleagues argue that judgments of external validity and effectiveness need to be causal (Cartwright 2010, 2012; Cartwright and Munro 2010; Cartwright and Hardie 2012). Among other causal considerations, assessing the effectiveness of an intervention requires an understanding of how the intervention works and confidence that what is required in the environment to support the intervention working is present in the context in which the intervention is to be employed. Despite some notable exceptions—specifically, Rothman et al. (2008, 129)—explicit causal reasoning is rarely emphasized in the assessment of external validity within medicine. The most common advice provided on assessing external validity in the medical literature is to judge the similarities of the randomized trial with clinical practice. For example:

External validity is matter of judgment and depends on the characteristics of the participants included in the trial, the trial setting, the treatment regimens tested, and the outcomes assessed.

(Moher et al. 2010, 20–21)

Assessing similarities between the experiment and routine practice will often be important, but it is not enough. What matters is whether or not the causal structure is sufficiently replicated (see Chapter 5, “Mechanisms in medicine,” and Chapter 6, “Causality and causal inference in medicine,” for further discussion).

Cartwright (2010, 60) argues that the standard way in which external validity is conceptualized risks obscuring the causal judgments that need to made based on an assessment of all the available evidence. Cartwright (2011) and Cartwright and Hardie (2012) focus on effectiveness arguments—the argument that supports a claim that an intervention will be effective—and articulate the assumptions present in a strong effectiveness argument.
Argument A:
1. \( x \) plays a causal role in the principle that governs \( y \)'s production [in the experimental setting].
2. \( x \) plays a causal role [in the clinical setting] as well as [in the experimental setting].
3. The support factors necessary for \( x \) to operate are present for some individuals [in the clinical setting].

Therefore, \( x \) plays a causal role [in the clinical setting] and the support factors necessary for it to operate are present for some individuals.

(Cartwright 2011, 222)

Here, \( x \) is the intervention and \( y \) is the outcome. Argument A must be sound for an intervention to have a causal effect in the clinical setting. A well-conducted randomized trial provides good evidence for premise 1, but more is needed to ground an effectiveness argument. Premises 2 and 3 also need to hold. The assessment of these premises is causal and will need to rely on evidence external to the randomized trial. Is the causal structure in the clinical setting similar enough to the causal structure present in the randomized trial? Are the factors that support \( x \)'s effect on \( y \) present in the clinical setting? Are there factors present in the clinical setting that reduce \( x \)'s effects or promote \( y \)'s harms? It is important to note that the conclusion of A is that the intervention will have a causal effect in the clinical setting—postulating that the causal effect will be positive on average (or in a particular patient) requires still further conditions to hold (see Cartwright 2011, 2012 for further discussion).

Cartwright’s argument schema better explicates the requirements on effectiveness arguments and judgments of external validity. This doesn’t make the difficult causal judgments any easier, but it is a step in the right direction for improving the assessment of external validity.

Conclusion

Well-conducted randomized trials reduce or eliminate more sources of error when assessing the efficacy of a treatment compared to alternative study designs. Randomized trials are neither infallible nor the only important source of evidence in medicine. Predicting whether the effects of a treatment observed in a randomized trial will be observed in routine clinical care is often challenging. The assessment of external validity relies on causal knowledge, which is provided by evidence from a range of sources.

References

THE RANDOMIZED CONTROLLED TRIAL


Further Reading


