3 Phases of Investigation

As discussed in Chapter 1, science is adversarial; scientists should consider *all* possible hypotheses, and only eliminate those that are inconsistent with the data. As illustrated in the scientist game and as practiced in clinical trials, this approach leads naturally to a sequence of experiments each of which are designed to limit the number of remaining hypotheses. Furthermore, the sequence of trials can be considered in advance so that the design of any particular trial fits logically into an overall program. The overarching objective for this sequence of trials is to take experimental therapies and identify those that are safe and effective for treatment of a specified indication. Trials are usually motivated by promising results in cell and animal studies. The initial trials seek to refine dosing and delivery methods to obtain a therapy that can be tolerated. Subsequent trials seek to establish safety. Efficacy is only evaluated after tolerability and safety are reasonably assured. In developing new drugs, the sequence is known by its phases. Phased approaches have also been developed for settings other than drug development. The phases of clinical trial are nearly codified by agencies charged with regulating the drug development process (reference ICH guidelines).

In addition to the phased approach to evaluation of new therapies, the goals of the trial are established by the clinical context. For example, a new treatment might have been developed to improve outcomes when compared with current standard of care. In such cases the trial goal is usually one of establishing the *superiority* of the new treatment when compared with placebo. In contrast, a new treatment that is known to be less toxic or is easier to deliver (e.g., pills versus IV infusions) would tend to replace current therapy as long as it can be shown to be *non-inferior* in its beneficial effects. The goal of *equivalence* trials is to establish if two treatments can be used interchangeably.

In this chapter we motivate and describe the phased approach to drug development. We also discuss various types of trial goals. In later chapters, the phase of the trial establishes a context that is useful in selecting and refining an appropriate endpoint for the trial (and vice versa: the endpoints often determine the phase). Trial phase may also determine the study population, type of comparison therapy, and duration of follow-up. The trial goal informs the specification of the statistical hypotheses which represent the formalization of the scientific question.

3.1 Trial Setting: Phases of clinical research

3.1.1 Motivation and definition

When teaching clinical trials, I often ask the class to stop and ponder the information that would be necessary before a particular new treatment could be used in general clinical practice. The ensuing discussion usually produces the key criteria that govern the phased approach to clinical research. These include:

- *Efficacy:* Before being adopted for routine clinical practice, a new treatment should be shown to have superior efficacy (or at least be non-inferior) to an established treatment. Furthermore it should have an acceptable safety/toxicity profile. Trials to fully evaluate safety and efficacy often need to be sufficiently large to provide enough information about various measures of efficacy along with (possibly rare) side effects.
- *Safety:* Prior to embarking on the ultimate (large) trial, it is usually scientifically efficient and often ethically important to first conduct a smaller trial to assure that the new treatment is not unsafe, and that there is some evidence that it might prove to be beneficial. Without this preliminary information, it may not be ethical to enroll large numbers of patients to a trial of an experimental treatment. It might also be difficult to spend the financial resources, to commit the time, or to invest human subjects in a large trial without preliminary evidence of safety/efficacy.
- *Tolerability*: Before either large scale trials or preliminary studies of safety/efficacy can be conducted, the specific dose, delivery schedule, and treatment duration must be determined. In truly new therapies, "first-in-human" trials must start from what is felt to be a very low does based on pre-clinical laboratory studies in animals. The dose and/or delivery schedule is then escalated in order to identify a therapeutic approach that is tolerated by patients and is felt to have the best opportunity for success.

The need for this logical sequence of information is dictated both by the global requirements for the evidence that is necessary for drug approval and by the requirement for preliminary information prior to launching large clinical trials. The classical phases of drug development are as follows:

- *Phase I:* A dose-finding study using safety or tolerability endpoints. A phase I trial is often the first time that a drug is used in humans or in a new patient population. Phase I studies are usually preceded by pre-clinical studies that examine drug effects on tumor tissue cells and/or on animals. Preclinical studies are used to demonstrate biological activity, to provide an indication of the type of toxicity that might be anticipated in humans, and to select a starting dose for first-in-human trials. Eligibility criteria for phase I trials often allow a heterogeneous mix of disease types and/or more advanced disease among patients who have exhausted other treatment options; in this way the phase I patient population is often different from the eventual target population for the drug. A classical design for a phase I trial would treat 3 patients at a dose level. If none of the 3 experiences a dose-limiting toxicity, then the next 3 patients are treated at a higher dose. The dose continues to escalate until one or more of the 3 patients experiences a dose-limiting toxicity. At this point, escalation may be halted and additional patients treated at this highest dose, or the dose may be reduced to a lower dose. The result is the selection of a ”Maximum Tolerated Dose” (MTD) for use in future trials. Alternative designs have been proposed for phase I trials in an attempt to improve their efficiency, however these designs have not experienced wide-spread use, probably because of a need for caution in escalating the dose of an unknown drug in human studies. Additional discussion of the design of phase I trials is given in Chapter 33.

- *Phase II:* Following identification of a drug dose and delivery schedule, the next phase evaluates safety in a larger sample on a potentially wider range of toxicity endpoints. The phase II trial is typically conducted in the intended target population for the therapy. It may or may not include a randomized concurrent control group. The phase II trial also begins to evaluate preliminary evidence for treatment efficacy. The trial may use a surrogate endpoint instead of the ultimate clinical endpoint (e.g., tumor response for survival, or cholesterol reduction for risk of myocardial infarct). The phase II trial must provide adequate information to evaluate whether or not a phase III trial is warranted; thus, the drug's safety profile must be such that it is ethical to give the drug to a larger number of patients. It must also provide as much reassurance as possible that the drug is likely to show benefit in a phase III trial.

- *Phase III:* The phase III trial is usually considered the gold standard for deciding whether or not a therapy should be approved for routine use in standard practice. The study population in a phase III trial should reflect the intended target population for its ultimate use. Potential adverse effects should be fully evaluated, including the potential for rare adverse effects for which the trial might have limited statistical power. The outcomes in a phase III trial should reflect the clinical targets for the therapy (e.g., cancer drugs should be shown to improve survival or lengthen the time to cancer recurrence). The sample size in a phase III trial should be sufficiently large to inform the decision about its ultimate use (see Chapter 9).

- *Phase IV:* A phase IV study evaluates the benefits of a drug after it has been approved for routine use. There is good evidence that the benefits of a new treatment as measured in a phase III controlled trial (usually called "efficacy") is larger than its effect in routine clinical use (usually called "effectiveness"). There are many reasons why this can occur including:

  - Differences between the types of people who volunteer for trials from those who do not.
  - Differences between dose and delivery in routine use from the carefully scheduled dosing algorithm that might be used in a phase III trial.
  - Once a drug is approved, it can be used "off-label" for other conditions that were not studied in the clinical trial.
  - Ancillary treatments and follow-up schedules are quite variable in routine use and are often carefully controlled in a clinical trial.

  A phase IV study is unlikely to be randomized; thus, it is often subject to biases that would not be accepted in a clinical trial. A randomized phase IV trial would look like a phase III trial with a much more general study population and with fewer restrictions on dosage and delivery schedules (i.e., evaluating effectiveness instead of efficacy). Randomized phase IV trials are uncommon, which argues for designing phase III trials that are less restrictive and are more likely to reflect the use of drugs once approved. The other side of this argument is that many of the restrictions in a phase III trial serve to reduce variability thereby improving trial power (requiring fewer patients in the trial). Thus, effectiveness is usually examined through uncontrolled post-market surveillance of drug effects in its routine use. There are many examples of important adverse effects that have emerged from post-marketing studies.

In the phased approach, the endpoints, length of follow-up, study population, treatments, comparison group, and sample size also progress across the phases (Table 1). Sample size increases with the phases because less information is needed to get an indication the effect on an intermediate biological marker than is required for evaluation of effects on clinical outcomes. Early phase endpoints tend to be designed to study immediate effects of the drug or biological processes that are targeted by the drug. Later phase studies focus on more distal effects of the drug that affect clinical outcomes of importance to patients. Early studies may not have a control group and the treatments may not reflect the ultimate delivery schedule or duration. Similarly, the study population in early phase trials may not reflect the ultimate target population for the drug. It is common for new drugs to first be used in patients with more advanced disease because the risk profile of the drug is not well known, and patients with advanced diseases often have more to gain (i.e., the risk:benefit ratio is acceptable). In later studies, risks are better understood, and may allow patients with less advanced disease.

	Phase I	Phase II	Phase III
Endpoints	Tolerance bioactivity	Safety/ safety	Efficacy/
Follow-up	Short	Moderate	Longer
Population	Non-target	Target	Target
Treatment	Dose finding	Study dose	Study dose
Comparison	Across doses	Historical/ randomized	Randomized
Sample Size	Very small	Moderate	Large

Table 1: Design elements from phase I-III

We note that although the above phased approach represents the classical construct, much of its utility is in providing a conceptual framework for developing an efficient sequence of trials that will ultimately produce the information necessary to decide if a new treatment should become part of routine clinical practice. We sometimes think that this phased approach has as many variants as there are drugs being developed. Thus, it is common to see a trial described as a phase I/II or II/III. In cancer therapeutics, phase IIa usually denotes a single-arm (uncontrolled) trial, and phase IIb denotes a randomized controlled trial which is smaller than the ultimate phase III trial and may measure efficacy by a surrogate for the clinical endpoint.

Other fields of clinical research have defined different phases. For example in surgery or device evaluation trials phase I allows changes to the procedure or device to refine the approach prior to larger trials; the remaining phases are similar to those for drug development. For development of diagnostic tests (ref Pepe) phase I involves pre-clinical exploration, phase II evaluates sensitivity and specificity in prevalent cases and control groups, phase III evaluates sensitivity and specificity in incident case-control studies, phase IV involves prospective screening for disease detection, and phase V involves evaluates whether or not screening reduces disease mortality in the population.

3.1.2 Efficiency and accuracy of the phased approach

In the previous section the phased approach is motivated using the sequence of information required to decide if a drug should be used in routine practice (tolerability before safety before efficacy). This way of motivating the phases of clinical research emphasizes the scientific and ethical requirements for establishing a sufficient information base in smaller studies prior to launching larger trials or to approving the therapy. With this motivation the phases become hurdles that must be passed by a particular drug before it can advance to the next stage.

The phased approach to drug development can also be viewed as a means of increasing the efficiency and accuracy of the process for identifying therapies that should become part of routine practice. This alternative motivation emphasizes a more global view of the drug development process in which tens of thousands of candidate drugs must be evaluated, and clinical trials are used to identify and adopt as many of the truly efficacious therapies as possible without incorrectly approving too many ineffective therapies. The FDA, patient advocate groups, and companies with many candidate drugs share this global perspective on the drug development process. From this viewpoint there is less concern over whether any particular drug is approved, and more concern over increasing the probability that approved drugs have a high probability of being efficacious. An investigator or company who is developing a single agent (or a participant in a particular trial) have the perspective of the previous section; there is relatively more concern over whether or not the particular drug under consideration wins approval, and relatively less concern over the global efficacy of all approved drugs.

From the global perspective, the clinical trial is exactly like a diagnostic test, although instead of diagnosing disease, a clinical trial diagnoses effective treatments. Recall that the accuracy of a diagnostic test is measured by its sensitivity (probability that the test is positive among people who have disease) and its specificity (probability the test is negative among people who do not have disease). In a clinical trial, the sensitivity is the probability that the trial will produce a positive (statistically significant) result if in fact the drug is efficacious. The probability is commonly known as the *power* of a the trial, and we denote this probability by β. The specificity of a trial is the probability that the results will be negative (not statistically significant) if in fact the treatment is not beneficial. This probability is $1 - \alpha$ where α denotes the probability of a type I statistical error (falsely rejecting the null hypothesis). We reserve a more rigorous definition of statistical operating characteristics for a later chapter, but use these concepts to illustrate the global perspective of the drug development process.

Of primary interest from the global perspective is the prevalence of truly efficacious drugs among all of the drugs that are approved. In a diagnostic test, this is known as the *positive predictive value (PPV)*; i.e., the probability that an individual has the disease given that the diagnostic test is positive. For diagnostic tests, the PPV is calculated from the sensitivity, the specificity, and prevalence using Bayes rule:

$$PPV \quad = \quad \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} \ + \ (1 - \text{specificity}) \times (1 - \ \text{prevalence})} \tag{1}$$

which for a clinical trial becomes:

$$= \quad \frac{\beta \times \text{prevalence}}{\beta \times \text{prevalence} + \alpha(1 - \text{prevalence})}$$

In diagnostic tests, "prevalence" represents the proportion of individuals with disease in the screened population. For rare (low prevalence) diseases, the PPV tends to be low even if the test has good sensitivity and specificity, and for common diseases PPV tends to be high even if the test does not have good sensitivity and specificity.

In the context of the drug development process, "prevalence" represents the proportion truly effective therapies among all therapies that are tested for approval. It is of course impossible to know the prevalence of truly effective therapies among all that might undergo clinical trials, however there is circumstantial evidence that the prevalence is in fact quite low. A review of the NCI drug development process (1970-1989) found that of approximately 350,000 compounds that were studied in the lab, 83 made it past pre-clinical and phase I testing, and 24 made it past phase II testing. Effective drugs are uncommon, and it is probably optimistic to assume that 10% of drugs in clinical trials are truly effective.

We now explore the application of the concepts from diagnostic testing to clinical trials and the drug approval process. Almost universally, the type I error rate for statistical tests is set at $\alpha = 0.05$ (specificity = 0.95). Phase III trials are generally large and have high power ($\beta > 0.9$). With sensitivity of $\beta = 0.9$, specificity of $1 - \alpha = 0.95$, and prevalence of 0.10, the PPV is 0.67; in words, 67% of treatments that are approved are truly effective and 33% of approved treatments are in fact ineffective. It is worrisome to think that the drug development process could allow such a high proportion of ineffective therapies to make it through the approval process. Furthermore, we note that this problem cannot be solved by conducting larger studies as long as the prevalence of effective therapies is low. In the above example increasing power to 0.975 only increases the PPV to 0.68.

There are two ways to solve this problem. The first is to reduce the alpha level (increase specificity); using $\alpha = 0.025$ increases the PPV to 81%. The second is to use smaller phase II studies to enrich the prevalence of truly effective studies that are then tested in phase III. For example, suppose that a phase II study has $\alpha = 0.05$ and $\beta = 0.15$ (i.e., a smaller sample size that gives 15% power), then the prevalence of truly positive drugs among those that are significant in a phase II study is 25% (i.e., PPV = 0.25). Then if moderate sized phase III studies with $\alpha = 0.05$ and $\beta = 0.8$ are conducted using drugs that are positive in phase II (i.e., on drugs with a 25% prevalence of truly effective treatments), then the overall PPV will increase to 84%.

As further illustration, suppose that 1,000,000 patients are available for enrollment in clinical trials for a particular disease or condition. Suppose further that investigators use these patients to conduct trials of 1000 drugs in large trials (1000 patients per trial) with power 0.975 and a false positive rate of $\alpha = 0.05$. Then if there are 900 ineffective drugs and 100 effective drugs, on average the evaluation process will yield 98 positive trials from the effective drugs ($0.975 \times 0.975 \approx 98$) and 45 positive trials from the ineffective drugs ($0.05 \times 900 = 45$). Thus, 98/142 (PPV = 69%) of the approved drugs actually work.

As an alternative, suppose that we conducted many smaller phase II trials in order to increase the prevalence of truly effective drugs prior to conducting larger phase III trials. Specifically, suppose we screened 12,500 drugs (1,250 effective and 11,250 ineffective) in 50 patient trials that each had 15% power. On average these trials would identify 187 effective drugs ($0.15 \times 1250 \approx 187$) and 563 ineffective drugs ($0.05 \times 11,250 \approx 563$). Thus, the phase II trials increased the prevalence of truly effective treatments to 25% (187/749). Now suppose that the 749 drugs from phase II are studied in phase III trials with 500 patients per trial and 80% power. These confirmatory trials would on average identify 150 of the 187 effective drugs ($0.8 \times 187 \approx 150$) and 28 of the 562 ineffective drugs ($0.05 \times 562 \approx 28$). The 2-phase development process results in screening of 12,500 drugs leading to approval of 178 drugs of which 150 (84%) are truly effective. In contrast the single-trial approach screens 1000 drugs and leads to approval of 142 of which 98 (69%) were truly positive.

3.2 Trial goals and objectives

The trial setting is determined by the current state of knowledge along with what that state should be upon trial completion. The phased approach described in section 3.1 is a convenient organization of the sequence of questions that follow from both efficiency considerations and from the ethical requirements for safety information prior to large efficacy trials. Although the trial setting helps to determine many of the trial characteristics (Table 1), the trial's goals and objectives are also determined by the disease setting, nature of the therapeutic intervention under investigation, and the nature of the current standard of care.

In this section we present a classification of trial's scientific goals and objectives. In combination with the trial setting, these goals will ultimately form the foundation for the statistical design including specification of the statistical hypotheses, trial decision criteria, and framework for evaluation of the trial power and sample size. In this discussion we denote the true underlying difference between treatments by the greek letter θ. In future chapters we develop a more rigorous definition of θ, but here it simply represents the true difference between therapeutic approaches. In what follows, trial goals and objectives are classified by two criteria: (1) the number of decisions (corresponding to the nature of the statistical hypotheses), and (2) the type of decision. We address each in turn:

1. *Number of decisions:*

- *Two-decision trials:* Trials whose goal is to choose between two actions: to either adopt (or continue studying) a new therapy, or do not adopt the new therapy. Thus, if treatment A denotes an active (new) treatment and C denotes its comparator, a two-decision trial either decides to use A or not to use A. A 2-decision trial is structured around one-sided statistical hypotheses. If θ is defined such that large values denote preference for treatment A, then a 2-decision trial will choose between two hypotheses. For example, a superiority trial chooses between:

$$\begin{aligned} H_0: &\quad \theta \leq \theta_0 \\ H_+: &\quad \theta \geq \theta_+ \end{aligned}$$

  with $\theta_0 < \theta_+$. Later in this section, we see that a decision in favor of treatment A corresponds to rejecting H_0 in favor of H_+, and a decision against A corresponds to rejecting H_+ in favor of H_0.

- *Three-decision trials:* Some trials are conducted to choose one of three decisions. Specifically, three decisions are implied whenever the goal of a trial is to see if two treatments are equivalent, and if not, then to choose the better of the two. Thus, if A and B denote two treatments, then a 3-decision trial will conclude either A is better than B, B is better than A, or either treatment can be used. Once again if large values of θ indicate A is better than B, then a three decision trial chooses between:

$$\begin{aligned} H_0: &\quad \theta = \theta_0 \\ H_+: &\quad \theta \geq \theta_+ \\ H_-: &\quad \theta \leq \theta_- \end{aligned}$$

  with $\theta_- < \theta_0 < \theta_+$, which corresponds to a two-sided statistical hypothesis test. The decision that A is better is B requires rejecting both H_- and H_0 in favor of H_+; the decision that B is better than A requires rejecting both H_0 and H_+ in favor of H_-; the decision that A and B are equivalent requires rejecting both H_- and H_+ in favor or H_0.

2. *Type of decision:*

- *Superiority:* The primary goal of a superiority trial is to show that a new experimental therapy is superior to the current standard of care. Evaluation of superiority is the primary objective in a placebo-controlled trial, and the placebo-controlled trial is the classic example of a superiority trial. A trial with an active control group (representing current standard of care) may or may not evaluate superiority as its primary objective, although if an active-controlled trial has a superiority objective, the ultimate goal would be to replace the current standard of care with the new therapy if proven superior. The superiority objective implies that the trial should reach one of two decisions: (1) the new therapy is *superior* and should replace its comparator, or (2) the new therapy is *not superior* and should not be used. Thus, a superiority trial is structured around one-sided statistical hypotheses. The sepsis trial is an example of a placebo-controlled superiority trial (see Chapter 4).

- *Inferiority:* The goal of an inferiority trial is to prove that an existing therapy is in fact harmful. It is clearly unethical to conduct a placebo-controlled inferiority trial since the goal would be to show that a new therapy was harmful. It is only ethical to conduct a trial to establish inferiority if the therapy under evaluation is part of routine care. As with the superiority trial, an inferiority objective implies one of two decisions are relevant (inferior or not inferior); it is therefore constructed around a 1-sided hypothesis test.
- *Bioequivalence:* The goal of a bioequivalence trial is to show that two therapies are equivalent; i.e., that the trial results have ruled out all clinically-important differences. These types of trials are used to establish that two different formulations of the same drug are equivalent. The objective would be to evaluate whether or not the two drugs could be used interchangeably. Unlike other types of trials, the bioequivalence trial can reach one of three decisions, and therefore is designed around a two-sided statistical test.
- *Non-inferiority:* The primary goal of a non-inferiority trial is to show that a new therapy is not inferior to the current standard of care; that is, the new therapy may not be superior to standard of care, but at least any important differences in the harmful direction have been ruled out. If shown to be non-inferior, then the new therapy could be used instead of the current standard treatment. For example, a non-inferiority trial would be used to evaluate a new therapy that is known to have a preferable toxicity profile or is easier to deliver than standard care; such a therapy might be used instead of the current standard as long as it was not inferior in terms of efficacy based on the primary clinical outcome. As with the superiority or inferiority trials, a non-inferiority trial reaches one of two decisions (non-inferior or inferior) and is therefore constructed around a one-sided hypothesis test.
- *Approximate equivalence:* Although not part of standard terminology, we use the term “approximate equivalence” to refer to a trial in which the goal is similar to the non-inferiority trial, but in addition we require that the point estimate from the trial must be in the beneficial direction. Thus, the goal of an approximate equivalence trial is to show that both (1) all clinically important harmful effects have been ruled out, and (2) the group receiving the new therapy actually did better than the comparison group. The additional requirement for a positive point estimate avoids ‘non-inferiority creep’ as discussed in section 3.2.3.

3.2.1 How trial goals define and structure statistical parameter space

Ultimately, the trial’s scientific and clinical goals form the basis for defining the statistical “parameter space” and decision criteria. By its very definition a statistical parameter space is an abstraction; e.g., θ represents the true difference between the outcomes of future medical care if the population receives a new treatment instead of a placebo. At the same time the statistical parameter space must capture the very concrete elements of the scientific and clinical questions for the specific trial at hand. We now discuss the conceptual elements of the relationship between the statistical parameter space and the trial's scientific and clinical goals. We return to this topic with greater detail from the statistical design perspective in Chapter 5.

The relationship between the abstract and concrete must be discussed in the context of a specific example. In this section we consider a recent trial of daptomycin for treatment of a severe bacterial infection (Fowler VG, *et.al.* NEJM 2006; 355(7); 653-65). (Note that the sepsis trial described in part VI of this book is a different example in a very similar clinical setting. As described in part VI, there are some interesting differences between these trials that help to illustrate the relationship between trial goals and the structure of the statistical parameter space.) The daptomycin trial focuses on bacterial infections caused by *Staphylococcus aureus*, which is a common bacterial species that inhabits our skin. As a skin organism it is not usually pathogenic, however by unknown mechanisms it can cause severe systemic infections and death. Historically, *S. aureus* infection could be treated with routine antibiotics, but over the years antibiotic resistant strains have emerged. The most worrisome strains are resistant to methicilin and are know as "methicilin resistant *S. aureus*" or MRSA. MRSA infections are usually treated with vancomycin, although outcomes are often suboptimal. The daptomycin trial examined the use of daptomycin for treatment of patients with systemic *S. aureus* infection and endocarditis (an inflammation of the interior lining of the heart chambers). The trial was a randomized comparison of daptomycin against standard therapy, although treatment was 'open-label' meaning that both the treating physician and patient knew which drug they received. The primary outcome was treatment success defined as the absence of adverse outcomes such as death, microbiologic failure, and/or clinical failure. In this setting, the statistical parameter, θ represents the true underlying difference between success rates. The data from the trial are used to estimate θ, and the confidence interval represents the range of true values for θ that are consistent with the observed trial estimate; i.e., the values of θ that cannot be ruled out by the results in the trial.

"Parameter space" refers to the possible values of θ. In this trial $\theta = \theta_1 - \theta_0$ where θ_1 denotes the success probability with daptomycin ($0 \leq \theta_1 \leq 1$) and θ_0 denotes the success probability with standard care ($0 \leq \theta_0 \leq 1$). It follows that parameter space is all values of θ between -1 and 1. To structure parameter space, we first note that $\theta = 0$ denotes equality of success rates between daptomycin and standard therapy; hence, $\theta < 0$ denotes inferiority of daptomycin, and $\theta > 0$ denotes superiority of daptomycin. Of course, the values of θ in either the superiority or inferiority regions are not all clinically equivalent; small differences are less important than large differences. Therefore, during the statistical design we further separate parameter space into regions of important inferiority ($\theta < \theta_-$) and regions of important superiority ($\theta > \theta_+$). Thus, parameter space is divided into regions of important superiority, important inferiority, and a third region of differences that are not as clinically important (Figure 1).

The daptomycin trial was designed to demonstrate non-inferiority. The trial objective was to show that either daptomycin-based therapy or standard vancomycin-based therapy could be used to treat *S. aureus* bacteremia and endocarditis. One motivation for this objective is that we need many different antibiotics when treating bacterial infections due to the ongoing risk of resistance.

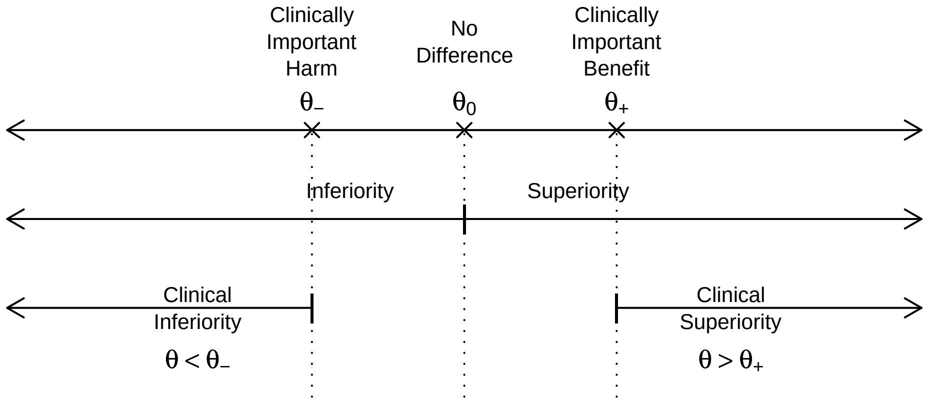

Figure 1: Structuring parameter space

Furthermore, there is evidence that the current standard of care is suboptimal. For this trial a non-inferiority boundary was set at -20%; that is, the trial would conclude that daptomycin was not inferior to standard care as long as the trial results rule out any decrease in the success rate greater than 20%. Based on this choice, parameter space for the daptomycin trial can be divided in regions of important inferiority $\theta < -0.2$, important superiority $\theta > 0.2$, and unimportant differences $(-0.2 < \theta < 0.2)$. (Note: it is not necessary that the boundaries between regions to be symmetric.)

3.2.2 How trial goals and parameter space determine decisions upon trial completion

Upon trial completion the interval estimate summarizes the hypotheses/parameters that can or cannot be ruled out (see also Chapter 1). In combination with the above structure on statistical parameter space, the location of the interval estimate will determine the action or decision regarding future use of the therapy under investigation. This relationship is commonly taught in introductory statistics courses in which the decision to reject the null hypothesis can be made if the null hypothesis is not included in the confidence interval. In a clinical trial we are also concerned with whether or not the confidence interval includes regions of important superiority or inferiority.

The structure of parameter space along with the confidence interval leads to decisions upon trial completion. Figure 2 combines the structured parameter space of Figure 1 with confidence intervals representing potential results upon trial completion. In this figure confidence intervals A and B would lead to superiority decisions (reject $\theta \leq 0$) and confidence intervals E and F lead to inferiority decisions (reject $\theta \geq 0$). Confidence intervals C and D do not reject $\theta = 0$), but reject any clinically important superiority $(\theta \geq \theta_+)$ or inferiority $(\theta \leq \theta_-)$, and therefore might lead to equivalence decisions. In terms of the types of decisions outlined at the beginning of this section, the structure of parameter space along with the trial results leads to the following decisions upon

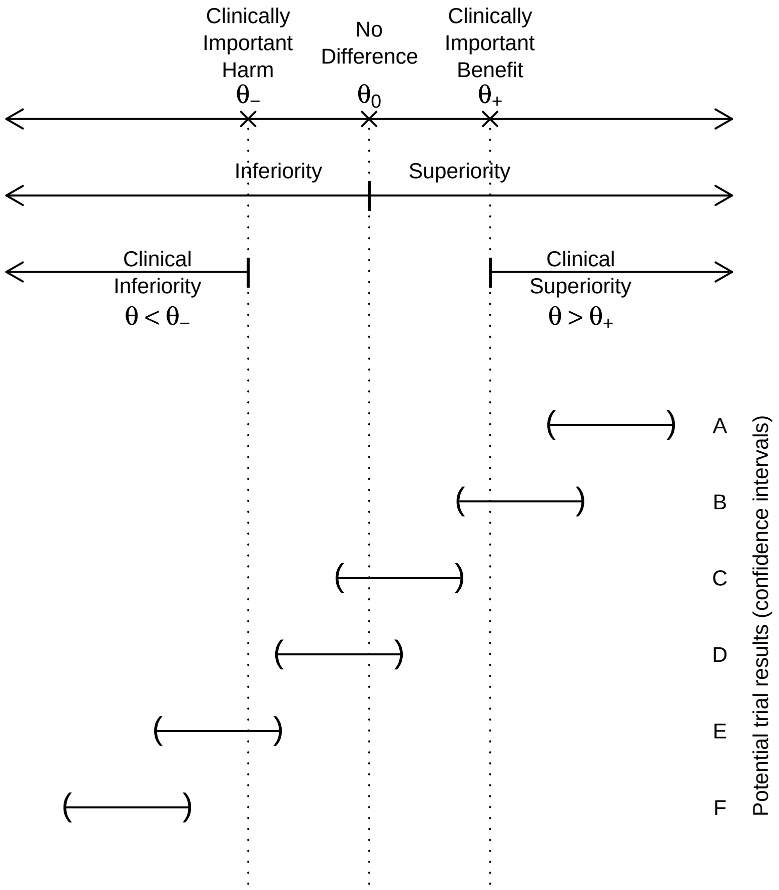

Figure 2: Potential trial results

trial completion:

- *Superiority:* Two decisions are of interest: superior or not superior. Confidence intervals A and B should lead to a conclusion that the new treatment is superior. Confidence intervals C-F all support a non-superiority decision. Thus, in a placebo-controlled trial A & B would support recommendation of the new therapy while C-F would not.

- *Inferiority:* Two decisions are of interest: inferior or not inferior. If the trial objective is to evaluate inferiority, then with an appropriately structured parameter space, intervals E & F would support an inferiority decision and the remaining intervals support a not-inferior decisions.

- *Bioequivalence:* In a bioequivalence trial, intervals C & D reject any clinically important differences between treatments, thereby supporting an equivalence decision. Intervals A & B support deciding that one treatment (call it W) is better than the other (call it V), whereas intervals E & F support the opposite conclusion (that V is better than W).

We reserve the discussion of non-inferiority and approximate equivalence trials for the next section.

Returning to the daptomycin example: $\theta_+ = 0.2$, $\theta_- = -0.2$, and $\theta_0 = 0$. Thus, confidence intervals A & B would support deciding that daptomycin is in fact superior to standard care, intervals C & D support deciding that daptomycin is no different from standard care, and intervals E & F support deciding that daptomycin is significantly worse than standard care. The goal of the daptomycin trial was to demonstrate non-inferiority; thus, the trial should recommend for the use of daptomycin as long as important inferiority is rejected. In this setting, intervals A-D would lead to non-inferiority decisions, and intervals E-F would lead to an inferiority decision.

Although when combined with the trial results through the confidence interval, a carefully-structured statistical parameter space must necessarily dictate the possible decisions upon trial completion, things are often not quite so simple. The following caveats must be considered:

- *Defining regions of importance:* Defining regions of unimportant and important difference (θ_- and θ_+) is inherently subjective. Furthermore, importance is usually a matter of degree rather than a specific threshold. The result is that the magnitude of the observed effect is considered in addition to which regions can be rejected. For example, intervals C & D might both be consistent with an equivalence decision, but the point estimate (center of the confidence interval) is positive in C and negative in D; thus, in a non-inferiority trial we might be much happier concluding non-inferiority for case C than for D.

- *Sample size and inclusive results:* In Figure 2 the width of the confidence intervals was selected so that it cannot overlap two regions. Since the confidence interval width is controlled by the sample size, a smaller study would produce a wider confidence interval and therefore runs the risk of including multiple regions such as both important superiority and inferiority. In such cases the trial could not reach a conclusive decision. We discuss how to choose the sample size to avoid inconclusive results in Chapter 9. Furthermore, since regions of importance are not objectively determined, it is possible that a conclusive result to one group of researchers may not convince another group. Differences in perspective can be incorporated into the statistical inference using a prior distribution in a Bayesian framework, although ultimately a mathematical representation cannot fully capture the subjective meaning of trial results.

- *Non-inferiority creep:* As noted above and discussed further below, interval D would meet the technical requirement for non-inferiority. However, because the point estimate is in the negative direction, a non-inferiority decision based on interval D would allow the use of a new therapy even though the outcome in the trial was modestly worse than the comparator treatment. If a sequence of non-inferiority trials is performed, each comparing a new agent to the non-inferior agent from a previous trial, then it is possible that each treatment will be slightly worse than its predecessor. This leads to the possibility that a sequence of non-inferiority trials would lead to the use of significantly inferior treatments. Remedies for this *non-inferiority creep* are discussed in the following section.

- *Other endpoints:* Although the trial design is almost always structured around a single primary endpoint, its results are ultimately interpreted within the context of treatment effects on several endpoints. Thus, the ultimate decision about the utility of a new therapy is not adequately captured by the conclusions on a single parameter.

We now examine some of these issues in greater detail through a discussion of the difference between non-inferiority and approximate equivalence trials.

3.2.3 Sample size, non-inferiority creep, and the equivalence decision

As described above, a non-inferiority trial must choose between inferiority (reject $\theta \geq \theta_0$) and non-inferiority (reject $\theta \leq \theta_-$). In essence, a non-inferiority trial is structured just like a superiority trial, but the hypotheses are reflected around $\theta = \theta_0$. A superiority design decides for non-superior when the confidence interval excludes $\theta \geq \theta_+$. Similarly, non-inferiority is decided when the confidence interval excludes $\theta \leq \theta_-$. In Figure 3, confidence interval A sits right at the threshold for finding non-inferiority: any shift to the left leads to an inferiority decision and any shift to the right leads to a non-inferiority decision. As described above, it is possible to decide non-inferiority when the point estimate (center of the confidence interval) is in the negative direction, which can lead to non-inferiority creep.

In order to avoid non-inferiority creep, it makes sense to require that the trial show a positive point estimate in addition to rejecting important inferiority. We refer to such a design as an "approximate equivalence" trial. Confidence interval B in Figure 3 illustrates a confidence interval that sits right at the threshold for an approximate equivalence decision. Note that the lower limit of this confidence interval is well above θ_-. Strictly speaking, given the non-inferiority limit of $\theta = \theta_-$, an approximate equivalence trial could get away with a smaller sample size than the analogous non-inferiority trial because we would only need to require that the lower limit of the confidence interval exclude $\theta \leq \theta_-$ when the point estimate was in a positive direction (Figure 3, confidence interval C). In practice we would usually choose a sample size that is intermediate to that illustrated by intervals B and C.

In the daptomycin trial the non-inferiority margin was set at $\theta_- = -0.2$, however the sample size was such that the final confidence interval would have a width of 0.25 instead of 0.2. Thus, the daptomycin trial allowed the possibility that the final confidence interval could include both θ_- and θ_0 (or θ_0 and θ_+). According to the authors' report, non-inferiority would be decided if the confidence interval was above -0.2 and if it included 0. Figure 3 shows the confidence interval at the non-inferiority threshold (interval D) along with the interval that was observed upon trial completion (interval E). The daptomycin trial provides an example in which the trial results (interval E) would be considered sufficient to conclude *approximate equivalence*, but the trial

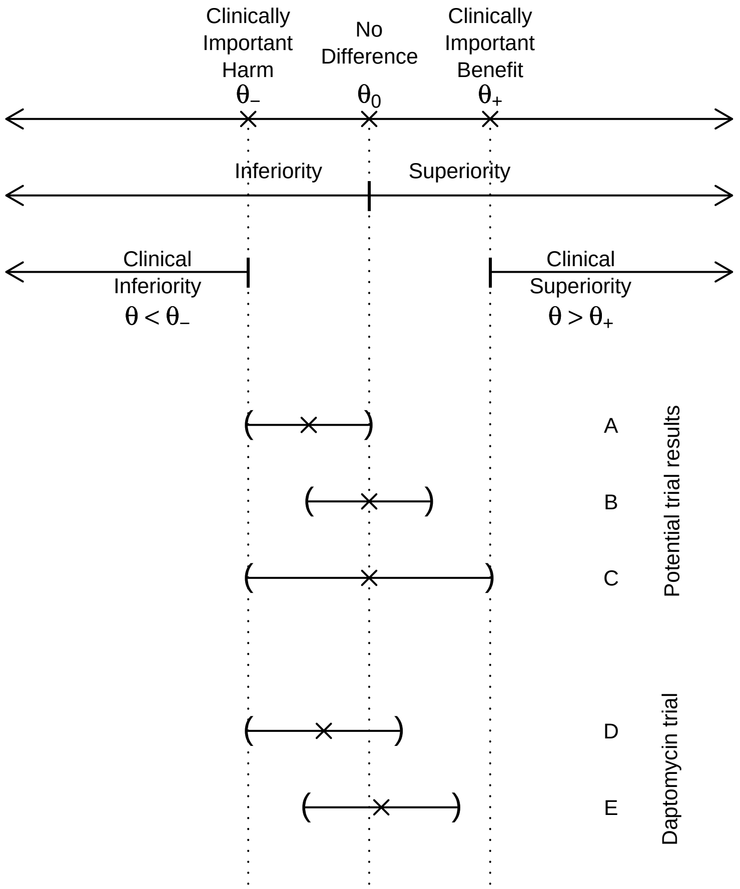

Figure 3: Potential criteria for a non-inferiority decision

was designed for *non-inferiority* and would have allowed a non-inferiority decision even though the point estimate indicated that daptomycin had a somewhat lower success rate than standard therapy (interval D).

The choice of objective and phase for a trial informs all other elements of the trial design including specification of the study population, selection of treatments and control group, choice of endpoints, study duration, and sample size, to name a few. Thus, the choices discussed in Part 1 of this book are often the first steps in trial design because they specify the key scientific, clinical, and ethical issues that underpin the statistical design of the trial. In the next part we describe the transition from this scientific setting to the foundations of the statistical design.