PhD degree in Foundations of the Life Sciences and their Ethical Consequences

European School of Molecular Medicine (SEMM) and University of Milan

Faculty of Medicine

Settore disciplinare: FIL-02

Statistics in Clinical Trials: Out of Condition

Some Problems of Unconditional Inference at the Crossroads of Methodology and Ethics

Cecilia Nardini

IFOM-IEO Campus, Milan

Matricola n. 821973

Supervisor: Prof. Giovanni Boniolo

IFOM-IEO Campus, Milan

External Co-supervisor: Prof. Jan Sprenger

TiLPS, Tilburg (NL)

Anno accademico 2011-2012

A special thanks goes to:
My family and my indispensable sisters;
My supervisors for their patience;
Folsatec students and alumni: who have been to me invaluable colleagues, friends, mentors;
and A., for having eventually permitted me to finish the PhD.

A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly concludes he has seen a mule
(Senn, 1997)

We balance probabilities and choose the most likely. It is the scientific use of the imagination
(Sherlock Holmes, in The Hound of the Baskervilles*, 1902)*

Contents

List of abbreviations

RCT Randomized Controlled Trial

NP Neyman-Pearson

LP Likelihood Principle

SRP Stopping Rule Principle

FDA Food and Drug Administration

EMA European Medicines Agency

WMA World Medical Association

NI Non-Inferiority

List of Figures

Abstract

Randomized controlled trials are experiments for the evaluation of a new treatment option, currently representing the “gold standard” in health care assessment. Clinical trials fulfill a double role of evidence production and of regulatory oversight in sanctioning new drugs’ approval into the drug market. For this reason trials are large and tightly regulated enterprises that have to comply with ethical requirements while at the same time maintaining high epistemic standards, in a balance that becomes increasingly difficult to strike as research questions become more and more sophisticated.

The statistical framework adopted for designing and analysing trials represents a relevant part of this architecture. Statististical methodology influences such aspects as which inferences are licensed on the basis of data and what is the degree of support granted to an hypothesis. Thus, statistics plays a fundamental role as a gatekeeper both in warranting the ethical permissibility of a trial, and in licensing conclusions about the most effective treatment.

Certain widely-accepted statistical principles have an impact on the way results from medical studies are evaluated. One such principle is *conditioning*, i.e. the possibility to incorporate an assessment of strength of evidence in inferential statements of confidence. Currently, conditioning is not part of the statistical method in use, although it is upheld by alternative statistical paradigms such as the Bayesian. In my thesis I analyze the impact of conditioning upon the ethical, epistemic and regulatory facets of trials and I suggest the possibility of incorporating conditioning within the current statistical paradigm of clinical research.

Introduction

In a 2007 Compass paper, John Worrall writes that “The logic of evidence as applied to medicine is [...] a new area where philosophers of science could have an enormous impact - both intellectual and (very unusually) practical” (2007a, p. 981). Indeed, this field provides a unique perspective to the philosopher of science, since the epistemic objectives in medicine are competing and sometimes conflicting with the ethical obligations that go with any form of experimentation with human subjects.

Randomized controlled trials (RCTs) are tightly controlled experiments for the evaluation of a new treatment option. Since its inception in the 1950s, the science of clinical trials has risen to a very sophisticated discipline. Nowadays clinical trials are large and tightly regulated enterprises that have to comply with ethical requirements while at the same time maintaining high epistemic standards, in a balance that becomes increasingly difficult as the research questions become more and more sophisticated. The statistical design currently in use for RCTs represents a relevant part of this architecture. Statistics plays a fundamental role as a gatekeeper both in warranting the ethical permissibility of a trial, and in licensing conclusions about the most effective treatment.

The statististical methodology adopted influences such aspects as which inferences are licensed on the basis of data or what is the degree of support granted to an hypothesis. Statistical approaches are generally distinguished along the fault line of a division between two major statistical schools. *Frequentist* statistics is based on an interpretation of probabilities as relative frequencies. Therefore frequentist inference relies on ensuring that the inferential procedure produces reliable conclusions in the sense that, if the test is repeated many times, the procedure will yield the correct conclusion most of the times. Frequentist inference is therefore focused on the control of the experimental design, in order to control the rate of erroneous conclusions. In virtue of the impartiality that this perspective appears especially suitted to warrant, frequentism is currently the framework that dominates the statistical treatment of trials. *Bayesian* inference is instead based on the idea that it is possible to establish probability as a degree of belief in a hypothesis on the basis of the available evidence. What

is relevant to Bayesian inference is therefore not the way the experiment is designed but only the prior belief in the hypothesis, if available, and the strength of the available evidence.

In my work I focus on the practical import that foundational questions about statistical inference have in application to evidence-generation through RCTs. A illustrative example is provided by the practice of monitoring of clinical trials. Monitoring refers to the analysis of trial data carried out as they accumulate, with the option of stopping the trial before the planned end, if the data indicate a conclusion with sufficient strength. This differs from vigilance over the manifestation of unexpected side effects: While the latter is an undispensible ethical safeguard, due to in-trial patients, the former is a strategy for achieving the epistemic objective of the trial with less resource expenditure. Monitoring clinical studies for early signs of benefit is a strategy that is becoming increasingly common. It meets a need, that has become pressing in the last decade, for an improvement in the cost-effectiveness of treatment evaluation through clinical trials. Since the trial can be stopped as soon as a convincing result is obtained, the valuable epistemic aim of the trial can be attained with less resource expenditure. In the meaning used through this work, monitoring revolves around a specific epistemic issue: At which point is the accumulating evidence sufficient to warrant a conclusion? The import of statistics, and the statistical method, to this question, is evident. What is possibly less evident is that different statistical schools may provide different answers to it.

As outlined above, Bayesian methods focus on assessing the plausibility of a specific conclusion, on the basis of available evidence and prior expectations, whereas classical methods focus on ensuring the general validity of the results produced with a specific method. Within a Bayesian framework, a result borrows its reliability directly from the strength of the observed evidence and, possibly, of prior information about it. For this reason, Bayesian inference is widely acknoledge to be more suited than the frequentist to appraisal of accumulating evidence from an ongoing trial. However, the greater epistemic efficiency of Bayesian methods rests on the incorporation of prior evidence in the inference. This has a downside, namely, that this evidence can be incomplete or misleading, as it has often been the case in the history of medicine; furthermore, the possibility of including external information can open the way to the most obvious perturbation of trial results, driven by financial interests.

Presently, the debate about which statistical framework fits the epistemic, regulative and ethical aims of trials best is deadlocked over the use of priors. While it is recognized that

Bayesian inference, through use of priors, can provide a more efficient and informative evaluation of trials'results, the concern remains that Bayesian methods may make medical research more bias-prone.

My work is motivated by the observation that the controversy over prior specification will not likely be overcome in a short time. Therefore, I take up a pragmatic perspective: I grant that conservatism towards frequentist inference reflects the conviction, among the main stakeholders, that this framework is overall adequate in serving the purpose, albeit in a suboptimal way. Taking this as my starting point, I find it fruitful to take a fresh look at the issue by proposing a different classification of statistical approaches, namely, the distinction between *conditional* and *unconditional* ones. In conditional approaches to inference, the conclusiveness of the inference depends on the strength of the observed evidence. This is not the case with unconditional approaches. The classification of inferential approaches as conditional or unconditional cuts across the fault line dividing the two schools. Indeed, while Bayesian approaches are all conditional and the classical frequentist approach is instead unconditional, it is nonetheless possible to develop a *conditional frequentist* approach to inference. Within this latter approach appraisal of results is still based upon measures of error rather than on posterior credibility. However, the error assessment is post-data rather than pre-data, as in the classical unconditional framework. The change in perspective that I propose offers remarkable advantages. As a first thing, it affords a way out of the deadlock because it identifies a strength of the Bayesian paradigm other than use of priors, namely the possibility ot condition on the observed strength of evidence. Furthermore it suggests how, by borrowing conditioning into an otherwise frequentist framework, it is possible to overcome some known limitations of classical frequentist inference, particularly in the context of trial monitoring. Ultimately, such approach affords a concrete opportunity to improve on the present framework while leaving the controversial issue of the use of priors out of the picture.

The thesis will be structured as follows. The first chapter will serve as an introduction to the problem and to the relevant concepts that I will use through the work. Afterwards, in the three core chapters, I will examine each of the three roles of statistical inference in clinical trials in turn. In Chapter II I will focus on the epistemic role of statistical inference, i.e. on the construction of medical knowledge from RCTs through the inferential procedure. In this chapter I will be focusing on the perspective of the medical consumers of statistics, physicians

and medical professionals, and their inferential needs. In Chapter III I will confront instead the ethical role of the inferential framework, i.e. its role of warranting the ethical acceptability of a trial. In the context of clinical trials, statistics represents the tip of the ethical balance, and this is all the more evident in relation to trial monitoring. Finally, in Chapter IV, I turn to the analysis of the regulatory role performed by statistical inference, in light of the fact that RCTs currently constitute the main gateway to market approval for medical interventions in the Western world. Hence, the evolution of the statistical framework for design and analysis of trials is deeply intertwined with the transformation of the needs and scope of the regulatory system. An outlook over the possible routes of this transformation will conclude the chapter and my work.

Chapter I:

Randomized Controlled Trials and Statistics

Randomized Controlled Trials (RCTs) represent the gold standard of evidence in medical research. RCTs are typically used to assess whether a newly proposed drug or device is more efficient than the established standard treatment for a certain condition. In the typical procedure for the evaluation of a treatment option, patients are assigned to either the experimental or the control group. Patients in the first group receive the new treatment that is being tested, while patients in the control group are instead administered the treatment that constitutes the benchmark for the evaluation of the newly proposed alternative, which can either be a placebo or the standard therapy for the condition under study. In order to avoid bias, allocation to the two groups is random and the investigators and physicians are generally blinded to group allocation. According to the health issue under study, different end-points will be taken into account. For anti-cancer interventions, for instance, the performance of the new treatment will most typically be evaluated with respect to time to recurrence and survival time. The experimental data of the RCT consist in the record of the value of the end-points of interest for all the patients involved in the trial. From this it is possible to estimate the action of the experimental and the standard treatment with respect to the end-point of interest – for instance, the average time-to-recurrence for patient on one or the other drug. The details of the analysis will actually vary depending on the details of the intervention under study and the nature of the end-point of interest –for instance, different kind of analysis are needed for data about binary events (e.g. no death vs.death or recurrence vs. no recurrence) and for time-to-event studies (e.g. recurrence-free survival). The trial is positive for the new treatment if there is a difference in event rates that favors the new treatment's arm. Clearly, not *any* difference will do. Events in one group may be higher than in the other just due to chance and we know that small differences, or fluctuations, are extremely common. The role of statistics in clinical trials is precisely that of separating minor trends that could depend on fluctuations from the conspicuous trends that point to genuine differences in outcome.

Statistics, however, is a full fledged scientific discipline rather than a mere set of scientific tools. Philosophical schools of statistics exist, which differ from one another by their respective take on fundamental principles of inductive inference. A outlook on the epistemology of clinical trials would miss an important aspect if it were to ignore that adherence to a particular statistical school, and thereby use of particular statistical methodologies in place of others, carries an epistemological commitment with it. As Mayo and Kruse (2001) observe, "Philosophers who appeal to principles and methods from statistical theory in tackling problems in philosophy of science need to recognize the consequences of the statistical theory they endorse". This work is concerned with the philosophical and ethical consequences that descend from adoption of a particular statistical methodology for the design and analysis of medical studies. Clinical trials are currently designed and analyzed under a classic interpretation of probability, endorsed by a school called frequentism. This is partly due to historical reasons, which I will examine in the first upcoming section. Afterwards I will introduce the school of statistics which is currently endorsed in clinical research together with its main adversary, the Bayesian school, in section one. In the third section I will describe in detail the methodology of the statistical treatment of trials. In section four and five, finally, I will introduce a situation where the conflict between the two schools has clear reverberations onto the practice. This is the case of monitoring and early stopping of trials. As I will discuss, philosophical differences between the two schools lead to conflicting methodological recommendations in this setting. Analysis of this conflict will conclude the first part of my work.

1.1 Setting the stage: The history and role of RCTs

Today, the RCT is above all the most praised means of evidence generation in medicine. For a large part of the history of modern medicine, cultural tradition, personal experience and expert judgment remained the only informing principles for clinical decision-making. The introduction of the controlled trial represented a revolution in medicine, providing what appeared to be a objective means of gathering evidence about the effect of a new treatment.

The birth of the RCT is generally marked with the trial of the antibiotics streptomycin in the treatment of pulmonary tuberculosis, carried out at the British Medical Research Council in 1947. What is so notable about this study is that it was the first trial to feature randomization

of patients between the treatment and the control arm. The trial was designed by epidemiologist and statistician Austin Bradford Hill, widely recognized as the founding father of modern medical statistics. Hill's work –especially the series of articles that appeared in 1937 in the *Lancet* on the topic of medical statistics– pioneered the application of the statistical methodology of the test of significance to the context of biomedical research, just a few years after Ronald Fisher's *Design of Experiments* (1935) had shaped the technique. This circumstance led to a shared perception that Hill simply adapted Fisher's method, originally developed for agrarian experiments, to trials on medical patients. This view is, however, quite reductive. Hill, an economist by education, came in contact with leading figures of statistical theory of his times such as Karl Pearson (Farewell and Johnson, 2010). Even though Hill did certainly corresponded with Fisher (Armitage, 2003), his ideas about the importance of randomization and control were oriented by practice. As defended also recently by Ian Chalmers (2009), it was not the theoretical import of randomization as defended by Fisher but rather the concept of a unbiased test, more fundamental and less technical, which was central to establishing these methodological principles in medicine. This suggests that the methodology of significance testing found its way in clinical trials not due to subscription of medical statisticians to a certain school of inference, but rather in virtue of certain perceived advantages of this framework. Among those, the most important was its being less discretionary than other means of appraisal of evidence at the time: This is because the interpretation of the experiment could rely on a set of fixed rules (Teira, 2011).

In the present days, the movement known as Evidence-Based Medicine (EBM) is playing a fundamental role in forwarding the epistemic indispensibility of RCTs. A methodological movement famously linked to the names of Archibald Cochrane and David Sackett, MDs, EBM was born in the latest decades of the twentieth century with the purpose to orient clinical-decision making. EBM proponents maintain that clinical decision-making has to be based on the best quality of evidence and they propose hierarchies that rank the methodological ability of different study designs to yield answers to certain types of research questions. For questions concerning the safety or effectiveness of an intervention, well-conducted RCTs and systematic reviews thereof are regarded as the gold standard of evidence. Actually, EBM is much more than a set of best practice standards. As Djulbegovic et al. (2009) put it, "[b]ecause EBM proposes a specific relationship between theory, evidence, and knowledge,

its theoretical basis can be understood as an epistemological system".

EBM's rejection of conventional wisdom in favor of evidence gathered through well-conducted RCTs is grounded on RCTs' superior epistemic virtue. According to Howick (2011), RCTs are valuable in that they provide a surrogate access to the counterfactual of interest in assessing the effectiveness of a medical treatment, which is: How would have the patient fared had she not received the treatment she got? Being it impossible to answer this question directly with regards to a single patient, researchers administer the treatment to one group of patients and they use a second group of patients to represent the second term in the counterfactual. Randomization is generally considered an epistemic device that makes the comparison meaningful, by ruling out relevant differences in the composition of the two groups. Whether randomization is or is not able to warrant this aspect of the epistemic adequacy of RCTs is subject of considerable philosophical controversy (Urbach, 1985; Kadane and Seidenfeld, 1990; Papineau, 1994; Worrall, 2007b). What is of interest to this work is however not so much the role played by randomization but rather a different aspect of the frequentist methodology of RCTs –namely, the epistemic value assigned to fixed, pre-experimental error rates in the test of hypotheses.

This connects with the foundational distinction between the two grand schools of statistics, the frequentist and the Bayesian.

1.2 Schools of statistics

The evaluation of effectiveness of a new treatment based on the result of a trial on a small number of patients is an instance of *probabilistic*, or *statistical*, inference. We observe the comparative effect of a new treatment in a sample population of patients, and from this we have to draw a conclusion about the comparative effectiveness of the treatment in the hypothetical population of all patients that may be treated with it, presently or in the future.

There are two main schools in statistics, which differ in the way they approach this problem. *Frequentist* statistics is based on an interpretation of probability as relative frequency of an event in a long series of repetitions. Therefore, frequentist inference relies on ensuring that the inferential procedure produces reliable conclusions in the sense that, if the test is repeated many times, the procedure will yield the correct conclusion most of the times. The

Bayesian school, instead, licenses interpretation of probability as a degree of belief in an hypothesis about the state of the world. Bayesian inference, therefore, consists in establishing the plausibility of a hypothesis in light of available evidence.

The main difference between the two schools lies in the fact that for a frequentist uncertainty applies to events, while for a Bayesian it applies to hypotheses. To make things clearer, it is helpful to go back to the case of clinical trials, where we observe a certain difference in patient performance between the two groups and we want to infer from this the tenability of a hypothesis about the treatment's effect. For a Bayesian, it is legitimate to think that different hypotheses derive a varying degree of support from the same data. Therefore the Bayesian will simply decide which value of the effect is more tenble in light of the observed data. For an adherent of the frequentist school, instead, it is all the observable values of the difference that have a different probability under a fixed hypothesis. Hence, the frequentist will set out from a certain hypothesis about the value of the treatment's effect, she will derive from it predictions about the probability of the outcome that was actually observed, and on the basis of this she will decide about the plausibility of the starting hypothesis.

Of the two schools, the more ancient is actually the Bayesian. The first comprehensive attempt at formalizing a solution to the problem of probabilistic inference was done by Thomas Bayes, a protestant minister and scholar, in the late 1700. His "Essay Towards Solving a Problem in the Doctrine of Chances" contains a foundational mathematical result, the theorem that bears his name. This theorem applies to events that are related. If a and b are two events, $P(a|b)$ is the *conditional* probability of a occurring, given that b is the case. If a and b are not independent, $P(a|b)$ will be different from $P(a)$: Knowledge about b will make a difference about the probability assigned to the occurrence of a. Bayes' theorem relates the two inverse probabilities $P(a|b)$ and $P(b|a)$, in order to learn about the one from knowledge of the other. To make an example, let us suppose that a is the event of someone having an appendicitis while b is the event of someone having a fever. From etiological and epidemiological data, we generally do know what are the chances of b occurring when a is the case. For instance, say we know that 56% of patients that have appendicitis develop a fever on top of abdominal pain[†]. Then we learn that b is the case: A patient with abdominal pain and a fever shows up in the emergency ward. What probability should we now assign to the patient having appendicitis

[†] I adapted this example from Westover et al. (2011)

–$P(a|b)$–, in light of the circumstances? Clearly, this probability will be higher than the general prevalence of appendicitis in the population, $P(a)$: After all, we know the patient is ill, so we think it more likely that she has an appendicitis as compared to her apparently healthy relative standing by. But how much higher? Bayes' theorem provides a solution to this problem.

$$P(a|b) = \frac{P(b|a)}{P(b)} \cdot P(a) \tag{1.1}$$

The *posterior* probability of appendicitis, $P(a|b)$, is obtained from the general prevalence of appendicitis $P(a)$ through a procedure known as *conditional updating* which involves the conditional probability $P(b|a)$ that a patient with appendicitis has also a fever, normalized by the chances $P(b)$ of observing fever in the general population (fever may be due to other causes, such as for instance a flu).

There is no question, between the two schools, about the validity of Bayes' Theorem (1.1). Both schools accept this result, likewise the basic rules of probability from which it descends, as uncontroversial. The point of divergence is however the *interpretation* that should be given to it. In the example used for introducing Bayes' theorem, all the quantities involved can be known from etiological –$P(b|a)$– or epidemiological data –$P(b)$ and $P(a)$. Determining $P(a|b)$ amounts then to a simple application of a mathematical procedure. However, Bayesians extend the scope of application of Bayes' Theorem, as they regard it as providing the fundamental rule for reasoning under uncertainty. For let us say that we are no longer concerned with relating the probability of two *events* such as a and b, but we want to relate the probability of making a certain *observation* e with the plausibility of a certain *hypothesis* H, when we know how likely e is produced under H. For instance, suppose a physician visits a patient who appears to be extremely ill. She prescribes some exams as she is in doubt that her patient may be suffering from a severe disease – this is hypothesis H. The day after, before the results arrive, she sees her patient again and this time the patient is up an feeling well – event e. Now, the physician knows that such a fast recovery is unlikely if the patient has the disease – in other words, $P(e|H)$ is very low. The physician may now attempt to use Bayes' theorem to update her belief that the patient has the disease with the new evidence e.

$$P(H|e) = \frac{P(e|H)}{P(e)} \cdot P(H) \tag{1.2}$$

Notice, however, the difference with (1.1). In the first example we introduced, all quantities appearing in the formula were known and objective quantities. In this case, instead, in order to determine $P(H|e)$ the physician would have to assign a value to $P(H)$, i.e. to her subjective belief that her original diagnosis was correct. This *prior* probability is at the heart of all the dispute. Indeed, critics of the Bayesian view point out that $P(H)$ cannot be established with objectivity: Since it reflects the subjective state of belief of the agent prior to observing e, factoring it in contaminates the inference to $P(H|e)$ with non-objective, epistemically suspect elements. The founding father of frequency statistics, Ronald Fisher, was motivated in his work by the desire to rid statistical inference from this perceived subjective interference and put it on a completely objective footing. Whether he actually succeeded in this endeavour is a question that will be examined later on.

For the moment, the contrasting interpretations of (1.2) serve to exemplify the two contrasting approaches to inference. For adherents to a subjectivist, or Bayesian, view, Bayes' theorem (1.2) dictates how a rational agent should update her state of belief on the basis of upcoming evidence (Howson and Urbach, 2006). Within a objectivist or frequentist interpretation, however, uncertainty applies to the occurrence of events, but hypotheses can only be either true or false. The idea that a hypothesis can have a changing degree of support does not have a legitimate interpretation. Therefore a frequentist would question the very interpretation of Bayes' theorem as a guide to updating an internal state of belief about credibility of a hypothesis. For a frequency-minded statistician, the only legitimate interpretation of Bayes' theorem is in terms of probabilities of related events, as in (1.1).

In most situations of interest, statistical inference involves learning about an unknown parameter θ, which determines the distribution of an observable variable X, from observation of data $X = x$. For a particular value of θ, the probability of observing the specific outcome x is called the *likelihood* of x under θ:

$$\ell(x, \theta) = P(X = x|\theta) \tag{1.3}$$

Since it summarizes the probabilistic relationship between the observed variable and the unknown parameter, the likelihood has a crucial role in both account of inference, the frequentist and the Bayesian. However, the interpretation of this quantity is different in the two cases. In

frequentist methods, (1.3) is regarded as a function of X for varying θ. This likelihood function is then used to calculate how likely different observations are under a fixed hypothesis. In the Bayesian interpretation, instead, (1.3) is interpreted as a function of θ for a given observation x. The likelihood function serves, in this framework, to measure the support lent by the observation to possible values of the parameter. It is important to note that, in the frequentist use, the function $\ell_\theta(X)$ is construed before performing the experiment and it is then used to infer, in an indirect way, the support lent by data to the postulated value θ_0. In Bayesian inference, instead, the likelihood function $\ell_x(\theta)$ is construed after $X = x$ has been observed and can then be interpreted as a comprehensive summary of the evidence coming from the experiment about the unknown parameter. Whether this interpretation is valid only within a Bayesian framework, or if it can be valid (and at what price) also within a frequentist framework, is the subject of the controversial issue around the likelihood principle. This problem and its implications will be confronted later on.

Another way to see the difference between the two methodologies is to identify two kinds of statistical problems: Problems of *testing* concern the validity of a statistical inference, while problems of *estimation* address the accuracy of the result of the inference. In the case of clinical trials, we distinguish between (a) the question of whether the trial result points to a genuine difference in performance between the two treatments and (b) the question of whether the estimate that results from the trial is close to the actual value of the difference in effect. Both questions are relevant to the issue of statistical inference. However, we may regard the two schools of inference as being more inclined towards meeting one or the other. Testing is more natural in frequentist statistics and is well established in the methods ascribed to Fisher (1935a, 1956) and Neyman and Pearson (1933). Bayesian methodology, on the other hand, is better suited to problems of estimation, because the inferential step incorporates an evaluation of which hypotheses about the value of the parameter are best supported by available evidence[†]. As I will expound in section 3, testing of hypotheses features prominently among the statistical procedures used for analysing the results of medical trials.

Recently Mayo (1996) and Mayo and Spanos (2006) have proposed a considered reconstruction and defense of frequentist ideas about testing, insisting in particular on the falsificationist rationale underlying it. The view expressed by Mayo and Spanos is that the frequentist

[†]See however Gelman and Shalizi (2011) for a reconstruction and a thought-provoking criticism of this traditional account

position, in a nutshell, consists in using probability "to assess the reliability of a test procedure to assess and control the frequency of errors in some (actual or hypothetical) series of application" (Mayo and Spanos, 2006, p. 324). This approach, termed *error-statistics*, provides a philosophical interpretation of the practice of frequentist testing, whereby credibility of an hypothesis depends on the stringency of the conditions under which it was tested. There is a large debate in philosophy of science regarding this account, and in particular the question of whether it captures the logic underlying scientific inference. Without engaging directly with this wider philosophical debate, I believe that there is something to say about this question while remaining in the more restricted scope of health-care assessment. Clinical trials represent a practical field of application of ideas about statistical inference where the import of such foundational issues can be substantial. Inference from a clinical trial to a statement about a new treatment's effectiveness is, on one hand, just an instance of statistical inference. On the other hand, though, this inference is embedded in a set of unique epistemic, ethical and social needs. This poses specific constraints to the choice of a statistical methodology. In order to proceed with this analysis, as a first thing, it is necessary to see more in detail how inference in clinical trials is performed.

1.3 Hypothesis test

As discussed in the introduction, testing of hypotheses is the technique typically used for trials. A textbook for students of biomedical sciences reads: "Most investigators run tests on the variables that are used to evaluate the success of medical treatments and report P values. For example, if a new treatment has been compared to a standard treatment, the readers of the results will want to know what the chance of making a mistake is if the new treatment is adopted. [...] In general, it is sensible to run tests when investigators have a rationale for the size of the P value that would lead them to reject the null hypothesis and when users of the results want to know their chance of making a type I error." (Dunn and Clark, 2009, p. 113)

I will discuss mainly two kind of testing scenarios. One, the *simple vs. simple* hypothesis test, is useful in situations where one wants to know which of two point-like alternatives is the case. For instance, we are testing a standard normal distribution with known variance $X \sim \mathcal{N}(\theta, \sigma)$ and we are interested in knowing which of two values θ_0 and θ_1 best approximates

the mean θ of the distribution. Therefore we test the two alternatives

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

Contrary to what may be thought, the simple hypotheses scenario is not particularly relevant in the context of clinical trials and the reason I will discuss it is merely its pedagogical value. In testing hypotheses about a new treatment's effectiveness, investigators are almost never interested in directly comparing two hypotheses about the treatment effectiveness. Since clinical trials are mainly conceived of as *scientific experiments*, the underlying statistical test is conducted by testing a null hypothesis of no effect –the usual $\theta = \theta_0$– not against a precise alternative, but against a unspecified alternative:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

or against a range of reasonable possibilities for the new treatment's effectiveness:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \in \Theta_1, \theta_0 \notin \Theta_1$$

The rationale behind the choice of the *simple vs. composite* hypothesis test is clear: we do not want the result of the test to be influenced by the specification of an alternative, since the test is set up precisely because the information about the treatment effect is lacking. The fact that clinical trials are mainly conducted as tests of a precise hypothesis against a composite alternatives, however, raises a number of issues which essentially form the subject of this chapter.

Trials are generally designed as one-sided hypothesis tests, meaning that the range of parameter values θ contemplated by the alternative H_1 lies wholly on one side of θ_0. This is adequate in most cases, since the trial comes typically at late stage in the research process. Most commonly earlier investigations suffice to test the expectation that the newly proposed treatment is at least as good as the standard. The use of one-sided tests has however been criticized on the basis that, since the one-sided test discounts the error in one direction, there is no error control in case an effect in an unexpected direction is found. Moyé and Tita (2002) note that "A one-tailed test designed exclusively to find benefit does not permit the assessment

of the role of sampling error in producing harm, a dangerous omission for a profession whose fundamental tenet is to first do no harm." For example in the case a new promising treatment turns out to be less efficient than the standard contrary to expectations, there is no warrant on the error associated to the conclusion and therefore a ethically problematic repetition of the trial may prove necessary. However, designing all clinical tests as two-sided is also an ethically problematic option, given that two-sided tests require on average a larger number of subjects to achieve equal confidence in the result. Owen (2007) for instance argues that "two-sided tests may expose research subjects to unnecessary risk, and are therefore unethical".

The data from the study can be represented with the random variable X, which distribution depends in some way on the unknown difference in effect. Two commonly used measurements for assessing the effectiveness of a treatment are the *odds ratio* and the *hazard ratio*. The odds ratio *OR* directly compares the probability of an event occurring under two different interventions. For instance, the event of interest can be death or disease recurrence. Defining with p_1 and p_2 its probability under the experimental treatment and the control, respectively, the odds ratio is:

$$\text{OR} = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} \tag{1.4}$$

an OR smaller than one implies a lower event rate under the new intervention, and therefore it favors the new.

The hazard ratio HR, on the other hand, is the quantity of choice for a different kind of statistical analysis, focusing on *survival* data. In building this variable, we are not interested in the probability of the disease event occurring but rather in the temporal rate of its occurrence. Does the new drug prolong survival, by making death or recurrence on average slower? In order to answer this question, one has to compare the hazard rate $h(t)$ –the istantaneous chance of dying or experiencing recurrence, provided survival until t– under the two interventions. The hazard ratio

$$\text{HR} = \frac{h_1(t)}{h_2(t)} \tag{1.5}$$

is generally assumed to be constant in time over the treatment period, in order to make inferences based on this variable.

The distribution of the data is assumed to have a parametric form, determined by the statistical model used and by the unknown value of the parameter θ. The use of a normal

approximation $X \sim \mathcal{N}(\theta, \sigma)$, where σ is a known or estimated variance, is often possible as a simplifying assumption. The study data are summarized in a *test statistics*,

$$Z(X) = (\overline{X} - \theta)\sigma \tag{1.6}$$

$\overline{X}$ is the sample mean and σ is the variance. The use of the test statistics Z provides a way to order the data and make them uniform. If $X \sim \mathcal{N}(\theta, \sigma^2)$, then Z is distributed according to $Z \sim \mathcal{N}(0, 1)$. Therefore, the probability of observing a value of X larger than x is given by

$$\begin{aligned} P(X \geq x) &= P\left(\frac{X - \theta_0}{\sigma} \geq \frac{x - \theta_0}{\sigma}\right) \\ &= P\left(Z \geq \frac{x - \theta_0}{\sigma}\right) = \Phi_0(-z(x)) \end{aligned} \tag{1.7}$$

In this expression, Φ_0 is the cumulative distribution function of Z under H_0. $\Phi_0(-z(x))$ sums up the values of the probability for all values of Z below $-z(x)$. This means that, for an observed value x, expression (1.7) measures the tail area at the right of $z(x)$, equal to the area at the left of $-z(x)$ due to the symmetry of the distribution, as in figure (1.a).

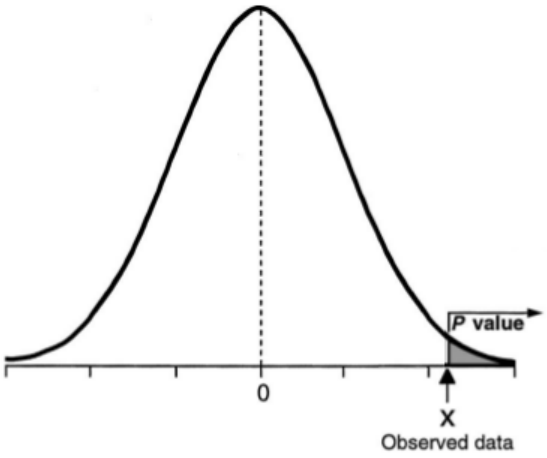

Figure 1.a: The *p*-value for observation x under a normal distribution. Image adapted from Goodman (1999a)

There exist different schools for testing, subscribing to either one or the other philosophy of inference. This means that the various schools differ not only in the way the test of the hypothesis is designed and conducted but most importantly in the interpretation to be given to the test result. Testing is more natural in frequentist statistics and is well established in the methods ascribed to Fisher (1935a, 1956) and Neyman and Pearson (1933). However,

Bayesian techniques for testing have been developed and I will describe them in due course.

1.3.1 Fisher

Ronald Fisher is the celebrated father of the methodology of significance testing, which he viewed as a key to resolve the long debated problem of induction (Fisher, 1935b). The test he developed is based on the idea of assessing the strength with which data speak against the null hypothesis. The crucial quantity in Fisher's test of significance is the p value

$$p = P(z(X) > z(x)|H_0) = \Phi_0(-z(x)) \tag{1.8}$$

In this formula, Φ_0 denotes the cumulative distribution function of Z under θ_0 as described in the previous section. The p-value measures the probability mass concentrated on values as or more extreme than the one observed, under the null distribution. In other words, the p value represents the probability, under the null hypothesis, of an observation at least as extreme as the one obtained.

Fisher's test makes no reference to an alternative hypothesis, because the underlying idea is to "give the facts a chance of disproving the null hypothesis" (1966, p.16). Fisher thought that hypotheses can only be disproved but that there is no way for inductively corroborating one particular hypothesis. In absence of other information, one has to turn to the probability distribution of the variable Z under the null hypothesis, the likelihood $\ell_0(Z) = \mathcal{N}(0, 1)$. Large values of Z are unlikely under the null hypothesis and, through (1.8), give rise to low p-values. The logic underlying the test of significance is the following, in Fisher's words: "There is the logical disjunction: Either an intrinsically improbable event will occur, or, the prediction will not be verified"(1959, p.43), with the latter occurrence implying that the null is not true. This *disjunctive argument* has been examined in depth by Spielman (1974). The philosophical basis of a Fisherian test is that of a proof by contradiction. The basis for suspecting a contradiction is actually observing data that are highly improbable under the null hypothesis. However, if the null hypothesis is not rejected, the best that one can say is that the data are consistent with it: Not rejecting does not entail any kind of corroboration of the null (Christensen, 2005).

Fisher's school of testing subscribes to a frequentist interpretation of probability, even though it is not ultimately an error-statistical account, as I will discuss briefly. In fact, Fisher

elaborated the methodology of significance test in opposition to the then prevalent school of inverse probability, based on Bayesian principles. According to Fisher, the choice of the significance level, i.e. the decision of which *p*-values are extreme enough to justify rejection of the null, should be a contextual one. This goes in contrast with the idea, proper of error statistical thinking, that we should judge the soundness of the probabilistic inference by the stringency of the error rates proper of the method. Fisher remarked "A man who rejects a hypothesis [...] when the significance is 1% or higher, will certainly be mistaken in not more than 1% such decision [...] However, the calculation is absurdly academic, for in fact no scientific worker has a fixed significance level at which [...] in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (1973, p. 44–45) In other words, Fisher's test focuses not so much on the probability of rejecting the null hypothesis but on the (a priori) probability of observing rare outcomes. In any case, the observed *p*-value should be reported since it represents an index of the strength of the evidence against the null hypothesis.

1.3.2 Neyman-Pearson

The test methodology developed by J. Neyman and E. Pearson (1933) is concerned with the *behavioral* aspects of testing. The objective is to control and minimize the chance of committing two kinds of error in a long run of repetitions of the same test procedure. A Type I error is committed when the null hypothesis is rejected when it is in fact true. A Type II error instead is committed when the null hypothesis is not rejected but it is in fact false.

The statistical test can be defined in terms of the test statistic Z (already defined at the beginning of the section) and of a critical region $\mathcal{R}$. If the test statistic falls within the critical region, $z \in \mathcal{R}$, the null hypothesis H_0 is rejected and the alternative H_A is accepted, if $z \notin \mathcal{R}$ the opposite occurs. Although the description could suggest that the test is symmetric between the alternatives H_0 and H_A, it is only H_0 that is actually tested against the data, similarly to what happens in Fisher's significance test. In other words, accepting H_A denotes a *statistical* acceptance, stemming from the fact that H_A was the proposed alternative to the null. Nothing is said about whether H_A is a reasonable alternative face the problem in the first place. However, in clinical trials, the choice of H_A can be regarded as setting a yardstick for the new treatment's effectiveness. The probability of a mistaken rejection is denoted α and it

is calculated as the probability that a sample from the null hypothesis falls within the rejection region:

$$\alpha = P(Z(X) \in \mathcal{R}|H_0) \tag{1.9}$$

instead β is the probability that a sample does not lead to rejection, while actually being taken from the alternative distribution:

$$\beta = P(Z(X) \notin \mathcal{R}|H_A) \tag{1.10}$$

A further important notion in Neyman-Pearson methodology is *power*, the capacity of the test to detect true superiority, that is equal to $\Pi = 1 - \beta$. The two quantities α and β are depicted in Figure 1.b, where it is possible to see how they are related.

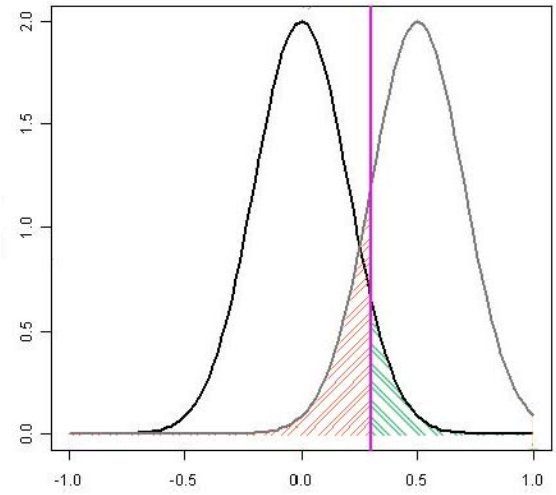

Figure 1.b: Error rates α (green) and β (orange). α is computed assuming the null hypothesis to be true (black curve) while β is calculated under the alternative (grey curve). As it is possible to see, decreasing α by decreasing the significance level entails a increase in β.

The size of the rejection region, α, is also the probability of mistakenly rejecting the null hypothesis in a long series of testing. This descends mathematically from the law of Large Numbers, assuming that the null hypothesis is true in all of the tested samples. In the Neyman-Pearson testing framework, the decision to accept or reject the hypothesis intrinsically depends on the costs associated with committing a Type I or Type II error. Typically, in the context of clinical trials, regulatory bodies and the medical community attach a high penalty to the possibility of mistakenly adopting a treatment that works less well than the standard.

Therefore the chance of committing a type I error is controlled to a lower level than is the other error rate. This is also due to the fact that the two error rates are related through the size of the rejection region (from 1.9 and 1.10), therefore the probability of one error cannot be decreased by choosing a smaller (larger) rejection region, without the other error rate increasing. It is important to note that the critical region of the test is generally the first design feature to be fixed and it is independent of the choice of the alternative. The alternative influences only the type II error rate and power: It follows that different choices for H_A can modify the error-statistical properties of the test but do not have an impact on the result of the inference. Once the alternative H_A is specified, Neyman-Pearson theory specifies the rejection region that yields the test with the specified level α and the largest power.

The most powerful Neyman-Pearson test for simple vs. simple hypotheses is based, by the Neyman-Pearson Lemma, on the ratio of the likelihoods (Lehmann, 1997). The test rejects H_0 in favour of H_A if the ratio of the likelihoods falls below a pre-determined level c_α

$$\mathcal{L}(x) = \frac{\ell(x, \theta_0)}{\ell(x, \theta_A)} \leq c_\alpha \tag{1.11}$$

where c_α is chosen so that

$$P(\mathcal{L}(X) \leq c_\alpha | H_0) = \alpha \tag{1.12}$$

In the case of simple vs. composite test, the power of the test is no longer fixed but it varies with θ in the alternative. However, it is still possible to define a test which is uniformly most powerful. For every fixed value of the parameter in H_A, this test is the most powerful among all alternatives. The test of level α rejects if

$$z(x) > z_\alpha = \Phi_0^{-1}(1 - \alpha) \tag{1.13}$$

and accepts otherwise. As it is possible to see, the rejection region is independent of θ. Φ_0 is the cumulative distribution function of X under H_0 and Φ_0^{-1} is its inverse. Thus, this test rejects the null hypothesis if the observation x falls in the tail of the probability distribution of Z right of the critical value z_α. The NP test in this situation is equivalent to the Fisherian test described above, in particular it corresponds to a test of significance that rejects when $p < \alpha$.

This resemblance has fuelled an inappropriate identification of the two methodologies.

This identification is inappropriate because of its consequences in terms of misinterpretation of test results, highlighted by Goodman (1999a), and because there are profound differences between the two accounts, as discussed by Hubbard and Bayarri (2003). To see the differences, suppose a test of level $\alpha = 0.05$ is conducted and the result is significant with $p = 0.001$. A p-value this low speaks against the null hypothesis, as both Fisher and Neyman would agree. Neither would accept that $p = 0.001$ represents the probability that H_0 has been falsely rejected in this case. However, Fisher would report the p-value as the most informative measure of evidence arising from the experiment, and would give it the evidential interpretation of data-related measure of strength of evidence against the null hypothesis. Neyman and Pearson, on the other hand, would merely report the fact that the results are significant at the 5% level, and therefore H_0 is rejected and H_A is accepted with probability of error equal to α. Indeed, the Neyman-Pearson testing framework permits one to draw a conclusion within the pre-determined error bounds, but contemplates no post-data measure for the assessment of strength of evidence.

1.3.3 Significance testing in clinical trials

Despite the differences just highlighted, the two approaches to testing just described have been integrated in a hybrid framework that is currently used in the analysis of clinical trials.

The results of a trial that enrolls n patients is summarized in the variable X that we assume to be normally distributed: $X \sim \mathcal{N}(\theta, \sigma^2/n)$. From this it is possible to derive the test statistics $z(x) = x\sqrt{n}/\sigma$. The test generally applied in this scenario is the simple vs. composite test described by (1.13). As already noted above, it is equivalent to a significance test that rejects if $\Phi_0(-z(x)) \leq \alpha$. The significance level is set in the phase of design typically at $\alpha = .05$ or $\alpha = .02$, corresponding to a Type I error rate of 5% and 2%, respectively. If the test is two-sided, this error probability should be divided evenly between the two tails of the distribution. The Type II error rate β is often chosen to be around $\beta = .2$, which yields an error probability of 20% and a power $1 - \beta = 80\%$. According to test (1.13), the null hypothesis will be rejected in favor of the alternative if $X > z_\alpha \sigma/\sqrt{n}$. The power Π is the probability of this occurrence when the alternative hypothesis is true, $\theta \in \Theta_A$. Hence:

$$\Pi = \Phi\left(\frac{\theta\sqrt{n}}{\sigma} + z_\alpha\right) \tag{1.14}$$

From formula (1.14) it is possible to determine the sample size n as a function of θ[†]. For any $\theta = \theta_A$

$$n = (\Phi^{-1}(\Pi) - z_\alpha)^2 \frac{\sigma^2}{\theta_A^2} \tag{1.15}$$

This formula yields the number of patients that it is necessary to enroll to detect an effect of at least θ_A with the desired power Π. Among the quantities appearing in this formula, Φ and α are set in the design phase as discussed. The standard deviation σ can be generally assigned a reasonable value based on natural variability among patient response. The choice of θ_A, the value to be assigned to the effect for the sake of determining n, is instead more critical. Clearly, it is desirable to minimize the number of patients, n, because each participant to the trial comes at an ethical and a practical cost. This would incentivize picking larger values for θ_A. However, remember that θ_A represents the minimum value of the effect size which can be detected with the desired power. If the true effect is smaller than θ_A, the test might not be able to detect it. If the observed x is within the rejection region, but yields an estimate for θ which lies below θ_A, the test result will be inconclusive. If this situation occurs, we say the test was *underpowered* for detecting the actual difference. Since underpowered tests expose patients to the risks of trial participation but fail to achieve the epistemic gain of a reliable conclusion, they are generally considered to be unethical (Halpern et al., 2002). Therefore, investigators have to choose a value of θ_A as close as possible to the true underlying effect, in order to minimize both the number of patients to be enrolled and the risk that the trial will be underpowered. Generally, θ_A is chosen with reference to two important pieces of information. On one hand there is the SCID (smallest clinically important delta) or the smallest value of the effect that would make the new treatment a viable therapeutic option. On the other hand there is the estimate of the efficacy of the new treatment coming from earlier phase trials (see Orloff et al. 2009). While the latter value provides a guidance about the true value of the effect, the SCID acts as a benchmark. If the trial fails to demonstrate existence of an effect at least as large as the SCID, the new treatment can be discarded because it does not fulfill clinical expectations.

The procedure just described, inspired by Neyman-Pearsonian principles of inductive behaviour can be regarded as a mechanism to adjust the design of the test in order to optimize the detection of an effect of magnitude at least θ_A. This fulfills an important role in the design

[†]See Spiegelhalter, Abrams and Myles (2004, eq. 2.38)

of clinical trials. Regulatory bodies such as the U.S. FDA put a special emphasis on careful trial design as a means to ensure the validity of trial results. For instance, in the FDA Guidance Document on drug evaluation (Food and Drug Administration, 1998) we read that "The inherent variability in biological systems may produce a positive trial result by chance alone. This possibility is acknowledged, and quantified to some extent, in the statistical evaluation of the result of a single efficacy trial". If the use of Neyman-Pearson ideas were limited to the optimization and control of experimental design, the hybridization so far described would be perfectly legitimate and it would pose no particular problem.

The methodological problems, however, arise because ideas from the two frameworks come to be mingled in the *interpretation* of the experiment. This is what Hubbard and Bayarri (2003) refer to when they denounce "the widespread nature of the anonymous mixing of Fisherian with Neyman-Pearson ideas in some statistics textbooks".

As discusses above, the Neyman-Pearson framework warrants the investigator that the test result is valid within certain pre-specified rates of error. This measure of validity is independent of the data that actually obtain: The investigator has to report the same error probability, whether the data are just at the significance boundary or far beyond it. However, once the data are at hand, the non-evidential nature of the Neyman-Pearson measures is evidently unsatisfying. Consequently, the use of *p*-values as measures of evidence, *à la* Fisher, has taken hold in the medical community. The problematic consequences of this usage have been analysed in depth by Goodman (1999a). For one thing, interpreting *p*-values in single (or ongoing) experiments is void of meaning in a Neyman-Pearson hypothesis testing context. Furthermore, *p*-values are often misinterpreted as post-experimental error probabilities, an interpretation which is by no means sanctioned by Fisher's theory of testing, besides being plainly wrong: The error probabilities associated with classical tests and estimation allow only for an interpretation in terms of long run frequencies.

The interpretational and methodological hurdles related to the significance testing methodology applied to trials will be largely the object of the present work, particularly as they arise in the context of monitoring. Before moving on with the discussion, however, I will conclude this section with a brief introduction to the Bayesian approach to testing.

1.3.4 Bayes

Likewise the Neyman-Pearson test, the Bayesian test (Cornfield, 1966a; Kass and Raftery, 1995; Goodman, 1999b) requires the specification of an alternative hypothesis. Indeed, the procedure of the two tests could superficially appear to be similar. If the two alternatives are specified, point-like hypotheses of the form $H_0 : \theta = \theta_0$ vs. $\theta = \theta_1$ are tested one against the other based on the *Bayes factor* which is simply the ratio of the likelihoods:

$$B(x) = \mathcal{L}(x) = \frac{\ell(x, \theta_0)}{\ell(x, \theta_A)} \tag{1.16}$$

Values of $B(x)$ lower than 1 express support for the alternative H_a, while values higher than 1 favor the null. From equation (1.11) we know that the NP test in the simple vs. simple case is also based on the ratio of the likelihoods. However, the likelihood ratio is confronted with the critical value c_α in NP test (1.11) while it is confronted with unity in the Bayesian test. This difference is crucially related to the two different interpretations underlying the two tests. In the case of the Bayesian test, the point is all about deciding from the experimental data x which hypothesis is to be assigned greater credibility. Clearly, values of $B(x)$ very close to unity lend weak support to either conclusion: However, Bayesian inference is based on the idea that the probability that a hypothesis is true represents a subjective state of knowledge. Hence there is no in principle problem with a result that does not license a conclusion with enough confidence*. The NP test, instead, sets off from a fixed level of confidence that we want the test to warrant. Therefore, the test only licenses a conclusion when the confidence in the result would equal the pre-specified level α, and the critical value c_α is chosen so as to ensure control on the *a priori* probability of error within the desired level of confidence.

The degree of support provided by the data can be quantified, in a Bayesian framework, by calculating the probability of error associated to the test result. In the Bayesian interpretation, the probability that the conclusion is erroneous is simply equal to the probability that the winning hypothesis is not true. Since Bayes' theorem can be used to calculate the probability that an hypothesis is true – simply equal to its posterior probability –, the calculation of the

*The contiguity of Bayesian analysis with formal decision theory makes it possible to use decision-theoretic principles, such as utility, in order to evaluate the appropriateness of accepting an hypothesis in conditions of severe uncertainty (cf. Lewis, Lipsky and Berry 2007). The development of Bayesian testing methodology along decision-theoretic lines seem, however, to be not particularly promising for the application to clinical trials, due to the complexity of modeling all the decision factors involved. These developments are, therefore, beyond the scope of this work.

error probability is straightforward. In case of the simple vs. simple test, the probability that acceptance of H_A is erroneous is equal to the probability that H_0 were true instead:

$$P(H_0|x) = \frac{B(x)}{B(x)+1} \tag{1.17}$$

This is the Bayesian version of the false positive probability or α. Conversely, the Bayesian equivalent to the false negative rate β is the probability that the alternative were true:

$$P(H_A|x) = \frac{1}{B(x)+1} \tag{1.18}$$

The important difference between the Bayesian error rates (1.17) and (1.18) and the NP α and β is that the former depend on the actually observed data x through the Bayes factor $B(x)$ while the latter are pre-specified in the design phase and constitute a property *of the test* rather than of the conclusion drawn in light of data. This fundamental difference stems, again, from the foundational divide between the two paradigms. The NP test can be interpreted as an attempt of putting statistical inference on an objective footing by grounding the plausibility of an hypothesis solely on the stringency of the test it survived. This is error-statistical philosophy in a nutshell (Mayo, 1996). In Bayesian inference, on the other hand, the focus is on assessing the plausibility of the conclusion in light of the data that were actually observed.

The application of the Bayesian approach to testing to cases beyond the simple vs. simple scenario requires an additional element. When the hypothesis to be tested against the null is a composite alternative, the quantity to be compared againts the likelihood of the null in the likelihood ratio has to be constructed as a *marginal* likelihood. This means averaging the likelihood function $\ell(x,\theta)$ over all the values of θ in the alternative. In this way the likelihood of the null hypothesis H_0 is compared to the support that data warrant to its competitor overall. The marginal likelihood is

$$\int_{\Theta_1} f(x|\theta)\pi(\theta)d\theta \tag{1.19}$$

In this formula, $\pi(\theta)$ is the prior density of probability over the alternative $\theta \in \Theta_1$. The idea is that evidence provided by x displaces the probability assignment to the possible values of θ from what it was originally assumed to be. It is possible to make an “objective” or “reference” choice for $\pi(\theta)$ (Jeffreys, 1961; Bernardo, 1979, 2009): in this way, the prior will spread out

the probability over the possible alternatives in a non-informative manner, without favoring any particular value of θ. This procedure tends however to be too conservative in most cases of practical interest (Spiegelhalter, Abrams and Myles, 2004): Since the probability mass is distributed among all components of the alternative, but it is concentrated on the point-like null on the other side, the ratio of the likelihood will tend to favor the null most of the times. This problem can be regarded as a consequence of the fact that Bayesian methodology is not well suited to the test of a point-like hypothesis.

For this reason, many Bayesian-inclined methodologists have proposed an alternative use of the Bayesian framework to the problem of health-care assessment[†]. As described in Section 1.2, the Bayesian philosophy of inference is specially suited for estimation. If input an accurate prior probability assignment $\pi(\theta)$ over the space of possible values for the parameter, application of Bayes' theorem (1.2) yields a posterior probability distribution over θ which directly expresses the posterior plausibility of the different values of θ, after seeing the data:

$$P(\theta|x) = \frac{P(x|\theta)}{P(x)}P(\theta) \tag{1.20}$$

In this formula, $P(x)$ is a normalization factor and its value is generally not of interest. The truly important part of the equation involves θ and it says that the posterior distribution is proportional to (has the same shape as) the product of the likelihood $P(x|\theta)$ and the prior:

$$P(\theta|x) \propto \ell_x(\theta)\pi(\theta) \tag{1.21}$$

The posterior probability $P(\theta|x)$ can be used to provide an interval of possible values of θ that are most credible in light of observation, or to make direct probability statements. Another possibility is that of testing the hypothesis that the parameter lies in a certain interval $\theta \in \Theta_*$, and accepting if the posterior probability mass in Θ_* is more than a specified threshold. Inferences based upon (1.21) are instances of a *full Bayesian approach*: This approach exploits the strenght of the Bayesian paradigm to its full extent. However, as it is easy to see, in this approach the problem of dependence upon the prior distribution $\pi(\theta)$ cuts even deeper than with the test based on the Bayes factor. The prior to be used in (1.21) can be based on previous data, it can be a non-informative distribution or it can be, as proposed by Spiegelhalter

[†]E.g. Berry (1993); for an overview of the method, see (Bolstad, 2007)

et al. (1996), a model of a prototypical skeptic or enthusiast opinion about the effect of a new treatment. In all these cases, there remains the unpalatable fact that the posterior distribution and hence the inference about θ will be influenced by the choice that was made for $\pi(\theta)$. This problem is of great concern to trial methodologists, as exemplified by Moyé (2008): "Without specific safeguards, use of Bayesian procedures will set the stage for the entry of non-fact-based information that, unable to make it through the 'evidence-based' front door, will sneak in through the back door of 'prior distributions'."

Why it is then that, notwithstanding these worries, still so many methodologists are convinced that Bayesian methods hold the potential for substantial contributions to the field of health-care assessment? The reason for this enthusiasm is the greater epistemic efficiency of this framework as compared to classical frequentist methods. Bayesian tests in most cases make it possible to arrive at a conclusion with less data and at a comparable level of confidence. In order to clarify and motivate this claim, it is useful to turn to an application of statistical methodology, where adherence to one or the other framework makes a substantial difference to the practice. This, the issue of trial monitoring and optional stopping, will be the subject of the remaining of the chapter.

1.4 Monitoring and sequential trials

Monitoring refers to the analysis of trial data carried out as they accumulate, with the option of stopping the trial before the planned conclusion, if the data indicate with sufficient strength a conclusion. According to the trial situation, this will mean either stopping before the planned number of patients has been recruited, or terminating the trial regime for in-trial patients before the planned number of events (e.g. deaths or disease recurrences) has been observed. Different scenarios are possible: the trial can be stopped because the new treatment outperforms the control (early stopping for benefit) or because it causes serious and unexpected side-effects (stopping for safety). Other possibilities are that the new treatment performs markedly worse than the standard (stopping for inefficacy) or finally that, given the current results, it is unlikely or impossible that the trial will demonstrate superiority of the experimental treatment (stopping for futility).

In all its declinations, data monitoring in clinical trials fulfills an ethical obligation towards

in-trial patients that ought not to receive a treatment known to be inferior. The characterization of this ethical obligation, while obvious in the case of safety monitoring, becomes less straightforward for the other two kinds of monitoring. This issue will be explored in depth in the third chapter. For what concerns the main focus of my work, namely monitoring for benefit, it suffices to note that there are independent pragmatic reasons that make it desirable regardless of whether it is also ethically beneficial for patients in trial. Data monitoring for benefit facilitates the rapid application of experimental findings to all patients in the population and it meets a need for improving cost-effectiveness of trials that has become pressing in the last decade. By allowing for stopping as soon as a convincing result is obtained, and before the full amount of information has accrued, monitoring makes it possible to attain the valuable epistemic aim of the trial with less resource expenditure.

A recent review (Montori et al., 2005) shows that the practice of trial monitoring is increasing, particularly in the fields of cardiology, oncology, and HIV/AIDS research. The proportion of all RCTs published in high-impact journals that were stopped early for benefit increased from 0.5% in 1990-1994 to 1.2% in 2000-2004. Typically, these were industry-funded drug trials. It is likely that these numbers have further climbed, due to issues of scarcity of research money (Allison, 2012) and of candidates for trial participation (Schroen et al., 2010).

In order to stop a trial early, investigators have to confront a particular form of inference. They do not only have to make a decision about the conclusion warranted by data, as is the case with traditional, fixed-sample trials; they also have to make a decision about when the data so far accumulated warrant a conclusion with sufficient confidence. For this kind of test, therefore, a *stopping rule* has to be specified together with the decision criteria. A stopping rule tells investigators when it is legitimate to stop the trial and draw a conclusion. A branch of statistics called *sequential analysis* provides the tools for confronting this problem.

1.4.1 Sequential analysis

The techniques of sequential analysis have been developed in order to make testing possible when data are still accruing. The importance of this possibility is highlighted by the fact that publication of the foundational work in sequential analysis by British statistician Abraham Wald was delayed until after 1947, due to the potential for war use of the technique he invented. The importance of the possibility offered by sequential monitoring in the context of clinical research

is likewise apparent. The first application of Wald's work to this field is due to Armitage (1975).

A sequential trial is defined as a trial in which the decision to continue or discontinue data collection depends on the information that accrued so far (Wald, 1947). In order to perform this test, pairs of subjects are formed by matching patients with respect to clinically relevant known confounders. When each pair enter the trial, one of the two patients is assigned at random to one arm of the trial and the other patient to the other. The outcome of the trial for each pair is transated to a preference for one of the two arms, according to which of the two patients responded first. The analysis is graphical. The outcomes of successive pair are inserted one after the other as successive steps in a random walk. The stopping and decision rules are represented by boundaries in this graph, V-shaped as in figure 1.c. The coordinates of the

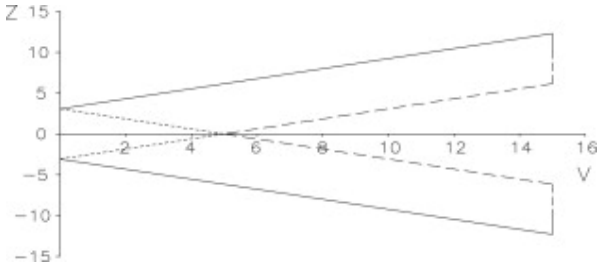

Figure 1.c: The graphical boundaries of Wald's test. Image taken from Todd et al. (2012)

graphical boundaries are determined by the selected type I and II error-rates (α and β) and the expected value for the treatment effect. If the line described by successive pairs results crosses one or the other boundary, the inference about the superiority of one or the other treatment can be drawn at a specified nominal level of significance (for instance, α = 0.05). A modification of this design, with an additional boundary across the open end of the V, enables monitoring for inefficacy. Upon crossing of this third boundary it can be concluded that there is no significant difference between the two treatments at a pre-specified type II error rate.

The strength of the matched-pairs technique is that it permits a continuous monitoring and at the same time it ensures that an inference will be reached with a predictable maximum number of paired subjects. This maximum size will typically be larger than for a fixed sample test with the same level and power, but the *average* sample size is lower, since the trial will generally be terminated before the maximum size is achieved. However, the sequential test based on matched pairs has serious practical limitations that greatly restrict its use in clinical trials. In order to apply this technique, there must be a large reservoir of eligible patients so that matching can be done, and the time from entry to outcome must be short compared

with the expected duration of the trial. These conditions rarely obtain in clinical trials. Even for large megatrials of diseases that have a fast response to treatment, matched sequential analysis requires that the registration of the accruing events be regular and continuous. This is difficult, especially if the trial is a multicenter one. For this reasons a different methodology, group sequential analysis, has been developed for application to the field of clinical .

1.4.2 Group sequential methods

The group sequential methodology has been developed by Pocock (1977). Further important methodological advances include the O'Brien-Fleming (1979) rule and the Haybittle-Peto rule (Haybittle, 1971; Peto et al., 1976).

The group sequential procedure is adequate for situations, fairly common in cancer trials, where events accumulate at an uneven rate and patients enrollment rate is unpredictable. In this procedure, patients do not enter the trial all at once but in groups. Each time two new groups are about to enter the trial (one in the active arm and one in the control), a statistical analysis is performed on the data to determine whether the results so far accrued indicate a significant difference; if this is the case, the trial can be stopped. The significance test is conducted at pre-defined time points using a significance level that is more stringent than the nominal α. This is because multiple subsequent analysis break the equivalence between rejection criteria and probability of error, that was explained in section 4.2. In fact, the significance level (for instance 0.05) corresponds to the chances of observing an effect when the null hypothesis is true instead. But if an investigator tests the data multiple times while they are still accumulating, she invites such an opportunity repeatedly. Therefore, the chance that the observed effect is actually a type I error will be considerably higher after the calculation of many significance tests than after a single one (McPherson, 1974). Asking for a lower *p*-value for stopping represents a way to "raise the bar" and make it more difficult for a false positive to mislead the researcher. This is a fairly intuitive criterion: It restricts the class of results that can lead to early stopping to only the results which are most extreme. In other words, in order for the investigators to accept the conclusion of superiority at a early stage and with only a small fraction of data at hand, the result has to be not just good but striking.

However, the group sequential procedures so far described represent more than a guideline for implementing this intuition: These stopping rules have a theoretical counterpart. As

I have described above, the problem in the setting of monitoring for benefit is the inflation of the possibility to commit a error of type I. At each interim analysis, part of the probability of committing this kind of error, which had been "set aside" as α in the design phase, is consumed. If researchers want to claim a certain level of confidence α for the analysis, the original amount has to be split up among all the interim analyses, and consumed at each of those in a controlled manner. The various group sequential procedures presented above differ precisely for the rate at which type I error rate is spent during the course of the trial. It is evident, then, that the objective of preserving the overall significance level can be achieved only by sticking to the number and frequency of looks at the data that were established in the design phase. In theory, if the formal sequential plan is not adhered to, a proper frequentist analysis is impossible. Actually, Lan and De Mets (1983, De Mets and Lan 1994) developed the notion of the adaptive α-spending function, that could be applied to unplanned interim analyses. This technique relies on a function that can be used to calculate how much of the type I error rate was spent by an unplanned look at the data. Thus, the procedure allows recalibration of the analysis following the unplanned look. However, Tsiatis and Mehta (2003) demonstrated that this approach is inefficient, in the sense that one can always construct a standard group-sequential test that has higher probability of reaching the correct conclusion earlier.

Unfortunately, the necessity to adhere to a rigid design makes monitoring unpractical in many real-world contexts, being often in clash with pragmatic considerations or unforeseen circumstances. The consequence is a low level of compliance to statistical sequential design in a large fraction of published trials. This was highlighted by Montori et al. (2005), who in their survey of 143 trials stopped early for benefit found that "In approximately one third (48/143) of RCTs stopped early for benefit, a statistical approach to monitoring the trial was either not used or not specified in the report". Also Mukherjee et al. (2011), in relation to oncology, observe that "there needs to be a higher degree of transparency by oncology trial investigators with respect to the number of interim analyses carried out, and the stopping rules that were applied to the trial should be explicitly stated". Overall, the presence of results with a low statistical quality contributes to undermine the credibility of the practice of monitoring.

1.4.3 The problem with stopping rules

There are, however, more substantial problems with the requirement to adhere to a fixed sample plan, as it has been pointed out since the early times of the framework's development by Anscombe (1963), Cornfield (1966b, 1976) and Berger and Berry (1988a). Indeed, the control over the pre-determined error rates requires that the significance level should be adjusted for the fact that there were interim looks at the data. If the trial is designed as sequential, the critical value in the final analysis needs to be lower than the nominal significance in order to preserve the advertised type I error rate α. This remains true even in case the trial was not stopped but instead it continued to the planned end. This leads to situations that often seem counterintuitive to the medical audience. For instance, a trial designed as sequential may yield a non-significant result for the new treatment, whereby the same result would have been significant if the trial had been conducted as a fixed size test (Jennison and Turnbull, 1990). Furthermore, an analysis based on the *p*-value depends on stopping rules also in a different manner. The calculation of the *p*-value depends on the set of more extreme outcomes that could have been observed, and this changes with the design adopted. In the case of sequential design, then, the calculated *p*-value would reflect the fact that interim testing was contemplated even if this additional testing was not carried out (see Berger and Berry 1988b). Thus, two statisticians faced with the same set of data may report different *p*-values because one entertained the possibility of stopping the trial early while the other did not. The following example, due to Royall (1997a), clarifies the problem. If in a matched pairs trial engaging 6 subject pairs we were to obtain 5 preferences for the experimental treatment and only 1 for the control, the analysis of the results would depend on the design and therefore on the intentions of the investigators. Did they plan to enroll just 6 pairs and then stop to analyse the results? Or had they decided to stop after obtaining at least one preference for the control? The outcome of the analysis will be very different in the two cases. In particular, the trial result would be significant in the second case but not in the first, due to the fact that the set of 'more extreme values' used to compute the *p*-value is different in the two scenarios described.

This example may seem artificial: Clearly, when the sample size increases the difference due to design will wane, to the point of being negligible for large trials. However, this kind of situations is by no means rare in contexts where monitoring is of its largest practical import, namely, when the illness under study is a severe one and available options are inefficient.

An instance of this occurrence is provided by the context of a trial comparing extracorporeal membrane oxygenation (ECMO) with conventional medical treatment in newborns with persistent pulmonary hypertension (O'Rourke et al., 1989). Since the condition had a fatality rate of 80% under conventional therapy, the investigators decided to halt the trial when the results amounted to four deaths among ten infants in the control arm versus none in the nine infants receiving ECMO. As the ECMO trial shows, in clinical trials it is often of the utmost importance to be able to assess the validity and reliability of a conclusion when just a small fraction of the data is at hand. Based on the example by Royall described above, one may wonder whether the stopping rule that was used should matter in such assessment. Biostatisticians Jennison and Turnbull (1990) put the question in the following terms: "should the statistical analysis of the data be affected by the knowledge that interim data reviews have been performed in the past or that further reviews might be undertaken in the future?"

1.5 The controversy about stopping rules

The question of the inferential relevance of stopping rules that I have briefly outlined in the previous section, far from being purely practical in nature, is at the roots of a heated debate in the Philosophy of Statistics. As Etzioni and Kadane (1995) observe, "Bayesian and classical approaches diverge completely on the question of how to monitor clinical trials, and the reasons for the divergence reach to the very foundations of the two paradigms". In fact, the inferential import of stopping rules represents a fault line dividing philosophies of inference. A crucial issue in the debate is represented by the *likelihood principle*, first introduced by Birnbaum (1962):

> **Likelihood Principle (LP)**: All the information about parameter θ obtainable from an experiment is contained in the likelihood function $\ell_x(\theta)$ for θ given x. Two likelihood functions for θ contain the same information about θ if they are proportional to one another (Berger and Wolpert, 1988, p. 19)

In other words, the likelihood principle states that the information about the parameter of interest should depend on the experiment only through the likelihood function and not through design parameters of the experiment. Once the data are at hand, knowledge of the observed

x is sufficient to draw conclusions about θ, regardless of the properties of the test that produced x. Statistical methods that acknowledge the validity of the likelihood principle are called *likehood-based*; among these, there are most Bayesian methods[†].

Birnbaum (1962) and Berger and Wolpert (1984), in introducing the LP, showed that the LP is formally entailed by the conjuction of two more primitive principles, the sufficiency principle and the conditionality principle.

1.5.1 A digression in statistical theory

In statistical theory, the term *statistic* is used to denote any attribute of a sample which may summarize statistical information about it, such as for instance the conventional sample mean. A statistic $T(X)$ is a function of the sample X. If $T(X)$ summarizes all statistically relevant information about the sample it is called *sufficient*. This means that in the case of a sufficient statistics the original sample provides no additional information that can be used to determine the probability distribution of the population from which the sample was drawn. In other words, if $T(X)$ is sufficient, its value contains all the information needed to compute any estimate of the parameter θ.

A quite straightforward consequence of the concept of sufficiency thus defined is the *sufficiency priciple*, which simply states that if a sufficient statistic exists for the problem at hand, then two observations x_1 and x_2 which yield the same value of the sufficient statistics $T(x_1) = T(x_2)$ license the same inference about the underlying parameter θ. The sufficiency principle is rather uncontroversial as both frequentists and Bayesians recognize it. This agreement is non-philosophical, being rather a consequence of mathematical –measure theoretic– considerations.

More controversial is the second principle invoked by Birnbaum in his famous proof, namely, the *conditionality principle*. In order to see what is the content of this principle, let us suppose that there is a panel of experiments that could be performed in order to learn about common parameter θ and that some random mechanism, independent of θ, is used to select which experiment will be performed. This setup is called a *mixture experiment*. The conditionality principle states that in a mixture experiment situation the inference upon θ should depend only on the experiment which was actually performed, and those experiments which

[†]Some objective Bayesian methods do violate the LP: see Bernardo (2011) and Sprenger (2013)

were not performed must be irrelevant to the inference.

The content of the principle appears to be intuitive. In order to grasp its full disruptive potential, however, it is useful to turn to an example proposed by Savage (1951). Suppose we are interested in testing θ, the unknown probability of heads for possibly biased coin $\mathcal{C}^1$. The null hypothesis is $H_0 = \theta = 1/2$. Prior to examining $\mathcal{C}^1$, however, we flip another, unbiased coin $\mathcal{C}^2$. If $\mathcal{C}^2$ lands heads, we will perform experiment E_1:

E_1 : toss $\mathcal{C}^1$ 12 times

If $\mathcal{C}^2$ lands tails, instead, we will perform experiment E_2 for learning about θ:

E_2 : toss $\mathcal{C}^1$ until 3 tails are observed

Now, the Conditionality Principle states that, since the probability that $\mathcal{C}^2$ will land heads is perfectly independent from θ, then if it does and E_1 is chosen, inference upon observation x_1 from E_1 should not depend upon the fact that we could have chosen E_2 instead. To see why this may be troubling for a frequentist, however, recall the Royall example discussed in section 5.3: It says that the *p*-value will be different in the two cases E_1 and E_2, because the set of more extreme values is different. On the other hand, though, denying the conditionality principle seems to go against common sense: If E_1 was performed, why should our evaluation of the result of it be affected by the fact that E_2 could have been performed instead?

The conditionality principle states that experiments that were not performed are irrelevant to the interpretation of the experiment that was performed. The likelihood principle, instead, states that *data* that were not observed are irrelevant to the interpretation of the data that was observed. The LP is therefore even more difficult for frequentists to accept, particularly due to one consequence of it, the *Stopping Rule Principle*:

> **Stopping Rule Principle (SRP).** In a sequential experiment with observed data $x = (x_1, ..., x_n)$, the information conveyed by the experiment about θ should not depend on the stopping rule τ that was used (Berger and Wolpert, 1988)

In other words, the SRP says that knowledge of the stopping rule that was used in an experiment should not affect the analysis and conclusions derived from the data. Supporters of the Bayesian view regard adherence to the SRP as an asset in statistical practice (Berry, 1987): Since the SRP entails that stopping rules are immaterial to the inference to be drawn from the

trial, analysis using likelihood-based methods can be performed on a sequential experiment even if the original sampling plan could not be adhered to due to unforeseen circumstances. This means that these methods do not suffer from the shortcomings described above in the context of monitoring.

On the other hand, however, frequency-minded statisticians deem the SRP unacceptable since this principle, in combination with classical testing at fixed significance level, leads to sure rejection of a true null hypothesis. It is easy enough to see how this can happen. If an experimenter is testing the null hypothesis at a fixed significance level α, she could use the stopping rule

SR1: "keep sampling until the sample mean $\bar{x}$ is z_α standard deviations from θ_0"

Then she is sure that she will reject H_0 upon stopping, because rejection at level α happens precisely when $\bar{x} - \theta_0/\sigma \geq z_\alpha$ (equation 1.13). In theoretical discussion, this problem goes under the name of sampling to foregone conclusion.

1.5.2 Sampling to a foregone conclusion

The problem connected with repeated sampling is described by Savage thus:

> If sequential properties of his experimental program are ignored, the persistent experimenter can arrive at data that nominally reject any null hypothesis to any significance level, when the null hypothesis is in fact true. (1962, p. 18)

This worrisome consequence of the LP has led many frequentist statisticians to reject the principle. This is apparently problematic given that the two generally accepted principles of sufficiency and conditionality formally entail the LP. Deborah Mayo (2006) has attempted a refutation of Birnbaum's original argument and in consequence claims that a frequentist statistician that subscribes to sufficiency and conditionality need not buy the LP. Pending the validity of Mayo's refutation of Birnbaum, doubts may remain that frequentist statistics needs to accept the LP, due to the controversial aspect of conditionality: Discussion of this issue is deferred to chapter 2.

A number of forceful arguments have been brought forward to counter the charge that likelihood-based and Bayesian methods are liable to sampling to a foregone conclusion. For instance, Cornfield (1966b) and Lindley (1972) take the problem of sampling to foregone

conclusion as a ground for rejecting, rather than the SRP, the classical practice of testing at fixed significance level. In fact, adherence to the SRP does not necessarily expose a *Bayesian* experimenter to the risk of sampling to a foregone conclusion: Bayesian inference is based on the posterior probability that the null hypothesis is true and, if assumptions on the prior are sufficiently reasonable, it may take an infinite time to have this probability fall below a specified level. In other words, as shown by Kadane, Schervish and Seidenfeld (1996a,b), a Bayesian cannot exploit the SRP to design an experiment that will lead to sure rejection of the null, if the null is indeed true.

1.5.3 Optional stopping and error control

Though the debate around the SRP has never really subsided in the statistical community, a criticism against the SRP has been recently rehearsed by Mayo and Kruse (2001). Rather than by the possibility of sampling to a foregone conclusion, Mayo and Kruse are concerned about the fact that ignoring the stopping rule used in testing has a consequence on the unconditional error rates.

The problem is that of testing a normal mean with known variance. Observations are available in N groups of size m, and after each group a two-sided test of level 0.05 is performed. The *actual* type I error is the probability that at least one of the N observations is significant when the mean is actually 0. This quantity increases with the number of observations N, and actually it approaches unity as N goes to infinity. This result is actually an old one, having been demonstrated by Armitage et al. (1969); Mayo and Kruse exploit it to disqualify the likelihood-based methods, due to these methods' adherence to the SRP. Armitage's result entails, in their view, that the appraisal of the inferential import of stopping rules is substantial to "our ability to control error and thereby the reliability and severity of our inferences and tests – generally regarded as important goals of science". As a consequence, "there is no obvious way in which approaches consistent with the LP can deliver these goods [reliability and severity]".

In the context of simple vs. simple hypothesis test, it is possible to demonstrate that the worry expressed by Mayo and Kruse need not be troubling for a supporter of the LP. Indeed, in this scenario it is possible to show that there are several bounds on the possibility of being misled that hold independently of the unconditional error rate. In particular, Savage (1962)

proved the existence of an upper bound on the probability that the Bayes Factor against the null (1.16) will rise above a specified threshold. Royall (2000) similarly showed that there exists an upper bound on the probability of observing evidence so extreme as to lead us to mistakenly reject a true null.

Confronting Mayo and Kruse's challenge in the simple vs. composite case has, however, proved more difficult. This is due essentially to the fact that the Bayes Factor in the simple vs. composite case depends on the marginal likelihood: It is proved that if a uniform, non-informative prior is used for θ (as would be the most sensible thing to do) the upper bound on the Bayes Factor ceases to hold and it becomes possible to sample to a foregone conclusion.

The discussion presented in this section may appear to be mostly of theoretical import. However, its consequences in relation to statistical practice are serious : If Mayo and Kruse are right that error control is a crucial goal in scientific testing, then methods that adhere to the likelihood principle such as Bayesian methods should be rejected on the grounds that they do not allow for control of error rates. However, there are reasons to believe that, at least in the context of clinical trials, the premise of this argument can be challenged. In fact, in the upcoming chapter I will lay out some problems with pre-experimental error rates, in the context of clinical trials and particularly of trial monitoring. These problems suggest that, in this context, control over unconditional (Neyman-Pearson) error rates may be of lesser importance than other inferential goals.

Chapter II:

Conditioning in Sequential Medical Trials

The current statistical framework for the design and analysis of clinical trials is grounded in frequentism. In the design phase, the control of pre-experimental error rates of type I and II guides the choice of all the relevant design parameters of the trial, including the number of patients to be enrolled. Once the trial is finished, *p*-values and confidence intervals are regarded as the principal summaries of evidence gathered through the trial.

An alternative to the classical statistical framework is represented by Bayesianism. The use of Bayesian principles in clinical research statistics has been supported by a growing advocacy movement in recent year. The differences between this framework and the frequentist are significant, to the point of possibly licensing diverging interpretations of the same clinical study. As a first thing, Bayesian inference obeys the likelihood principle–introduced in the preceding chapter–, to the effect that both the decision and the reported error rates are independent from the sampling plan. Furthermore, Bayesian inference consists in updating a prior state of knowledge about a parameter by incorporating the result of fresh observation. The inclusion of prior information in the inference has advantages and shortcomings. On one side, this allows the meaningful placing of unexpected results within a context of prior knowledge. On the other side, though, this external element introduced into the inferential step is seen as jeopardizing the reliability of Bayesian methods.

In the current chapter I will examine a different categorization of the competing statistical approaches, one that does not rely on the genealogy of the methods but upon a feature the different methods may –or may not– possess. The categorization I will examine divides statistical approaches in *conditional* and *unconditional* ones. The term “conditional” identifies statistical procedures that quantify the conclusiveness of a test result by conditioning on the data actually observed. Determination of the posterior probability distribution in a Bayesian framework is an example of such a procedure. This conditioning is instead absent in approaches identified as unconditional: For instance, the error rates in a Neyman-Pearson test

are set pre-experimentally and do not depend on the data that were actually observed.

The classification of statistical approaches as conditional or unconditional turns out to be more fruitful that the often invoked division between frequentist and Bayesian approaches. First and foremost, this division cuts across the fault line between the two schools. Currently, frequentism is characterized by the use of unconditional procedures while most of the proposed Bayesian approaches are conditional. Nonetheless, conditioning is a possibility in frequentist statistics too, and one that bears the potential to overcome some known limitations of the classical framework.

The structure of the chapter will be as follows. The first part of the chapter will be strictly theoretical: I will introduce the distinction between conditional and unconditional approaches in greater formal detail, I will discuss the shortcomings of unconditional procedures from the theoretical point of view, and I will introduce an approach that is aimed at modifying frequentist inference in a conditional sense. In the second part of the chapter I will move to application: I will discuss problems of an unconditional framework as they arise in the context of clinical trials, and I will discuss how moving to a conditional framework may alleviate these.

2.1 Conditional and unconditional procedures

In the previous chapter I have introduced the statistical methodology currently used for design and analysis of clinical trials. Trials are designed following Neyman-Pearson principles of minimization of error in a long run of experiments. By following these principles, the decision to reject or accept the null hypothesis –declare the new treatment effective or not– is associated with a certain statement of confidence, provided by the pre-experimental error coverages α and β respectively. These error coverages, however, only apply to the bipartite conclusion to reject or accept. This implies that strength of evidence –the data-based evaluation of the tenability of the null hypothesis or its rejection– cannot be properly assessed in this framework. The test will report the same, pre-experimental error probabilities no matter how extreme the data are, whether deep into the rejection region or close to its boundary. The focus of the Neyman-Pearson test is on the long-run performance of the procedure and its logic is purely deductive.

The inferential toolkit of the frequentist paradigm, however, is not limited to the hypothe-

sis test. As I described in the previous chapter, the current methodology complements the Neyman-Pearson test with other measures of evidence, which are sensitive to the observed data while at the same time being compatible with frequentist testing principles. The most important such measure is the *p*-value. More recently, discontent with some limitations of the *p*-value as a measure of evidence has prompted interest towards an alternative measure, the confidence interval, which is more adequately described as an estimation procedure. However, both the *p*-value and the confidence interval share the unconditional nature of the NP error rates, and I am going to argue that this makes them suboptimal as assessments of the strength of evidence. After an overview of these quantities, I will discuss what is wrong with them being unconditional.

2.1.1 P-values

The definition of the *p*-value has been introduced in the previous chapter. This quantity measures the probability of an outcome as or more extreme than the one observed, under the null hypothesis. From a statistical point of view, the *p*-value is a measure of evidence with much more dependence on the actual observation than mere rejection at the 5% level. However, *p*-values have been repeatedly denounced as an inadequate measure. As a first thing, interpretational problems due to misapprehension of *p*-values are fairly common in the medical literature (Goodman, 1999a). *P*-values are misinterpreted as the post-experimental probability of the null hypothesis, as an analogous of the Bayesian posterior probability. Alternatively, they are misinterpreted as a post-data probability that the rejection of the null hypothesis is erroneous, a sort of post-data measure of type I error (Hubbard and Bayarri, 2003). Needless to say that neither of these interpretations is sanctioned by frequentist theory. Although not an outright argument against *p*-values, the interpretational difficulty has been taken to testify the counterintuitive character of *p*-values as measures of evidence (Sterne and Davey Smith, 2001; Lee, 2010). A similar point has been raised in the field of Psychology, where misinterpretations of the significance test are equally common, for instance by Wagenmakers (2007) and Fidler (2012). The seminal contribution of Edwards, Lindman and Savage (1963) to the debate about classical significance testing also had its origin in this field.

Turning to more principled arguments against the use of *p*-values, the main reason why *p*-values are regarded as a bad measure of evidence is the fact that the same *p*-value in dif-

ferent experiments does not speak with the same force against the null hypothesis (Cornfield, 1976), due to the *p*-values dependence on the size of the sample. This dependence is evident by considering the definition: The *p*-value is the measure of the tail area of the distribution. As the size of the sample increases, the probability distribution function becomes more peaked. Consequently, the tail area associated with a certain point $\overline{x}$ will become smaller and smaller. It is intuitively plausible that extreme data should speak more forcefully against the null hypothesis when the size of the sample is small than when it is large. The problem is, however, that the indirect character of *p*-values' dependence on sample size makes it impossible to meaningfully compare *p*-values across experiments. Fisher was aware of this property of *p*-values, and indeed he insisted that evaluation of significance should be contextual. Royall (1986) has pointed out a further, apparently paradoxical, behavior of a inferential strategy based on *p*-values: When we know the exact significance level of a result, its weight of evidence against H_0 decreases with sample size, for the reason discussed above. But if we only know that a result achieved significance at a fixed level, the evidence against H_0 *increases* with sample size (Peto et al., 1976).

The evidential shortcomings of *p*-values become apparent by a comparison with the Bayesian measure of strength of evidence, the Bayes Factor (1.16). An important difference between the two quantities is that the *p* value can be calculated by making reference to the null hypothesis alone, while the BF requires the specification of an alternative. On the other hand, though, the *p*-value's independence from alternatives is offset by this quantity's dependence on data that were not observed. Jeffreys (1961) highlights the paradoxical consequence this has on inferential behavior: "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred" (1961: p. 385).

In comparison with Bayes Factors, *p*-values have a tendency to over-estimate the evidence against the null (Dickey, 1977). This is due to the fact, mentioned above, that when the sample size is large even a small discrepancy from the null is sufficient to yield a low *p*-value. In the simple vs. simple setting, the Bayes Factor is simply equal to the likelihood ratio of the null vs. the alternative, therefore it depends only on the data that were observed and it is unaffected by any prior probabilities. In this setting the Bayes Factor has the intuitive behavior of rejecting H_0 when the sample mean is closer to the mean of the alternative and

accepting if it is closer to the mean of the null instead. The *p*-value instead will lead to rejection of H_0 regardless of the distance from the alternative mean, provided that the sample is large enough. In other words, a low *p*-value is the sign of a statistical discrepancy, which may be of no practical importance. Reliance on *p*-values has been criticized in the medical literature, for instance by Ocana and Tannock (2011), due to the misleading character of this feature.

In the simple vs. composite case matters are complicated by the fact that the Bayes factor in this scenario requires specification of a prior distribution over the composite alternative. It is in this context that the notorious Lindley's paradox arises (Lindley, 1957). The paradox states that, given hypothesis H_0 and significance level α, there always exist an experimental outcome $\overline{x}$ which yields simultaneously a *p*-value which is marginally significant *against* H_0, and a Bayes Factor $BF(\overline{x})$ which instead *favors* H_0. The paradox registers a divergence of the two measures of evidence as the sample size increases. As already discussed, the *p*-value in this setting decreases with sample size increase, thus triggering rejection for values which are even closer to the null. The Bayes factor does not show the same dependence from the sample size: The likelihood for the alternative is spread out among all values in the composite, so the Bayesian measure sticks to the null, essentially as the lesser of two evils (Spiegelhalter, Abrams and Myles, 2004, sec. 4.4.5). However, the paradox just points out a divergence of the two measures: It does not adjudicate which one is correct. Indeed, frequentists typically point out that conservatism towards the null of the Bayesian measure may be an undesirable quality (Mayo and Kruse, 2001).

Casella and Berger (1987) show that it is possible to reconcile frequentist evidence expressed in terms of the *p*-value with a second Bayesian measure, the posterior probability that H_0 is true. This accordance does not persist in the two-sided testing of a point-like alternative, as discussed by Berger and Sellke (1987) and Berger and Delampady (1987). In particular, Berger and Sellke make the very important point that the discrepancy between the magnitude of *p* and the magnitude of the post-data evidence against the null as expressed by the posterior distribution is essentially due to a failure to condition on available evidence. Indeed, in order to determine *p*, the knowledge that $X = x$ is replaced with the knowledge that X lies somewhere in the tail of the distribution: $X \in A = \{y : z(y) \geq z(x)\}$. Due to its use of A instead of the actual x, "the frequentist calculation may cause a substantial overevaluation of

the evidence against H_0" (Berger and Sellke, 1987, p. 114).

2.1.2 Confidence intervals

As I have anticipated, use of significance tests and *p*-values has been strongly criticized in recent years in several contexts of application. Use of confidence intervals (CIs) as a way to bypass the shortcomings of *p*-values has been advocated forcefully (Gardner and Altman, 1986; Jennison and Turnbull, 1990). CIs represent an improvement over the use of *p*-values, because unlike these they provide an assessment of the effect size. Thus, CIs allow the appraisal of practical significance of a result, as opposed to *p*-values that allow to gauge statistical significance alone.

The procedure for the construction of a $(1-\alpha)$% CI is as follow: First, for every possible value of θ, the set D_θ of observations consistent with θ at the $(1-\alpha)$% level is identified. This corresponds to the set of observations which would not lead to reject θ at α% significance. Next, one considers the set that is formed by taking all values of θ, and intersects it with the observed data x to obtain the CI

$$C_{1-\alpha}(x) = \{\theta | x \in D_\theta(x)\} \tag{2.1}$$

Thus $C_{1-\alpha}(x)$ contains all values of θ that would not be rejected by a significance test at α% upon observing x.

For a normal distribution, the two-sided confidence interval at 95% level is centered around the sample mean. Its width depends on the standard deviation and on the confidence level α = 5%:

$$C_{95} = [\overline{x}_n - z_5\sigma/\sqrt{n}, \overline{x}_n + z_5\sigma/\sqrt{n}] \tag{2.2}$$

The simple form of the confidence interval suggests that the interval represents the region of the parameter space where the true value of the parameter lies with 95% probability. However, this interpretation is not correct. A 95% CI has the meaning that, if we were to repeat the trial many times and every time construct a 95% confidence interval around the result, then 95% of the times the true value of the effect would be fall within one of the intervals. This, however, does not mean that the particular confidence interval constructed from the result of the one trial that was performed contains the true value with 95% probability, because any particular

CI either contains the true value or it does not. A confidence interval should be interpreted as a "consistency interval": It identifies the set of parameter values θ that are *consistent* with observation $\overline{x}$ at a specified level of significance.

The set which contains the true value of θ with a specified probability can be constructed from the observation $\overline{x}$ relying on the Bayesian posterior probability distribution over the parameter θ. This is called a *credible interval*; it can be equivalent with the classical interval, if the prior distribution invoked in calculation of the posterior is uniform. The practical equivalence of the Bayesian and the frequentist credential sets in this case should not be misleading about the true nature of classical CIs. Even though it is constructed after data were observed, a CI is a pre-experimental measure. When determining the CI for the mean θ of a normal distribution as in (2.2), we would expect the confidence of a genuinely post-data set to be higher when the variance of the sample is low than when it is high. Classical CIs do not behave according to this intuition: The coverage remains the same, but the change in the variance is reflected in a change in the width of the interval (Kiefer, 1977). Furthermore, as shown by Seidenfeld (1981), there are examples of confidence intervals that include the entire sample space with an advertised coverage of less than 100%, raising the question of whether the entire notion is a misnomer. These behaviors depend on the CI being an unconditional measure. As Feinstein (1998) commented, *p*-values and confidence intervals are the "two sides of the same unsatisfactory coin". We turn now to analyzing what is that is unsatisfactory about unconditional measures.

2.2 Some problems with unconditional measures

In a seminal paper entitled "Some problems connected with statistical inference", Cox (1958) highlighted the counterintuitive behavior of some traditional procedures –the hypothesis test and the confidence interval– in situations where conditioning is involved. I introduced the concept of conditionality in the previous chapter, in discussing the likelihood principle. The situation where this concept arises is that of the mixture experiment: A composite experiment in which the toss of a coin is used to decide between two different experiments to be performed, two apparatuses to be used, or the like. In the situation of the mixture experiment, two choices are possible: Optimizing the test parameters for the overall experiment –i.e. in-

cluding the coin toss and its outcome– or waiting for the coin toss and then optimizing the test *conditional* upon the experiment that will actually be performed. As I will show by means of two examples well discussed in the literature, conditioning appears as the more sensible procedure but it typically comes to conflict with the objective to optimize the error-parameters of the test.

As observed by Lehmann (1993), this conflict is totally internal to the frequentist school, since it marks the divergence between the two theories of testing of Fisher and of Neyman and Pearson. Fisher held that conditioning was necessary every time a relevant subset was recognizable in the sample space, in order to avoid situations where probability takes a different value on a set and one of its subsets. He regarded this as necessary in light of a requirement of *total evidence* (Carnap, 1962). Neyman and Pearson, on the other hand, maintained that considerations about power should always take precedence over conditioning. Lehmann makes it very clear that the divergence of the two theories happens because different considerations take precedence. In Neyman-Pearson's theory, the error-statistical properties are paramount because the focus is not on individual cases but on the long-run frequency of errors. Fisher's sole concern, instead, was for situations inherent to scientific inquiry where the circumstances of testing are important, while the long run performance isn't. "This issue –Lehmann observes– seems to lie at the heart of the cases in which the two theories disagree on specific tests". Two such cases will be presented shortly. A doubt may arise at this point, that what is presented here is a purely theoretical issue. After all, how often do scientists toss a coin for deciding which experiment to perform, as in the situation exemplified by the mixture experiment? Actually, this theoretical controversy has instead substantial reverberations in the practice, as I will discuss in the second part of the chapter with application to healthcare assessment.

The first example (Kiefer, 1977; Berger, 2003) revolves around confidence intervals. Let us suppose that we take two observations X_1 and X_2 with probability law

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2 \end{cases}$$

If we now construct a confidence interval for θ, then the interval $C_\theta(\cdot,\cdot)$ defined by

$$C_\theta(X_1, X_2) := \begin{cases} X_1 + 1 & \text{if } X_1 = X_2 \\ (X_1 + X_2)/2 & \text{if } X_1 \neq X_2 \end{cases} \tag{2.3}$$

has an unconditional coverage of 75%. Yet, this does not seem to be a sensible conclusion regarding the *confidence* that the data warrant with respect to the true value of θ. Indeed, if we observe $X_1 = X_2$ we are equally uncertain as to whether θ equals the common value plus one or the common value minus one: So the (a posteriori) confidence of the confidence set is actually only 50%. But if we observe $X_1 \neq X_2$, then θ equals their mean with certainty and we can assign 100% confidence to this conclusions. The unconditional coverage of 75% neglects that, after learning the value of $|X_1 - X_2|$, we are in a much better position to assess the confidence that the data grant to our inference. If we stick to the goal of maximizing the overall confidence, we should go for interval (2.3): However, this interval is not efficient in learning from the data.

A slightly more complicated example, discussed most notably by Cox (1958) and Royall (1997a), can further illuminate this distinction and help us introduce conditional procedures. The situation is that of testing the hypothesis about the value μ of the mean of a normal distribution $\mathcal{N}(\mu, \sigma)$, with variance σ^2 known: $H_0 : \mathcal{N}(0, \sigma^2)$ against $H_1 : \mathcal{N}(1, \sigma^2)$. This is a classical situation of testing of a simple null against a simple alternative, described exhaustively in the first chapter. What is peculiar about the situation however is that the toss of a fair coin is to decide whether we will draw $N = 1$ or $N = 100$ observations. We are confronted with a mixture experiment. Conventional reasoning about good design principles would lead us to choose the rejection boundaries that would make the resulting test more powerful at a fixed (5%) significance level. One possibility is that of choosing the test that maximizes power over the two possible occurrences: drawing $N = 1$ or drawing $N = 100$ observations. This test is

$$T_u(X) = \begin{cases} \text{if } N = 1 & \text{reject } H_0 \text{ for } x_1/\sigma^2 > 1.282 \\ \text{if } N = 100 & \text{reject } H_0 \text{ for } \bar{x}/\sigma^2 > 0.508 \end{cases}$$

The test $T_u(X)$ is unconditionally most powerful. Qualitatively, the idea behind test $T_u(X)$ is to exploit the fact that discrimination between H_0 and H_1 is easier if $N = 100$, in order to lower

the significance in the other, equiprobable scenario. It says that we can afford a (relatively) high type I error rate to an observation taken in the the $N = 1$ scenario, because, if we were to repeat the experiment, we might get a different distribution where discrimination would be easier. The two components of the experiment have opposing properties: The $N = 1$ component has high significance level (high type I error rate) and hence high power. The $N = 100$ component has low significance, and lower power. The overall test $T_u(X)$ combines these properties over the region of interest. It borrows high power from its $N = 1$ component and low significance from its $N = 100$ component. It is intuitively clear, but it can be verified by calculation, that the resulting power is higher than that of the test $T_c(X)$ which rejects H_0 for a fixed value of X, namely $\bar{x}/(\sigma^2 \cdot N) > 1.64$. $T_c(X)$ maximizes power and size conditionally on the distribution that is known to have been sampled, but is not the unconditionally more powerful test. However the test which is most powerful in the overall mixture experiment may not be the best choice once we know which experiment is actually performed. Considerations of which test maximizes power in the mixture scenario appear to be irrelevant when we set out to interpret an observation that we know came, say, from the $N = 1$ scenario. As an aside, it should be noted that use of the test $T_c(X)$ corresponds to a direct application of the Conditionality Principle, discussed in the previous chapter. As in the previous example, also in this case we see that the choice to to conditionalize is problematic, because it conflicts with the objective to optimize the unconditional error-statistical properties of the test.

Turning to the field of application of health-care assessment, the conflict presented above seems to settle against conditioning. Indeed, as I have presented in the previous chapter, control over unconditional error rates of the test procedure is generally perceived as the main warrant upon reliability of clinical trials results. Then, if this objective is jeopardized by conditioning, it seems that we should be satisfied with an unconditional result with good error properties rather than turning to the conditional test, even if the latter appears the more sensible choice. I propose now a reason why reliance upon the unconditional error statistical properties should not be regarded as the most important objective, and question this conclusion. This example will also make the practical import of this discussion more evident.

2.2.1 Against unconditional error assessment: The base rate fallacy

The base-rate fallacy (BRF) is an error in reasoning that results from not taking in due account the size of population subgroups when drawing conclusions on the frequency of certain events in the subgroup. The fallacy is easily understood in the context of Bayesian reasoning, where it can be formalized as follows: In the assessment of hypothesis H based on evidence E, the fallacy is committed if the posterior probability of H being true $P(H|E)$ is calculated on the basis of $P(E|H)$ without taking into account the prior probability of H being true, $P(H)$, as Bayes' theorem would instead mandate. The prior probability $P(H)$, however, needs not reflect a degree of belief in H in the Bayesian sense; it could represent the objective frequency of true H in the population: its *base rate*, hence the name.

In many practical contexts, the fallacy is caused by neglecting this kind of objective frequencies. A 'classical' example is the medical test described by Psillos (2007, p. 17–18):

> A test for the presence of a disease has two outcomes, 'positive' and 'negative' (call them + and $-$). Let a subject (Joan) take the test. Let H be the hypothesis that Joan has the disease and $\neg H$ the hypothesis that Joan doesn't have the disease. The test is highly reliable: it has zero false negative rate. That is, the likelihood that the subject tested negative given that she has the disease is zero. The test has a small false positive rate: the likelihood that Joan is tested positive though she doesn't have the disease is, say, 5 per cent.

If the disease is extremely rare –that is, it has a low base rate $P(H)$– the probability that Joan does not have the disease although she tested positive will be quite high and, in particular, it will be higher than the advertised false positive rate of the test. In other terms, taking the FPR rate of the test $P(+|H)$ as a faithful indication of whether the subject tested has the disease $P(H|+)$ without taking into account the base rate $P(H)$ leads to the base rate fallacy and to grossly overestimating Joan's risk.

Now the point is that it is possible to establish a straightforward analogy between the medical test just described and the significance test as applied to clinical trials. As I will briefly outline in this section, this suggests that reliance on error properties of the test can be epistemically dangerous in this case also. Furthermore, this concern seems to be supported by recent findings, as I will detail in due time.

A medical test like the one described by Psillos can be straightforwardly mapped into a statistical test upon a real-valued variable of the kind used for clinical trials and described in chapter I. Indeed, most of the medical tests rely on the detection of some biological parameter, for example of blood glucose level, body temperature, presence of a particular antigene and so on. For the sake of concreteness, let us think of some biological parameter that is found in the blood stream, so that the medical test physically consists in a routine blood analysis. The patient is healthy if the level θ of the marker in her blood is less than a threshold value θ_0 signaling disease. Therefore, the medical test can be modeled as a standard one-sided test, with null hypothesis H_0 that the biological parameter θ is less than a threshold value θ_0 –hence the patient is healthy. The alternative hypothesis H_A will be that the parameter is above the threshold, hence the patient has the disease.

$$H_0 : \theta \leq \theta_0 \qquad \text{(corresponding to } \neg H\text{)}$$

$$H_A : \theta > \theta_0 \qquad \text{(corresponding to } H\text{)}$$

As the underlying statistical model, it is reasonable to consider the distribution of the random variable W, representing the level of the biological marker within the analysed blood sample. In order to simplify the problem, let us assume that W is distributed normally, with known variance: $W \sim \mathcal{N}(\theta, \sigma)$. The medical test is thus all about deciding, from the observation w -the sample taken from Joan- whether the population mean θ that represents the level of the biological marker in Joan's blood, is greater or less that the value θ_0 signalling disease. Thus, it is possible to see that the medical test represents a device needed to turn statistical information about a real-value parameter into a two-tiered choice among two options, precisely as it is the case with hypothesis test in clinical trials[†].

The classical test is equipped with unconditional error probabilities, the Type I and II error rates. Such error rates play a fundamental role in assessment of results within the error-statistical perspective, which is the most comprehensive philosophical underpinning of frequentist testing. The type I error rate, the false positive rate of the NP test, is calculated assuming the null hypothesis to be true. It is therefore indicative of the false positive rate that obtains in repeated testing, but only if an assumption is made about the underlying propor-

[†]The analogy between the two problems has been proposed also by Westover et al. (2011)

tion of true nulls. Without information on this underlying proportion, there is no guarantee that the Type I error rate provides a faithful estimate of the probability that the result we have just accepted is a false positive. Indeed, as Spielman (1974) observes, "if all the hypotheses tested by a researcher are true, all of his rejections will be erroneous, no matter how small his level of significance is". Spanos (2010) has recently endeavored to defend the error-statistical perspective against the threat posed by the BRF. However, his result is limited to showing that the consistency of the error-statistical approach is not undermined by the BRF. What Spanos' analysis leaves instead open is the possibility that the consequences of the fallacy may indeed be serious in fields of scientific practice, such as clinical research, where the error-properties of the Neyman-Pearson test are employed for the evaluation of results.

The weight of such consequences can be appreciated by taking into consideration a recent provocative paper by Ioannidis (2005). In this article, Ioannidis discusses a form of the base-rate fallacy in which –in his opinion- most of the scientific community is caught. He claims that the error properties of classical Neyman-Pearson test are often misunderstood, taking the fixed significance level α to correspond to the probability that the hypothesis that passed the test is a false positive. Based on this fact, scientists in the different fields expect a very high proportion of the relations that achieved formal significance to be true. However, as Ioannidis notes, after a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. This is a post-data figure that depends not only on the error properties of the test, but also on the underlying ratio of true relationships to no relationships. For instance, if 95% of all nulls were true a priori and α = 0.05, then about 50% of the nulls will be true given $p \leq \alpha$: This is the true fraction of false findings[†].

Ioannidis' conclusion uncovers a clear instantiation of the base-rate fallacy. In particular, he considers the field of epidemiology, where relationships that are being tested have generally a very low prior probability. In this context, even good error statistical properties of the test (optimality in Neyman-Pearson terms) are not sufficient to warrant high plausibility to the hypotheses that passed the test. Based on the very high fraction of true nulls that reasonably exists in epidemiological research, Ioannidis can provocatively claim that "most published research findings are false". Even though determining a plausible value for the

[†]Johnson and Gastwirth (1991) make the same calculation in the context of a medical test screening for HIV

percentage of true nulls in drug research is far more challenging, the risk posed by the BRF should discourage exclusive reliance on the error properties of the test for evaluating trial results. As pointed out by D. G. Altman in a unfortunately neglected methodological paper, "[w]e obviously do not know whether the null hypothesis is true, so the probability of rejecting it in error is also unknown, although this clearly reduces as [the significance level] reduces" (Altman, 1980, p. 1612). Together with the problem with false positives here described there is an analogous problem with false negative claims and type II error rates, which has been discussed by Greenland (2011).

Clearly, Ioannidis' point about the BFR in interpretation of epidemiological findings does not necessarily carry over to interpretation of results from clinical trials. Typically, incorporation of such results within a body of shared medical knowledge does not rely upon a –albeit imprecise– knowledge of the true proportion of ineffective drugs that are tested. Rather, physicians are in most cases able to meaningfully place the results of a new trial, by interpreting such results in light of everything that has been learned about a treatment both within and outside the trial. As I will describe in the next section, however, the picture is complicated in the context of monitoring of an ongoing trial. In this situation unconditional error rates prove inadequate to guide proper appraisal of results.

2.3 Monitoring in clinical trials: Problems of an unconditional approach

As I have described in the previous chapter, the practice of monitoring trials for early signs of effectiveness has become increasingly common in the last decade, as a means for improving the cost-effectiveness of the trial process. In the case of monitoring for benefit, the trial can be stopped early if a large treatment effect is found before the end of the study. Stopping early means stopping with an amount of statistical information smaller than what was initially planned. According to the trial typology this will mean either stopping before the planned number of patients has been recruited, or terminating the trial regime for in-trial patients before the planned number of events (e.g. deaths or disease recurrences) has been observed. As I have already described, monitoring in the current framework involves a sequential plan which specifies the number and timing of interim analyses. At each analysis, the significance level

that can prompt stopping is lower than the nominal significance level of the test: This means that, for an analysis at 5% significance, the trial will be stopped early if the *p*-value is lower than 0.02, 0.01 or the like.

This procedure is *prima facie* sensible: We accept to stop the trial early only if the result appears to be exceptional. However, as we saw in the discussion about *p*-values in section 2.1.1, what is *statistically* relevant may not be of real practical importance. Furthermore, adopting the classical sequential procedure for guiding monitoring and early stopping introduces a problem concerning the correct appraisal of results of trials that stop early. The practical relevance of this issue is testified by a series of articles appeared recently in the main medical journals (Montori et al., 2005; Wilcox et al., 2008; Bassler et al., 2008, 2010). These studies found that trials stopped early for benefit show implausibly large treatment effects, relative to what the medical community would be inclined to expect. In a review of 134 trials stopped early for benefit, Montori et al. (2005) point out an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. Bassler et al. (2010), based on a comparison between results from early stopped trials and trials on the same research question that were continued to the planned end, found that estimates from the truncated RCTs are systematically larger than the those from the non-truncated studies.

Such claims of bias supported by empirical evidence severely discredit the reliability of trials that were stopped early, as well as their reputation in the medical community. For instance, Montori and colleagues suggest that "clinicians should view the results of such [foreshortened] trials with skepticism" (Montori et al., 2005, p. 2209), while Bassler and colleagues conclude that clinicians should assume "appreciable overestimates of effect in trials stopped early" (2010, p. 1187). This attitude of mistrust, in turn, threatens to nullify the possible advantages of monitoring.

The overestimation described by Montori, Bassler and their colleagues is actually inevitable under the classic significance test, as discussed by several methodologists, especially those that are familiar with a Bayesian framework (Goodman, 2007, 2009; Goodman, Berry and Wittes, 2010; Berry, Carlin and Connor, 2010; Ellenberg, DeMets and Fleming, 2010). In fact, as the classical sequential methodology mandates stopping under extremely low *p*, the class of early stopped trials is enriched in results that have an uncommonly high

value of the difference in performance between the new treatment and the control. Thus, estimates from such truncated trials will on average overstate the true extent of the benefit. The bias discussed in the medical literature affects only the point estimate of treatment effect, and it is an unavoidable consequence of early stopping under fixed significance level. This problem should not be perceived as a deficiency of the frequentist method: The technique of sequential analysis is optimized with respect to its primary objective, which is the control over the error rates in repeated testing. The appraisal of results should simply not rely on the face value arising from the trial.

Jennison and Turnbull (2000) have proposed to correct the known shortcomings of the classical procedure through recourse to a sequential test based on confidence intervals. From what I have described in section 2.1.2, it should be clear that a methodology based on confidence intervals is surely better able to cope with the estimation problem due to sequential analysis *vis-à-vis* a *p*-value based one. On the other hand, though, this approach is liable to the same criticism as a strategy based on the *p*-value, since confidence intervals are built upon a unconditional level of significance. For one thing, this entails the same problems of rigidity of the sampling plan that I have already described in the previous chapter: proper sequential plans with legitimate stopping rules often cannot be adhered to, and this determines a low level of compliance to statistical sequential design in a large fraction of published trials. However, the problem with unconditional procedures is more profound.

When invested with the decision to stop a trial early for apparent benefit, medical investigators are concerned about the possibility of promoting a treatment that is actually less efficacious. For instance, Mueller et al. (2007) report a case of two leukemia treatments where interim analyses suggested a high relative risk reduction (53% and 45%) in a particular chemotherapy regimen, but the assessment had to be reversed after completion of the trial. From what I have discussed so far, it should be clear that the classical unconditional procedure does not provide a way to discriminate this kind of situations efficiently. The unconditional procedure for sequential analysis poses constraints on the decision to stop which are aimed at the control of the error rate in the long run. As I have discussed in section 2.2.1, however, the unconditional error rate does not provide a useful guidance about the probability that a particular inference is wrong. The fact that the conditions for stopping warrant that we will not be too often wrong in the long run, is not a good warrant that we are not making

a mistake in this particular trial that we are deciding to stop early. Furthermore, as I have discussed in section 2.1, it is a matter of contention that unconditional measures such as the *p*-value are able to gauge appropriately the strength of evidence in the data. The idea that such measures can adequately inform the decision to stop is therefore equally problematic. In the upcoming section I will discuss different attempts to find a more adequate solution than the classical sequential procedure to the problem of monitoring. As I will eventually show, conditioning is a most interesting element common to all these alternative approaches.

2.4 Reforming statistics for monitoring

In the early 1980s, when the practical problem with sequential analysis, the rigidity of the sampling plan, was starting to gain the attention of investigators and methodologists, some attempts were made at reforming sequential analysis within the boundaries of the accepted theory of hypothesis testing. The concept of *stochastic curtailment* was introduced, as a way to make unplanned termination of trials possible. The idea behind this approach is to stop the trial as soon as the available data make a conclusion –towards either H_0 or H_A– certain or determined with very high probability. In other words, one should stop when further data, whatever they are, do not have the capacity to revert the current trend. For instance, Ware et al. (1985) proposed an approach along these lines based on a *futility index* which could be used to establish a point in the trial whence achieving a significant result would no longer be possible.

In seminal work conducted over the years with different co-authors, K. K. G. Lan pioneered the approach to model the process of the sequential test as a random walk with unknown drift (Lan, Simon and Halperin, 1982; Lan and Wittes, 1988; Lan and Zucker, 1993). Lan showed that it is possible to exploit this mathematical equivalence in order to extrapolate to the end of the study based the current value of the estimate. As an instance, the concept of *conditional power* introduced in Lan, Simon and Halperin (1982) represents the probability of rejecting the null hypothesis based on the current data. The projected outcome of the trial is determined by combining the data already observed up to the moment of the analysis, with the future observations that are predicted using the null value θ_0 for θ. Thus, the conditional power approach is "only half conditional": it uses the originally hypothesized value, and not

the current estimate, to make the prediction about future observations which is needed to draw the inference ahead of completion.

This procedure can be criticized on the grounds that the value θ_0 may not be a plausible one in light of data that are available at the moment of the interim analysis (Spiegelhalter, Freedman and Blackburn, 1986). In practice, the conditional power approach ignores the knowledge about θ that accumulated so far in the trial. We can easily recognize, in this criticism, an echo of the theoretical arguments agains unconditional procedures presented in section 2.2. However, this time the problem can be gauged in its practical import.

2.4.1 Bayesian advocacy

Turning to Bayesian statistics has been repeatedly proposed as a way to substantially improve the effectiveness and adequacy of the practice of monitoring. This advocacy has come from both biostatisticians –Donald Berry, Steven Goodman, Joseph Kadane, David Spiegelhalter among the most active– and methodologically aware physicians –George Diamond and Sanjay Kaul in heart and circulation, among the most recent–. Indeed, Jerome Cornfield had been advocating a Bayesian outlook in clinical trials ever since the 1960s (Cornfield, 1966a, 1969, 1976). According to Keating and Cambrosio (2012), the beginnings of the RCT methodology were marked by a coexistence of the two statistical schools, with researchers in the UK favoring frequentist methods and researchers on the other side of the Atlantic pushing Bayesian methods. After less than two decades, however, standardization was achieved and frequentist methods prevailed. Among the reasons which hindered the spread of Bayesian statistics in the beginning was the computationally-intensive character of Bayesian procedures in most real-world applications. In the beginning of the 1990s, however, a computational breakthrough was achieved though development of the technique known as Markov Chain Monte-Carlo (MCMC: Gilks et al. 1996). This paved the way to considerable expansion of Bayesian methods to applied problems in the medical arena in the following years (Ashby, 2006), particularly after creation of a computer package for applied statisticians (BUGS: Thomas et al. 1992).

But what does ground the conviction that Bayesian methods could provide particular advantages in the context of monitoring? From what we have seen in first chapter, Bayesian inference is based on Bayes formula, which guides update of a state of knowledge in light of

new evidence. This makes Bayesian inference especially suited for a continuous update in a flow of accruing data. Parmar et al. (1994), Spiegelhalter et al. (1994) and Parmar et al. (2001) gave a practical demonstration of this feature with a prospective Bayesian analysis of the CHART trial, investigating a new radiotherapy technique in both non-small-cell lung cancer and head and neck cancer. This trial was conducted at the Medical Research Council in the UK. The trial was designed and performed according to traditional frequentist principles; however, a Bayesian analysis was conducted contextually and its results were presented at each meeting of the statistical committee monitoring the trial. Two priors were used in the Bayesian analysis: Both were obtained by formalizing in probabilistic terms the expectations of participating clinicians about the new treatment. One prior collected together the more skeptical opinions and the other one included the opinions more favorable to the new treatment. In this way, the two priors constituted useful reference positions: a considered skeptic and a considered enthusiast. The Bayesian analysis consisted in updating these priors based on the likelihood of incoming trial data. With this kind of analysis in play, a trial can be monitored by using the posterior distribution in order to check whether the evidence accumulated up to a certain point is sufficient to significantly displace the initial assignment of probability or, in other words, to persuade a reasonable skeptic or a reasonable enthusiast to change their mind.

A different approach for constructing priors is the so-called *empirical* choice. In this case, the prior probability distribution for the parameter is constructed so as to summarize the current state of knowledge about a treatment effect. This approach applies any time empirical results –such as earlier stages of the trial or of other trials of the same treatment– are available about the treatment under study. In this way the available information can be incorporated formally into the inferential process. Thus, unexpected results can be positioned meaningfully within a context of prior knowledge, so as to mitigate the impact of implausible findings. Spiegelhalter, Abrams and Myles (2004) give a demonstration of this method by applying it to a trial comparing two drugs for dissolving clots in occluded coronary arteries following a myocardial infarction. The trial in question (GUSTO; Migrino et al. 1997) was analyzed retrospectively, using a prior based on the data from two previous studies that investigated the same compounds. The GUSTO trial concluded with a statistically significant result under the frequentist analysis. However, the previous studies (GISSI-2 and ISIS-3) had found minimal

difference in the primary endpoint. The Bayesian analysis incorporating these results suggested that the observed statistical significance may not translate into a practically relevant difference between the two treatments.

In the context of monitoring, this approach to prior choice can indeed provide some guidance when it comes to deciding whether an early positive finding is to be considered the effect of mere chance or of a truly effective medication. For example, previous studies of the same treatment that had a positive outcome will make a positive result for the current trial more expected and therefore will support the decision to stop early, whereas previous indecisive results would downplay early bursts in performance. It is important to note that this kind of reasoning is routinely used in the decision about early stopping of trials, albeit only in an informal way.

An important thing to note about Bayesian methods used in monitoring is that Bayesian inference obeys the likelihood principle, presented in chapter I. As reviewed there, this means that the analysis on the data is unaffected by the decision to stop. When a Bayesian method is used for monitoring a trial, the time-points of the analyses need not be specified in advance, unlike in frequentist analysis. Furthermore, the overestimation associated to trials that are stopped early in the traditional framework (Montori et al., 2005) is not present if the trial is stopped following a Bayesian analysis: This is because the decision to stop is grounded on the current level of tenability of the null hypothesis face the alternative, while in classical analysis it is directly determined by the occurrence of a exceptionally rare observation (Berry, 1987; Goodman, 2007).

Among philosophers of science, Teira (2011) and Stanev (2012) have critically evaluated these claims. Both downplay the enthusiasm towards Bayesian methods. Even though it seem uncontroversial that Bayesian treatment can make the trial process more efficient, when it comes to assessing which statistical framework to adopt for health-care evaluation, the possibility for monitoring represents just one among the aspects that are relevant and possibly not even a crucial one. The need to specify priors, unescapable in a Bayesian approach, is perceived to be in conflict with the values of objectivity and external validity which are asked from clinical research. Medical professionals, in particular, are concerned about the possibly distorting effect of prior opinions of physicians participating in a clinical trial. This worry is well expressed by Moyé (2008), who warns his colleagues against “non-fact-based information

that, unable to make it through the 'evidence-based' front door, will sneak in through the back door of 'prior distributions"' (p. 476). Most Bayesians would contest that such position is overstating the objectivity warrant provided by frequentist statistics *vis-à-vis* the Bayesian. Nonetheless, Moyé's remark reflects a genuine uneasiness in the medical community about including priors in the evaluation of evidence from trials, since this would require a profound revision of the criteria for this which are firmy entrenched.

In the last chapter I will return to this issue and discuss it more in detail from the perspective of the regulation of medical research. For the moment I point out that there is an interest in reaping the advantages of monitoring if this is possible without completely overthrowing the present framework. A possibility in this sense is offered by theoretical attempts at reconciling the two frameworks, motivated by the intention to exploit the respective strengths of the two paradigms in order to integrate them in a superior statistical framework. The reconciliation takes two possible routes. One is that of complementing Bayesian methods with statistical tools for model evaluation, the so-called *calibrated Bayesian* approach (Little, 2006). A second route consists in complementing frequentist statistics with evidential measures of strength of evidence, using Bayesian insight. This is the *conditional frequentist* proposal (Kiefer, 1977; Berger, 2003). The merit and feasibility of such reconciliation approaches is currently a hotly debated topic in the Philosophy of Statistics. Evaluating the merits of reconciliation may however be easier in applicative scenarios such as testing in clinical trials. In the context of trials conditional frequentism is, in my opinion, the most promising option. In what remains of this chapter, I will introduce the conditional frequentist approach and explore its application to RCTs and the context of monitoring in particular.

2.5 The conditional frequentist test

The idea behind conditional frequentist methods is that of identifying a statistics S indicating strength of evidence in the data and then calculate the error probabilities associated with the inference conditional on the value s that S happens to have in the experimental sample:

$$\alpha(s) = P_0(\text{reject } H_0 | S = s) \tag{2.4}$$

$$\beta(s) = P_A(\text{accept } H_0 | S = s) \tag{2.5}$$

The role played by S is that of partitioning the sample space, according to the conclusiveness of acceptance or rejection when data are in a particular region of the sample space. In other words, S has the function of ordering the possible outcomes, based on the strength of evidence that is associated to them.

The choice of the conditioning statistics S is far from straightforward and the quest for a good conditioning statistics has considerably slowed down developments of this approach. In 1994 Berger, Brown and Wolpert proposed, for the simple vs. simple testing scenario, a conditioning statistics based on Bayesian considerations, using the Bayes factor (1.16) as a measure of strength of evidence.

$$S(X) = \min\{B(X), \Psi^{-1}[B(X)]\} \tag{2.6}$$

Ψ is a function defined on the basis of the cumulative distribution functions of $B(X)$ under H_0 and H_A respectively[†]. Basically, $S(X)$ defined as in (2.6) matches points in the sample space that have the same strength of evidence against H_0. The two hypotheses are treated symmetrically. So the points matched according to the support they provide to the rejection of H_0, provide the same support to the acceptance of H_1, and vice versa. It is important to note that the ordering function performed by the statistics S is entirely pre-data: It corresponds to deciding, before the experiment is performed, which data would lead to more confident rejection (vs. acceptance) than others. No assignment of probability to hypotheses, the contested hallmark of Bayesian theory, is involved. The Bayesian insight is only exploited in what constitutes a strength of Bayesian paradigm, namely the task of identifying a sensible measure of strength of evidence.

The conditional test can now be defined as:

$$T_C(X) = \begin{cases} \text{Reject } H_0 & \text{if } B(X) < c \\ \text{Accept } H_0 & \text{if } B(X) \geq c \end{cases} \tag{2.7}$$

[†]The cdf relative to the Bayes Factor can be written:

$$F_0(c) = P(B(X) \leq c|H_0)$$
$$F_1(c) = P(B(X) \leq c|H_1)$$

then, Ψ

$$\Psi(b) = F_0(b)^{-1}(1 - F_1(b))$$

and for observed $B(x) = s$, we report *conditional error probabilities*

$$\alpha(s) = P_{H_0}(\text{reject}\, H_0 | S = s) = \frac{s}{1+s} \tag{2.8}$$

$$\beta(s) = P_{H_1}(\text{accept}\, H_0 | S = s) = \frac{1}{1+s} \tag{2.9}$$

where the latter inequalities have been proven by Berger, Brown and Wolpert (1994, Theorem 1) The critical value c of the test can be thought of as analogous to the value c_α in the Neyman-Person test (1.11)[†]. In the conditional test by Berger, Brown and Wolpert, the inference relies upon a pre-set rejection region like the classical NP test. The error probabilities associated to the inference are instead calculated conditionally. The fact that the Bayes factor is used should not be misleading about the frequentist nature of the test: in the simple vs. simple hypothesis test, $B(X)$ is simply equal to the ratio of the likelihood, and from (1.11) we know that the most powerful Neyman-Pearson test in this setting is based on the same quantity. Furthermore, Kiefer (1977) and Brown (1978) established that conditional error probabilities such as (2.8) have a legitimate frequentist interpretation, i.e. a valid interpretation in terms of error rates in repeated sampling.

Thus, T_C is a valid (conditional) frequentist test. But T_C is valid in the Bayesian context, too, since the conditional error probabilities (2.8) have a valid interpretation in Bayesian terms. Indeed, if we assume that the two hypotheses have equal prior probability, $P(H_0) = P(H_1) = 1/2$, the posterior probability of H_0 and H_1 can be written as

$$P(H_0|x) = \left(1 + B(x)^{-1}\right)^{-1} = \frac{B(x)}{1+B(x)}$$

$$P(H_1|x) = (1 + B(x))^{-1} = \frac{1}{1+B(x)}$$

Thus, we see that the posterior probabilities of H_0 and H_1 correspond to the conditional error probabilities for rejecting H_0 and H_1 respectively, if a neutral prior is assumed. Indeed, the decision to accept H_0 will be wrong whenever H_1 is actually true, that is, with probability $1/1 + B(x)$. Thus T_C is simultaneously a conditional frequentist and a Bayesian test.

Berger (2003) further proved that conditioning on (2.6) is equivalent to the choice of the

[†] It is the value that satisfies the equality $F_0(c) = 1 - F_1(c)$

following statistics:

$$S'(X) = \max\{p_0, p_1\} \tag{2.10}$$

where p_0 is the usual p-value calculated under the null hypothesis while p_1 is the p-value calculated under H_1. According to Berger, the equivalence between S and S' is an extremely important fact because it brings the conditional frequentist framework that uses S as the conditioning statistics even closer to the goal of unification. In fact using S implies, by this result, making use of p-values as measures of strength of evidence, much in the spirit of Fisher's testing; however, through conditioning and the determination of data-specific error rates (2.8), the conditional test yields a measure of evidence that has a legitimate interpretation in both the Bayesian and the Neyman-Pearson framework, unlike p-values. Unification is surely a matter of theoretical importance. However, its consequence for the practice are far more interesting. It means that Bayesians and frequentists can conduct the same (conditional) test and obtain the same numerical conclusions. As long as there is methodological agreement on procedures, philosophical questions about the interpretation of probability can remain in the background: Practitioners do not have to decide for either camp (Berger, 2003). As discussed through the chapter, in the present framework the statistical evaluation of trial results relies on unconditional error rates, on one side, and indirect indications of strength of evidence conveyed by p-values, on the other. Conditional error probabilities fit into this picture because they are consistent with an evaluation of trial results from an error-based perspective. However, by exploiting Bayesian insight, the conditional error incorporates an assessment of strength of evidence. Thus, the conditional inference represents a way to make error-based inference more informative while not betraying its committment to evaluating results on the basis of associated error, rather than their posterior probability. On the other hand, Bayesians may appreciate the fact that the conditional test avoids many of the pitfalls of classical frequentist inference, such as assigning the credibility of a result on the basis of data that could have been observed but were not.

Berger, Boukai and Wang (1997) have extended conditional tests to the simple vs. composite testing scenario. Clearly enough, this requires assigning a prior distribution over the composite alternative in order to calculate the Bayes factor. This may raise the concern that the attempt to evade the issue of prior specification through the conditional test fails in this other scenario, which is surely of greater practical interest. However, it can be argued that

the role played by priors in the conditional test is intrinsically different from that played in traditional Bayesian analysis. In a test based on the Bayes Factor such as described in Section 1.3.4, assigning a prior over the alternative has the precise meaning of assigning degrees of plausibility to the values in the alternative. The prior to be used for calculation of the Bayes factor for the purpose of conducting a conditional frequentist test need not take up this meaning. Hence, a non-committal, reference prior is the more adequate choice for application to the conditional frequentist test.

2.5.1 The conditional test in sequential setting

The use of a fully conditional plan for monitoring of trials is possible using a result by Berger, Brown and Wolpert (1994), who developed the sequential version of the conditional test, building on Wald's landmark Sequential Probability Ratio Test (cf. section 1.4.1). Berger, Boukai and Wang (1999) confronted the properties of the two tests and they found substantial agreement between the two, in the sense that the tests mandate stopping in similar situations. Discrepancies are limited to situation where the boundaries for the classical test include significant overshoot –that is, if the boundaries were designed under the alternative hypothesis of a smaller effect than was actually observed. In this case, however, it seems more reasonable to rely on the error probabilities provided by T_C, once the data are at hand.

One important difference is the existence of a no-decision region in the conditional test. If the data happen to be in the no-decision region, the test is inconclusive as to whether H_0 should be accepted or rejected. This happens because, in this region, data would lend anyway weak support to either hypothesis. The existence of the no-decision region is not necessarily a drawback of conditional inference. As a first thing, it seems just reasonable to think that sometimes data just do not allow us to discriminate effectively. Additionally, as shown by Paulo (2002), the (pre-data) probability of ending up in the no-decision region is never higher than than the largest unconditional error rate. An important thing to observe about the conditional error probabilities in the sequential version of the test is that they obey the stopping rule principle, since they depend only on the likelihood ratio $B(X)$. This means that, in conducting the sequential version of T_C, the error probabilities can be calculated even if the stopping rule was not fully specified. Clearly, investigators may prefer to abide by carefully designed sequential plans for regulatory purposes; what is important is that, if

for some unforeseeable occurrence the plan could not be adhered to, analysis using the conditional test can be legitimately conducted.

The advantage of turning to the conditional sequential test are easily seen. The decision to stop an ongoing trial is a complex one and a number of factor have to be taken into account. A relevant consideration is the evaluation of whether the data that have so far accrued would be sufficient to convince the medical community of the result. At present, the statistical tool that is used to guide this evaluation is the interim *p*-value in the context of a sequential stopping plan. As I have described in this chapter, however, the interim *p*-value is not a reliable indicator of whether the result would be accepted in the medical community; in fact, since a low *p*-value is associated with a likely over-estimation of treatment effect, it can even be said to be misleading for this purpose. A conditional measure of error instead is directly related to the credibility of the hypothesis and could therefore provide a sounder guidance in deciding whether the evidence accumulated so far is sufficient to ground the decision to stop. Given the material presented so far, there are good reasons to think that the conditional test would represent a substantial improvement in the conduct of RCTs.

However, since the conditional test obeys the stopping rule principle, the test T_C is liable to the objection presented in section 1.5.3: When observations are taken without a formal stopping plan, the unconditional error rates cannot be properly controlled and are actually inflated far beyond acceptable. Should one be worried about the possibly bad frequency properties of the conditional test? Answering this question proves particularly difficult, because it requires committing to a supposition about the elements that are important within the medical community for evaluating medical evidence from trials. The support warranted by pre-experimental error properties can indeed be lost by the conditional test, just as shown by Mayo and Kruse (2001). This is only a problem for the conditional test, however, if the reliance upon pre-experimental error rates of the trial procedure represents an important part of the evaluation of trial results. It seems reasonable to maintain that medical investigators are more concerned with the actual probability of drawing the wrong inference than with the absolute (unconditional) error rate of the testing procedure and, in this respect, post-experimental error rates prove certainly more adequate. Similarly, Sprenger (2009) has argued from a decision-theoretic perspective, based on Schervish, Kadane and Seidenfeld (2003), that for a knowing agent stopping rules have a role pre-experimentally but they become irrelevant

post-experimentally. On the other hand, though, pre-experimental error assessment certainly represents an important tool for the institutional consumers of trial statistics, the regulatory authorities. Thus, it is also possible that the proposal of such a substantial revision of the statistical framework in use for trials may encounter considerable resistance. I delay discussion of this issue to Chapter IV, and I close this chapter by suggesting a possible solution for exploiting the valuable aspects of conditioning without the need to compromise on classical stopping rules.

2.6 Conditional error rates in reporting

As long as the the criterion for stopping early is based on the observation of a low p-value, results of trials stopped early will likely overestimate the extent of the new treatment's benefit. Thus, the overestimation discussed in section 2.3 is really not eliminable in the present framework. However it may be possible to correct the most problematic consequence of this overestimation, namely, the unjustified skepticism towards trials that stop early, by using the conditional assessment of error associated to the conditional test (Nardini, 2013; Nardini and Sprenger, 2013). When a trial stops early for benefit, the claim of superiority is less strong, since it is based on a smaller set of data than it was originally planned. However, in the present framework there is no methodologically sound way to make sense of this intuitive truth. A trial stopped earlier than planned has prima facie the same reliability as if the trial had been carried to the planned end. In fact the sequential procedure has the objective of maintaining the pre-planned error rate even if the experiment was stopped early. Instead, the conditional statements of error associated to T_C are sensitive to evidence contained in the data and also to the fact that the trial was stopped early. Since conditional error rates can legitimately be interpreted as expressions of the probability that the conclusion is erroneous, they would facilitate medical readers in their judgment about the reliability of the magnitude of a trial result, be the trial foreshortened or not.

In practice, the solution here envisaged would be extremely easy to implement: a trial can be designed as usual, in particular with the usual strategies in place for the control of unconditional error rates. When the decision to halt the trial is made, the current result can be used to compute the conditional error. The test would yield the same, possibly inflated, estimate

for the treatment under study as does the current methodology. However, the estimate would now be associated with a conditional statement of credibility of the conclusion. In order to see the potential of this proposal, it is useful to consider an illustrative example. The example involves a trial for adjuvant therapy in resectable hepatocellular carcinoma (Lau et al., 1999). The trial was stopped early based on interim findings, but additional data were available after the decision to stop was taken. Pocock and White (1999) describe the situation in detail:

> At the planned interim analysis, the local disease recurrence rates for the active treatment (intra-arterial lipiodol-iodine-131) and control (no adjuvant treatment) groups were three/14 (21%) and 11/16 (69%) respectively (p = 0.01). According to the predefined stopping rule, $p < 0.029$ was sufficient for early stopping. The non-significant overall survival difference, three versus six deaths, was in the same direction. Thus, the investigators decided to stop the trial. [However], 13 more patients were randomised before the trial was stopped, and the investigators also decided to postpone analysis while patients already randomised were followed up. Hence, the report (18 months after the trial was stopped) reveals updated recurrence rates of six/21 (29%) and 13/22 (59%), respectively (p = 0.04). Thus the absolute difference in recurrence rates shrank from 48% to 30% during the interval between stoppage and publication. (p. 944)

Such shrinkage of the estimated benefit between the interim and the final analysis is precisely what physicians dread about early stopping. There is a widespread view that early stopping captures trial results which are on a "random high": If the trial were to continue, the outcome would eventually "regress to the truth" similarly to what happened with the trial of Lau and colleagues.

In this situation conditional error rates can provide real guidance, as I am going to show. In order to calculate the error rates using (2.8), it is necessary to set up a alternative hypothesis. This is possible using Lau et al.'s expectations that "131I-lipiodol would reduce the rate of recurrence [postulated to be 50%] by 50% and double the disease-free survival rate" (1999: 798). Using this value in calculation yields a Bayes factor $B(x)$ = 0.09 and a type I error rate at the interim analysis of α^* = 9%. This value should be contrasted with the unconditional error rate, which is always α = 5% since the trial was stopped following a proper group sequential rule, and with the inadequate informational content provided by p = 0.01% which

is just indicative of an unexpectedly high performance. The conditional error, instead, reflects the greater statistical uncertainty associated with the small sample when the decision to stop the trial was taken. When a trial stops early, the conditional error will be generally higher than for a fully completed one, unless its result is strikingly good. Hence, assessment of error after data collection through conditional error probabilities offers the possibility to fine-grain the decision about the reliability of results from trials that are stopped early. Making a post-data measure of error available on medical journals would enable medical readers to judge single trials that stop early and to evaluate each in its own right. Eventually, this would dispel the mounting skepticism about all trials that stopped early and would thus facilitate the appraisal of the benefits of monitoring and early stopping. However, the conditional error does not only reflect the amount of statistical information concurring to a conclusion but also the strength of the evidence. This is shown by the conditional error probability associated to the final analysis: The calculation based on the same assumptions as the preceding yields $B(x) = 0.16$ and $\alpha^* = 14\%$, higher than at the interim analysis. The increased chance that the conclusion is erroneous reflects the fact that the final data speak less forcefully for the efficacy of the new treatment. Thus, even though no statistical method can eliminate situations where a promising treatment reveals itself as a fluke, we see that the conditional assessment of error enables a much better appreciation of the results of a trial, be it foreshortened or not.

A possible criticism that could be raised against this proposal is that, by making early stopping more widely accepted, it may incite unwise usage of this technique. Indeed, the possibility of licensing drugs with fewer data appeals to pharmaceutical companies for reasons that are all too obvious. Shouldn't we be worried of a technique that makes it easier to implement monitoring and early stopping, and possibly compromise RCTs' capacity to act as a gatekeeper against ineffective or harmful compounds? I think that this position has a standing, however it should not be regarded as a conterargument to the use of conditional methods. The question of whether stopping a trial early is a good or a bad thing hinges on a value judgement that is not for the statistical framework to adjudicate. It involves different perspectives, the ethical side of the bedside of participating patients and the view of the regulator which is entrusted with the protection of social interests in medical research. In the next two chapters of the thesis I will engage with these two perspectives in turn. I anticipate my conclusion, which is that, from both of these points of view, there are reasons that make

flexibility in monitoring a desirable quality for a statistical framework.

Chapter III:

Monitoring with Equipoise

Clinical trials rest on a delicate ethical balance. On one hand, the aim of the trial is to forward medical knowledge and possibly to establish a new, more effective treatment option. Currently, RCTs represent the "gold standard" of health care assessment; the epistemic objective of a clinical trial constitutes, therefore, a collective benefit that society is to gain from the trial. On the other hand, though, it is clearly unacceptable according to our ethical standards that this benefit comes from an exploitation or harm to the individual patients that are involved in a ongoing trial. This implies that clinical trials are ethically acceptable under the requirement that patients participating in the trial are not receiving a treatment that is *known* to be inferior.

In the previous chapters of the thesis aspects connected to the ethics of trials have been mentioned several times. However, until now the term "ethics" has been used in a rather vague manner to simply denote the protection of the well-being and the interests of the patients participating in the trial. Such limited wealth of conceptual instruments proves insufficient for confronting the question that was left open in the previos chapter, about the ethical desirability of stopping a trial as soon as there are signs of manifest superiority of one treatment. Is stopping early a beneficial thing for patients involved in clinical experimentation? Is it mandated by an obligation to protect such patients? Is stopping early compatible with the *social* role of clinical research in forwarding medical progress? In order to provide a sensible answer, these questions have to be confronted from within a coherent and sound ethical framework. Claims about the ethical virtue of different statistical approaches–and of Bayesian ones particularly– abound in the biostatistical literature. Unfortunately however, as I am going to discuss in this chapter, the ethical underpinning of the most part of such claims is presently too underdeveloped. Thus, my objective in this chapter will be that of using the tools and concepts of current clinical research ethics in order to assess the ethics of early stopping and to discuss whether any statistical approach can be deemed the most ethical in this respect.

The chapter will be structured as follows. As a first thing, I will elucidate the ethical dilemma

that underlies randomized trials and the practice of trial monitoring in particular. Then I will discuss an ethical framework that has achieved a certain popularity in the biostatistical literature for evaluating the adequacy of different statistical and methodological solutions in tackling this tension. This framework is the *individual-collective ethics* dichotomy. However, as I will discuss, the dichotomy framework has been deemed inadequate in the characterization of its ethical foundations. Therefore, I will turn to describe the most widely accepted framework for the ethical evaluation of RCTs: Namely, the concept of *clinical equipoise* introduced by Freedman (1987). Equipoise, or relevant clinical uncertainty, is widely reputed to be able to dissolve the inherent tension between collective and individual interest involved in clinical research. In the context of monitoring and optional stopping, however, equipoise encounters problems which appear to undermine its adequacy. Being convinced of the value of this framework, in the final part of the chapter I will attempt to rescue equipoise by redefining it in its aspects which are related to the epistemic side – most notably, the characterization of clinical uncertainty in statistical terms. Finally, I turn to confronting the original question about the ethics of trial monitoring from within the revised equipoise framework.

3.1 The dilemma of clinical research ethics

The aim of clinical research is producing reliable and generalizable medical knowledge. Therefore, clinical experimentation is forwarding first and foremost what could be described as a collective interest in producing the knowledge needed to guide future treatment choices. On the other hand, however, this collective interest cannot be allowed to overweight the interests and well-being of the individual patients involved in research. Guidelines for the ethics of research on human subjects like the Helsinki Declaration emphasize the discounting of collective interest that has to take place in order to make clinical research ethical: “Concern for the interest of the subject must always prevail over the interest of science and society” (WMA, 2008). It is true that, in recent years, this statement contained in the Declaration has been exposed as an overprotection that would, if applied strictly, deem a significant part of current clinical research as unethical (Giordano, 2010). Nonetheless, it remains a fundamental tenet of research on human subjects that the interests of participants cannot be consistently overridden. Were this the case, indiscriminate appeal to other-regarding considerations would make

it possible to perform whatever extremely dangerous research on living persons, provided that the expected benefits for future patients or for society are sufficiently high.

In other words, two distinct sets of interests are identifiable–the collective interest in novel medical evidence and the individual interest of participating patients–, and clinical research should be forwarding both. However, reconciling these two sets of interests appears hard at first glance, at least in some cases. In particular, randomized controlled trials pose an important ethical dilemma, in that they represent, at the same time, extremely reliable tools for the production of medical knowledge and ethically delicate experiments with human beings. RCTs entail specific procedures, like randomization and blinding, that may conflict with the individual interest of the participating patients. In particular, two different ethical issues are attached to randomization. As a first thing, randomization entails that the choice of the treatment that will be administered to the participating patient is out of the hands of both the patient and her treating physician. This is problematic both because it represents a violation of the participating patient's autonomy, and because it prevents the treating physician to take individualized, considered decisions regarding her patient's care, a duty arising from the fiduciary patient-doctor relationship. As a second thing, randomization entails that, by entering trial, the patient may receive the treatment that will eventually turn out to be inferior. This is generally conceived of as a violation of therapeutic beneficence and therapeutic non-maleficence towards the patient, besides going against the physician's therapeutic obligation as already mentioned[†].

The ethical tension in RCTs is often described as a conflict between research and therapy: Between allegiance to the scientific method and therapeutic obligation (Royall, 1991); between the two roles of the physician as a scientific investigator or as a care-provider (Hellman and Hellman, 1991); or, on more consequentialist grounds, as a conflict between the interest of future patients and the interest of trial participants (Joffe and Truog, 2008). This tension has all but increased with the rise of the Evidence-Based Medicine movement, as most EBM theorists place evidence from RCTs at the very top of the hierarchy of medical evidence (Howick, 2011). Since RCTs cannot be renounced to as a tool for acquiring medical knowledge, it follows that some ethical instruments must be found to tackle the tension.

Informed consent is (when obtainable) a necessary condition for the ethical conduct of

[†]The ethically relevant consequences of blinding are similar, but they emerge particularly in monitoring

clinical trials. However, informed consent alone does not provide sufficient protection against the violation of therapeutic beneficence that RCTs, as we have seen, seem to entail: Treating a patient with an intervention which is less effective than the available standard remains a unethical act even in case the patient consents. The view that is currently prevalent in the ethical literature is that *equipoise*, denoting a epistemic state of indifference between two treatments, is able to provide a relief of the ethical tension inherent in RCTs. A fundamentally epistemic criterion, equipoise identifies the condition under which it is legitimate to conduct a trial. The idea underlying equipoise is that competent uncertainty of medical experts about which treatment is superior entails that patients are not harmed by the offer of randomization between the two treatments. This uncertainty warrants that trial entry represents "an equal bet in prospect" for patients (Edwards et al., 1998). Hill, the celebrated father of the RCT methodology in clinical practice, was referring to a similar idea when he observed "Only if, in his state of ignorance, [the doctor] believes the treatment given to be a matter of indifference can he accept a random distribution of the patients to different groups" (Hill, 1963). The concept of equipoise will be explored in depth in the forthcoming discussion.

A radically different solution is the one proposed by Miller and Brody (2003, 2007). According to Miller and Brody, most medical ethics literature on clinical research has been caught in a misguided "similarity position", the notion that "the ethics of clinical trials rest on the same moral considerations that underlie the ethics of clinical medicine" (2003, p. 20). Instead, the authors claim this is not the case: While clinical medicine entails a commitment to the well-being of patients, clinical research entails no such commitment. Rather, clinical researchers have an obligation to *not exploit* research subjects. But the therapeutic misconceptions that clinical research is aimed at the benefit of participating patients may actually cover exploitation of participants. Consequently, Miller and Brody argue that clinical trials should be regulated by principles appropriate to the ethical evaluation of research, including a principled assessment of the ethical features of the trial as a research project, especially with regards to the potential for participants' exploitation. This proposal can however be challenged on consequentialist grounds, as done for instance by Weijer and Miller (2003). The solution proposed by Miller and Brody requires, in fact, to explicitely disclaim the therapeutic obligation towards present patients, including many with serious illness: This, in turn, is likely to affect in a negative manner both recruitment into trials and societal trust in research. Furthermore, Weijer and

Paul Miller observe that it is unclear how Miller and Brody's solution could apply to research on vulnerable subjects who cannot autonomously renounce to their status as patients, such as children or patients in the emergency room. Finding such objections to Miller and Brody's view to be sound, I will not discuss their proposal any further.

3.1.1 The dilemma in monitoring

The ethical tension between individual and collective interests in trials emerges with particular force in the context of an ongoing trial. As the evidence keeps accumulating, investigators may be concerned about the ethics of keeping patients on the treatment which is performing worse albeit not conclusively so. On the other hand, crossing all patients over to the best-performing arm may irreparably compromise the possibility of achieving the epistemic goal of the study. The practice of data monitoring was introduced in the previous chapters. As already discussed, there are different reasons to stop a trial early and consequently different kinds of monitoring. The trial can be stopped because the new treatment outperforms the control (early stopping for benefit) or because it causes serious and unexpected side-effects (stopping for safety). Other possibilities are that the new treatment performs markedly worse than the standard (stopping for inefficacy) or finally that, given the current results, it is unlikely or impossible that the trial will demonstrate superiority of the experimental treatment (stopping for futility). Often, the term 'inefficacy' is used to denote both the latter two reasons for stopping.

Safety monitoring, which is done in order to promptly detect unexpected harmful side effects in the active arm, is virtually indispensible in any trial of a novel treatment. Since its necessity is universally recognized, this kind of monitoring is the least controversial, also because it does not partake the peculiar epistemic problem that characterize other kinds of monitoring. In fact, a very low amount of evidence is generally sufficient to trigger stopping for safety, while the epistemic standards for drawing conclusions on the treatment effect ahead of time –such as is done with efficacy or inefficacy stopping– are much more demanding.

The other two kinds of monitoring, efficacy and inefficacy monitoring, are becoming increasingly common essentially as a means of optimizing research expenditure. Korn and Freidlin (2011) recently advocated a higher commitment in inefficacy monitoring on the part of investigators and monitoring bodies. They argue that "in terms of protecting the patients

enrolled on the trial, inefficacy monitoring is more important than superiority [efficacy] monitoring. This is because inefficacy monitoring protects patients from receiving an experimental treatment that may be worse than the standard treatment they would have received if they had not participated in the trial" (p.2). Efficacy monitoring, on the contrary, seems to be more ethically controversial: On the one hand, the advantages of an early termination to in-trial patients seem less marked while, on the other hand, the commercial interest of pharmaceutical companies on the possibility of a shorter trial is all too evident.

Also the statistical requirements are different in the two cases. As Pocock (2006) illustrates "It is widely recognized that any statistical boundary for benefit needs to be sufficiently tough [...] so that they relate well to the public health implications of a decision to stop the trial early" (p. 513). On the other hand, "there is usually an asymmetry in the statistical stopping boundary whereby one requires less extreme evidence of a treatment difference to stop early if such a difference is in 'the wrong direction' i.e., new treatment looking inferior" (p. 517). The focus is on maintaining the planned significance level in the case of efficacy monitoring, since the risk involved is that of committing a type I error. In the context of inefficacy monitoring, instead, the trial is concluded with a rejection of the alternative hypothesis, and therefore considerations about the type I error do not bear relevance. What matters in this case is that the test maintains sufficient power to not let a existing effect go unnoticed. For the sake of concreteness, all of the following discussion will be focused on one of the two cases, namely on efficacy monitoring.

When setting up a clinical trial, investigators are confronted with an ethical dilemma: On the one hand, they cannot ethically treat patients with an inferior remedy; on the other hand, in order to achieve a conclusion of superiority of one of the two treatments, they need to administer to part of the participants the treatment which will turn out to be inferior. In the context of monitoring, the problem is even more marked. To see why this is the case, let us consider the matched pair sequential design introduced in chapter 1. In this kind of study patients are enrolled in pairs as the trial proceeds, with one member of the pair chosen at random receiving the experimental treatment and the other receiving the control. Enrollment continues until the interim results cross a pre-set boundary of statistical significance. At first glance, this design would appear more ethical than the classical fixed sample trial were patients get enrolled all at once, for the reason that it is more epistemically efficient: The number

of patients that get exposed to the risk of receiving an inferior treatment is strictly limited to the number needed to achieve a scientifically valid result. However, Lellouch and Schwartz (1971) identify an ethical issue in enrolling patients in sequential trials, which they describe thus: “near the end of the trial, [...] one of the patients among the last pairs necessary to draw conclusive findings will undergo a therapy that, at the time, will seem considerably inferior to the other treatment”. In other words, as soon as some evidence in the ongoing trial begins to favor one treatment over the other, even if not conclusively so, there is an apparent violation of therapeutic beneficence towards patients that are offered entry into the arm which is performing worse. Lellouch and Schwartz refer to this issue as the *accumulating-data* problem. The problem does not only affect entry in a sequential trial, but as it is easily seen it affects all trial situations where interim findings are made available while the trial is still ongoing. In this case, the question is whether it is ethically legitimate that patients are kept on the treatment which is performing worse. Monitoring emphasizes the ethical tension inherent in the conduct of clinical trials because, while trial entry is a decision which is taken once and for all, the decision to stop or continue – the decision to keep patients on the inferior treatment or not– keeps re-proposing itself as the trial proceeds. The problem with accumulating data has prompted the development of an ethical framework, which has gained some popularity especially among biostatisticians: The *dichotomy view*. Unfortunately, as I will discuss, the dichotomy view is inadequate not only as a general framework for clinical research ethics, but also for resolving the particular problem it was meant to solve.

3.2 Solving the dilemma: The ‘dichotomy view’

The ‘individual–collective ethics’ dichotomy was originally introduced by Lellouch and Schwartz (1971) in an effort to tackle the ethical difficulties of sequential RCTs. According to this framework, the ethical choice in clinical trials dichotomizes into, respectively, doing what is best for current subjects in the trial versus doing what is best for future patients who stand to benefit from the trial’s results. Pocock (1993) points out that “each clinical trial involves a balance between individual and collective ethics” and that such a balance is never simple but complex. In this case, ‘individual ethics’ refers to the interest of patients participating in trials while ‘collective ethics’ refers to the scientific and societal interest in the generalizable medical knowledge

that can be obtained through the trial.

It is evident that the conflict expressed in the dichotomy coincides with the dilemma of clinical research expressed in the previous section. However, in the dichotomy view, the interests embodied by individual and collective ethics are in an uneasy coexistence at all times in clinical research. This differs profoundly from the equipoise view, whereby instead clinical uncertainty or equipoise warrants that the tension is absent in all cases where trials are ethically acceptable. The two 'ethics', individual and collective, are supposed to represent conflicting interest that have the same stance, so that one of the two has every time to be sacrificed for the other. For instance, Palmer (2002) proposes to categorize trials according to the severity and prevalence of the illness involved. The idea is to identify in this manner situations in which it is acceptable to forgo the individual interest of participating patients in order to forward the collective good, versus situations in which the protection of participants must take precedence[†].

The conflict between individual and collective ethics defines the boundaries of acceptable trial design. By the dichotomy view, different trial designs can be evaluated according to the extent they warrant individual or collective ethics. Most of the ethical discussion around RCTs in the statistical literature has been carried out in terms of the dichotomy view: For instance Palmer (1993); Palmer and Rosenberger (1999); Pocock (1993); Pullman and Wang (2001); Pocock (2006). Possibly this is due to the fact, mentioned by Palmer (2002), that the dichotomy approach is accessible to non-specialists in Ethics due to its simplicity. A more adequate explanation of the appeal of the dichotomy view may be the one provided by Heilig and Weijer, who identify in this framework "an effort to cast specific statistical methodologies as solutions to ethical problems".

The accumulating-data problem, in particular, has been cast in terms of a balance between collective ethics and individual ethics. According to this view, monitoring entails a conflict between stopping early in the interest of present patients and continuing in order to forward the collective interest in a more conclusive result. As I have already described, the dichotomy view entails that the interest of individual participants and the collective interest in reliable trial results are traded off one against the other during the course of the trial. For instance, in relation to HIV-AIDS research, Pocock (1993) observes that "especially in the United States,

[†]Palmer's project can arguably be said to represent an informal risk-benefit profile assessment.

the push towards individual ethics at the expense of collective ethics has been detrimental to determine the most effective therapeutic policies"(p. 1466). In fact, Pocock continues, a trial that stops early "does not represent strong evidence for the superiority" of one treatment and therefore "provides little scope for making reliable judgments on the benefits of this treatment for universal use" (p. 1460). The conclusion that follows seems to be that the only way to minimize controversies about early stopped trials, is to avoid early stopping altogether: "Our overall recommendation is that very convincing evidence of treatment benefit based on a large number of patients is required to stop and publish a clinical trial ahead of its preplanned completion." (p. 1466)

The dichotomy view strikes a balance between two equally valuable things and tells us that we have to choose: On the one hand we have the protection of patients that are kept on what seems an inferior treatment, on the other hand we have the possibility to conduct a RCT to its planned end and thus gain convincing medical evidence. Albeit this idea may appear convincing, there are actually several problems with the idea that patient protection and attainment of an epistemic goal are to be negotiated one against the other. Taken together, these problems hopelessly undermine the possibility to use the dichotomy view for adjudicating the ethical dilemmas of clinical research.

3.2.1 Some problems with the dichotomy view

Heilig and Weijer (2005) have thoroughly reviewed the framework of individual and collective ethics and they have identified several problems with it. As they observe: "First, the contrast has over-simplified the stakes. Secondly, [it has] not resolved the accumulating-data problem [it was designed to solve]. Thirdly, [...] the concept has not inspired a rigorous incorporation of ethics in trial technology [...] appearing in some instances to encourage superficiality" (p.249).

Heilig and Weijer conclude that the dichotomy approach does not represent a workable ethical framework for the discussion of clinical research. One reason is that the dichotomy view entails a forceful oversimplification which ends up hiding some relevant details. A more important point is that the dichotomy view has failed to tie the concepts of collective and individual ethics, and the interest that they refer to, to a rigorous ethical foundation. As Heilig and Weijer describe it, "over time, these concepts have metamorphosed into normative, often ambiguous claims about conflicting duties to individuals and to society with no prescribed

means for arbitrating that conflict" (p. 252). This is what happens, for instance, with an account like Palmer's (2002): It is unclear where the interests that are labeled as individual and collective derive their normative force from; whether it is ethically acceptable in a clinical trial that these conflicting interests are traded off one against the other; and finally, which principles should guide the weighting and adjudication of this tradeoff.

While the criticisms provided by Heilig and Weijer question the adequacy of the framework from the ethical point of view, the failure of the dichotomy view in the context of the accumulating-data problem is characterized epistemically. When a trial is being monitored, two choices are available: stopping the trial or continuing it. According to the dichotomy view, these two choices forward the individual or the collective interest, respectively. However, this contrast is problematic, because the question of whether stopping the trial goes in the interest of participating patients can only be adjudicated once a conclusion is drawn from the trial. Indeed, the treatment which *seems* to be the better performing option may not eventually prove to be such. If a trial stops early on a faulty conclusion of superiority, patients that were in the trial will be harmed by the stopping of the trial rather than by its continuation, because their future therapy will be based on the faulty conclusion. The dichotomy view seems to imply that the interest of in-trial patients can be forwarded independently from the attainment of the epistemic good that the trial was designed to provide, but this is clearly not the case. The interest of both in-trial patients and collectivity is forwarded in the same way: by having the trial reach the correct conclusion.

This observation reveals the fundamentally *epistemic* nature of the monitoring dilemma. This aspect is recognized by Freedman and Shapiro (1994), who criticize the dichotomy view precisely on this ground:

> Discussions about the need to sacrifice individual ethics on the altar of collective ethics [...] seem to us misguided. The dilemma seems to presuppose that some time in the course of the [trial] we learn which is the preferred treatment, but do not stop the study because we seek a result that will be convincing to potential skeptics as well as to us [...] The problem however only arises because lurking in the back of the mind is the belief that any difference [...] between treatments [...] is relevant. But, of course, it is not. Until convincing evidence of a clinically relevant difference between treatments has been reached (i.e., prior [to] disturbing clinical

equipoise) the trial has not accomplished its task

As Freedman and Shapiro rightly point out, the dichotomy view presupposes a double standard of evidence: A standard for the participating physician (who is able to "learn which is the preferred treatment" at some point during the trial) and a different one for the medical community (the reason why the investigators continue the trial in order to "seek a result that will be convincing to potential skeptics"). This double standard is difficult to defend. In the upcoming section we will see that the epistemic characterization of the monitoring dilemma is inadequate also in the presently accepted ethical framework, the one based on the concept of equipoise. Before discussing this aspect it is, however, necessary to introduce the current framework in detail.

3.3 Equipoise

As I anticipated in Section 3.1, competent uncertainty of medical professionals is, when present, held able to reconcile the individual and collective interests at stake in clinical trials. If equipoise is present at the beginning of a trial, participation in the trial represents an equal bet in prospect for patients and therefore no violation of therapeutic beneficence or non-maleficence occurs. What is less straightforward, however, is the question of whether equipoise *is* present in the context of a particular trial. As a first thing, a number of definitions of equipoise exist. But even if one abides by one specific definition, as I will do, adjudicating the question of whether equipoise is present remains difficult, particularly in the context of ongoing trials.

3.3.1 Definitions of equipoise

The physician-patient relationship and its fiducial nature has been perceived as the primary locus where the ethical tension of RCTs arise. Accordingly, the concept of *individual* equipoise is the first form under which the notion of equipoise entered the modern ethical literature, in Fried's (1974) landmark work. Concerned mainly about the fact that trial-specific procedures –randomization and blinding– hinder the physician's ability to take individualized, considered decisions regarding her patient's welfare, Fried held that a physician who personally favors one treatment cannot ethically offer to her patients enrollment in a trial were the patient has

just the same probability to receive the other treatment. He therefore maintained that only if the physician finds herself in a state of indifference or equipoise she can ethically propose trial entry to her patients. Through Fried's work, equipoise was established as a necessary condition in clinical research; Fried's notion applies to the individual physician and for this reason it is referred to as *individual* or also *theoretical* equipoise.

Fried himself regarded theoretical equipoise as a condition that rarely obtains in practice, not only because doctors typically have individual views and preferences about treatments, but also because, when considering to offer trial entry to a particular patient, they are bound to consider the unique circumstances and values attached to that patient. This makes theoretical equipoise a principle of almost no practical use for the regulation of clinical research. In consideration of this fact Freedman (1987), in a equally foundational article, dismissed the requirement for individual equipoise as unpractical and proposed to adopt *clinical* or *collective* equipoise in its place. Freedman observes that the fact that physicians are rarely in a state of personal equipoise poses "nearly insuperable obstacles to the ethical commencement or completion of a controlled trial and may also contribute to the termination of trials because of the failure to enroll enough patients" (p.141). In order to overcome these issues, that make the concept of equipoise unworkable as the ethical justification for RCTs, Freedman proposes a redefinition of the concept. He observes that "the basic reason for conducting clinical trials [is that] there is a current or imminent conflict in the clinical community over what treatment is preferred for patients in a defined population" (p. 143). Accordingly, trials are justifiable when there exists *clinical equipoise*, or a state of "honest, professional disagreement among expert clinicians about the preferred treatment". The important aspect of Freedman's proposal is that, according to the requirement of clinical equipoise, it is the standard of the medical community rather than the individual physician's inclination that should determine whether participation in an RCT is ethically acceptable. A similar view, relying on epistemic standards shared by the scientific community, was proposed by Levine (1988) who argued that a trial is ethical if "there is no scientifically validated reason to predict that therapy A will be superior to therapy B".

Some authors have argued that this account of the moral landscape is incomplete, and that patients' own values and judgements should be taken into account when deciding about the moral soundness of a proposed RCTs. The notion of *patient equipoise* has been intro-

duced following the argument that it is patients, and not clinicians, who should be indifferent among the various treatment options when enrolling in a trial. This view has been proposed, for instance, by Veatch (2002) and Lilford (2003) and it has a force in some special cases. For instance, it is clear enough that a patient with prostate cancer may not desire to enter a randomized controlled trial in which he stands just the same chance of receiving radiotherapy or a radical prostectomy: As Lilford (2003) notes, “A man with early prostate cancer who wants a child may place a higher value on preservation of fertility than someone who has no such aspirations”. The delicate trade-off between prevention and side-effects that is evident in a case like this cannot be adjudicated without taking the personal values of the concerned individuals into account. Karlawisk and Lantos (1997) have argued that the locus of morally relevant disagreement should be extended to include patients and their representatives together with medical professionals. Therefore they have proposed the concept of *community equipoise*.

These positions remain, however, marginal or restricted to special cases and clinical equipoise is accepted, at least in the U.S., as the appropriate principle for providing ethical justification. In Europe an alternative principle is favored, referred to as the *uncertainty principle*. Part of the difference between the uncertainty principle and equipoise is that the former lacks the semantic commitment expressed in the word ‘equipoise’ for an equal balance between the two treatments in trial: The uncertainty principle merely expresses a lack of knowledge, while “Equipoise is different from simply ‘not knowing’ or being ‘uncertain’ because it implies that we have no rational preference whatever” (Lilford and Jackson, 1995). The main difference, however, lies in the fact that the uncertainty principle brings the moral focus back to the physician-patient relationship: For instance, according to Peto and Baigent (1998) “A patient can be entered if, and only if, the responsible clinician is substantially uncertain which of the trial treatments would be most appropriate for that particular patient”.

Supporters of the uncertainty principle claim that clinical equipoise or collective uncertainty among the body of clinicians should not be taken as the ethical defence of RCTs. For instance, Enkin (2000) maintains that “If we grant moral authority to the medical community as a whole, we devalue the responsibility of individual clinicians”, and this is undesirable because the medical community may get stuck in a faulty and difficult to shake consensus (“the complacent collective certainty”). However important the judgment of individual clinicians may

be, I maintain that it is preferable to turn to community standards for the ethical justification of trials. A first reason for this is the consequentialist worry already explored by Freedman: a framework that leaves the moral onus of the decision to participating physicians can lead to low participation and consequently to difficulties in recruiting. More in general, though, it seems that the focus on individual perspective entailed by this kind of framework overlooks the perspective of society and the fact that there is a collective interest in trials as a means to respond to a clinical and scientific question. For this reason, a framework like the uncertainty principle seems to be ill equipped to account for the principled version of the clinical research dilemma that I have described at the beginning of this chapter. Clinical equipoise as introduced by Freedman, instead, takes the value of RCTs for resolving a conflict or an uncertainty in the medical community explicitly into account. It is, therefore, better suited to truly provide a reconciliation between the clinical and scientific obligations that exist in medical research.

Miller and Weijer (2003) argue that the two versions of equipoise may be compatible: "FE [Fried's Equipoise] articulates conditions under which the fiduciary duties of physician to patient may be upheld in the conduct of research. CE [Clinical Equipoise] sets out a standard for the social approval of research by institutional review boards. Viewed in this way, FE and CE are not necessarily competing notions, but rather address complementary moral concerns" (p. 93). I am but tangentially interested in this divide as I intend to concentrate primarily on clinical equipoise and the way it is affected by epistemic standards adopted in the scientific community. However I observe that it is problematic to maintain that the two criteria can coexist, since situations where they give rise to contrasting indications would pose significant problems, both moral and pragmatic. Consider, for instance, a situation in which a trial can be commenced since it is consistent with the criterion of social acceptability (CE) but no physicians are willing to offer participation to their patients (not in a state of FE). The strength of Freedman's proposal is precisely that of identifying the epistemic state that is relevant with that of the scientific community, thus reducing the moral tension for the individual treating physician. In other words, the physician is not in the position to offer the best option to her patient because her knowledge of which is the best option is conditional on the trial obtaining a conclusive result.

Besides, the fact that the conditions for a trial to be ethical ultimately rest upon an epistemic criterion–namely, the competent uncertainty of the community of medical experts– reveals

the most fascinating aspect of this debate. For the reasons discussed so far, the *epistemic* adequacy of an RCT is part and parcel with its ethical adequacy. Thus, RCTs constitute a most interesting interplay of ethics and epistemology (Worrall, 2008), as the debate on trial monitoring stands to testify.

3.3.2 Equipoise and monitoring

Equipoise reveals its weakness in the context of monitoring of an ongoing trial. If equipoise is conceived of as a perfect balance of alternatives, it is clear that the slightest element of accumulating information will be able to tip the balance, thereby rendering the trial unethical. Critics of equipoise argue that the notion of equipoise is useless because, once the trial is started, equipoise cannot be invoked any longer to justify trial continuation (Miller and Joffe, 2011).

As we have discussed so far, equipoise is considered a necessary condition for a clinical trial to be ethical. A trial can be instated if the dedicated surveillance body -the Institutional Review Board or IRB- establishes that a state of equipoise is indeed present concerning the trial's research question. However, as soon as the study has started and data begin to accumulate, equipoise seems to be no longer sufficient to adjudicate the ethical acceptability of the trial. The problem is twofold. On the one hand, there is a violation of therapeutic beneficence that equipoise seems incapable to justify. In fact, small differences in effect will inevitably appear between the two arms of the trial and the epistemic state of uncertainty that made the trial ethical in the first place will cease to hold. At that point therefore we cannot invoke the epistemic state of the medical community to justify the fact that patients in the inferior arm are kept on a treatment that is performing worse. In late phases of a large trial, when the balance of evidence has shifted considerably in favor of one treatment, patients in the inferior arm are being offered a treatment that is less effective, even if not conclusively so. The impression that these patients are being harmed is even more forceful. The problem is described by Joffe and Truog (2008) in the following terms, as the "consequentialist concern that randomization may require assignment of some participants to therapy that is likely to be inferior, even though the preliminary evidence supporting that judgment falls short of conventional standards of methodological rigor" (p.247). The cognate issue has to do with informed consent, and in particular with the fact that the condition of substantial uncertainty

under which consent was obtained starts to change as soon as trial data begin to accumulate. Should the investigators rely on the original consent, even though it does arguably not cover the new situation? Or should they ask for a renewal of consent each time new results become available? In either case, there seems to be no use in relying on equipoise, since this principle does not support in a clear manner neither of the possible courses of action. The problem is that, once trial data begin to accumulate, a difference in efficacy will emerge, even though it is not yet marked enough to displace the epistemic indifference that existed in the clinical community. Continuing the trial in this situation seems then to violate therapeutic beneficence towards patients that are on the inferior arm, and it seems that this violation can no longer be justified by appealing to a state of uncertainty.

Is this objection to equipoise serious enough to jeopardize the adequacy of this framework? On one hand, this objection captures the fact that equipoise is mostly conceived of in the literature as a static concept –the image of a balance or scale is frequent in the literature in reference to this concept. At the moment of the deciding the trials' acceptability, the IRB has to look at the scientific question that the trial is meant to answer. Only the current epistemic state is relevant to the IRB's decision about whether equipoise is present. However, once the trial has started, it begins to produce information that has a moral import. Equipoise in the present conception is not well suited to account for this dynamic perspective on the ongoing trial.

On the other hand, though, I believe that the objection to equipoise based upon its behavior in the context of monitoring doesn't actually cut as deep as it may appear. This is because this objection is founded upon methodologically naïve assumptions. The first of these assumptions was denounced already by Freedman in his original paper about clinical equipoise (1987, p. 141):

> Late in the study — when P values are between 0.05 and 0.06 — the moral issue of equipoise is most readily apparent, but the same problem arises when the earliest comparative results are analyzed. Within the closed statistical universe of the clinical trial, each result that demonstrates a difference between the arms of the trial contributes exactly as much to the statistical conclusion that a difference exists as does any other. The contribution of the last pair of cases in the trial is no greater than that of the first. If, therefore, equipoise is a condition that reflects equivalent

> evidence for alternative hypotheses, it is jeopardized by the first pair of cases as much as by the last. The investigator who is concerned about the ethics of recruitment after the penultimate pair must logically be concerned after the first pair as well.

As a second thing, the position that trial continuation violates beneficence towards in-trial patients is liable to the same kind of criticisms that were raised against the dichotomy view, namely, that interim results of a trial do not generally represent a valid guidance for treatment choice. A sentence like “Interim findings should be presented to patients in order to ensure that no one receives what seems an inferior treatment” (Hellman and Hellman, 1991, p. 1587) entails that, when a treatment ‘seems’ inferior, we should act under the presumption that it is. However, relying on incomplete evidence for grounding our ethical conduct seems an unwise course of action to take, since the interim result could well represent a transient fluctuation in the data. This objection is crucial in that it brings the close connection between equipoise and the epistemic standard to the spotlight.

In routine settings, this connection is codified into a set of procedural rules. The decision to stop a trial is overseen by a dedicated body, the Data Monitoring Committee (DMC), also referred to as Data and Safety Monitoring Board. The DMC is entrusted with the responsibility to monitor the accumulating data and to take the decision about early stopping. But what is exactly the extent of the moral responsibility of DMCs? Some authors argue that the role of DMCs is that of concealing equipoise-breaking results from the public until the result can be declared scientifically valid. This view is held, for instance, by Joffe and Miller (2012) who denounce “The practice of routinely withholding information from the expert clinical community to preserve clinical equipoise”. There is an alternative view of the moral responsibility of monitoring committees, which I personally find more sensible. This view acknowledges that the individual equipoise of participating physicians is liable to be dispelled by minor trends; the role of the DMC then is that of guaranteeing that the trial would go on until the trend is strong enough to dispel collective (clinical) equipoise. This reading is supported by Beauchamp and Childress (2009): “A [Data Monitoring] Committee will likely decide that clinical equipoise must have been eliminated from the perspective of the expert medical community” (p.322) and later on “A data and safety monitoring committee will [...] end the trial when statistically significant data displace clinical equipoise” (p.323). Adopting this view, the DMC’s role is that

of deciding when the accumulating evidence is sufficient as to consider equipoise dispelled by it.

The reason why I consider this view particularly interesting is that it sanctions the identification of the equipoise-breaking result with a result that leads to early stopping at a predefined level of significance. The statistical criterion of significance is used in this context as a safeguard against the possibility of stopping the trial on a random fluctuation and therefore faultily concluding superiority. Of course, the criterion does not warrant certainty to the claim of superiority, but at least it gives a grounded presumption that can be acted upon. More importantly, significance represents the epistemic criterion shared in the medical community which can be used to decide unambiguously whether equipoise is still present or not.

As I have discussed in chapters I and II, trial monitoring is made possible in frequentist terms by the technique of sequential analysis. This methodology creates the possibility of achieving a statistically significant result before the planned end of the study. Therefore, when a trial is monitored using a sequential plan, equipoise can be broken at every step of the sequential analysis. From the point of view of the ethical conduction of an ongoing trial, this means that monitoring multiplies the chances to break equipoise. As Edwards et al. (1998) put it, "any trial has the potential to destroy the equipoise that it relies on to be ethical". When considering the decision to stop, DMCs rely on statistical stopping rules and significance stopping boundaries as guidelines. As discussed in the previous chapter, the methodology for efficacy monitoring requires a low *p*-value –lower than the advertised significance level– as a condition for stopping early. The reason behind this requirement is twofold: on the one hand, in this way the analysis can maintain the advertised significance level and, more importantly, the declared value of the type I error rate. The second rationale is more behavioral than methodological: a 'tough margin for stopping' like that invoked by Pocock (1993) is considered by most frequentist biostatisticians to warrant results that are also convincing to the medical consumers of statistics. Pocock, for instance, claims that "*p*-values [as high as] 0.05 are simply not persuasive enough to make a new treatment widely accepted" (1992). Most (frequentist) biostatisticians maintain, like Pocock, that a low enough *p*-value makes for a convincing result and therefore is a guarantee of acceptance of results in the medical community.

In the upcoming section I am going to discuss a problem, the monitoring paradox, which

reveals this view as misguided. Furthermore, the paradox reveals a genuine weakness of equipoise in the context of monitoring. Actually, as I will eventually claim, the monitoring paradox does not go so much against the notion of equipoise in itself: rather, it signals a fault in the *epistemic* connotation of equipoise as clinical uncertainty defined in narrow statistical terms.

3.3.3 The monitoring paradox

Monitoring multiplies the chances to break equipoise before the planned end of the trial. When a trial stops early due to manifest superiority of one treatment, practicing physicians have a reason to endorse the new treatment, in light of the large benefit that has expectedly prompted stopping. In recent years, however, a series of articles published in top medical journals (Montori et al., 2005; Wilcox et al., 2008; Bassler et al., 2008, 2010) have cast a shadow of doubt on the reliability of results from early stopped studies. I have already introduced these studies in the previous chapter, section 2.3. These papers present empirical work in support of the claim that results from trials stopped early for benefit provide biased estimates for the treatment effect; consequently, they suggest that these trials are unreliable. For instance, Montori and colleagues suggest that "clinicians should view the results of such [foreshortened] trials with skepticism" (p.2209), while Bassler and colleagues conclude that clinicians should presuppose "appreciable overestimates of effect in trials stopped early" (2010, p. 1187). The point raised by these studies is of extreme relevance for the medical community because it questions in a direct way an established source of medical knowledge. Empirical proofs of bias like those contained in the studies mentioned propagate the idea that results from early-stopped trials are unreliable because they overestimate the treatment effect. Indeed, all the studies mentioned go on to suggest that medical readers should treat result from these trials with skepticism. This skepticism is, however, at the source of a problematic situation.

Let us suppose that at some point in the monitoring of an ongoing trial, the Data Monitoring Committee decides that the information accrued is enough to break equipoise and consequently it would be unethical to maintain in-trial patients on the treatment that is performing worse. Once the trial result is published, however, it is regarded with skepticism. If medical professionals do not regard evidence from the trial as reliable, this means that the state of uncertainty in the medical community remains unresolved. Does this imply that the

DMC was wrong in claiming equipoise to be broken? Or is equipoise in fact dispelled, and the practicing physicians are behaving unethically in denying the new treatment to their patients? The bottom line is that trials stopped early because equipoise was broken may in fact fail to remove the very uncertainty that is at the basis of the accepted definition of equipoise. I will refer to a situation of this sort as the *monitoring paradox*.

The paradox is implicit in a methodological paper on monitoring (Fayers et al., 1997):

> If the early results [of the trial] provide reasonably conclusive evidence of an advantage in favour of one of the treatments, it may be considered unethical to continue recruiting patients. Superficially, therefore, it might appear that clinical trials should be terminated as soon as there is a convincing and statistically significant difference between the treatments. Nevertheless, there is an opposing school of thought which argues persuasively that the role of a clinical trial is to influence clinical opinion and clinical practice. Thus if a clinical trial detects a large treatment effect after half the patients have been entered, and as a consequence is terminated early, that trial may be received with considerable scepticism by clinicians; despite any significant p-values that are cited, many clinicians may still remain unconvinced by the weight of evidence that has been produced. (p. 1414)

The main underlying cause of the paradox may be traced back to methodology because, if we adopt the identification of statistical significance with equipoise breaking that has been described above, there is no way of making sense of it. It seems that the physicians in the community who refuse to endorse the significant result are simply behaving irrationally. Actually, however, statistics is less binding for DMCs than what many bioethicists think: As Ellenberg et al. (2002) describe, "[t]he process of making such judgments is rarely straightforward; the observation of a low p-value or a sequential boundary crossed represents only the beginning of this process, not the end" (p.42). On the other hand, if we endorse a more general definition of clinical equipoise as the epistemic state that prevails in the community, the fact that the trial achieved significance becomes irrelevant. This, however, implies that in the situation envisaged in the monitoring paradox the DMC's evaluation was simply wrong. At first sight, then, the monitoring paradox cuts both ways and it makes equipoise untenable.

3.4 Redefining equipoise: The role of statistics

The monitoring paradox represents a problem for the notion of equipoise because it reveals an apparent inconsistency in the definition of this concept. Indeed, as we have seen, clinical equipoise can be defined either as a state of reasonable disagreement in the community or, more formally, as the lack of a statistically significant result that can adjudicate the disagreement. In the situation which leads to the paradox, these two definitions turn out to be incoherent with each other, giving rise to contradicting ethical indications. However, this need not be the case. In this section I put in place a twofold strategy aimed at dissipating the paradox and at establishing equipoise more firmly as a workable principle in the context of trial monitoring. Firstly, I will argue that equipoise is better understood as a mid-level principle in ethical theory (Rachels, 2009). This means that equipoise should not be conceived of as a formally defined condition that must apply for clinical research to be ethical. Rather, it is more adequately understood as a ideal principle that should guide researchers, similar to the 'do no harm' precept of medical ethics. As a second thing, and more importantly, I will propose that it is necessary to revise the definition of equipoise in probabilistic terms: The identification of a statistically significant result with a equipoise-breaking result is too tenuous to hold.

The monitoring paradox highlights first and foremost a difficulty with the delimitation of the concept of equipoise which exists in the medical literature. For instance, Sackett (2000) denounces equipoise as "lacking clinical reality" and "still lacking a precise definition". Indeed, the monitoring paradox stems from a conflict between two coexisting definitions of equipoise: a narrow an procedural definition in terms of lack of a statistically significant result, and a more comprehensive account in terms of the epistemic attitude in the community. In my opinion, however, insistence that equipoise should be nailed down to a precise and procedural definition is misguided, because equipoise is more adequately described as a mid-level principle. 'Mid-level' denotes that a moral principle is situated in an intermediate plane in between more fundamental ethical considerations, such as the Utility Principle or the Categorical Imperative in Kantian ethics, and more practical moral rules. Mid-level principles are valuable in ethical theory because different normative systems can converge on mid-level principles. Such principles can then guide and adjudicate moral conduct even in situations where different normative theories appear irreconcilable (Krom, 2011).

Conceptualizing equipoise in these terms goes a long way towards addressing the various criticisms of equipoise considered in section 2.2. What is relevant in relation to the monitoring paradox is the fact that a mid-level principle is not liable to a definition in terms of a criterion or a procedure, just as much as respect for the autonomy of a patient is not *defined* in terms of the consent form she has signed. Thus, invoking the status of a mid-level principle for equipoise means that it need not be tied to the particular epistemic state of any of the actors involved –the DMC, or the medical community. In the context of trial monitoring, this means that even though different agents may choose different criteria as appropriate for identifying equipoise, the *principle* of equipoise remains the unequivocal identifier of ethical trial conduct. A definition of equipoise (as, say, "trials are justifiable when there exists a state of honest, professional disagreement among expert clinicians about the preferred treatment") might clarify the general norm for instating a RCT, but it would not narrowly determine its defensibility in the light of the inconclusive evidence of ongoing trials. Thus, the monitoring paradox cannot be used as an argument against equipoise conceived of as a mid-level principle.

A mid-level principle, however, is incomplete as long as the norms and methodologies needed to implement it in the practice remain ill-specified. I claim this is the situation at present for clinical equipoise. Indeed, a conclusion that can be derived from the monitoring paradox at the methodological level is that the narrow identification of a equipoise-breaking result with a statistically significant result would not work. There is the need for a more refined representation, in probabilistic terms, of the relevant clinical uncertainty. Recognizing equipoise as a mid-level principle entails that there is space for methodological work aimed at redefining the notion of uncertainty at the basis of equipoise.

3.4.1 Redefining clinical uncertainty

In a complex and thorough article, Ashcroft (1999) reviews the ethical importance of clinical uncertainty or indifference for trial defensibility, and points out the importance of value-laden aspects of the evaluation process. Through the paper Ashcroft analyzes the worth of objective Bayesian and of theoretic approaches to the problem of clinical uncertainty and dismisses all of them as inadequate. The inadequacy of what he identifies as "attempts to replace political judgment with inductive logic" stems in Ashcroft's view from the fact that any decision theoretic approach is ultimately unable to effectively incorporate the political dimension of the

many judgments that are involved. For instance, statistical accounts of clinical uncertainty based on different weights assigned to the probability of committing various types of error reveal themselves as inadequate, in Ashcroft's view, as there exists no objective account of seriousness of error.

What Ashcroft fails in my opinion to recognize is that the rational uncertainty he refers to is already at the present state being given a formulation in terms of probability, as I have discussed in the first part of this Chapter. Avoiding the use of a statistical or probabilistic concept of clinical uncertainty is simply not an option, as indifference between competing treatments is itself characterized epistemically. For this reason, I follow Hansson (2006) in attempting a probabilistic definition of the relevant clinical uncertainty underlying equipoise.

Hansson presents a situation where there are two treatments: The well-established standard therapy *a* and the newly proposed *b*. It is reasonable to maintain that low uncertainty surrounds the effect of treatment *a*, the standard which is already in place, while relatively large uncertainty surrounds the true value of the effectiveness of treatment *b* that is the experimental drug that has yet to be tested. According to Hansson, the clinical uncertainty that can justify a trial comparing *b* with *a* exists to the extent that the possibility that *b* is a better therapeutic option than *a* can reasonably be entertained in light of the state of knowledge before the trial. This uncertainty can be represented as the overlap between the two probability distribution functions for the value of the effect, respectively, of the experimental treatment and the control, as depicted in Fig. 3.a.

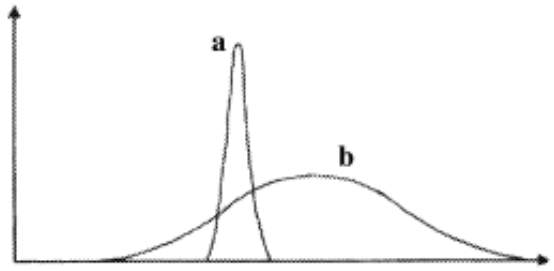

Figure 3.a: Statistical definition of equipoise in Hansson (2006): The overlap in the probability distribution functions for the effect of treatments *a* and *b* represents the relevant clinical uncertainty that makes a trial comparing the two justifiable. From Hansson (2006, p. 157)

The framework proposed by Hansson makes it very intuitive to connect the notion of clinical uncertainty with our present state of knowledge concerning the two treatments. However, in my opinion, Hansson's account as it stands is inadequate to deal with the issue of clinical

uncertainty in the context of monitoring which is the main concern of this chapter. This is due chiefly to two reasons. As a first thing, Hansson's definition applies to the equipoise that may or may not be present at the beginning of a trial. Therefore his definition of relevant clinical uncertainty is meant to represent a static quantity, and Hansson does not provide an explicit treatment of situations where this quantity is allowed to change over the course of a trial. Furthermore, Hansson defines clinical uncertainty starting from the two probability distributions relative to the effect size of the two treatments. In Hansson's reconstruction, clinical uncertainty can be identified as the overlap of the two probability distributions. This reconstruction is incomplete in that it is unclear, if we take uncertainty to be expressed in those terms, how it can be put in relation with the quantities that are relevant at present in constructing medical evidence from trials: the *p*-value, the significance level and the rates of type I and II error in the Neyman-Pearson hypothesis test. Ultimately, we see that this difficulty is due to the fact that Hansson relies on a *subjective* interpretation of probabilities while we have seen that the statistical treatment of trials is dominated by frequentism. Indeed, Hansson observes (p. 155):

> For every specified treatment (and patient group) there is, in principle, a "true" probability distribution that corresponds to ideal knowledge about the relative frequencies of different treatment outcomes. In practice, we only have access to estimates (subjective probability judgments) that can be based on anything from extensive previous clinical trials to uncertain interpretations of animal experiments.

Clearly, the first interpretation referred to by Hansson is a frequentist interpretation, while the second one –which he adopts– is Bayesian. In order to constructively rely upon Hansson's insightful proposal, it is then necessary to translate the probabilistic notion of the relevant clinical uncertainty it in frequentist terms. Only in this way it is then possible to confront it with relevant quantities in frequentist analysis, and eventually make sense of the fact that significance tests are able to dissipate this uncertainty.

The conventional way to assess which between two treatments is superior is through a statistical test of significance upon a variable which expresses the difference in performance between the experimental and the control group. This variable, θ, may be for instance the odds ratio for the two treatments in question. The odds ratio (OR) has been introduced in chapter I and it is essentially a measure of how more (or how less) likely it is for a patient to

experience a certain outcome –for example, disease recurrence– under the new intervention, with respect to the control.

As it was discussed in chapter I, uncertainty in frequentist terms can apply to possible outcomes of the trial given a value for the difference in effect, but it cannot apply to the value of the difference itself. In order to understand how this uncertainty may be characterized in univocal terms in order to determine whether equipoise is present or not, it is useful to observe how uncertainty is resolved by means of the significance test. As we have seen in previous sections of this thesis, clinical trials are designed as frequentist hypothesis tests to decide between two candidates: One, H_0 or the null hypothesis that there is no difference in effect between the two treatments, and the other, H_A, that there is a difference equal or greater in magnitude than a certain pre-set extent that is of clinical relevance. In terms of the previously defined parameter, the odds ratio, the two hypotheses can be defined as $H_0 : OR = 1$ and $H_A : OR \geq \delta$ where $\delta > 1$ is a certain pre-designated, clinically relevant amount. This situation is depicted in Figure 3.b. Notice that the two distributions H_0 and H_A have different

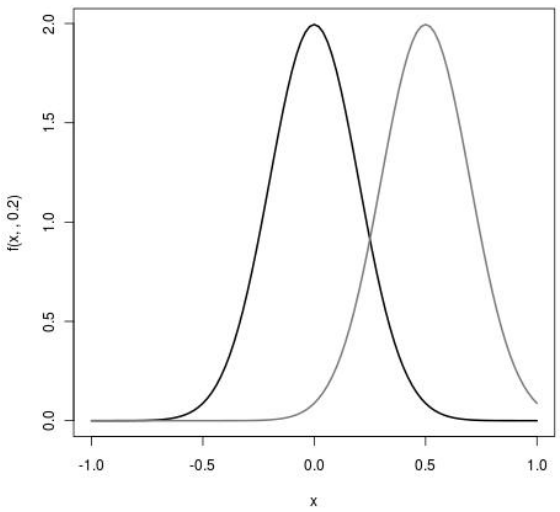

Figure 3.b: Clinical uncertainty in frequentist terms: The two probability distributions of H_0 (black) and H_A (grey)

means but they have the same width. This is because they represent different objects from *a* and *b* in figure 3.a. In figure 3.a, *a* and *b* summarize our state of knowledge about the value of the effect: We are more or less certain about this value according to the amount of information about it provided by prior knowledge about the treatment. As Hansson notes,

"The new evidence obtained in a clinical trial will typically justify a probability distribution for treatment outcomes that is narrower than the previous estimate". Therefore the distributions depicted in figure 3.a are likely to get progressively narrower as the trial proceeds.

Instead, H_0 and H_A in figure 3.b represent the distribution of the probability of observing various trial outcomes, under the hypothesis of no difference in treatment, or of superiority of the new, respectively. The information that accrues as the trial proceeds is not going to modify the way H_0 and H_A look like: Rather, this information will determine a definitive choice between the two. Let us bracket for a moment the issue with the correct interpretation of a frequentist test and let us concentrate on the test itself as a means of deciding, with a certain control on the possibility of committing a mistake, which one of two alternatives is the case. It becomes then clear that the clinical uncertainty that we are interested in, and that must exist for the trial to be ethical, can be expressed as uncertainty between the two alternatives H_A and H_0. Note that this differs from Hansson's proposal in that the uncertainty is about a choice between two probability distributions *over possible observations* rather than being represented as the overlap of two probability distributions *over the value of the effect*. Clinical uncertainty is present at the beginning of a clinical trial because the the comparison index between the two treatments is spread out and therefore we are not able to distinguish epistemically which of the two alternatives is the case. Clinical uncertainty is dispelled by a clinical trial when the data enable us to conclude for one of the alternatives.

We are now ready to see the connection between clinical uncertainty thus defined and confidence in the result expressed via traditional frequentist measures. As I have discussed at length in previous sections of the thesis, the frequentist test is calibrated to decide between H_0 and H_A with the desired pre-specified probability of committing a mistake. The confidence in adjudicating uncertainty is evaluated via the classical pre-experimental error rates α and β. But then, if the results of clinical trials are considered a valid means of breaking equipoise by dispelling the underlying uncertainty, this means that conventional error rates must play a relevant part in qualifying a result as equipoise-breaking.

In other words, the characterization of a equipoise-breaking result requires not only statistical significance but also a measure of credibility. When using the classical test this is implicit, because results which are significant come with a certain statement of confidence. However, as I am going to argue in the concluding section, classical, unconditional error rates remain

liable to the monitoring paradox: Therefore, they do not represent adequate tools for defining the relevant clinical uncertainty at the basis of equipoise.

3.5 Clinical equipoise, statistical uncertainty and the monitoring paradox

As discussed through this chapter, what makes the debate about the ethics of trial monitoring so complex and so interesting is the tight intertwining of ethics and epistemology which this debate brings to light. The ethics of stopping a trial earlier than planned cannot be easily adjudicated because the scientific validity of the trial result at any stage is a relevant factor in deciding whether the trial can be ethically allowed to continue.

The dichotomy view reviewed in this chapter has been often taken as grounds for arguing for the ethical superiority of Bayesian methods in the design of clinical trials (Palmer, 1993; Berry, 2004). The position is that Bayesianism is more ethical in that it allows a more efficient use of the data arising from patients participating in the trial. This has been illustrated also in the second chapter, in discussing the issue of monitoring. However the dichotomy framework overlooks the complexity just mentioned because it rests on the false premise that stopping early is always in the best interest of participating patients. In other words, the dichotomy view wrongly sanctions the idea that stopping a trial early is a value *per se* and that, therefore, Bayesian methods are more ethical because they can afford shorter trials.

The concept of clinical equipoise provides instead, as I have argued, the adequate ethical tools to tackle the ethics-epistemology interplay that comes to the spotlight in the context of trial monitoring. Equipoise establishes an identification of the ethical criterion that makes a trial acceptable with the epistemic criterion of (lack of) competent consensus in the scientific community. Thus, it becomes clear that some element of *credibility* or *confidence* in the conclusion must enter the picture. This consideration reveals that the evaluation of the ethical adequacy of a statistical framework is a more complex affair than what the uphelders of the dichotomy view maintained. Thus, the judgement extends beyond the framework's eventual efficiency to arrive at a result quickly, to include some measure of confidence in the result it commends.

The current (unconditional) frequentist framework associates a fixed measure of confi-

dence to all results that were achieved following stopping rules specified pre-data. However such measure of confidence does not seem to capture the epistemic standard that is shared by the medical community, as the monitoring paradox testifies. As we have seen, in order to dispel the relevant clinical uncertainty that made the trial ethical, a trial result must possess two features. Not only it must tip the balance in favor of one of the hypothesis, but it must also do so with a weight of evidence that offers adequate confidence in the conclusion. In the case of monitoring through a sequential plan, the situation which gives rise to the paradox, the stopping rule of the sequential design warrants that both conditions will be verified. In fact, it will only sanction stopping in case the nominal significance level is preserved, which means warranting the pre-defined level of confidence. However, stopping under these conditions leads to likely overestimation of the true effect of the treatment, for the reasons that I have discussed in chapter II, section 3. When medical readers are confronted with a trial that was stopped early, they see a result which is likely an overestimate, and which furthermore is arising from a patient sample which is smaller than was initially planned. It is reasonable to maintain that, in this situation, medical professionals may not be satisfied with the standard confidence level. Viewed in these terms, it seems that the uneasiness in the medical community is not so much about stopping early, but about trials with implausibly large effects – effects which require, in the words of Mueller et al. (2007), "astute clinicians" to make an appropriate interpretation of the results.

In the previous chapter I have described how conditional –as opposed to unconditional– measures of error provide a better guidance for interpreting results from trials stopped early. Conditional error rates are sensitive to the strength of evidence in the data and at the same time they will generally reflect the fact that the conclusion is based on a reduced amount of data. On these grounds, I have concluded that the adoption of conditional error rates can dissipate the skepticism about early-stopped trials. The conclusion drawn in chapter II was however restricted to the epistemic attitude of the medical community toward trials that stop early. Here, in closing this chapter, I will examine this conclusion with regards to its ethical import on the issue of monitoring under equipoise.

First of all it is necessary to elaborate the *conditional* version of the statistical definition of clinical uncertainty considered above. The conditional test is based, likewise the classical frequentist test, on the comparison of two alternatives. The only difference which is relevant

to the present purpose is that the probability of error, or in other words the confidence in the result, will also be reported conditionally. This difference is, however, crucial. Indeed, as we have seen in the previous chapter, conditional error rates can greatly differ from the unconditional ones: In the example presented in section 2.5.1, the post-data error probability at two subsequent analysis of a ongoing trials were 7% and 14%, respectively. These probabilities have to be confronted with the nominal 5% confidence level warranted by the classical analysis. The discrepancy between conditional and unconditional error probabilities has an import in terms of clinical uncertainty – and therefore on equipoise.

This is evident once we consider the definition, given in the previous section, of clinical uncertainty as lack of a result which can warrant discrimination between two alternative hypotheses with a desired level of confidence. In the case just mentioned, if we consider a confidence of 5% to be the threshold, uncertainty is resolved at the end of the trial or it is not according to whether we consider the unconditional confidence level (5%) or the conditional error rate (7%). Clearly, the monitoring paradox is more complex than what this simple example pretends to illustrate. The paradox involves clinical disagreement more than mere uncertainty, and therefore it entails a behavioral dimension which cannot be ignored and can't be reduced to a purely statistical formulation. However, as far as the epistemic dimension of the paradox is concerned, we see that we can make sense of the fact that clinicians outside the trial are unconvinced by the result, even though it is statistically significant: Possibly, they are just evaluating their confidence in the conclusion *conditionally*.

Clearly, as long as conditional error assessment does not become familiar to medical practitioners, there is no way to ground this claim. My point is, however, that a definition of clinical uncertainty which relies on conditional quantities instead of unconditional ones may be less easily liable to the monitoring paradox. This is because, if we agree that the skeptical attitude towards trials that stop early is caused by distrust towards extreme effect and small sizes, a conditional measure is able to take these into account; a unconditional one isn't. Thus a conditional criterion for defining clinical uncertainty stands a better chance of resolving the monitoring paradox and of providing an adequate methodological grounding to the principle of equipoise.

A further thing to be noted is that, while the unconditional confidence level is a fixed quantity that is only warranted if the data cross a global or sequential boundary of significance, the

conditional quantity varies as the data accrue and it can be calculated at any point in the trial. This is because the conditional test obeys the likelihood principle and is therefore independent of the stopping rule, as discussed in section 2.5. For this reason, a definition of clinical uncertainty in terms of conditional confidence is also better able to account for the dynamical dimension of equipoise that we have seen to be badly represented by the unconditional definition.

A final remark concerns the possible opportunities afforded by a conditional interpretation of the clinical evdence that can dispel equipoise. In the present unconditional framework the decision to stop a trial early concerns stopping with a fixed level of confidence and hence it is an "all or none" decision. Instead, evaluating confidence conditionally makes it possible to envisage a scenario where the acceptable level of confidence is *variable* across trials, depending on factors such as the severity of the condition or the pace of the response to treatment. Such contextual connotation of the notion of equipoise is intriguing, because it seems to capture an intuition which is underlying the dichotomy framework – that what level of uncertainty is acceptable varies with disease- or treatment-specific circumstances. Though we need to conduct trials on human patients for obtaining evidence needed to guide treatment choices, the possibility to negotiate such need for evidence down to the level where clinical trials are conducted provides space for moral reasoning and thus, potentially, a way to alleviate the most painful situations that may arise in the context of clinical research. Thus, conditional inference represents a promising and feasible possibility for more ethical clinical trials.

The objective of this chapter was that of examining the ethics of randomized trials in its connection with statistical methodology. When discussing the merits of the alternative statistical schools, the claim is often made that Bayesianism is more ethical in that it allows a more efficient use of the data arising from patients participating in the trial. This position is, however, grounded upon an inadequate ethical framework, the dichotomy view. In the first part of the current chapter I have argued that the ethics of trials should not be represented as a tension between conflicting interests where efficacy of the statistical methodology is paramount; rather, it should be clear that *uncertainty* about the preferred treatment is the more appropriate grounding for the ethics of RCTs. Thus, the discussion contained in this chapter clarifies the role of statistics as a gatekeeper in sanctioning the ethical permissibility of a clinical trial. Rather than balancing the conflicting interests of current and future patients as envisaged in

the dichotomy view, the ethical role of statistics is better understood as adjudicating whether uncertainty exists at the beginning of a trial, and deciding whether and when it is dissolved.

Within the ethical framework that is currently endorsed uncertainty about the preferred treatment, under the name of equipoise, plays a fundamental role in the identification of the conditions that make trial initiation and continuation ethical. The *methodological* connotation of this uncertainty has however been rather sloppy, especially in relation to the statistical dimension. Along this dimension, a superficial identification of equipoise with lack of a statistically significant result has prevailed. This identification is however problematic, particularly in the context of trial monitoring where the situation labeled as the monitoring paradox arises. As I have discussed in the second part of the chapter, the monitoring paradox surely stems in part from the confusion surrounding the definition and the proper understanding of the concept of equipoise. On the other hand, though, the shortcomings of unconditional inference, which I have already confronted in the previous chapter, play a relevant part in the genesis of the paradox. Unconditional inference complicates the appraisal of the strength of the conclusion and thus it fuels skepticism towards results of trials that stop early. The conclusion of this discussion is that, even though the ethical framework based on equipoise does not mandate a specific statistical treatment for defining the underlying relevant clinical uncertainty, a *conditional* framework should be preferable. But is the idea of implementing such a framework a realizable possibility, within the current milieu of medical research? In the upcoming and final chapter I am going to analyze the outlook for trial methodology in light of some of the most recent trends in drug research, paying particular attention to the perspective of regulatory bodies.

Chapter IV:

Perspectives: Moving past Unconditional Trial Design

In the course of this work I have analysed some issues connected with the choice of the statistical framework in use for clinical trials. As reviewed already in chapter I, over the past decades RCTs prevailed over clinical judgement, case report and observational studies as evidential standards in medicine. What is more important, during the same time frame, RCTs became a crucial part of the regulatory process whereby a new therapeutic can gain access to the drug market. The testing of drugs through clinical trials represents a fairly recent addition to the approval process. The need itself for pre-market approval of new compounds did not exist until the 1960s and as far as the 1970s in several countries decisions about approval were essentially taken on a case-by-case basis. The approval pipeline as we know it, with its elements of objectivity and universality enshrined in a fixed set of procedures, is a fairly recent construct.

As I will discuss in this chapter, it is the social role played by RCTs within the context of this construct which places the strongest constraints upon the choice of the statistical framework. Indeed, the epistemic role ascribed to RCTs does not prescribe any particular requirements at the level of the statistical principles involved in inference. This is proven by the fact that, as far as the *discovery* phase of clinical research is concerned, usage of Bayesian methodologies is unproblematic and it has even become the routine in the context of specific problems such as phylogenetic inference (Ronquist and Huelsenbeck, 2003).

In the stage of drug testing, however, Bayesian methods encounters still a lot of difficulties to get established. This is clearly due to the fact that in drug testing the stakes, in terms of both patients and public protection, are clearly higher than in basic biomolecular research. There is a shared perception that frequentist methods grant objectivity and impartiality at the level needed for clinical research.

In this final chapter I will confront this position with particular reference to the problem of trial monitoring, which has been a constant theme through the thesis. Monitoring and early stopping are particularly tricky from the regulatory point of view, giving the plain commercial interest of the pharmaceutical industry in shorter trials. After having introduced the current state of the regulatory landscape, I will present what are in my opinion the most important driving forces that are shaping the landscape of clinical research. Grounding upon these, I will defend the claim that early stopping, and even more extreme forms of flexibility in trial conduct, may be a necessity in the future landscape of clinical research regulation. Thus, in concluding the chapter, I will discuss the consequences of this conclusion for the subscription to a particular form of statistical inference in clinical research.

4.1 Regulators, objectivity and statistics

Nowadays we consider it natural that new drugs must gain regulatory approval, i.e. have an independent body assess their safety and efficacy, before being allowed to be purchased or prescribed. However, this idea established itself only quite recently. The drug market is characterized by an informational asymmetry: In most instances, drugs consumers are not in a position to make decisions about which drugs to use and to weigh potential benefits against risks. For this reason some form of regulation is required. However for most of modern medicine it was thought that having physicians supervise drug administration was sufficient in order to have this regulatory need fulfilled. The watershed between this way of thinking and the tightly regulated drug market of today is arguably represented by the Thalidomide scandal.

Thalidomide was a sedative and hypnotic that first went on sale in Western Germany in 1956 and in the following years was introduced in 46 different countries worldwide. In the same years a dreadful increase in frequency of phocomelia and other deformities in newborns was observed. Correlation between the two was proved in 1961 by Widukind Lenz, a German physician, prompting withdrawal of the drug from the markets of all the nations involved. It is estimated that until the drug's withdrawal, as much as 10'000 babies were born with dysmelia as a result of their mothers taking thalidomide during pregnancy. The drug had been presented as a remedy to morning sickness and therefore it was naturally prescribed

to thousands of pregnant women to relieve their symptoms. Despite knowing this, the manufacturers had not taken sufficient action to examine the drug's effects on the developing fetus. This tragic scandal had a huge impact on the public and it drove the reshaping of the regulatory system in several countries into the form we know it today.

In the U.S. the regulation of the pharmaceutical market is entrusted to a governmental agency, the Food and Drug Administration (FDA). In European countries, instead, regulation of drugs proceeded along heterogeneous trajectories in the different national states until as far as the 1990s. EC Regulation No. 2309/93 set up a communitarian body, the European Agency for the Evaluation of Medicinal Products, later renamed to the European Medicines Agency (EMEA or EMA – EC Regulation No. 726/2004). Nowadays, member states of the EU have national agencies overseeing the national market, but EMEA has responsibility over a large number of drug approvals through a centralized procedure. Even though EMEA is not an FDA-like independent regulatory agency, its decisions have considerable authority and impact at the national level. Differences, even critical, between the policies and view of EMEA and FDA exist, however they can largely be regarded as tangential to the topic of the present work. Recently, a international coordination initiative was launched with the objective to standardise requirements for marketing applications submitted in Europe, Japan and the USA. The name of the initiative is International Conference on Harmonization (IHC).

In order to understand how the RCTs assumed the current role in the regulatory arrangement, the trajectory of the American system turns out to be more relevant. FDA, the federal body appointed for the oversight of medicine industry, was formally established in 1927. However, the agency's authority over the drug market was established only in 1938 through the *Food, Drugs and Cosmetics Act* (FDC) of 1938. This legislation introduced mandatory premarket approval for new drugs, that was to be regulated by the FDA. In 1962, a significant expansion of FDA authority over drugs occurred as the Congress enacted the Kefauver-Harris amendments to the FDC. The so-called *Drug Amendments* of 1962 were a reaction to the thalidomide crisis: The fact that FDA had not granted approval to thalidomide in the US at the time of the scandal's outbreak induced a positive perception in the public towards a tight regulation in the drug market. As a result of the 1962 Amendments, the additional burden of proof of establishing a new drug's efficacy –and not only its safety– was put on drug companies. Furthermore, the FDA was granted greater autority on the oversight of clinical trials

for new drugs. If before 1962 clinical trials were essentially just a device needed to generate evidence that could guide decision-making in the clinics about a new therapeutic, after this date trials –and Phase III RCTs in particular– became an essential step in gaining approval for the new drug.

4.1.1 Statistics and early stopping of trials

The process of testing a new compound goes through several phases. The first stages, phase I and II trials, are experiment-setting trials. Phase I trials are needed to rule out unexpected toxicity of the new drug and phase II trials have the aim to optimize dosage. Decision about regulatory approval depends instead the data of phase III trials. Phase III trials are conducted on a large population of patients and in tightly controlled conditions, ideally under double or triple blinding and with randomization of participants between the arms of the study. Phase I and II, on the contrary, do not necessarily feature patient randomization and a control group. The higher stakes involved in phase III RCTs entail a higher level of methodological standardization and strictness. What is possibly less obvious is that this strictness extends to the level of the statistical framework which should be adopted. A review of regulatory guidelines of the main national and supranational Agencies (White et al., 2000) found most of the guidelines to explicitely endorse frequentist principles of testing; in cases where no explicit statement of the inferential framework was made, classical hypothesis testing approach was still implicitely assumed.

Different documents issued by the FDA, the EMA and the IHC do not give specific recommendations on which statistical methods to use. However, they do have specific requirements on what statistical information should be documented and presented. For instance, the FDA recently decided that trials for regulatory approval of medical devices –diagnostic or therapeutic devices such as RX apparatuses or prostheses– can be presented with data analysis that use Bayesian methods[†]. However information on the frequency properties of the design remains crucial: The recommendation to investigators is that “in adherence to regulatory practice, FDA recommends you provide the type I and II error rates of your proposed Bayesian

[†]This openness towards Bayesian methods in the case of medical devices has more to do with the technology than with statistics. As pointed out by Spiegelhalter (2004), medical devices are typically developed in incremental steps and a large body of relevant evidence is usually available. On the other hand, proper assessment of the quality and relevance of prior information is much harder in the case of drug trials. This aspect will be confronted more in detail later on

analysis plan" (FDA, 2010b, p. 29). A few lines afterwards we read "FDA considers type I error, along with other operating characteristics of the trial design, in evaluating submissions. We strive for reasonable control of the type I error rate".

The focus on pre-experimental quantities on the part of the regulators is perfectly understandable. The classical Neyman-Pearson methodology of hypothesis testing was conceived with an eye to the problem of quality control in the industry, as described in chapter I. The pre-experimental error rates inform us about the probability of a mistake in several application of the same testing criterion. Regulatory agencies need to decide about the approval of hundreds of new drug candidates each year. It is then clear that the kind of information provided by the unconditional error rates is well-suited to the perspective of the regulator. As we have seen through Chapter II, however, the classical frequentist measures reveal inadequate for the purpose of properly evaluating trials that were stopped before their planned conclusion. This is a significant issue, given the clear interest of pharmaceutical companies in the possibiity of grounding claims of benefit upon shorter, less expensive trials. It is well known that the agenda of pharmaceutical companies does not always reflect the set of priorities and needs of society Psaty and Kronmal (2008). This mismatch is explainable in terms of the different and potentially conflicting sets of ends for which society and private corporations pursue clinical research. While the primary aim of society is that of obtaining generalizable knowledge to benefit future patients, the primary (and legitimate) aim of pharmaceutical industries is, instead, that of making profits and to increase their market share. The role of regulators is that of ensuring that the legitimate mismatch of interest does not result in an illegitimate divergence, whereby drug companies pursue treatments that are of no real value for society (Huskamp, 2006). "Me-too" drugs –drugs that are commercially but not pharmacologically different from already known and proven treatments– provide a clear illustration of the kind of drug research that should be discouraged.

On these grounds, the doubt might arise that early stopping should be discouraged because of the possible misuse of this technique. In the upcoming section, however, I will provide some material that points to a different direction.

4.2 Drivers of change

So far the regulatory construct embedding the unconditional inferential measures has surely done its job of protecting patients and payers from harmful or ineffective compounds. However new elements are in sight which could significantly affect the drug regulation landscape that I have just described. The first of these elements is pragmatic: A relevant public pressure has mounted in recent years around the regulatory architecture, calling for a greater efficiency in the trial process. The second aspect, instead, is epistemic in content, being related with the emergence of the concept of *personalized medicine*. Let us review these factors in turn.

4.2.1 The valley of death

In a spotlight 2008 commentary, Butler denounced the "valley of death" gaping in between basic biomedical research and clinical applications of it. There is, according to Butler, "a growing perception that the enormous resources being put into biomedical research, and the huge strides made in understanding disease mechanisms, are not resulting in commensurate gains in new treatments, diagnostics and prevention." (Butler, 2008, p.840).

Butler's article voices a concern that is shared by many. This so-called translational gap is surely a multifaceted problem. The pre-clinical process, i.e. the process that lies in between a compound behaving promisingly *in vivo* and a drug which can be administered to human beings, is fraught with obstacles: The compount may not be easily synthesizable, or its action may be different in human patients than what it was on the animal model, and so on. Finally, once the development process is through, the drug must still undergo regulatory approval. The cost-efficacy of a tight regulatory system such as the U.S. one has been a matter of contention since the 1970s. Less a decade had elapsed since the 1962 Amendments when scholars holding libertarian views began to challenge the regulatory action, in many cases supporting their claims with empirical data about the drug market and health statistics in different countries. Among the most notable of such criticisms, Peltzman (1973) pointed out that the high costs that pharmaceutical firms have to bear in order to complete the pre-market approval process affects the rate of innovation and the availability of new pharmaceuticals. The idea is that climbing R&D costs discourage companies in making investments in the research of new therapeutics, while high costs of the approval process creates a bottleneck

whereby fewer drug candidates can afford to be tested. Peltzman's argument has not lost its force over the years: The existence of such 'regulatory bottleneck' continues to be blamed for the progressive shrinking of investments from the private sector (Cressey, 2011). A second line of criticism concerns the *drug lag* (Wardell and Lasagna, 1975; Kaitin and Brown, 1995) induced by tighter regulation, i.e. the observation that new medicines become available more quickly in drug markets less regulated than the U.S. one. Even though to an extent the drug lag is inevitable if some form of third party oversight over the drug market is desired, critics argue that the costs in terms of number of lives that could be saved by making the new drug quickly available far exceeds the benefit in terms of deaths prevented by regulatory action (Gieringer, 1985). Stewart et al. (2010) re-estimated this cost-benefit ratio using current data about life-expectancy and average improvement in cure-rates of advanced lung cancer and they concluded that "If it were conservatively estimated that cured patients have a life expectancy of just 5 years, then the 5-year delay in the advance would mean 54,860 life-years lost in the United States attributable to patients who could have been cured but were not" (2010: p. 2928). This figure should be compared, according to Stewart and colleagues, with a mere 16.3 life-years gain that results from preventive effect of regulatory action. In evaluating the impact of such claims, it should be considered that big players in the pharmaceutical industry are not the only stakeholders who have an interest in a leaner approval pathway. Patient advocacy organizations, who have a comprehensible desire for new drugs to be approved rapidly, can represent as powerful an interest group, as the history of the antiretroviral drug AZT in HIV-AIDS treatment (Epstein, 1996) stands to testify. Carpenter (2004) observes that "the rise in patient advocacy has led to a balancing of the visibility of Type II versus Type I errors", by creating public pressure on the FDA around new drugs that are being reviewed slowly.

Debates about the adequacy of the regulatory process have been raging for decades now. What is new about it is, however, the fact that as of late the methodology of RCT itself is coming under scrutiny as a part of the regulatory process which is not as efficient as desired (Jones, 2005). The starting point of the controversy are the data on attrition rate of new drug candidates. The attrition rate of the approval process is defined as the difference between the number of compounds that enter the regulatory pipeline and those that successfully gain approval. A recent report (Peck, 2007) describes the steep rise witnessed in the last

years: “Increasing attrition rates are a significant contributor to increasing R&D costs, with a recent estimate from the FDA suggesting that only 8% of the molecules that enter clinical development are being successfully registered compared with 14% ten years ago. The latest surveys confirm that success rates from the first study in humans to launch are now <10%”. Drug attrition rates in cancer treatment are even higher: “Only 5% of agents that have anti-cancer activity in preclinical development are licensed after demonstrating sufficient efficacy in phase III testing, which is much lower than, for example, 20% for cardiovascular disease” (Hutchinson and Kirk, 2011). What is considered to be crucial is the fact that attrition rate is high particularly in the later phases of the approval process. The review mentioned earlier on (Peck, 2007) reveal that attrition rate in phase II is now over 70% and, what is worse, almost one in three molecules fail in phase III. Having candidate drugs progress through early phases and then fail in the latter, more burdensome stages of the process, is seriously sub-optimal. It means that the current system is not good at screening off useless compounds at an early stage, when dropping them would be less painful. A commentary published in 2009 in *Nature Reviews* concerning “declining pharmaceutical industry productivity, which is well recognized by drug developers, regulatory authorities and patient groups” denounces as a “key part of the problem” the fact that “clinical studies are increasingly expensive, driven by the rising costs of conducting Phase II and III trials” (Orloff et al., 2009, p. 1). Thus, at least part of the dissatisfaction surrounding the regulatory process stems from a perceived *inefficiency* of the current trial architecture. In order to focus limited research resources on the most promising approaches, it is asked that trials will fail ineffective drugs more quickly, and exploit all sources of relevant information.

Attempts at addressing these demands and improving on the present state of affairs have been made by regulatory agencies both through direct action and by encouraging initiatives and partnerships between key players both public and private. As for direct regulatory acts, in 2004 the FDA has launched the Critical Path Initiative, calling for a national effort to modernize scientific and technical tools as well as harness information technology (FDA, 2004). Then, in 2012, the agency has launched a program for speeding approval of seemingly promising new drugs called “Expedited Drug Development Pathway” (FDA, 2012). Among non-institutional initiatives PACE (Pragmatic Approaches to Comparative Effectiveness) is a U.S.-based collaborative that was created, as stated in the initiative’s page, “in the belief that whereas the

comparative effectiveness national agenda must include comparative and pragmatic trials, traditional approaches to designing and conducting such trials are too costly, take too much time and are commonly not answering real world needs" (PACE, 2008). The collaborative includes a number of different professional profiles –clinicians, policy-makers, biostatisticians–, in order to be able to confront the different facets of the problem, such as the economic arrangement of the trial process (Luce and Claxton, 1999), the guidelines and reporting system (Zwarenstein et al., 2008), effective alternatives for trial design (Getz et al., 2008) and for trials statistical analysis (Spiegelhalter et al., 2000; Berry, 2006).

Efforts to modernize trial conduct appear even more timely in light of the second element I am about to discuss: The rise of *personalized medicine*.

4.2.2 Personalized medicine: Targeted cancer therapies

The term "personalized medicine" refers to a new concept of therapy that stemmed from the achievement of the Human Genome Project (HGP) in 2003. Prior to this watershed, the guiding idea in medical research was that of identifying treatments that worked best on a large statistical basis. The HGP spurred the promise of precise cognition of genetic mechanisms of disease and response. This created the possibility to proceed from the understanding of the characteristic features of the target in order to design and administer situationally the least harmful and most effective treatment.

In the field of oncological research, in particular, the idea of personalized medical treatment took on a specific meaning, due to the impressive molecular heterogeneity of tumor lesions. Common tumors, once thought to be single entities, are now recognized as a mixture of different molecular profiles, a discovery made possible by genomic analyses. Thus, novel tumor therapies developed in light of this knowledge are "personalized" in the sense of being modulated to the molecular profile of a tumour.

The discovery of the heterogeneity of tumor gene profiles, together with an increasing availability of compounds that target specific molecular pathways, heralded a new perspective on cancer drug research: Investigators started to look for *molecularly targeted* drug agents. Traditional therapies for cancer are based on cytotoxic drugs that attack, in a non-specific manner, all fastly dividing cells. In contrast, molecularly targeted agents act in a selective manner on precise nodes of cellular pathways that are mutated or dysregulated in cancer

cells.

The two most renowned of these compounds are probably Gleevec (imatinib) in chronic myelogenous leukemia (CML) and Herceptin (trastuzumab) in breast cancers characterized by overexpression of HER2 receptor. A review of the role and discovery path that led to these landmark drug agents can be found in two short monographies about “milestones in personalized medicine” that appeared in *The Lancet Oncology*, Gambacorti-Passerini (2008) and Gelmon (2008) respectively. Targeted drugs can act against the tumor by means of different mechanisms. Some agents, like trastuzumab, are antibodies that recognize and bind a molecule which is overexpressed by the tumor cells. Antibodies that recognize tumor cells specifically can be exploited either to elicit the patient’s immune response against the tumor, or as probes, in order to direct onto the malignant cell toxic compounds that will kill it (Ledford, 2011). A second mode of action of targeted drug therapies is direct interference with cellular mechanisms involved in tumor growth and progression. The drug compound would interfere with cell growth signaling or tumor blood vessel development, or promote the specific death of cancer cells. Imatinib represents an instance of this approach. In CML, the tyrosine kinase enzyme ABL in white blood cells is locked in its activated form due to a chromosomal mutation and it speeds up cell division. Following the discovery of this genomic mechanism, investigators started screening chemical libraries to find a drug that would inhibit the enzyme. The protein thus identified, a tyrosine kinase inhibitor, was later developed into the drug Gleevec.

There is significant hype around the promise of targeted cancer therapy, due to the spectacular results of some of these drugs. Imatinib, for instance, has essentially turned CML from a fatal disease into a chronic manageable condition. A further relevant feature of targeted agents is their specificity: The action of antibodies like trastuzumab or selective inhibitors like imatinib affects in a specific manner the cells in the tumor. For this reason targeted agents have typically less harmful side-effects than conventional chemotherapy that instead attacks healthy and malignant cells alike.

Being sometimes described in the media as “wonder drugs”, the reputation of targeted agents in cancer treatment may be something of an overstatement. As a first thing, sometimes the prospective target cannot be attacked because it is expressed by an important set of healthy cells in their normal physiology, therefore attacking it would severely harm the patient;

in other cases, even when the target is known and unique, developing an agent with the needed sensitivity and specificity for it proves technically impossible. Furthermore, tumors typically do not rely on a single molecular pathway for growing and spreading: For this reason targeted therapies are rarely used alone but rather in combination with other targeted agents or even with traditional chemotherapy. In this latter case the quality-of-life benefit of these drugs is largely lost. More importantly, even the most successful targeted agents can lose their effect due to onset of resistance. Malignant cells in most tumors eventually find a way to get around blocked pathways and give rise to drug-resistant variants –or, in other cases, pre-existing variants are selected during therapy. Even the brightest success stories are spoiled by the fact that a however small percentage of patients will eventually develop resistance to the drug. As an instance, vemurafenib, an inhibitor that interferes with the mutated version of an enzyme called B-Raf, roused great excitement for breathtaking results in its preliminary phase I trial in late stage melanoma. About 80% of the terminal patients receiving the drug showed partial to complete remission of the disease, an unheard-of result. However, the regression only lasted 2 to 18 months (Flaherty et al., 2010). This means that vemurafenib can add some months to the life of melanoma patients, but it will not cure them. This is still to be considered a significant achievement give the dire prospects of these patients, but no wonder-drug yet.

These qualifications notwithstanding, targeted agents are at present the major way forward in cancer research (Majewski and Bernards, 2011; Chin et al., 2011). The efficacy of a traditional chemotherapeutic, a cytotoxic agent, is balanced on a knife's edge with its toxicity. The former cannot be augmented over that of currently available treatments, without the latter becoming unbearable. Targeted agents, by their specificity against the cells of the tumor, appear as the only option for improving over the present safety/effectiveness deadlock. It is expected that "[t]he application of specially designed combinational therapies, dependent on the unique characteristics of the individual patient and his or her malignancy, may become the standard therapeutic strategy in patients with incurable solid tumors" (Wells and Nevins, 2004).

A crucial aspect of the fragmentation of disease and therapy underlying targeted cancer treatments is the variability of patients' response. When patients that have the same kind of tumor, but harboring different molecular lesions, are exposed to a targeted compound, the

response can vary dramatically to the point that not only the magnitude, but also the direction of the treatment effect may be different across molecularly identified subgroups. For instance, treatment with trastuzumab is conditional on the level of expression of the drug's target, the HER2 receptor. This has dramatic consequences on conduct of trials for these treatments.

As a first thing, phase III trials are typically designed to include a large number of participants in order to and generate accurate and reliable results, given that so much of the approval process and judgment on the efficacy of the experimental treatment depends on them. This model however is generally not applicable to molecularly targeted agents, due to the small size of the sub-population of patients potentially sensitive to the targeted treatment. For instance, Tursz et al. (2011) in relation to breast cancer note that the population of patients exhibiting both mutations that are predictive of response to a particular molecular agent "accounts for around 0.4% of breast cancer. The feasibility of large clinical trials in this population is questionable, as this equates to 250 patients overall per year in France, when the total number of newly diagnosed breast cancer cases in the country is 50 000 per year". Even large, specialized cancer centers are in most cases insufficient to provide enough patients to start a trial. Therefore, large multi-national collaborations are formed for this purpose. An instance is *Lungscape*, a European initiative in personalized treatment of non-small-cell lung cancer (NSCLC; McIntyre 2012). Lungscape is a collaboration of 16 research hospitals around Europe which are pooling clinical data about their lung cancer patients. The aim of the collaborative is that of defining small subgroups of patients which are likely to respond differently to treatment depending on the molecular alteration that their tumour harbors. The network is planning to launch a prospective trial within the next two years.

Even when investigators are able to enroll a sufficient number of subjects in the trial, however, the traditional RCT proves inadequate to the task of adequately testing the efficacy of a molecular agent. This is because the beneficial effect of the targeted agent is often restricted to a small class of the initially eligible patients, as previously noted, but the identifier of the class of patients that would benefit often cannot be determined prior to beginning the study. This issue entails that a trial for this kind of agents must fulfill specific ethical and epistemic requirements.

4.2.3 Adaptive trials

"Adaptive design" is a label for all approaches to trial design that allow for adaptations of the study protocol, while the trial is still ongoing, to information generated during the trial. Modifications can include changes in group size, adjustments in medication dosage, or even changes in the treatments being compared. This feature makes adaptive trials especially suited for the assessment of targeted agents, that for the reasons outlined above require flexibility in the design and conduct of the trial. Two groundbreaking trials that are pioneering the adaptive approach can serve as a illustration of the concepts here described. These are large collaborative efforts which are testing multiple agents and at the same time looking for predictive response markers.

BATTLE (Biomarkers Integrated Approaches of Targeted Therapy for Lung Cancer Elimination) is a phase II trial in non-small cell lung carcinoma, testing a panel of four different drugs in several small cohorts of patients, differentiated according to the biological characteristics of their tumor (Kim et al., 2011). The use of adaptive principles in design allowed investigators to test a panel of associations in a limited time, using a relatively small sample of patients. The first patients entering trials were randomized equally among the four arms of the studies. As results from the first patients became available, however, researchers used the accruing information to assign subsequent patients to therapies more likely to work for their particular tumor type, identified on the basis of biological markers. An adaptive randomization such as it has been used in BATTLE clearly improves epistemic efficiency in detecting possible associations between patient profile and therapy. The randomization imbalance has the effect of amplifying possible differences in effect between the arms of the treatment. An adaptive design, the randomized play-the-winner rule, was famously pioneered by Marvin Zelen in the first American trial of ECMO (Barlett et al., 1985), in order to figure out as soon as possible the appropriate treatment for severely ill newborns.

The I-SPY2 study (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And MoLecular Analysis 2) is a phase II trial for the identification of new adjuvant agents in breast cancer therapy (Barker et al., 2009). The trial, launched in March 2010, is planned to evaluate 12 different drugs and to follow multiple biological markers as possible predictors of response. Investigators are able to drop or add new compounds to the study while the trial is ongoing. Donald Berry, a leading statistician at M.D. Anderson Cancer Care

Center and co-principal investigator of I-SPY2, describes the unique features of this trial: “At any given point during the I-SPY2 trial there are up to six treatment arms, including control. Randomization is adaptive across the arms within biomarker subtypes, with arms that are performing better within a subtype being assigned with greater probability to patients having that subtype. A consequence of this adaptive randomization is that better therapies move through the process faster, and have greater exposure to responding subtypes, potentially resulting in more-accurate and faster drug development. Agents may be replaced as they graduate from the trial to the phase III confirmatory stage or are dropped for futility” (Berry, 2012, p. 205). In other words, I-SPY2 leverages the opportunities offered by adaptiveness in order to identify improved treatment regimens that could work better for patients with certain molecular characteristics of their disease.

Among the key features of adaptive design, as they can be gathered from the two examples described, are adaptive allocation of patients among arms of the trial, flexibility of the study protocol and the possibility to add or drop hypotheses to the panel that is being tested. An adaptive design can feature just some of these strategies, or all of them. It is clear that continuous inflow of real-time patient data is a key element in the implementation of adaptive monitoring techniques. This replaces the fixed schedules and rigid adherence to predefined plans used in traditional studies.

The advantages of adaptive designs are manifold. First of all, adaptive approaches can reduce the sample size, cost and time needed to arrive at decision-relevant information, and they enable a timely appraisal of which lines of investigation are more promising. This has clear advantages in improving the cost-efficiency of the trial process. A second aspect is represented by the value of these approaches in terms of ethics. Adaptive randomization minimizes the chance for a patient to be exposed to a therapy that is not effective for him or her. It is true, though, that the ethics of adaptive randomization may not to be adjudicated so straightforwardly: As Joffe and Truog (2008) note, “[a]lthough unbalanced randomization and play-the-winner strategies may reduce the number of subjects exposed to the inferior intervention, this advantage may be more than offset by the problem of justifying to those assigned to the nonpreferred arm why it is ethical for them to be recruited into the trial”.

Finally, adaptive designs have a sharp pragmatic edge. Targeted agents typically represent a tough market for pharmaceutical companies, due to the fact that the potential con-

sumers represent a very small fraction of the population of cancer patients. Adaptive trials are more efficient and certainly faster. Not only this, but "[a] goal [of using adaptive designs] is to enable small phase III trials that focus on patients who are most likely to respond to the therapy, which should, in turn, shorten drug development" (Berry, 2012). In other words, adaptive designs have the potential to make research in targeted compounds interesting for companies despite the paucity of the prospective market, if this can be offset by a leaner and less expensive approval path.

There is however an important point. BATTLE and I-SPY2 are phase II trials: Agents that graduated from these studies will still need to go through a large, traditional phase III trial in order to gain approval[†]. However, there are reasons to believe that this stumbling block for adaptive approaches may not be unyielding because, as it is being increasingly pointed out by medical researchers, phase III trials may be *unnecessary* for evaluating targeted drugs. The statistical rationale behind the requirement of large samples for phase III trials is to allow for the detection of an effect that can be small with a sufficiently low error rate. Large samples have indeed already been deemed unnecessary to provide evidence of dramatic therapeutic effects in well-known cases such as penicillin for bacterial infections, smallpox vaccination and insulin in insulin-dependent diabetes (Black, 1996). On the other hand, most molecular targets are expected to show a dramatic effect on the class of patients that harbor the sensible mutation, and this is indeed the reason for interest in them. Of primary importance for targeted agents is to assess that the molecular mechanism of action works within the human body and that, by interfering with the targeted tumor pathway, it can improve patient-relevant outcomes. Small studies on a highly selected sample of patients, when combined with laboratory findings, can suffice to prove this. It follows that the pivotal stage of assessment for targeted drugs moves from phase III to phases I and II, designed as small-scale comparative studies. Once the early-phase trials have proven that the agent is effective through the hypothesized mechanism, the rationale – both pragmatic and ethical – for conducting large phase III trials is questionable (Sharma and Schilsky, 2012). Nardini, Annoni and Schiavone (2012) have further argued that personalized medicine constitutes a paradigm of evidence-generation in medicine which can be regarded as distinct from and complementary to EBM. This grounds

[†]Indeed, one of the drugs that graduated from BATTLE, Bayer's Nexavar (sorafenib), has just failed its phase III trial in lung cancer as of May, 2012: http://www.onyx.com/view.cfm/599/phase-3-mission-trial-of-nexavar-sorafenib-in-patients-with-non-small-cell-lung-cancer-did-not-meet-primary-endpoint-of-improving-overall-survival

the claim that personalized medicine has distinctive evidential needs that are not accounted for by the classical paradigm of statistically significant effects in large populations.

There have been recent cases in which the substantial benefit provided by the targeted agent, in clinical settings in which effective options are scarce, has led the regulator to accelerate approval, without waiting for the published results of phase III trials and in some cases, even before phase III trials were initiated. One paradigmatic case is represented by Gleevec, the celebrated targeted drug imatinib, which was approved to treat chronic myelogenous leukemia in May 2001 on the basis of dramatically positive results in three short, early-phase clinical trials (Keating and Cambrosio, 2012). Notably, the FDA provides a fast track for molecular targets highly likely to benefit patients with life-threatening diseases compared with available treatments: This is the Accelerated approval programme Subpart H, launched in 1992 (FDA, 1992).

For what concerns adaptive trials more specifically, regulatory authorities both in the US and in Europe have recently published "Guidance Documents" on the use and implementation of these designs. The European Medicines Agency (EMEA) led the way with a "Reflection Paper on Methodological Issues In Confirmatory Clinical Trials Planned With An Adaptive Design" (2008). The US FDA released in 2010 a draft guidance for the use of adaptive methodology: "Draft Guidance – Adaptive Design Clinical Trials for Drugs and Biologics"[†]. Expressing the view of the regulator, Hung et al. (2006) warrant that "[a]daptive designs probably are useful for Phase III confirmatory trials in some situations" even though "much more careful planning is needed for such trials than in traditional non-adaptive trials". According to Hung and colleagues, regulators will accept trials designed adaptively, insofar as there is a sound rationale for choosing this design and the trial is designed in close coordination with the regulator. At first sight, this seems to limit the scope of adaptive approaches to a restricted number of trials set up by large, specialized clinical research centers. Indeed, a large fraction of adaptive trials worldwide that are ongoing or recently completed, including the I-SPY and BATTLE studies mentioned in the previous section, were set up in a single research hospital, the M.D. Anderson Cancer Care Center (Houston, TX).

On the other hand, though, the appreciation of adaptive design principles coming from regulatory bodies clearly has a strong pragmatic rationale underlying it. Adaptive clinical

[†]The contextual issuance by the FDA of guidelines on the use of Bayesian statistics in medical device trials (FDA, 2010b) could be seen as motivated by the same principles. This will become clearer in the next section

trials represent a way for regulators to meet the two pressing demands described at the beginning of this section: The public demand for streamlining the approval process, and the demand for more informative trials that stems from the peculiar epistemic features of targeted drug evaluation. As White et al. (2000) observe, "[t]he two-group randomised controlled trial (RCT) frequently employed in medical research is actually a very simple experimental design conceptually, though with special challenges, notably sequential accrual of subjects. More complex designs are employed within the pharmaceutical industry for preclinical research. As health technology assessment develops, it is likely to need more complex studies, and hence these methods to answer questions about packages of interventions and interactions between them" (p. 3). This comment seem to apply particularly well to the possible spread of adaptive designs, as of now limited to early phases of drug research, to the central stage of drug assessment. In the final section of this chapter I will discuss the consequences and implications for the statistics of clinical trials.

4.3 Statistics: Why change? Why not?

As discussed through in the first part of the chapter, the current evidential paradigm underlying the action of regulatory bodies is grounded in unconditional frequentist inference. Unconditional inference evaluates the reliability of a result on the basis of the design features of the test which produced it. It is clear then that, to the extent that regulators subscribe to this inferential framework, they are bound to consider statistical significance from large trials as the "gold standard" of proofs of therapeutic effectiveness. While I have no intention to deny the value of adequately powered, well conducted trials, the material I presented in this chapter suggests that this model of treatment evaluation may be no longer viable, as the evidential and regulatory scenario in which trials are embedded is fastly changing. Large trials add to the drug lag and to the soaring costs of drug research, therefore contributing to the perceived unsustainability of the drug approval process. Furthermore, the novel emerging field of personalized medicine brings about a paradigm of evidence that appears to be inadequately captured by the traditional notion of 'statistical significance from large RCTs'.

Adaptive trials, reviewed in the previous section, constitute an interesting alternative to large RCTs for treatment evaluation. However, fully adaptive trials are particularly difficult to

design and analyse in the classical frequentist framework. I have already described, in previous chapters, the issues arising in frequentist analysis as an effect of early stopping based on interim data. Other features of adaptive designs, such as imbalances in the proportion of group allocation or introduction or dropping of hypotheses on-the-run, raise similar problems in view of the fact that design features have an import in frequentist inference. Considering the import of design features of the trial for classical frequentist inference, it is to be expected that this framework is ill-equipped to deal with situations where a number of design parameters are allowed to vary based on the available results. Classical frequentist treatment of adaptive trials, though technically possible, is fraught with difficulty. The Bayesian approach, on the other hand, is ideally suited for building adaptive trials. Bayesian inferential measures, which are the posterior probability of an unknown parameter and the predictive probability of future observations, can be updated as information is accrued in the trial. This process is actually the most natural version of Bayesian inference: An application of Bayes' rule upon the current probability distribution, to update it in light of new evidence. Furthermore, as described in Chapter I, Bayesian inference obeys the Likelihood Principle (sec. 1.5). This entails that the Bayesian analytic treatment of the trial results is unaffected by modifications to the study design such as those required by an adaptive design. This link between adaptive approaches and Bayesian design is confirmed by numbers: According to a recent review (Chevret, 2012), more than 1/4 of articles on Bayesian clinical trials published since 2002 were concerning adaptive Bayesian trials.

What this analysis suggests is that we may be on the verge of witnessing the decade-long dominance of the unconditional frequentist paradigm over regulatory standards give way. In order to motivate this claim, in the remaining of the chapter I will discuss the specific strengths that are generally acknowledged to the paradigm over its competitors and I will propose that they are becoming less cogent as medical research evolves. As a first thing, there exists the shared perception that the current paradigm is able to secure objectivity and impartiality of the evaluation process, by providing fixed rules for the interpretation of an experiment. Secondly, the current paradigm provides medical investigators and regulators with a well-structured and standardized methodology. Thirdly, it seems that pre-experimental error control it provides fits an important need of the regulator.

4.3.1 Objectivity

As I have discussed in the first chapter, the RCT was welcomed in medical research as a liberation from overpowering, and often mistaken, expert judgement. Thus, the main reason for the endorsement of the classical test of significance is the shared perception that this method is an objective and impartial assessment of a trial's result. This commonly held view is however overstating the objectivity that frequentist methods may be able to provide. As famously discussed by Berger and Berry (1988b), reaching sensible conclusions from any statistical analysis inevitably requires some subjective input. In Bayesian analysis this is incorporated in the prior; in frequentist analysis, more subtly, the investigator's knowledge as well as her expectations about the phenomenon of interest shape the design parameters of the study.

Furthermore, when evaluating the methodological adequacy of current RCTs for fulfilling the regulatory role that is asked from them, impartiality with respect to prior convictions of medical practitioners is only part of the story. An equally important aspect is impartiality with respect to commercial interest of pharmaceutical companies. In this respect, classical statistics seems to provide less safeguard than is currently thought. An interesting example that can substantiate this claim is represented by the problem of *design bias* (Montori et al., 2004; Sismondo, 2008). This term refers to the observed association between industry sponsorship and pro-industry result of a trial. What is more interesting is the possible origin of this bias. The fact that sponsored trials are less likely to fail is not necessarily a sign of data manipulation on the side of drug companies: On the contrary, surveillance on trial conduct and impartiality of the analysis is generally higher in this kind of trials, because of the obvious commercial interests involved (Djulbegovic et al., 2000). A more interesting hypothesis is that there exists the possibility of trimming some aspects of the test procedure, based on prior information, so to make a positive result more likely. According to Fries and Krishnan (2004, p. 250):

> From an industry perspective the drug development process must involve 'designing for success'. In a well established set of procedures company consultants and staff debate what is known about the drug, its competitors, its potential advantages in terms of toxicity or efficacy, and the potential disease indications [...] Then, trials are designed that include the patients, dosages, study duration, end-points, and

comparators that are likely to provide a positive result for the sponsor and one that is acceptable to the US FDA.

Some of the factors mentioned by Fries and Krishnan, which are involved in "design for success" of candidate drugs trials, are clearly involved in the statistical design of the study, such as study duration and setting of the end-points. In light of design bias, frequentist statistics reveals itself as much less impervious to manipulation than what is currently thought. There is, however, an example which shows this even more clearly: The case of *non-inferiority* drug trials.

Non-inferiority (NI) trials are designed so to allow a direct comparison between an experimental intervention and the standard of care. Often these active control studies are difficult to design as superiority trials, especially if the difference between treatment effectiveness is small. In this case, the sample size required to demonstrate superiority with sufficient confidence generally makes the trial unfeasible (Shapinn, 2000). The standard superiority study enables investigators to conclude with a certain confidence level (e.g. 95%) that the new treatment is better than the control. In a NI trial, instead, the objective is not to show that the new treatment has a positive advantage over the control, but rather that its *inferiority* with respect to the new does not extend below a certain limit, $-\Delta$. In conventional superiority trials, the purpose of the trial is to prove that the new drug is superior to the standard. In order to prove it, investigators start with the assumption that there is no difference between the two drugs, the null hypothesis $H_0 : \theta - \theta_0 \leq 0$. In a NI trial, instead, investigators aim at proving that there is no relevant difference between the two drugs. However, to prove it scientists need to start with the assumption that the new drug is worse: The null hypothesis is $H_0 : \theta - \theta_0 \leq -\Delta$. It is important to note that zero, the value indicating no difference, is part of the null hypothesis in traditional superiority designs, but is is part of the alternative in the case of NI trials. In other words, while in superiority trials the starting hypothesis is formulated in terms of the 'neutral' value of no difference, in NI trials the starting point is the presumption of a certain value of the difference in effect. This value Δ is the non-inferiority margin: If the trial disproves the hypothesis that the difference between the new treatment and the active control is as low as the NI margin, the new treatment will be considered non-inferior. The clear problem with this procedure is that the choice of the margin can influence the conclusion that are drawn from the trial. Clearly, a sloppy margin will make a positive verdict all to probable.

Furthermore, demonstrating non-inferiority under a sloppy margin also requires less patients than under a tight one. This is exactly what happens in superiority studies, where a large imputed difference between treatment and control requires a small sample size for achieving a desired power for the test. However, in a superiority test, choice of a large margin makes the test more *difficult* to pass for the new treatment if its actual superiority is less or approximately equal to the margin. Instead, a large margin makes a NI trial *easier* to pass for a compound that is less efficacious than the control.

There are two possible reasons that justify the choice to design a trial as a NI rather than as superiority. A first reason is that we may be confronted with a new therapeutic option which has secondary advantages with respect to the current standard of care –for instance, it is less expensive, more tolerable or it consists in a less invasive procedure. Therefore, we are interested in proving that the new treatment is therapeutically roughly equivalent to the standard. Another possible reason is the need to identify new compounds which are equivalent to the standard of care, in order to have therapeutic alternatives in case of resistance to the existing drugs. Both reasons provide sound justification to the necessity to conduct NI trials. Nonetheless, what has been said above suggests that ensuring objectivity to NI trials is a substantial challenge because of the critical import of the choice of the NI margin Δ. Both the U.S. FDA and the EMA have issued guidelines for the conduction of NI trials (FDA, 2010a, draft guidance, and EMA, 2000; EMA, 2005), with particular attention to the problem of margin-setting. Typically, the margin Δ is formulated in terms of a percentage of the active control's effect, which represents the maximum amount of the standard's effect that it would be acceptable to give up. As described in detail by Schumi and Wittes (2011), this acceptable value can be chosen through a deliberative process, by involving physicians and patients representatives, or through a technical evaluation. The deliberative approach is preferred by the EMA while the technical approach is favored by the FDA. In any case it should be evident that in the case of NI trials, due to arbitrariness in the choice of the margin, the significance test cannot be regarded as a perfectly objective device to disprove an impartial null hypothesis in light of data.

4.3.2 Standardization

In the course of this thesis, especially in chapters I and II, I have described the unsatisfactory nature of the classical frequentist criteria and measures of evidence. In particular, an aspect that has been often condemned in the literature is the fixity of the pre-experimental error rates α and β. The fact that conventional values are chosen for these rates, regardless of many possibly relevant contextual factors, has been denounced as unreasonable and arbitrary. For instance, Healy (1994) asks "Why the invariable 5% for α? Conditional on this, why the larger 10% or even 20% for β? Is it really more important not to make a fool of yourself than it is to discover something new?"

A more accurate representation of reality is that policy makers and experts in the regulatory agencies are often aware of the simplistic nature of these criteria. However, they consider oversimplification a price that is worth paying for avoiding a pluralism of criteria and methods that would be impossible to manage effectively. Robert Temple, Deputy Director at the FDA, informally remarked "The alternative to adopting a standard is to actually determine a criterion for success on the spot for each new case. That is my idea of a nightmare. So, we use a foolish, if you like, simplification [...] I don't want to have to have a symposium for every new trial to decide on an acceptable level of evidence" (Berry, Goodman and Louis, 2005, p. 303). Clearly, standardization is valuable in that it provides the regulator with the conditions to take quick and more certain decisions. Standardizing Bayesian methods proves far more difficult. Different schools of Bayesianism exist, differing mainly in the criteria guiding the choice of the prior. Empirical Bayesian methods base the prior on empirical data such as other trials or systematic reviews. Proper (or subjective) Bayesianism relies on those kind of data as a basis for the prior, but also on expert's opinion through the technique of prior elicitation. Both these approaches however face difficulties in situations where prior information is lacking or in cases when subjective commitment is undesirable. Reference Bayesianism, finally, is an attempt to ground Bayesian inference on a completely objective footing (Bernardo, 1979, 2009) by constructing *reference*, non-informative priors. This strand of Bayesianism, however, goes rather close to frequentist positions.

Ultimately, while the frequentist framework is well-established in clinical research, and it provides a set of reliable and widely known methodologies, the Bayesian community lacks at present a well-articulated perspective on research discipline. The Bayesian methodological

tools are surely extremely advanced and, as discussed through this thesis, they are more powerful than their frequentist counterpart in many applications of practical interest. The Bayesian approach, however, does not present itself as coherent and unified. On these grounds, David Teira (2011) has recently argued that the Bayesian camp is at present too underdeveloped to successfully take up the regulatory role. Furthermore, even provided that one single Bayesian perspective could prevail and be endorsed, standardizing the operation of constructing a prior upon available evidence remains a challenge. This is connected with the approved usage of Bayesian methods for trials of medical devices that was mentioned in section 4.1. As pointed out by Spiegelhalter (2004), medical devices are typically developed in incremental steps and a large body of relevant evidence is usually available. This can be formalized into a prior distribution in a rather uncontroversial manner. On the contrary, drug trials take place in an altogether different epistemic landscape. Prior expectations about the performance of a new treatment encompass the results of laboratory research, possibly of previous trials on the new drug and of previous phase trials. Deciding what among this evidence is relevant, and how it should be weighted, is far from straightforward.

However, the case of non-inferiority trials described in the previous section shows that use of a well-worn procedure cannot avoid the need for a case-by-case evaluation in many circumstances. In the case of non-inferiority trials, the conclusion at the end of the trial depends on the choice of the non-inferiority margin. For this reason, regulators need to embark in a painstaking evaluation in order to decide, in the case of each new drug proposed for approval through a NI trial, whether the margin that has been chosen is acceptable. This way of proceeding is remarkably similar to what has been proposed for circumventing the problem of possible arbitrariness of the prior in using Bayesian methods for drug trials. As Berry (2006) observes, “[i]n a regulatory setting, it is important for sponsors and regulators to agree in advance as to the prior distribution(s) that will be used”. An advantage that is generally recognized to Bayesian methods is that, once agreement on priors is achieved, the procedure which yields posterior and predictive distributions based on the prior and the trial data is mechanical, and it is actually completely automatized through use of software like the already mentioned BUGS. In other words, the only aspect of a Bayesian trial which is really in need of regulatory oversight is the choice of the prior. This seems to suggest that an alternative approach to drug assessment, where regulators focus on careful assessment of

the prior instead of the choice of the design parameters, is at least possible to envisage. As again Berry (2006) notes “Regulators are appropriately concerned about the choice of prior, but this no longer seems to be a stumbling block to using a Bayesian approach”.

4.3.3 Error control

Indeed, prior input is not the only distinction between frequentist and Bayesian inference. As I have reviewed in chapter I, Bayesian inference obeys the likelihood principle, while classical frequentism doesn't. This aspect is tangential to reliance on priors: An inferential approach like Royall's (1997b) is founded on the likelihood principle, but it rejects assignment of priors to hypotheses. Adherence to the likelihood principle appears to be, at first sight, as problematic from a regulatory point of view as reliance on untested priors is. Regulatory authorities rely massively on pre-experimental error rates for their evaluation of the reliability of a study's result. As I have discussed in Chapter I, reliance on pre-experimental error rates and the likelihood principle are incompatible.

The problem of controlling the unconditional error rates can be seen at play in the case of adaptive trials, described in the section 2.3. As a first thing, adaptive designs massively rely, as we have seen, on real-time monitoring of accruing data. As discussed in chapter I, multiple looks at the data inflate the type I error rate and are generally not acceptable in frequentist terms. Additionally, one feature of adaptive designs is the possibility to add or drop study arms, which in statistical terms corresponds to adding or dropping hypotheses to the panel being tested. A frequentist study is designed in order to have the type I error rate bound below a certain value. However, when a new hypothesis is added to the panel, the probability of committing a type I error increases. Therefore, the error rate needs to be adjusted. On the other hand, when a hypothesis is dropped from the comparison, the part of type I error pertaining to it could be re-distributed among remaining hypotheses. However, this is not always straightforward. The analysis of a trial in which predictive biomarkers are identified prospectively, then, presents even more difficulties. The classical hypothesis test is not optimized for identifying correlated variables out of a large set of possible interactions. The problem here is similar to the one encountered in epidemiology, and described by Ioannidis (2005). When a large panel of hypotheses is tested at the same time, and all have low prior probability of being true, the false positive rate gets essentially decoupled from the type I error

rate. This effect was described in chapter II under the name of *base rate fallacy*.

In some cases the problem, more than the difficulty to control the conventional pre-data error rates, is the difficulty to even define them meaningfully. For instance Hung et al. (2006) describe a trial, very similar to the already described I-SPY2, in which results from earlier stage of the trial are used for modulating patient accrual to the various arms of the later stage of the trial. Hung and colleagues observe "[t]he challenge is whether the data from the early (exploratory) stages can be combined with the data of the later (confirmatory) stage for evaluation of the relevant clinical endpoint [...] The concept of overall type I error rate for the study in this case can be unclear".

However, the fact that the pre-experimental error rates cannot be fixed through the choice of design parameters does not imply that, in these alternative designs, there is no bound on the error or no way to check, pre-data, the frequency properties of the test. Indeed, pre-experimental error rates can be obtained through simulation, by using computers to simulate the result –and the proportion of false-positives and false-negatives– of the test under variable conditions. Indeed, in the FDA Guidance Document for Bayesian designs in medical device trials (FDA, 2010b), this methodology is explicitly requested for evaluating the prospective frequency properties of the trial:

> FDA recommends you provide tables of the probability of satisfying the study claim, given various 'true' parameter values (e.g., event rates) and various sample sizes for the new trial. This table will also provide an estimate of the probability of a type I error in the case where the true parameter values are consistent with the null hypothesis, or power in the case where the true parameter values are consistent with the alternative. In some simple cases (e.g. a single arm trial with a binomial outcome) these probabilities can be calculated directly. If the study design is complex it is usually necessary to use simulation to compute these probabilities. In general, the simulation should reflect the study design.

What this excerpt makes evident is the fact that the FDA has already shown openness to the possibility of evaluating the degree of error control warranted by the statistical procedure via simulations on the statistical model rather than via knowledge of the design parameters of the test. This fact is of extreme relevance for the discussion of unconditional vs. conditional appraisal of error that I have already confronted in Chapter II. There we mentioned a

potential counterargument to the use of Bayesian and conditional approaches: The fact that such approaches, by obeying the likelihood principles, may fail to provide proper control upon the pre-experimental error rates. On the basis of the material just provided it is instead possible to conclude that this objection is unlikely to be crucial for the evaluation of conditional procedures by regulators. Indeed, even though regulators remain rightly concerned with the unconditional, pre-data properties of the test procedure, they seem ready enough to move this requirement to the background for the sake of other, equally relevant, inferential qualities. For instance, in the case of trials for medical devices and of adaptive trials in oncology, the epistemic advantage provided by Bayesian methods seems qualified to offset the loss of the possibility to specify the error properties directly through the design.

4.4 Conclusion: An outlook

In this chapter I have analyzed the regulatory choice of adherence to a particular statistical framework in the context of the fastly moving stage of medical research, with particular attention to the field of oncology. I have described two emerging needs in the medical arena that in my opinion are most likely to play an important part in the future shaping of the regulatory process. On one hand there is a strong public call for streamlining the approval process through an improvement of the efficiency of clinical trials. On the other hand, on the side of medical science, the emerging paradigm of personalized medicine brings about a shift in the way trials should be conducted, fostering a reappraisal of evidence from small comparative studies face statistical evidence from large RCTs. Accordingly, some of the arguments traditionally upheld in favor of the current framework are losing force. Objectivity, impartiality and pre-experimental error control remain important for the regulator but they come to clash with the objective of improving the epistemic efficiency of the statistical framework. Interestingly enough, this may be regarded an applied version of the “conditioning dilemma” which was proposed in Chapter II.

However I propose that the dilemma remains such only as long as it is evaluated according to the current statistical framework. Within an unconditional framework ‘sound evidence’ may only be the evidence coming from large trials, because confidence in the result is determined by the design features of the study. This entails diminishing the value of evidence from

small trials, trials that stop early, or adaptive trials, while instead these constitute the most promising way of streamlining evidence production in medicine. But if the present framework does not provide the tools to assess the confidence in results from foreshortened or adaptive trials properly, this may not be impossible if a different framework is used. As Berry (2012) remarked, “Asking more questions in a single clinical trial makes drug development faster and more accurate. However, asking many questions requires a wholly different way of thinking about research. In particular, not every question that is asked in a trial can be answered with high statistical power” (p.201). In other words, making the clinical research process more efficient requires us to think differently about statistical evidence: To dismiss the idol of statistical significance from large trials, and to stipulate alternative criteria for evaluating conclusiveness of a result.

Conditional procedures, both Bayesian and frequentist, can provide such new criteria. As I have described in chapter II, while the unconditional test sanctions a decision at a fixed level of confidence and error, the conditional test lets the confidence in the decision vary with the strength of evidence in the data. Another way to see the issue is by observing that sequential analysis is about drawing inference with an amount of statistical information which is inferior to what was originally planned. Clearly this has a price in terms of confidence. Unconditional procedures set out with a pre-defined level of confidence that we want to achieve, and then license stopping only when this level can be warranted. Conditional procedures, instead, conform to our decision to stop and inform us about the loss of confidence. Thus, appraisal of conditional error rates affords a fine-grained discrimination of trial results according to their size but also of the strength of the conclusion they commend. Ultimately, conditional evaluation of evidence offers a platform for weighting the evidence coming from trials, both large and small, on a par. Thus, endorsement of a conditional appraisal of strength of evidence represents an effective way of fostering the needed change in the way medical evidence is build from trials.

While the discussion so far conducted is valid for Bayesians as well as for conditional frequentist procedures, the way to actual endorsment of Bayesian procedures by regulators appears to be still a long and winding one. Side by side with the promising signs of openness from the regulators that I have presented towards the end of this chapter, there remains in fact the issue that full adoption of the Bayesian framework ultimately requires a radical change in

the scope of the regulatory action: From the oversight of the design features of the trial an assessment of the appropriateness of the prior used in the analysis. In this context I suggest that the adoption of a conditional frequentist framework can provide an effective transition solution. Conditional evidence is fully compatible with an evaluation of results in terms of frequentist error rates. At the same time, however, this framework cures the most serious deficiencies of unconditional error reporting. This approach is actually very much in line with a suggestion by eminent biostatistician Stephen Senn, who noted: "[I]t is profitable to look at things in Bayesian and frequentist ways [...] If a reluctance to use 'subjective' information persists, Bayesian inspiration for different approaches to analysis can be found." (Senn, 2001, p. 154)

More importantly, though, the conditional frequentist framework can make the medical audience more familiar with the valuable consequences of conditioning, such as the Likelihood Principle, and thus facilitate the adoption of a full Bayesian framework. In conclusion, endorsment of a conditional frequentist framework can free medical research from the tyranny of the large and costly RCT by providing the tools to properly evaluate small, truncated, or adaptive trials. Eventually, this will bring regulators closer to the needs of current medical research and pave the way to a fruitful reform of the regulatory system.

Conclusion

Randomized clinical trials have an ever-increasing importance for evidence generation in modern medicine. Nonetheless, or maybe precisely for this reason, the epistemic adequacy of RCTs has been in latter years called into question. A particular issue within this debate revolves around the statistical methodology in use for design and analysis of clinical trials. There are two main schools of statistical inference, having equal methodological standing but differing in the choice of foundational principles they subscribe to.

Frequentist inference is based upon the interpretation of probabilities as long-run frequencies of events. Consequently, credibility of a hypothesis is based on the pre-specified error properties of the procedure that was used to put it to test. These error probabilities are determined by the choice of the design parameters of the study. Strength of evidence is measured indirectly, through *p*-values, as the probability of obtaining a resut as or more extreme than the one observed, under a fixed hypothesis. Currently, design and analysis of trials are grounded in frequentism and they exploit the whole frequentist toolkit of fixed confidence levels, pre-experimental error rates and *p*-values.

The Bayesian framework of inference is based upon assigning degrees of credibility to hypotheses in light of the available evidence. According to this inferential school, the reliability of a conclusion is in direct relation with its prior credibility and with the strength of evidence in the data supporting it. The two distinctive elements of Bayesian inference are therefore reliance on priors, on one hand, and adherence to the likelihood principle, on the other. The likelihood principle formalizes the dynamic nature of Bayesian inference, whereby the conclusion depends only on the strength of evidence in the data and not on design features of the trial.

These two features which distinguish the Bayesian framework from the frequentist counterpart are at the roots of the greater epistemic efficiency of Bayesian methods as compared to frequentist ones. As a first thing, the inferential step incorporates relevant prior information, thus requiring less novel evidence to tip the balance in favor of one or the other hypothesis. Furthermore, thanks to the likelihood principle, Bayesian inference is independent from the

experimental design and, in particular, from the stopping rule. In this way, Bayesian methodology is able to make a more efficient use of the evidence generated by the trial, even as it is still incomplete.

This comes to light particularly well in considering a specific issue: The monitoring of ongoing trials for detecting conclusive evidence ahead of the planned end. This practice, which is becoming increasingly common in later years, bears potential of great advantages both in terms of protection of in-trial patients and of a more cost-effective use of resources in medical research. From the statistical point of view, this practice requires to draw a conclusion on the basis of the accumulating data before the full amount of information that was initially planned has accrued. Bayesian statistics is more effective than frequentist statistics for what concerns monitoring, due to the features described earlier on. On the one hand, reliance on priors allows a better appraisal of interim results because it allows placing them meaningfully in a context of prior knowledge. On the other hand, the decision to stop does not affect the Bayesian analysis, due to adherence to the likelihood principle. Frequentist analyses, instead, have to be adjusted for the fact that the analysis was undertaken at a different time-point than what had initially been planned. In recent years the evolution of the context of medical research has brought about an increase in the need for flexibility in trial conduct. An instance of this is represented by the emergence of adaptive design for trials. Within an adaptive study, design features are adjusted in response to emerging trends in the data while the trial is still ongoing. Adaptive trials exasperate the contrast between the two schools that is brought to light by trial monitoring.

In a context of growing scarcity of resources like the one we are witnessing at present in medical research, inferential efficiency as expressed by Bayesian methods is clearly an asset. However, the shared consensus is that frequentist statistics is more adequate for the design and analysis of trials, as far as principles are concerned. Indeed, clinical trials fulfill a double role in the current system: They represent the gold standard for the generation of medical evidence and, at the same time, they constitute the main gateway to market approval for medical interventions in the Western world. Frequentist statistics is perceived as better suited to trial analysis because it seems better able to warrant two goals which are extremely important face the double role of RCTs, namely control over the possibility of mistaken conclusions and objectivity of the framework. Bayesian methods are held to be less capable

than frequentist ones of providing those goods: A position which remains well entrenched and difficult to shake.

Objections to Bayesian inference in the context of medical trials revolve mainly around two points. As a first thing, reliance of priors is considered a fragility in that it leaves unchecked an opportunity for bias and manipulation of trial results. In this context, frequentist methods appear to provide a higher warrant of objectivity. Secondly, the independence of Bayesian inference from design features of the trial compromises the possibility of controlling the error rates of the test procedure through design choices. This seems to leave Bayesian methods without a proper safeguard against the possibility of a mistaken conclusion.

Even though these objections have for long been considered crucial, they are losing force in recent years. With regards to objectivity, awareness is growing about the fact that no statistical method can be absolutely impervious to manipulability: In the case of frequentist methods, this is proven by the phenomenon of design bias and by the difficult appraisal of results from non-inferiority studies. Indeed, frequentist inference does make use of prior information in the phase of planning the experiment: This is precisely what Neyman-Pearson theory of optimal design is about. For what concerns error rates, statistical theory tells us that control over pre-experimental error rates is often obtained at the price of proper appraisal of individual results. Thus, while the objective of controlling the pre-set error of the procedure remains an important goal especially from the perspective of regulators, the drive towards improving the efficiency of the trial opens the way to alternative considerations. In particular, error control can be obtained otherwise than through design choices, and regulators and practitioners alike are showing openness to this possibility.

Nonetheless, implementing Bayesian methods within the regulatory context would require a radical rethinking of the statistical framework that is currently in place. Namely, it would imply a shift in focus from considering the appropriateness of the design choices for the trial, to assessing the appropriateness of the prior used in the analysis. Furthermore, the Bayesian school is at present still lacking a coherent and unified methodology for application to clinical trials, particularly for what concerns a unified perspective on prior construction. This problem, more than the classical objections to the use of priors in medical research, may hinder the spread of Bayesian methods in clinical trials in the close future.

The good news is that there is a third way for safeguarding the advantage of flexibility from

within the current frequentist paradigm of inference. This alternative is provided by conditional approaches to inference. Conditional procedures quantify the degree of confidence that can be assigned to a conclusion as a function of the observed evidence. Such assessment of the strength of evidence is absent from classical procedures, which are for this reason denoted as unconditional. Conditional frequentist methods, in particular, provide a reconciliation between the two opposing schools of statistical inference, by supplementing frequentist procedures with a post-data assessment of strength of evidence and probability of error. These conditional measures are constructed with Bayesian insight and obey the likelihood principle. However they involve no assignment of prior credibility to hypotheses, the contested hallmark of Bayesian theory. Thus, the conditional frequentist approach is able to improve the epistemic efficiency of trial conduct without introducing subjective probabilities and remaining firmly within the frequentist approach to inference. From the statistical point of view, the value of conditional frequentism is that of fostering unification at the foundational level and of providing a common ground for practitioners. For Bayesians, insisting on the virtues of conditional error-assessment could represent a way to forward the advantages of the full Bayesian method, while at the same time downplaying the import of the most contested aspect of Bayesianism, the use of priors. For frequentists, adopting a conditional approach would create the possibility to make trials more efficient, as is required with mounting urgency, without the need to compromise with their core inferential principles.

The scope of the examination of the statistical methodology undertaken in this thesis is however wider than the realm of statistical theory and practice, as this examination ultimately provides a novel and fruitful perspective on medical research itself. Indeed, the different inferential principles that have been discussed through this work reflect different priorities in inference as they are held by the contrasting schools of statistics –or also by different approaches internal to a single school, as the fault line between unconditional and conditional approaches stands to testify. Thus, it can be seen that the problem of choosing one particular inferential framework for the design and analysis of trials is so complex precisely because the differences among the various statistical approaches reflect the plurality of values and goals that exists in medical research.

In the past, the tension between these different values has been mostly identified as a contrast between an obligation of care and protection towards participating patients *vis-á-*

vis the recognized objective of achieving a reliable and generalizable scientific conclusion. This picture is however, as I have described, overly simplistic and inadequate to represent the stakes that are presently involved in medical research. The real conflict turns out to be the one involving the proper characterization of a reliable conclusion. Traditionally, the focus of investigators and regulators has been on controlling the long-run performance of the test procedure and on keeping a bound on unconditional error rates as a way to secure a warrant upon the objectivity and credibility of statistical conclusions. Under this view, trials needed to generate strong evidence must necessarily involve a large number of patients and follow strict design planning. This perspective is however in contrast with the idea that small, flexible trials can be more effective in discriminating credible results and thus lead to better medical decisions.

For a long time it has been thought that unconditional inferential goals correspond more adequately to the shared goals of the different actors involved in medical research. The conclusion I arrive to in my work is a different one. True, the need to comply with traditional requirements, strongly influenced by the classical unconditional inferential approach, still exists. However, this need is likely to get eventually offset by a pressing demand for inferential efficiency which is growing in the changing landscape of medical research. Flexible, adaptive and efficient trials are increasingly regarded as a promising way forward. Such trials are able to meet the epistemic needs of medical research, including the newly emerging exigencies of personalized medicine, all in ensuring adequate principia for the protection of participating patients. However, only a proper appraisal of strength of evidence in each single case, as made possible by conditioning, can afford the opportunity for this scenario to develop. Indeed, only conditional approaches to inference can warrant that these goods are achieved without losing safeguard on the validity of trial results.

The proposal of using conditional frequentist methods for clinical trials is, at the current stage, slightly less than a blueprint. The work collected in this thesis should be understood as aimed at paving the way to the possibility of developing this approach in a working set of procedures, by showing its potential from the epistemological, ethical and regulatory point of view.

Bibliography

Allison, M. (2012). Reinventing clinical trials. *Nature Biotechnology 30*(1), 41–49.

Altman, D. G. (1980). Statistics and ethics in medical research – Interpreting results. *British Medical Journal 281*, 1612–4.

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association 58*, 365–83.

Armitage, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell.

Armitage, P. (2003). Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology 32*, 925–8.

Armitage, P., C. K. McPherson and B. C. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society 132*, 235–44.

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine 25*, 3589–631.

Ashcroft, R. (1999). Equipoise, knowledge and ethics in clinical research and practice. *Bioethics 13*(3/4), 314–326.

Barker, A. D., C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry and L. J. Esserman (2009). I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology & Therapeutics 86*, 97–100.

Barlett, R. H., D. Roloff, R. G. Cornell, A. F. Andrews, P. W. Dillon and J. B. Zwischenberger (1985). Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics 76*, 479–87.

Bassler, D., M. Briel, V. M. Montori, M. Lane, P. Glasziou, Q. Zhou, D. Heels-Ansdell, S. D. Walter, G. H. Guyatt, D. Flynn et al. (2010). Stopping randomized trials early for benefit

and estimation of treatment effects. *Journal of the American Medical Association 303*(12), 1180–1187.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou and G. Guyatt (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of Clinical Epidemiology 61*(3), 241–246.

Beauchamp, T. L. and J. F. Childress (2009). *Principles of Biomedical Ethics, 6th Ed.* New York (NY): Oxford University Press.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? (with discussion). *Statistical Science 18*(1), 1–12.

Berger, J. O. and D. A. Berry (1988a). The relevance of stopping rules in statistical inference (with discussion). In S. Gupta and J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics IV*, pp. 29–72. New York: Springer.

Berger, J. O. and D. A. Berry (1988b). Statistical analysis and the illusion of objectivity. *American Scientist 76*, 159–165.

Berger, J. O., B. Boukai and Y. Wang (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science 12*(3), 133–160.

Berger, J. O., B. Boukai and Y. Wang (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika 86*(1), 79.

Berger, J. O., L. D. Brown and R. L. Wolpert (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics 22*(4), 1787–1807.

Berger, J. O. and M. Delampady (1987). Testing precise hypotheses (with discussion). *Statistical Science 2*, 317–52.

Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconciliability of P-values and evidence (with discussion). *Journal of the American Statistical Association 82*, 112–39.

Berger, J. O. and R. L. Wolpert (1984). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

Berger, J. O. and R. L. Wolpert (1988). *The Likelihood Principle. 2nd ed.* Hayward, CA: Institute of Mathematical Statistics.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological) 41*, 113–147.

Bernardo, J. M. (2009). Modern Bayesian inference: Foundations and objective methods. In P. S. Bandyopadhyay and M. R. Forster (Eds.), *Philosophy of Statistics*. Amsterdam: North Holland.

Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. In *Bayesian Statistics 9. Proceedings from the 9th Valencia International Meeting*, pp. 1–68.

Berry, D. A. (1987). Interim analysis in clinical trials: The role of the likelihood principle. *The American Statistician 41*(2), 117–122.

Berry, D. A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine 12*, 1377–93.

Berry, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science 19*, 175–187.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery 5*, 27–36.

Berry, D. A. (2012). Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology 9*, 199–207.

Berry, D. A., S. N. Goodman and T. Louis (2005). Floor discussion. *Clinical Trials 2*, 301–4.

Berry, S. M., B. P. Carlin and J. Connor (2010). Bias and trials stopped early for benefit. *Journal of the American Medical Association 304*, 156.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*, 269–306.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal 312*, 1215–18.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics*, Volume 2. New York (NY): Wiley-Interscience.

Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *The Annals of Statistics 6*, 59–71.

Butler, D. (2008). Crossing the valley of death. *Nature 453*, 840–2.

Carnap, R. (1962). *Logical Foundations of Probability (2nd ed.)*. Chicago (IL): University of Chicago Press.

Carpenter, D. P. (2004). The political economy of FDA drug review: processing, politics, and lessons for policy. *Health Affairs 23*(1), 52–63.

Casella, G. and R. L. Berger (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association 82*, 106–11.

Chalmers, I. (2009). Explaining the unbiased creation of treatment comparison groups. *Lancet 374*, 1670–1.

Chevret, S. (2012). Bayesian adaptive clinical trials: a dream for statisticians only? *Statistics in Medicine 31*(11-12), 1002–1013.

Chin, L., J. N. Andersen and P. A. Futreal (2011). Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 297–303. doi: 10.1038/nm.2323.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson and Bayes. *The American Statistician 59*(2), 121–126.

Cornfield, J. (1966a). A Bayesian test of some classical hypotheses with applications to sequential clinical trials. *Journal of the American Statistical Association 61*, 577–94.

Cornfield, J. (1966b). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician 20*(2), 18–23.

Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics 25*, 617–57.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. *American Journal of Epidemiology 104*, 408–21.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics 29*, 357–372.

Cressey, D. (2011). Traditional drug-discovery model ripe for reform. *Nature 471*, 17.

De Mets, D. L. and K. K. G. Lan (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine 13*, 1341–52.

Dickey, J. M. (1977). Is the tail area useful as an approximate Bayes Factor? *Journal of the American Statistical Association 72*, 138–142.

Djulbegovic, B., G. H. Guyatt and R. E. Ashcroft (2009). Epistemologic inquiries in Evidence-Based Medicine. *Cancer Control 16*(2), 158–168.

Djulbegovic, B., M. Lacevic, A. Cantor, K. Fields, C. Bennett, J. Adams, N. Kuderer and G. Lyman (2000). The Uncertainty Principle and industry-sponsored research. *The Lancet 356*(9230), 635–38.

Dunn, O. J. and V. A. Clark (2009). *Basic Statistics. A Primer for the Biomedical Sciences. 4th ed.* Hoboken (NJ): Wiley.

Edwards, S. J., R. J. Lilford, D. A. Braunholtz, J. C. Jackson, J. Hewison and J. Thornton (1998). Ethical issues in the design and conduct of randomised controlled trials. *Health Technology Assessment 2*(15).

Edwards, W., H. Lindman and L. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review 70*(3), 193.

Ellenberg, S., T. Fleming and D. L. DeMets (2002). *Data Monitoring Committees in clinical trials. A practical perspective*. Hoboken, N.J.: John Wiley.

Ellenberg, S. S., D. L. DeMets and T. R. Fleming (2010). Bias and trials stopped early for benefit. *Journal of the American Medical Association 304*, 158.

Enkin, M. W. (2000). Against: Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *British Medical Journal 321*, 757–8.

Epstein, S. (1996). *Impure Science: AIDS, Activism, and the Politics of Knowledge*. Berkeley (CA): University of California Press.

Etzioni, R. D. and J. B. Kadane (1995). Bayesian statistical methods in public health and medicine. *Annual Review of Public Health 16*(1), 23–41.

European Medicine Agency (2000). Points to consider in switching between superiority ad non-inferiority. Available at www.ema.europa.eu/pdfs/human/ewp/048299en.pdf. Last access 04/01/2013.

European Medicine Agency (2005). Guideline on the choice of the non-inferiority margin. Available at www.ema.europa.eu/pdfs/human/ewp/215899en.pdf. Last access 04/01/2013.

Farewell, V. and T. Johnson (2010). Woods and Russell, Hill, and the emergence of medical statistics. *Statistics in Medicine 29*, 1459–76.

Fayers, P. M., D. Ashby, M. K. B. Parmar et al. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in medicine 16*(12), 1413–30.

Feinstein, A. R. (1998). P-values and confidence intervals: Two sides of the same unsatisfactory coin. *Journal of clinical epidemiology 51*(4), 355–360.

Fidler, F. (2012). *From statistical significance to effect estimation: Statistical reform in Psychology, Medicine and Ecology*. Routledge.

Fisher, R. A. (1935a). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society 98*, 39–54.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York: Hafner Publishing Co.

Fisher, R. A. (1959). *Statistical methods and scientific inference. 2nd ed.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1966). *The Design of Experiments. 8th ed*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical method and scientific inference. 3rd ed.* London: Collins and Macmillan.

Flaherty, K. T., I. Puzanov, K. B. Kim, A. Ribas, G. A. McArthur, J. A. Sosman, P. J. O'Dwyer, R. J. Lee, J. F. Grippo, K. Nolop and P. B. Chapman (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *New England Journal of Medicine 363*(9), 809–819.

Food and Drug Administration (1992). New drug, antibiotic, and biological drug product regulations; accelerated approval. Final rule. Fed. Regist. 57, 58942–58960.

Food and Drug Administration (1998). Guidance for industry. Providing clinical evidence of effectiveness for human drug and biological products. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm078749.pdf. Last accessed 07/27/2012.

Food and Drug Administration (2004). Innovation or stagnation? – Challenge and opportunity on the critical path to new medical products. Available at http://www.fda.gov/oc/initiatives/criticalpath/. Last accessed 17/02/2013.

Food and Drug Administration (2010a). Guidance for industry: Non-inferiority clinical trials. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Last access 04/01/2013.

Food and Drug Administration (2010b). Guidance for the use of Bayesian statistics in medical device clinical trials. Available at http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm. Last access 26/01/2012.

Food and Drug Administration (2012). Expedited drug development pathway. Available at http://www.fda.gov/AboutFDA/ReportsManualsForms/Reports/ucm274441.htm. Last access 01/11/2012.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine 317*, 141–5.

Freedman, B. and S. H. Shapiro (1994). Ethics and statistics in clinical research: Towards a more comprehensive examination. *Journal of statistical planning and inference 42*, 223–240.

Fried, C. (1974). *Medical Experimentation: Personal Integrity and Social Policy*. New York, N.Y.: American Elsevier Publishing.

Fries, J. and E. Krishnan (2004). Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development. *Arthritis Research and Therapy 6*(3), R250–5.

Gambacorti-Passerini, C. (2008). Part I: Milestones in personalised medicine—imatinib. *The Lancet Oncology 9*(6), 600 –.

Gardner, M. J. and D. G. Altman (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal 292*, 746–750.

Gelman, A. and C. R. Shalizi (2011). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/j.2044-8317.2011.02037.x.

Gelmon, K. (2008). Part II: Milestones in personalised medicine—trastuzumab. *The Lancet Oncology 9*(7), 698.

Getz, K. A., J. Wenger, R. A. Campo, E. S. Seguine and K. I. Kaitin (2008). Assessing the impact of protocol design changes on clinical trial performance. *American Journal of Therapeutics 15*, 450–7.

Gieringer, D. H. (1985). The safety and efficacy of new drug approval. *Cato Journal 5*, 177.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Giordano, S. (2010). The 2008 Declaration of Helsinki: Some reflections. *Journal of Medical Ethics 36*, 598–603.

Goodman, S. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of internal medicine 146*, 882.

Goodman, S. (2009). Stopping trials for efficacy: An almost unbiased view. *Clinical Trials 6*, 133–135.

Goodman, S. J., D. Berry and J. Wittes (2010). Bias and trials stopped early for benefit. *Journal of the American Medical Association 304*, 157.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine 130*(12), 995–1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine 130*(12), 1005–13.

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine 53*, 225–8.

Halpern, S. D., J. H. T. Karlawish and J. A. Berlin (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association 288*, 358–62.

Hansson, S. O. (2006). Uncertainty and the ethics of clinical trials. *Theorethical Medicine and Bioethics 27*, 149–167.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology 44*, 793–97.

Healy, M. J. R. (1994). Probability and decisions. *American Journal of Diseases in Children 71*, 90–4.

Heilig, C. M. and C. Weijer (2005). A critical history of individual and collective ethics in the lineage of Lellouch and Schwartz. *Clinical Trials 2*, 244–253.

Hellman, S. and D. S. Hellman (1991). Of mice but not men: Problems of the randomized clinical trial. *The New England Journal of Medicine 324*(22), 1585–9.

Hill, A. B. (1963). Medical ethics and controlled trials. *British Medical Journal*, 1043–9.

Howick, J. (2011). *The Philosophy of Evidence-Based Medicine*. London: Wiley-Blackwell.

Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach (3rd ed.)*. Chicago (IL): Open Court.

Hubbard, R. and M. J. Bayarri (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician 57*(3), 171–178.

Hung, H. M. J., R. T. O'Neill, S.-J. Wang and J. Lawrence (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal 4*, 565–73.

Huskamp, H. (2006). Prices, profits, and innovation: Examining criticisms of new psychotropic drugs' value. *Health Affairs 25*, 635–44.

Hutchinson, L. and R. Kirk (2011). High drug attrition rates—where are we going wrong? *Nature Reviews Clinical Oncology 8*, 189–90.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.

Jennison, C. and B. W. Turnbull (1990). Interim monitoring in medical trials. *Statistical Science 5*(3), 299–317.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods With Applications to Clinical Trials*. Boca Raton (FL): Chapman and Hall.

Joffe, S. and F. G. Miller (2012). Equipoise: asking the right questions for clinical trial design. *Nature Reviews Clinical Oncology 9*, 230–235.

Joffe, S. and R. D. Truog (2008). Equipoise and Randomization. In E. Emanuel, C. Grady, R. Crouch, R. Lie, F. Miller and D. Wendler (Eds.), *The Oxford Textbook of Clinical Research Ethics*, pp. 245–60. Oxford: Oxford University Press.

Johnson, W. O. and J. L. Gastwirth (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of Acquired Immune Deficiency Syndrome. *Journal of the Royal Statistical Society Series B 53*, 427–39.

Jones, T. C. (2005). A call to restructure the drug development process: government over-regulation and non-innovative late stage (Phase III) clinical trials are major obstacles to advances in health care. *Science and engineering ethics 11*(4), 575–587.

Kadane, J., M. Schervish and T. Seidenfeld (1996a). When several Bayesians agree that there will be no reasoning to a foregone conclusion. *Philosophy of Science 63*, 281–9.

Kadane, J. B., M. J. Schervish and T. Seidenfeld (1996b). Reasoning to a foregone conclusion. *Journal of the American Statistical Association 91*(435), 1228–35.

Kadane, J. B. and T. Seidenfeld (1990). Randomization in a Bayesian perspective. *Journal of Statistical Planning and Inference 25*, 329–45.

Kaitin, K. I. and J. S. Brown (1995). A drug lag update. *Drug Information Journal 29*, 361–73.

Karlawisk, J. T. and J. Lantos (1997). Community equipoise and the architecture of clinical research. *Cambridge Quarterly of Healthcare Ethics 6*, 385–396.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Keating, P. and A. Cambrosio (2012). *Cancer on Trial: Oncology as a New Style of Practice*. Chicago: The University of Chicago Press.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 789–808.

Kim, E. S., R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, C. M. Alden, S. Liu, X. Tang, F. R. Khuri, H. T. Tran, B. E. Johnson, J. V. Heymach, L. Mao, F. Fossella, M. S. Kies, V. Papadimitrakopoulou, S. E. Davis, S. M. Lippman and W. K. Hong (2011). The BATTLE Trial: Personalizing Therapy for Lung Cancer. *Cancer Discovery 1*(1), 44–53.

Korn, E. L. and B. Freidlin (2011). Inefficacy interim monitoring procedures in randomized clinical trials: The need to report. *The American Journal of Bioethics 11*(3), 2–10.

Krom, A. (2011). The harm principle as a mid-level principle? Three problems from the context of infectious disease control. *Bioethics 25*, 437–44.

Lan, K. K. G. and D. L. De Mets (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*, 659–663.

Lan, K. K. G., R. Simon and M. Halperin (1982). Stochastically curtailed sampling in long-term clinical trials. *Communications in Statistics C 1*, 207–19.

Lan, K. K. G. and J. Wittes (1988). The B-value: A tool for monitoring data. *Biometrics 44*, 579–585.

Lan, K. K. G. and D. M. Zucker (1993). Sequential monitoring in clinical trials: The role of information and Brownian motion. *Statistics in Medicine 12*, 753–65.

Lau, W. Y., T. W. T. Leung, S. K. W. Ho, M. Chan, D. Machin, J. Lau, A. T. C. Chan, W. Yeo, T. S. K. Mok, S. C. H. Yu et al. (1999). Adjuvant intra-arterial lipiodol-Iodine-131 for resectable hepatocellular carcinoma: A prospective randomised trial. *The Lancet 353*(9155), 797–801.

Ledford, H. (2011). Toxic antibodies blitz tumours. *Nature 476*, 380–1.

Lee, J. J. (2010). Demystify statistical significance–Time to move on from the P-value to Bayesian analysis. *Journal of the National Cancer Institute 103*, 1–2.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association 88*(424), 1242–1249.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses. 2nd ed.* New York (NY): Springer.

Lellouch, J. and D. Schwartz (1971). L'essai thérapeutique: éthique individuelle ou éthique collective? *Revue de l'Institut International de Statistique 39*, 127–136.

Levine, R. J. (1988). *Ethics and Regulation of Clinical Research, 2nd ed.* New Haven (CT): Yale University Press.

Lewis, R. J., A. M. Lipsky and D. A. Berry (2007). Bayesian decision-theoretic group sequential clinical trial design based on quadratic loss function: a frequentist evaluation. *Clinical Trials 4*, 5–14.

Lilford, R. J. (2003). Ethics of clinical trials from a Bayesian and decision-analitic perspective: Whose equipoise is it anyway? *British Medical Journal 326*, 980–1.

Lilford, R. J. and J. Jackson (1995). Equipoise and the ethics of randomization. *Journal of the Royal Society of Medicine 88*, 552–9.

Lindley, D. (1972). *Bayesian statistics: A review*. Philadelphia (PA): SIAM.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 40*(1/2), 187–192.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician 60*, 213–23.

Luce, B. and K. Claxton (1999). Redefining the analytical approach to pharmacoeconomics. *Health Economics 8*, 187–9.

Majewski, I. J. and R. Bernards (2011). Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nature Medicine*, 304–312. doi: 10.1038/nm.2311.

Mayo, D. and D. Cox (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (Ed.), *Optimality: The Second Erich L. Lehmann Symposium*, pp. 77–97. Beachwood (OH): Institute of Mathematical Statistics (IMS).

Mayo, D. and M. Kruse (2001). Principles of Inference and their Consequences. In D. Corfield and J. Williamson (Eds.), *Foundations of Bayesianism*, pp. 381–403. Dordrecht: Kluwer.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago (IL): University of Chicago Press.

Mayo, D. M. and A. Spanos (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science 57*, 323–357.

McIntyre, P. (2012). Lungscape: A living lung laboratory. *CancerWorld*, 44–47.

McPherson, K. (1974). Statistics: The problem of examining accumulating data more than once. *The New England Journal of Medicine*, 501–2.

Migrino, R., J. Young, S. Ellis, H. White, C. Lundergan, D. Miller, C. Granger, A. Ross, R. Califf and E. Topol (1997). End-systolic volume index at 90 to 180 minutes into reperfusion therapy for acute myocardial infarction is a strong predictor of early and late mortality. The Global Utilization of Streptokinase and t-PA for Occluded Coronary Arteries (GUSTO)-I Angiographic Investigators. *Circulation 96*(1), 116–121.

Miller, F. G. and H. Brody (2003). A critique of clinical equipoise: Therapeutic misconception in the ethics of clinical trials. *Hastings Center Report 33*(3), 19–28.

Miller, F. G. and H. Brody (2007). Clinical equipoise and the incoherence of research ethics. *The Journal of Medicine and Philosophy 32*(2), 151–65.

Miller, F. G. and S. Joffe (2011). Equipoise and the dilemma of randomized clinical trials. *New England Journal of Medicine 364*(5), 476–80.

Miller, P. B. and C. Weijer (2003). Rehabilitating equipoise. *Kennedy Institute of Ethics Journal 13*(2), 93–118.

Montori, V., R. Jaeschke, H. Schünemann, M. Bhandari, J. Brozek, P. Devereaux and G. Guyatt (2004). Users' guide to detecting misleading claims in clinical research reports. *Britisk Medican Journal 329*(7474), 1093.

Montori, V. M., P. J. Devereaux, N. K. J. Adhikari, K. E. A. Burns, C. H. Eggert, M. Briel, C. Lacchetti, T. W. Leung, E. Darling, D. M. Bryant et al. (2005). Randomized trials stopped early for benefit: a systematic review. *Journal of the American Medical Association 294*(17), 2203.

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine 27*, 469–82.

Moyé, L. A. and A. T. N. Tita (2002). Defending the rationale for the two-tailed test in clinical research. *Circulation 105*, 3062–3065.

Mueller, P., V. Montori, D. Bassler, B. Koenig and G. Guyatt (2007). Ethical issues in stopping randomized trials early because of apparent benefit. *Statistics in Medicine 146*, 878–81.

Mukherjee, S. D., J. R. Goffin, V. Taylor, K. K. Anderson and G. R. Pond (2011). Early stopping rules in oncology: Considerations for clinicians. *European Journal of Cancer 47*, 2381–2386.

Nardini, C. (2013). Monitoring in clinical trials: Benefit or bias? *Theoretical Medicine and Bioethics*. Forthcoming.

Nardini, C., M. Annoni and G. Schiavone (2012). Mechanistic understanding in clinical practice: complementing evidence-based medicine with personalized medicine. *Journal of Evaluation in Clinical Practice 18*(5), 1000–1005.

Nardini, C. and J. Sprenger (2013). Bias and conditioning in sequential medical trials. *Philosophy of Science*. Forthcoming.

Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society. Series A 231*, 298–337.

O'Brien, P. and T. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics 35*, 549–56.

Ocana, A. and I. Tannock (2011). When are "positive" clinical trials in oncology truly positive? *Journal of the National Cancer Institute 103*(1), 16.

Orloff, J., F. Douglas, J. Pinheiro, S. Levinson, M. Branson, P. Chaturvedi, E. Ette, P. Gallo, G. Hirsch, C. Mehta et al. (2009). The future of drug development: advancing clinical trial design. *Nature Reviews Drug Discovery 8*(12), 949–957.

Owen, A. (2007). The ethics of two-and one-sided hypothesis tests for clinical trials. *Clinical Ethics 2*(2), 100–102.

O'Rourke, P. P., R. K. Crone, J. P. Vacante, J. H. Ware, C. W. Lillikei, R. B. Parad et al. (1989). Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the new born: A prospective randomized study. *Pediatrics 84*, 957–963.

PACE collaborative group (2008). The PACE initiative: Pragmatic Approaches to Comparative Effectiveness. Available at http://www.paceinitiative.org/index.html. Last accessed 18/10/2012.

Palmer, C. R. (1993). Ethics and statistical methodology in clinical trials. *Journal of medical ethics 19*, 219–22.

Palmer, C. R. (2002). Ethics, data-dependent designs, and the strategy of clinical trials: Time to start learning-as-we-go? *Statistical Methods in Medical Resarch 11*, 381–402.

Palmer, C. R. and W. F. Rosenberger (1999). Ethics and practice: Alternative designs for phase III Randomized Clinical Trials. *Controlled Clinical Trials 20*, 172–86.

Papineau, D. (1994). The Virtues of Randomization. *British Journal for the Philosophy of Science 45*, 437–50.

Parmar, M. K. B., G. O. Griffiths, D. J. Spiegelhalter, R. L. Souhami, D. G. Altman and E. van der Scheuren (2001). Monitoring of large randomised clinical trials: A new approach with Bayesian methods. *The Lancet 358*, 375–81.

Parmar, M. K. B., D. J. Spiegelhalter and L. S. Freedman (1994). The CHART trials: Bayesian design and monitoring in practice. *Statistics in Medicine 13*, 1297–312.

Paulo, R. (2002). *Problems at the Bayesian/Frequentist Interface*. Ph. D. thesis, Duke University, Department of Statistical Science. Available at http://www.isds.duke.edu/people/theses/rui.pdf. Last accessed 03/01/13.

Peck, R. W. (2007). Driving earlier clinical attrition: If you want to find the needle, burn down the haystack. Considerations for biomarker development. *Drug Discovery Today 12*, 289–94.

Peltzman, S. (1973). An evaluation of consumer protection legislation: The 1962 Drug Amendments. *The Journal of Political Economy*, 1049–1091.

Peto, R. and C. Baigent (1998). Trials: The next 50 years. Large scale randomised evidence of moderate benefits. *British Medical Journal 317*, 1170–1.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto and P. G. Smith (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer 34*, 585–612.

Pocock, S. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*, 191–99.

Pocock, S. (1992). When to stop a clinical trial. *British Medical Journal 305*, 235–40.

Pocock, S. and I. White (1999). Trials stopped early: Too good to be true? *The Lancet 353*, 943–4.

Pocock, S. J. (1993). Statistical and ethical issues in monitoring clinical trials. *Statistics in Medicine 12*, 1459–69.

Pocock, S. J. (2006). Current controversies in data monitoring for clinical trials. *Clinical Trials 3*, 513–21.

Psaty, B. M. and R. A. Kronmal (2008). Reporting mortality findings in trials of rofecoxib for Alzheimer disease or cognitive impairment. *Journal of the American Medical Association 299*(15), 1813–1817.

Psillos, S. (2007). *Philosophy of Science A–Z*. Edinburgh: Edinburgh University Press.

Pullman, D. and X. K. Wang (2001). Adaptive designs, informed consent, and the ethics of research. *Controlled Clinical Trials 22*, 203–10.

Rachels, J. (2009). Ethical theory and bioethics. In H. Kuhse and P. Singer (Eds.), *A companion to Bioethics 2nd ed.* London: Wiley-Blackwell.

Ronquist, F. and J. Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*(12), 1572–1574.

Royall, R. (1997a). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Royall, R. (1997b). *Statistical evidence: a likelihood paradigm*. London: Chapman & Hall.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association 95*, 760–768.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician 40*(4), 313–315.

Royall, R. M. (1991). Ethics and statistics in randomized clinical trials. *Statistical Science 6*(1), 52–8.

Sackett, D. L. (2000). Why randomized controlled trials fail but needn't: 1. failure to gain 'coal-face' commitment and to use the uncertainty principle. *Canadian Medical Association Journal 162*(9), 1311–14.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association 46*, 55–67.

Savage, L. J. (1962). *The Foundations of Statistical Inference*. London: Methuen. With prepared contributions from Bartlett, M.S., Barnard, G.A., Cox, D.R., Pearson, E.S. and Smith, C.A.B.

Schervish, M. J., J. B. Kadane and T. Seidenfeld (2003). Measures of incoherence: How not to gamble if you must. In J. Bernardo et al. (Ed.), *Bayesian Statistics 7: Proceedings of the 7th Valencia Conference on Bayesian Statistics*, pp. 385–402. Oxford: Oxford University Press.

Schroen, A. T., G. R. Petroni, H. Wang et al. (2010). Preliminary evaluation of factors associated with premature trial closure and feasibility of accrual benchmarks in phase iii oncology trials. *Clinical Trials 7*(4), 312–321.

Schumi, J. and J. Wittes (2011). Through the looking glass: Understanding non-inferiority. *Trials 12*, 106.

Seidenfeld, T. (1981). On after-trial properties of best Neyman-Pearson confidence intervals. *Philosophy of Science 48*(2), 281–291.

Senn, S. (1997). Statistical basis of public policy – present remembrance of priors past is not the same as a true prior. *British Medical Journal 314*.

Senn, S. (2001). Consensus and controversy in pharmaceutical statistics. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(2), 135–176.

Shapinn, S. M. (2000). Non-inferiority trials. *Current Controlled Trials in Cardiovascular Medicine 1*, 19–21.

Sharma, M. R. and R. L. Schilsky (2012). Role of randomized phase iii trials in an era of effective targeted therapies. *Nature Reviews Clinical Oncology 9*(4), 208–214.

Sismondo, S. (2008). How pharmaceutical industry funding affects trial outcomes: causal structures and responses. *Social Science & Medicine 66*(9), 1909–14.

Spanos, A. (2010). Is Frequentist Testing Vulnerable to the Base-Rate Fallacy? *Philosophy of Science 77*(4), 565–583.

Spiegelhalter, D. J. (2004). Incorporating Bayesian ideas into health-care evaluation. *Statistical Science*, 156–174.

Spiegelhalter, D. J., K. R. Abrams and J. P. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Wiley.

Spiegelhalter, D. J., L. S. Freedman and P. R. Blackburn (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials 7*, 8–17.

Spiegelhalter, D. J., L. S. Freedman and M. K. B. Parmar (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society, Series A 157*, 357–416.

Spiegelhalter, D. J., L. S. Freedman and M. K. B. Parmar (1996). Bayesian approaches to randomized trials. In D. A. Berry and D. K. Stangl (Eds.), *Bayesian Biostatistics*, pp. 67–108. New York (NY):Marcel Dekker, Inc.

Spiegelhalter, D. J., J. P. Myles, D. R. Jones and K. R. Abrams (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment 4*(38), 1–130.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science 41*(3), 211–226.

Sprenger, J. (2009). Evidence and experimental design in sequential trials. *Philosophy of Science 76*, 637–49.

Sprenger, J. (2013). Testing a precise null hypothesis: The case of Lindley's Paradox. *Philosophy of Science*. Forthcoming.

Stanev, R. (2012). Stopping rules and data monitoring in clinical trials. In *EPSA Philosophy of Science: Amsterdam 2009*, pp. 375–386.

Sterne, J. A. C. and G. Davey Smith (2001). Sifting the evidence – what's wrong with significance tests? *BMJ 322*, 226–31.

Stewart, D. J., S. N. Whitney and R. Kurzrock (2010). Equipoise lost: Ethics, costs, and the regulation of cancer clinical research. *Journal of Clinical Oncology 28*, 2925–2935.

Teira, D. (2011). Frequentist versus Bayesian clinical trials. In F. Gifford (Ed.), *Philosophy of Medicine*. North Holland.

Thomas, A., S. D. J. and W. R. Gilks (1992). BUGS: a program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics, vol. 4*, pp. 837–42. Oxford: Oxford University Press.

Todd, S., M. F. Baksha and J. Whitehead (2012). Sequential methods for pharmacogenetic studies. *Computational Statistics & Data Analysis 56*, 501–2.

Tsiatis, A. A. and C. Mehta (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika 90*, 367–78.

Tursz, T., F. Andre, V. Lazar, L. Lacroix and J. C. Soria (2011). Implications of personalized medicine—perspective from a cancer center. *Nature Reviews Clinical Oncology 8*(3), 177–183.

Urbach, P. (1985). Randomization and the Design of Experiments. *Philosophy of Science 52*, 256–73.

Veatch, R. M. (2002). Indifference of subjects: An alternative to equipoise in randomised clinical trials. *Social Philosophy and Policy 19*, 295–323.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review 14*(5), 779–804.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wardell, W. M. and L. Lasagna (1975). *Regulation and Drug Development*. Washington (DC): American Enterprise Institute for Public Policy Research.

Ware, J. H., J. E. Muller and E. Braunwald (1985). The Futility Index: An approach to the cost-effective termination of randomized clinical trials. *The Americn Journal of Medicine 78*, 635–43.

Weijer, C. and P. Miller (2003). Therapeutic obligation in clinical research. *Hastings Center Report 33*(3), 3.

Wells, S. A. and J. R. Nevins (2004). Evolving strategies for targeted cancer therapy–Past, present, and future. *Journal of the National Cancer Institute 96*(13), 980–981.

Westover, M. B., K. D. Westover and M. T. Bianchi (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine 9*, 20. http://www.biomedcentral.com/1741-7015/9/20.

White, S. J., D. Ashby and P. J. Brown (2000). An introduction to statistical methods for health technology assessment. *Health Technology Assessment 4*(8).

Wilcox, R. A., B. Djulbegovic, G. H. Guyatt and V. M. Montori (2008). Randomized trials in oncology stopped early for benefit. *Journal of Clinical Oncology 26*(1), 18–19.

World Medical Association (2008). Declaration of Helsinki, 6th revision.

Worrall, J. (2007a). Evidence in Medicine and evidence-based Medicine. *Philosophy Compass 6/7*, 981–1022.

Worrall, J. (2007b). Why there is no cause to randomize. *British Journal for the Philosophy of Science 58*, 451–488.

Worrall, J. (2008). Evidence and ethics in medicine. *Perspectives in Biology and Medicine 51*, 418–31.

Zwarenstein, M., S. Treweek, J. J. Gagnier et al. (2008). Improving the reporting of pragmatic trials: An extension of the CONSORT statement. *British Medical Journal 337*, a2390.