FREQUENTIST VERSUS BAYESIAN CLINICAL TRIALS

David Teira

INTRODUCTION

Stuart Pocock [1983] defined clinical trials as any planned experiments, involving patients with a given medical condition, which are designed to elucidate the most appropriate treatment for future cases. The canonical example of experiments of this sort is the drug trial, which is usually divided into four phases[1]. Phase I focuses on finding the appropriate dosage in a small group of healthy subjects (20-80); thus such trials examine the toxicity and other pharmacological properties of the drug. In phase II, between 100 and 200 patients are closely monitored to verify the treatment effects. If the results are positive, a third phase, involving a substantial number of patients, begins, in which the drug is compared to the standard treatment. If the new drug constitutes an improvement over the existing therapies and the pharmaceutical authorities approve its commercial use, phase IV trials are begun, wherein adverse effects are monitored and morbidity and mortality studies are undertaken.

This paper focuses on phase III drug trials. The standard experimental design for these trials currently involves a randomised allocation of treatments to patients. Hence the acronym RCTs, standing for randomised clinical (or sometimes controlled) trials[2]. The statistical methodology for planning and interpreting the results of RCTs is grounded in the principles established by Ronald Fisher, Jerzy Neyman and Egon Pearson in the 1920s and 1930s. A hypothesis is made about the value of a given parameter (e.g., the survival rate) in a population of eligible patients taking part in the trial. The hypothesis is tested against an alternative hypothesis; this requires administering the drug and the control treatment to two groups of patients. Once the end point for the evaluation of the treatment is reached, the interpretation of the collected data determines whether or not we should accept our hypothesis about the effectiveness of the drug, assigning a certain probability to this judgment.

[1] Clinical trials can be set to analyse many different types of treatment: not only drugs, but also medical devices, surgery, alternative medicine therapies, etc. The characteristics of these types of trials are quite different; so, for the sake of simplicity, I will only deal here with drug testing.
[2] In this paper, for the sake of simplicity, 'RCTs' will refer to standard frequentist trials. Notice though that randomization may well feature in the design of a Bayesian clinical trial.

This statistical methodology is based on a specific view of probability – called the frequestist approach -- according to which probabilities are (finite or infinite) relative frequencies of empirical events: here these are treatment effects in a given experimental setting. However, there are alternative interpretations of the axiomatic definition of probability and it is possible to construct clinical trials from at least one of these: Bayesianism. In the Bayesian approach, probabilities are conceived as degrees of belief. Hence, for instance, these probabilities can be calculated on the basis of whatever information is available and are not tied to a particular trial design [Berry, 2005]. Unlike in the case of standard RCTs, we can calculate these probabilities with or without randomisation and with any number of treated patients. Hence, depending on the conception of probability we adopt, clinical trials can be designed and their results interpreted in different manners, not always convergent.

The first clinical trial planned and performed following a frequentist standard was the test of an anti-tuberculosis treatment, streptomycin. It was conducted in Britain and published in 1948. Over the following decades, RCTs would be adopted as a testing standard by the international medical community and by pharmaceutical regulatory agencies all over the world. Today, RCTs constitute the mainstream approach to drug testing and, through evidence-based medicine, they even ground standards of medical care. The 1980s brought a boom in Bayesian statistics, with many practical implementations in medicine, as well as in other disciplines. As soon as the computing power required by Bayesian calculations became available, increasingly sophisticated Bayesian trials were designed and implemented. It has been argued that these trials may be more efficient and more ethical than a frequentist RCT: e.g., reaching a cogent conclusion about the efficacy of a treatment may require fewer participants, minimising the number of patients exposed to the risks of the experiment. Today, there is debate about whether regulatory agencies, and in particular the FDA, should accept evidence from Bayesian trials as proof of the safety and efficacy of a drug. If (or, rather, *when*) this happens, frequentism may lose its commanding position in the field of medical experiments. The question is whether the grounds for this change are in fact sound.

The aim of this paper is to provide an overview of the philosophical debate on frequentist versus Bayesian clinical trials. This has been an ongoing discussion over the

last thirty years and it is certainly not closed. The comparison between these approaches has focused on two main dimensions: the epistemology of the statistical tools (e.g., *p*-values vs. prior and posterior probabilities) and the ethics of the different features in each experimental design (e.g., randomisation). As of today, the mainstream view among philosophers (certainly not among biostatiscians) is that RCTs are epistemically and ethically problematic and a Bayesian alternative would be welcome. I would like to add a third dimension of comparison, so far neglected in this debate: the advantages of each approach as a regulatory yardstick. I contend that a fair comparison between these two approaches should simultaneously consider three dimensions: epistemological, ethical and regulatory. Philosophers and statisticians care deeply about the epistemological issues. Physicians and patients are equally concerned about the ethical issues. But we all care, as citizens, about the regulatory issues. There is a trade-off between these three different dimensions and the perfect trial that would satisfy all the concerned parties may well not exist.

Most of the conflicts created by RCTs derive from the regulatory constraints imposed on medical experimentation. In a world where clinicians and patients were free to negotiate which testing standard was more mutually suitable for their goals in research and care, it is likely that the frequentist and Bayesian trials would both flourish. Yet for the last 100 years we have lived in a regulated world in which we want state agencies to conduct trials in order to determine whether treatments are safe and effective enough to warrant authorisation of their commercial distribution. RCTs were adopted as a testing standard by many of these regulatory agencies and, despite their epistemic and ethical flaws, they seem to have done a good job in keeping harmful compounds off pharmaceutical markets. As long as we want this type of regulatory supervision, we should be willing to accept certain constraints on our testing methodologies (be these frequentist or Bayesian) whenever we conduct experiments in order to gain regulatory approval.

I will open the first part of this paper by trying to elucidate the frequentist foundations of RCTs. I will then present a number of methodological objections against the viability of these inferential principles in the conduct of actual clinical trials. In the following section, I will explore the main ethical issues in frequentist trials, namely those related to randomisation and the use of stopping rules. In the final section of the first part, I will

analyse why RCTs were accepted for regulatory purposes. I contend that their main virtue, from a regulatory viewpoint, is their impartiality, which is grounded in randomisation and fixed rules for the interpretation of the experiment.

Thus the question will be whether Bayesian trials can match or exceed the achievements of frequentist RCTs in all these respects. In the second part of the paper, I will first present a quick glimpse of the introduction of Bayesianism in the field of medical experiments, followed by a summary presentation of the basic tenets of a Bayesian trial. The point here is to show that there is no such thing as "a" Bayesian trial. Bayesianism can ground many different approaches to medical experiments and we should assess their respective virtues separately. Thus I present two actual trials, planned with different goals in mind, and assess their respective epistemic, ethical and regulatory merits. In a tentative conclusion, I contend that, given the constraints imposed by our current regulatory framework, impartiality should preside over the design of clinical trials, even at the expense of many of their inferential and ethical virtues.

1.1 In what sense are RCTs grounded in frequentism?

Running a phase III clinical trial is a manifold task, which goes far beyond its statistical underpinnings. The credibility (and feasibility) of a trial is conditional on a complete preplanning of every aspect of the experiment. This plan is formally stated in the study protocol. The following items should feature in the protocol, according again to Pocock [1983, p. 30]:

Background and general aims	Patient consent
Specific objectives	Required size of study
Patient selection criteria	Monitoring of trial progress
Treatment schedules	Forms and data handling
Methods of patient evaluation	Protocol deviations
Trial design	Plans for statistical analysis
Registration and randomisation of patients	Administrative responsibilities

The aim of a trial is to test a hypothesis about the comparative efficacy of an experimental treatment (be it with the standard alternative or a placebo). Leaving aside

for a moment the statistical design of the test, first it is necessary to define which patients are eligible for the study (e.g., they should be representative of the disease under investigation); how to create an experimental group and a control group; how to administer treatment to each of them; and what the end-points for the evaluation of their responses are. During the course of the trial, an *interim analysis* is usually performed in order to monitor the accumulating results, since reaching the number of patients specified in the design may take months or years and because the information gleaned from such an interim analysis may in fact warrant some action such as terminating the trial early. Once the trial is completed, the hypothesis about the comparative efficacy of the treatment will be either accepted or rejected and the results published. Depending on the disease and the planned sample size, this may add several years to the time taken up by the two previous trial phases. Thus the development of a new drug may well take a decade before it is approved for public use by the pharmaceutical regulatory agency[3].

In this section, we will focus only on those aspects of the trial more directly connected to the frequentist view that have been more broadly discussed in the medical literature: namely, randomisation as a treatment allocation mechanism, on the one hand, and the use of significance testing and confidence intervals in the analysis of the results of the trial, on the other. The goal of this section will be limited to showing how these concepts are related to the frequentist interpretation of probability. It is important to clarify them in order to show the real scope of the Bayesian alternative: paradoxically, *p*-values and confidence intervals are often understood as if they measured some kind of posterior probability -- i.e., as if they were measuring Bayesian degrees of belief for certain events rather than frequencies.

Let us start with randomisation. Once a patient is deemed eligible (according to the trial's protocol) and recruited, the informed consent form signed and the log sheet with her identification details filled out, the treatment is assigned at random. Depending on the arrangement of the trial (number of treatments, whether or not it is double blinded, whether or not it is multi-centre), randomisation may be implemented in different ways. The general principle is that each patient should have an equal probability of receiving

[3] For an updated account of the practical arrangements involved in a trial, including compliance with the current regulatory constraints, see [Hackshaw, 2009, pp. 157-201]. The book provides a concise overview of every dimension of a clinical trial today. For a thorough overview, see [Piantadosi, 2005].

each treatment. If it is convenient to control the allocation of treatments according to patient characteristics, in order to prevent imbalances, randomisation can be *stratified*. What is the statistical rationale of this procedure?

Let us informally sketch Fisher's original argument for randomisation (as reconstructed in [Basu, 1980]). In order to study the response differences between the two treatments in trial patients, we need a test statistic with a known distribution: for instance, $T=\Sigma d_i$, where d_i is the response difference. Assuming the hypothesis that there is no difference between treatments, suppose we observe a positive difference d_i between treatments in a given pair of patients who received them at random. Assuming that our hypothesis is true, this difference must have been caused by a nuisance factor. If we kept this factor (and all other factors) constant and repeated the experiment, the absolute value of $|d_i|$ would be the same with the same sign, if the treatments were identically allocated; it will be reversed if the allocation had been different. The sample space of T will be the set of 2^n vectors $R = \{(\pm d_1, \pm d_2, \ldots, \pm d_n)\}$

Randomisation warrants that all these vectors will have an equal probability. If d_i is positive for all i, we will observe another n response differences d'_i equal or bigger than d_i if and only if $d'_i = d_i$. The probability of observing this response is $(½)^n$, the significance level of the observed differences, as we will see below. This probability was, for Fisher, the frequency of observing this difference in an infinite series of repetitions of the experiment. And we will need it in order to calculate how exceptional the results of our experiment have been.

This statistical rationale for randomisation is usually skipped in medical textbooks, where random allocations are usually justified through the following two arguments. First, randomisation prevents selection bias: it prevents investigators from assigning (consciously or unconsciously) patients with, say, a given prognosis to any one of the treatments. For instance, an investigator might allocate the experimental treatment to the healthier patients, if she wants the trial to be positive, or to the patients with a worse prognosis, if she thinks they will benefit more. This is an argument that never fails to appear in medical textbooks and, as we will see below, it was extremely influential in the acceptance of clinical trials by the medical profession, at least in the United Kingdom and the United States. A second argument that is often cited in medical

textbooks to justify randomisation can be traced back to Fisher's famous *tea tasting* experiment. In clinical trials, randomisation would allow control over unknown prognostic factors, since, over time, their effect would be distributed in balance between the two groups. Bayesians and frequentists usually accept the first argument –we will see more about this in the second part of this paper. But, as we will see in the following section, there is more disagreement about the second argument within both approaches. However, neither of these two arguments presuppose a particular conception of probability, so we will not develop them at more length here.

Let us focus instead on the statistical interpretation of test results. The aim is to evaluate how *significant* they are under a number of probabilistic assumptions. Again, it is often the case that their statistical rationale is only partially explained in medical textbooks, giving rise to great confusion about what clinical trials actually mean. So let us revisit once more the original rationales for significance levels, because, as we will see, the medical community (as it is often the case in many social sciences) uses a combination of them.

The use of significance tests certainly predates Fisher. Leaving aside previous uses in astronomy, Karl Pearson was already using them to measure the discrepancy between a theoretical distribution of probability and a curve of empirical frequencies, using χ^2 as a test of the "goodness of fit" [Cowles & Davis, 1982]. If the probability of observing a given value of χ^2 was below 0.1, Pearson considered the goodness of fit "a little improbable". But this implied nothing about the truth or falsity of any hypothesis – being a committed positivist, Pearson viewed curves just as summaries of observations. W. S. Gosset made a more precise estimate of significance levels, arguing they should be "three times the *probable error* in the normal curve": the odds of such an observation were approximately 30 to 1, which was usually rounded to 0.5. In the 1920s Fisher restated the concept within his own statistical framework. He was a frequentist for whom any probability judgment should be theoretically verifiable to any chosen degree of approximation by sampling its reference set. However, Fisher admitted various ways to represent our uncertainty depending on the extent of our prior knowledge[4].

Significance tests will better assess the plausibility of a given hypothesis (the *null hypothesis*) about which not much is previously known. It should allow us to specify a

[4] Fisher's positions is certainly simplified here. For a brief comparative discussion see [Lehmann, 1993].

unique distribution function for the *statistic* that we will use to test it. But there may be many different such statistics available. With this function, we can calculate the probability of each possible value of the statistic on the assumption that the hypothesis is true. Once the experiment is run and actual data provide the observed value of the statistic, we can also calculate how likely it is, assuming the truth of the hypothesis, to obtain a result with less or equal probability than the observed one: this is the *p-value*. In other words, the *p*-value is the proportion of an infinite series of repetitions of an experiment, all conducted assuming the truth of the null hypothesis, that would yield data contradicting it as strongly as or more so than the observed result. Therefore, the *p*-value is a probability of observed and unobserved results which is tied to the design of the experiment and cannot be properly interpreted without it.

Suppose the probability of observing a result within this tail area is less than 0.5: if one such result occurs in the experiment, Fisher would interpret it as a serious deviation from what we would have expected, were the hypothesis true. Such a result would make the hypothesis "implausible": either an exceptionally rare chance has occurred or the hypothesis is not true. But the data alone cannot establish whether the former or the latter is the case (or whether both are).

Fisher was careful (usually, but not always) to remark that a single experiment did not provide solid enough grounds to demonstrate any natural phenomenon. Only when an experiment is repeated and delivers results that systematically deviate from the hypotheses tested can we judge the latter to be implausible. However, the truth of the hypothesis can never be established with significance testing: it is just assumed. Neyman and Pearson developed a different rationale for the testing of hypotheses: instead of assessing the plausibility of a single (null) hypothesis, we should have a criterion for choosing between alternative exclusive hypotheses, with a known probability of making the wrong choice in the long run. Errors could be of two kinds: rejecting the null hypothesis when it is true or accepting it when it is false.

Using the probability distribution of the statistic, we define a rejection region R: if the observed value of the statistic falls within R, the null hypothesis (H_0) should be rejected and the alternative hypothesis (H_1) accepted; if the observed value falls outside R, H_0 should be accepted and H_1 rejected. The probability α of making an error of the first kind (accepting H_1 when it is false) is called the *size* of the test; given the probability β of making an error of the second kind (rejecting H_1 when it is true), the *power* of the test

amounts to 1-β. In order to construct the test, we should decide which hypothesis would be the null, in order to minimize the probability α of an erroneous rejection. We then choose a rejection region with the desired probability α that maximises the power of the test. Achieving this power implies a certain sample size (a given number of patients in a trial).

In the Neyman-Pearson approach, instead of measuring how implausible the observed result makes H_0 (without any actual probability value), α gives us the probability of incorrectly rejecting H_0 in a hypothetical long run of repeated experiments. Again, nothing is concluded about the truth of H_0: accepting a hypothesis implies, at most, acting as if it were true. Whereas Fisher wanted significance to ground an inductive inference (repeated experiments would make H_0 implausible), Neyman calculated probabilities (size and power) for a test, trying to minimize their epistemic import. For Neyman, we cannot know that H_0 will be incorrectly rejected in only a given number of instances: we can only *decide* to believe it[5].

In its more widespread interpretation, the Fisherian *p*-value would somehow express the inductive support that a hypothesis receives from certain experimental data: given a certain observation, and assuming the hypothesis is true, it is the probability of observing it or a more extreme result[6]. The Neymanite significance level α is a deductively established probability of making type I errors in a series of experiments, before observing any particular result.

Fisher was extremely unhappy with the approach advanced by Neyman and Pearson. Leaving aside technical objections, Fisher considered Neyman's behaviouristic tests as an industrial procedure aimed at cutting experimental costs, not at solving inferential problems. However, as Gigerenzer et al. [1989] put it, their respective views were merged in a sort of "hybrid theory" that textbooks popularised over the second half of the 20th century. Neyman's behaviourism was dropped and error probabilities were given an epistemic interpretation: the *p*-value became an observed α, a post trial error rate that measured the inductive evidence for a hypothesis. This is what Steve Goodman [1999a] calls the *p-value fallacy*. In a similar vein, a confidence interval is often simply

[5] In Neyman's [1957, p. 12] own words, this is "an act of will to behave in the future (perhaps until new experiments are performed) in a particular manner, conforming with the outcome of the experiment". From a Bayesian perspective, this inductive behaviour is just decisión-making without loss functions.

[6] As Donald Gillies made me notice, after 1930 Fisher himself would have preferred the fiducial argument by way of inductive measure: see [Gillies, 1973; Seidenfeld, 1979] for an analysis.

understood as a range within which the true outcome measure is likely to lie, without any mention of error probabilities in the long run. As we will see in the second part of this paper, such misinterpretations would correspond more to Bayesian posterior probabilities than to the original frequentist definition.

It is an open question to what extent these sorts of misconceptions have actual consequences on the assessment of the safety and efficacy of drugs. Perhaps a better understanding of the scope of *p*-values and confidence intervals would contribute to reducing the confusion generated by so many trials with apparently mutually contradictory results[7]. However, this confusion may well have other sources, such as, for instance, the publishing practices inspired by pharmaceutical companies [Sismondo 2009]. For the time being, I hope the previous clarification is enough to clarify in what sense RCTs are conceptually grounded in the frequentist paradigm. In the following section we will examine a number of objections concerning the possibility of implementing RCTs according to this very demanding standard.

1.2 METHODOLOGICAL ISSUES

The controversy over the foundations of statistics between frequentists and Bayesians is too long and deep to be summarised here. Equally beyond the scope of this paper is a discussion of the technical objections addressed by each party against their respective approaches to clinical trials[8]. I will focus instead on the philosophical debate on the flaws of frequentist RCTs, presenting a number of arguments that hold independently of any conception of probability[9]. These objections, listed below, have been developed over the last thirty years, mainly by Peter Urbach and John Worrall, without much response so far. The reader may now judge to what extent they are conclusive.

Objection #1: which population?

[7] Statistical mistakes of this sort were soon denounced in the medical literature: see, for instance, Mainland 1960. For an update, see, e.g., [Sterne and Smith, 2001].
[8] The interested reader can catch a glimpse of this debate in the special issue on this topic of the journal *Statistics in Medicine* 12: 1373-1533, 1993.
[9] A connected but separate issue that I will not address here either is the scope of RCTs in causal inference, which has also received some philosophical attention: see, e.g., [Cartwright, 2007; Papineau, 1994] and also Dan Steel's paper in this volume.

In a clinical trial there is no real random sampling of patients, since the population random samples should be drawn from remains usually undefined: there is no reference population, just criteria of patient eligibility in the trial protocol. Generalizing from the individuals entered into the study to any broader group of people seems ungrounded [Urbach, 1993].

Objection #2: significant events may not be that rare

A positive result in a significance test is interpreted as an index that H_0 is false. Were it true, such result would be an "exceptionally rare chance". It would be exceptional because a randomised allocation of treatments would ideally exclude any alternative explanation: uncontrolled factors would be evenly distributed between groups in a series of random allocations. However, it would not be "exceptionally rare" that the treatment was effective in the case where it had been allocated to the healthier patients alone, to those with best prognoses or to any group of patients that for whatever reason could differentially benefit from the treatment.

Colin Howson, among others, has argued that randomisation as such does not guarantee that the occurrence of such unbalanced allocation *in a particular trial* is rare: it may be produced by uncontrolled factors. As Worrall [2007, pp. 1000-01] puts it, "randomisation does not free us from having to think about alternative explanations for particular trial outcomes and from assessing the plausibility of these in the light of 'background knowledge'". This further assessment cannot be formally incorporated, as it should be, into the methodology of significance testing. Hence, we cannot ground our conclusions on this methodology alone.

Objection #3: post randomisation selection

By sheer chance, a random allocation may yield an unbalanced distribution of the two treatments, i.e., the test groups may differ substantially in their relevant prognostic factors (these are called *baseline imbalances*). This difference may bias the comparison between treatments and spoil the experiment. If one such distribution is observed, the customary solution is to randomise again seeking a more balanced allocation. However, argues Urbach [1985], the methodology of significance testing forbids any choice between random allocations: if they are adequately generated, any allocation should be

as good as any other. Hypotheses should be accepted or rejected on the basis of the experiment alone, without incorporating our personal assessment of the generated data (justified though it may be).

It is usually assumed that with a high number of enrolled patients, it is very unlikely that randomisation generates unbalanced groups. Urbach argues that we cannot quantify this probability and much less discard it. At best, a clinical trial provides an estimation of the efficacy of a treatment, but there is no direct connection between this result and the balance of the two groups. The conclusions of the trial can be spoiled by the following two objections.

Objection #4: unknown nuisance variables after randomisation

Even after randomising, uncontrolled factors may differentially influence the performance of a treatment in one of the groups. Further randomisations at each step in the administration of the treatment (e.g., which nurse should administer it today?) may avoid such interferences, but this is quite an impractical solution. Declaring such disturbances negligible, as many experimenters do, lacks any internal justification in the statistical methodology assumed [Urbach, 1985; Worrall, 2007].

Objection #5: known nuisance variables

It has been argued that randomisation can at least solve the problem created by known perturbing factors that are difficult to control for. These could be at least randomised out. Following Levi [1982], Urbach [1985, p. 267] argues that since we know of no phenomena correlated to these confounding factors, "there is no reason to think that they would balance out more effectively between groups by using a physical randomising device rather than employing any other method".

Objection #6: RCTs do not necessarily perform better than observational studies

Despite all these objections, it is often claimed that RCTs are more reliable than non-randomised "observational" studies such as, for instance, case-control studies, where retrospective samples of cases and controls matched for known risk factors are

compared. Cohort studies or registry databases may also provide information about comparative interventions. In the 1970s and 1980s analyses of randomised and non-randomised trials of a given treatment showed that the estimated effects were higher in the latter than in the former[10]. If we assume that RCTs provide the more reliable estimation of the true effect of a treatment, we can conclude that the observational studies indeed "exaggerated" the effects. However, a recent wave of analyses concerning the quantitative bias of observational studies shows that there might not be such overestimation. In view of all these, Worrall [2007, p. 1013] concludes that there is "no solid independent reason for thinking that randomisation has automatic extra epistemic weight": if we do not commit *ex ante* to RCTs as the gold standard to provide the estimation of the effects of a treatment, the comparison is not necessarily unfavourable to observational studies.

These six objections are sound, in my view. Even if, over the last 50 years, RCTs have certainly succeeded in identifying effective and ineffective treatments, their *a priori* epistemic grounds are not as flawless as you might think if you just relied on the standard biomedical literature. There is certainly room for competing alternatives, as we will have the occasion to discuss in the second part of this paper. However, let me close this section now noting that there is quite a general agreement, even among Bayesian critics, about one argument *for* randomisation: it offers protection against selection biases. As I already mentioned, the medical profession has always appreciated this epistemic virtue of randomisation, perhaps because there has been a clear awareness of biases of this sort in the medical literature for at least a hundred years if not more. Allocating treatments at random prevents any manipulation and guarantees a fair comparison. This argument for randomisation is also independent of any particular view of probability[11] and, as we will see below, played a central role in the acceptance of RCTs as regulatory standards, as I will discuss in section 1.5 below. But let us now examine a different source of objections against frequentism in clinical trials: the ethical dilemmas it leads to.

1.3 ETHICAL ISSUES

[10] See [Worrall, 2007, pp. 1009-1013] for a discussion.

[11] See [Berry and Kadane, 1997] for a nice decision-theoretic argument for randomisation developed from a Bayesian perspective. See also how the impossibility of manipulation provides a very good defence for observational studies in [Vandenbroucke, 2004].

Randomised clinical trials are, obviously, experiments with human subjects. As such, they are usually conducted under external supervision according to the ethical principles approved in the Nuremberg code, the Helsinki declaration, and other national and international guidelines[12]. Trials are conducted for research purposes, and the design of the experiment often imposes constraints on the standards of care that patients may receive. Many ethical dilemmas arise therein. In this section I will only focus on the conflicts more directly related to the frequentist foundations of RCTs: namely, the ethical issues involved in randomisation (as a treatment allocation procedure) and in the stopping rules that may close a trial before it reaches the targeted sample size[13].

There is a common stance regarding the ethics of randomisation: it is only acceptable when there is genuine uncertainty in the medical community about which one among the allocated treatment is most beneficial for a patient (in the population determined by the study's eligibility criteria)[14]. This is often referred to as *clinical equipoise*. Ideally, this would be reflected in the null hypothesis adopted (no difference between treatments) and the trial should eliminate this uncertainty. It is open to discussion whether there is a sound ethical justification for random assignment rather than patient or doctor choice whenever clinical equipoise obtains, or whether this is just an *ad hoc* ethical principle to justify the random allocation of treatments required by significance testing. There has long been evidence that individual clinicians have preferences about the best treatment for their patients, in particular when the illness is serious and the risks and possible benefits are not negligible[15]. This could be interpreted as resistance to treat them as the indeterminate members of a statistical population, as required in the statistical design of the experiment.

But even if there were genuine equipoise, why would it be ethical to allocate treatments at random? The standard argument for justifying the participation of patients in clinical trials draws on the general normative principles usually applied in bioethics after the

[12] However, in developing countries the regulation of clinical trials is significantly softer and this creates a clear incentive for the industry to conduct their tests there: for an overview and discussion see [Macklin, 2004].
[13] For a general overview of bioethics with particular attention to clinical trials, see [Beauchamp and Childress, 2001] and [Levine, 1998].
[14] For a critique, see, for instance, [Gifford, 1986 and 1995].
[15] E.g., [Taylor et al., 1984].

Belmont Report: autonomy, beneficence or non-maleficence, and justice. Autonomy is granted if the patients consent to receive their treatment at random after being properly informed about the clinical equipoise of both treatments and the research design of the trial. If the equipoise is genuine, then random allocation is consistent with the *expected* effect being as good as possible. As for justice, if there finally were a difference between treatments despite the initial equipoise, those who received the less effective one did so at random, which doesn't seem intuitively unfair. The principles of autonomy and justice bear a more direct connection to the statistical assumptions of the trial, so let us discuss them in more detail.

For all practical purposes, the autonomy of every patient in a trial is grounded in the informed consent she gives to participate in the experiment, signing a formal agreement before it starts. This is a legal requirement in many countries and, in addition, an Institutional Review Board usually oversees the process. There are different standards concerning the information that the patient should receive before giving consent, but it should certainly include the fact that the trial is for research purposes, the fact that participation is voluntary, and an explanation of the procedures to be followed. In RCTs, there is at least a paragraph about the random allocation of treatments, stated in a non-technical language[16]. However, there is qualitative evidence that patients often misunderstand these paragraphs, making their informed consent to randomisation dubious. Moreover, there is also evidence that clarifying this confusion is often difficult, if not expensive[17].

Various surveys of the patients' motivation to take part in trials (e.g., [Edwards et al., 1998]) point out that a randomised allocation of treatments is at odds with their goals: they are expecting to benefit personally from the treatment and the more information there is about the different effects of each drug, the more reluctant they are to a random assignment. It is often cited in this context how AIDS activists subverted research protocols in the early 1980s trials: among other things, they exchanged treatments after randomisation in order to increase their probabilities of receiving the experimental drug

[16] E.g., "You will be randomised into one of the study groups described below. Randomisation means that you are put into a group by chance. It is like flipping a coin. Which group you are put in is done by a computer. Neither you nor the researcher will choose what group you will be in. You will have an EQUAL/ONE IN THREE/ETC. chance of being placed in any group" (From the informed consent template developed by the American national Cancer Institute in 1998, included as an appendix in [Hartnett, 2000])

[17] See, e.g., [Featherstone and Donovan, 1998; Flory and Emanuel, 2004].

[Epstein, 1996]. They vindicated their autonomy to bear the risks of receiving untested treatments and succeeded in gaining access to the first antiretroviral drug, AZT, before any trial was completed. If patients perceive any difference in the treatments that they can benefit from, they may well prefer to dispense with randomisation. Such differences exist: equality among treatments may refer just to the single quantified outcome measured in the trial, but the quality of life that each treatment yields may significantly differ.

The question of the benefits that a patient can expect from a trial is also relevant for the discussion of the justice of randomisation. Intuitively, patients can perceive randomisation as a fair lottery. However, lotteries are considered fair procedures when the good allocated is scarce. This was sometimes the case in clinical trials, but usually there are doses of the experimental treatment for every patient in the experiment, even if only half of them receive it. What should be distributed are the potential benefits and burdens of the test, which are a priori unknown. The fairness of such a distribution does not rely on the outcome (some may win and some may lose, none of them deserving it), but rather on the impartiality of the allocation. No patient can claim that the allocation was intended to favour one person over another.

The best formulation for the view of justice intuitively captured in the idea of a fair lottery is probably a contractarian one [Stone, 2007]. If the participants in a trial acknowledge that, all of them being equally eligible, all of them have equally strong claims to receive the potential benefits and burdens of the trial and, on the other hand, no other consideration is taken into account, then it seems plausible that they would agree to use an equiprobable lottery in order to distribute whatever comes out of the treatments. However, if we adopt a different approach to justice, the fairness of randomisation can be questioned. In a utilitarian perspective, for instance, the allocation of treatments would be fair if it maximised the social utility of the participants in the trial (or, perhaps, society as a whole). There is no a priori reason to presume that a randomised allocation would achieve this. E.g., if equipoise fails concerning the comparison of these treatments, there may be differences in the expected utility that each treatment may yield to each participant. Hence certain non-random allocations may yield a superior average expected utility superior and be ethically preferable from a purely utilitarian perspective.

To sum up, from an ethical perspective, randomisation is quite controversial, and it seems clear that if it were possible to avoid it, this would bring ethical gains. Alternative approaches to randomisation may fare better in some respects, as we will see in more detail in the second part of the paper. Yet if we want to interpret the results of the trial through significance tests, some sort of randomisation is necessary.

The second ethically contentious topic regarding frequentism in clinical trials is that concerning stopping rules[18]. It may happen that before completing the trial, a larger than expected treatment effect, either beneficial or harmful to the patients, is observed in the experimental arm. Or, alternatively, it may appear that the experimental therapy is having no effect. Hence the trial might be terminated early due to the very bleak prospect of demonstrating any effect, at the risk of not reaching the statistical power initially envisaged in the protocol, which is tied to the sample size (i.e., the number of patients treated in the trial). In order to justify this decision, certain factors should be considered. Namely, the plausibility of the observed effect, the number of patients already recruited, the number of interim analyses performed, and the monitoring method applied. If the trial takes a long time to be completed, the protocol will specify a number of *interim analyses* (e.g., according to certain clinical endpoints). At each stage, there will be a stopping rule providing a criterion for whether or not to continue the trial. The patients' interests are usually considered in the choice of the interim endpoints. As mentioned above, a common view about the ethics of trial interruption nowadays is that this should happen as soon as the evidence accumulated contradicts the initial assumption of clinical equipoise. However, if the effect of the experimental drug is, at that point, positive, should we stop the trial and use it on other patients without conducting an additional trial in full? [19]

Our views on this question will depend on the epistemic standard we adopt. We might choose the standard view in evidence-based medicine, namely that only accomplished RCTs with a given statistical power count as proper evidence of the safety and efficacy

[18] See [Baum et al., 1994; Cannistra, 2004] for a general discussion. See also [Mueller et al., 2007; Goodman, 2007] for a discussion incorporating the Bayesian perspective.
[19] This is what Gifford [2000, p. 400] calls the RCT dilemma: if trials are stopped as soon as clinical equipoise vanishes, but before we reach their predefined statistical endpoints, there would be no point in designing the experiment in search of a certain level or significance or power.

of a treatment. In this view, it is unethical to administer the experimental drug to a patient without completing a trial[20]. Cases in which patients have been injured after receiving an improperly tested drug are cited for this position: thalidomide provides a good example. Under the trade name *Contergan*, one million West Germans consumed thalidomide as a sedative in the early 1960s –and subsequently many more people around the world. Reports were soon published in medical journals showing an association between the drug and peripheral neuropathy and, later, between the drug serious birth defects when consumed by pregnant women. Only then did the manufacturer withdraw the drug from European markets, but eight thousand babies had been already born with severe deformities. At that point, there was no clear regulatory standard about the safety of a compound, neither in the United States nor in Europe. As we will see in the following section, the thalidomide scandal prompted the approval of more strict regulations, leading to the current prevalence of RCTs. However, there are cases in which lives were lost in additional trials for a treatment whose efficacy was seemingly clearly evident, but not statistically grounded in a proper RCT: e.g. the ECMO trials, as analysed by John Worrall [2008].

A recent systematic review shows that the number of trials that are being stopped early for apparent benefit is gradually increasing [Montori et al., 2005]. This decision is usually not well justified in the ensuing reports: the treatment effects are often too large to be plausible, given the number of events recorded. Again, this is open to various interpretations: trials may have been stopped out of genuine concern for the patients' welfare, but less altruistic motivations could have also played a role (e.g., pressure from the funding body or the urgency of an impact publication). Yet, this review [Montori et al., 2005] depends on the evidentiary standard we adopt: if we only consider credible the evidence originating from properly powered RCTs, we should be sceptical about the results of early stopping trials. However, if we accept alternative sources of evidence, as Worrall suggests, we may accept some of the results from these trials as legitimate.

Just as it happened with randomisation, the problem is whether there is any alternative standard for judging clinical evidence which is at least as epistemically strong as RCTs,

[20] Therefore stopping rules should be calibrated depending on the trade-off between benefits for the participants in a trial and benefits for future patients in order to minimize the loss of information if the trial has to be interrupted. See [Buchanan and Muller, 2005] for a discussion.

but causes less ethical trouble. Part of the strength of the Bayesian approaches we will discuss in the second part of this paper is that, in principle, they can solve these problems. On the one hand, randomisation is not strictly necessary for inference (even if it can be defended on other grounds); hence perhaps following a Bayesian approach will allow us to avoid it. On the other hand, a trial may be stopped at any point without disturbing the statistical validity of the results: the conclusions will be as strong as the evidence gathered so far.

But before we get to examine this alternative, we should first consider just what the original alternatives to RCTs were and why the latter succeeded so quickly. Relatedly, it is worth noticing here that the ethical dilemmas we have just discussed are not created by RCTs as such, but rather by our current regulatory framework, in which RCTs feature prominently as a testing standard. As I anticipated in the introduction, a fair comparison between RCTs and any alternative approach to clinical trials should take into account not only the inferential and ethical merits of each option, but also their respective soundness as a regulatory standard.

To this end, I will now present in some detail the different approaches to drug testing implemented over the 20th century, considering also their regulatory impact. This will show that we demand from clinical trials not only certain inferential virtues and ethical foundations, but also certain warrants of impartiality that vary according to each social context.

1.4 Regulatory issues

From the 1950s on, RCTs have been adopted in many countries as a regulatory standard to decide whether a drug is suitable for commercial distribution: a properly conducted phase III trial would decide about its safety and efficacy. As I mentioned in the introduction, this regulatory dimension of RCTs is not usually considered in their philosophical discussion, despite the attention it receives from sociologists and historians. However, the epistemic merits of RCTs as a regulatory yardstick should be considered together with their methodological and ethical foundations, if only because these merits were certainly considered by the agencies that adopted them as their testing standard. This adoption poses an interesting philosophical problem: assuming that the

civil officers at these agencies were statistical novices, what sort of arguments convinced them of the superiority of RCTs as opposed to other testing methods? Were they justified in accepting these arguments, or did they blindly follow the advice of the statistical experts who recommended RCTs?

The standard sociological account of statistical expertise provides the following picture of how, in modern democracies, it came to replace non-mathematical forms of expert advice. [Porter, 1995] An increasing pressure for public accountability made politicians choose statistical advisors. Statistical figures were perceived as the outcome of impersonal rules and calculations that exclude bias and personal preferences. Hence weak professional groups adopted statistical methods in order to strengthen their credentials as experts. In this account, trust in numbers is somehow blind: if there is no external check, the mere appearance of impartiality makes quite a poor epistemic justification. This approach grounds nowadays the best historical accounts of the introduction of RCTs for regulatory purposes.

In this section, I will provide an overview of the regulatory uses of RCTs, discussing the main alternatives considered for drug testing in different countries. In choosing between these alternatives it seems as if the regulatory bodies were driven by an epistemic concern: they wanted their testing standard to be impartial, i.e., the result of the test should be independent of the interests of any of the concerned parties (patients, clinicians, the pharmaceutical industry, and the regulator itself). Historians and sociologists claimed that the adoption of RCTs as an impartial testing standard was blind, because their frequentist foundations were never well understood either by the medical profession or by the regulators. This claim may be true, but I contend that they understood quite well in what sense randomisation and significance testing provided insurance against testing biases, independently of their statistical underpinnings. In this respect, their adoption was clearly justified. By way of conclusion, I will briefly consider what our regulatory dilemmas are today and to what extent this impartiality request is still valid today.

Between 1900 and 1950 expert clinical judgment was the main criterion in the assessment of the properties of pharmaceutical compounds, both in Britain and the United States. An experienced clinician would administer the drug to a series of patients

he considered more apt to benefit from it. His conclusions would be presented as a case report, informing of the details of each patient's reaction to the treatment. The alternatives were first laboratory experiments and then controlled clinical trials (from which RCTs would later emerge). Laboratory experiments would proceed either *in vitro* or *in vivo* (on animals and patients) and they were considered superior by clinicians with a scientific background. Yet their scope was usually restricted to safety considerations. It soon gave way to comparative trials, in which two treatments were alternated on the same patient or administered in two groups of patients (simultaneously or not). The arrangements to secure the comparability of the two treatments were the *controls,* and these adopted different forms: among the most prominent features were a clear statement of eligibility criteria to enter the trial, alternation and then randomisation in the allocation of treatments, uniformity in their administration and *blinding* (concealing the administered treatment from the patients and sometimes the doctors). These controls were not necessarily used all at once. Descriptive statistical reports from these trials conveyed their results with different degrees of sophistication. Significance testing features only occasionally in the medical literature before 1950[21].

The regulatory authorities in Britain and the United States arranged official drug testing depending on the standards adopted by the research community within their respective medical professions. In both cases, and all throughout the 20th century, regulators were concerned about *impartiality*, here understood as independence from the financial interests of the pharmaceutical industry. Tests sponsored by manufacturers for advertising purposes were considered suspicious by consumers in both countries and this prompted, in different ways, the development of public pharmaceutical agencies to conduct or supervise the tests. However, most clinical researchers considered themselves impervious to biases from non-financial sources and impartial enough to conduct clinical tests without bias-proof mechanisms. Until the 1960s, regulatory decisions were fundamentally based on expert judgments of this sort. Expert judgment came only to be discredited in the United States because in the late 1950s a group of methodologically-minded pharmacologists imposed their views on the superiority of RCTs at the Food and Drug Administration[22]. However, as Iain Chalmers and Harry Marks have often argued, even for this enlightened minority the inferential power of

[21] For excellent overviews see [Edwards, 2007; Marks, 1997; Toth, 1997].
[22] For an illustration of these points see [Marks, 2000, Carpenter and Moore, 2007].

RCTs and its statistical foundations were not the primary reason to adopt them: randomisation and significance testing were understood as impersonal rules for allocating treatments and interpreting trial results which warranted the impartiality of the assessment.

During the 1960s and 1970s, RCTs became mandatory for regulatory decisions in different degrees. In the United States, before the 1960s, the Food and Drug Administration was only entitled to test the safety but not the efficacy of pharmaceutical compounds. In the late 1950s there were voices in the FDA demanding stricter testing standards linking safety and efficacy, under increasing public mistrust in the pharmaceutical industry. The thalidomide scandal gave them the opportunity to project their views on the 1962 Drug efficacy amendment to the Food, Drug and Cosmetics Act. It required from the applicant "adequate and well-controlled clinical studies" for proof of efficacy and safety (although the definition of a well-controlled investigation would not be clarified until 1969, when it was formally quantified as two well-controlled clinical trials plus one previous or posterior confirmatory trial). Carpenter and Moore [2007] are correct, in my view, when they claim that this set of regulations created the modern clinical trial industry. In the following three decades, pharmaceutical funding would boost the conduct of RCTs in the United States and abroad.

In the United Kingdom, the Medical Research Council (MRC) acted as a consulting body to the Ministry of Health in pharmaceutical issues from the 1920s on. Unlike the FDA, the MRC did not supervise *ex officio* the British drug market: when its Therapeutic Trials Committee started testing compounds in the 1930s, it was always at the request of the manufacturer. The MRC trials were undertaken in support of the British pharmaceutical industry, with a view to foster its international competitiveness and domestic reputation. Until the thalidomide scandal in the 1960s, the commercialisation of a drug in the UK did not formally require any sort of clinical test for either safety or efficacy. The thalidomide scandal prompted the creation of the Committee on Safety of Drugs (CSD) within the Ministry of Health, with a subcommittee in charge of clinical trials and therapeutic evidence. However, neither the Ministry of Health nor the CSD could legally prevent the commercialisation of new drugs. The industry informally agreed to get CSD approval for their trials and inform

them about the toxicity of their products: it was a non-compulsory licensing system established on the basis of safety alone, not efficacy. This voluntary arrangement operated smoothly for almost a decade (1964-1971). A statutory system came to replace it as a result of the 1968 Medicines Act, which gave to the Ministry of Health the licensing authority, with a Medicines Commission acting as advisory body. The industry was now required to present evidence regarding safety and efficacy, but clinical trials were defined in the 1968 Act in a way general enough to encompass all the testing procedures mentioned above (from expert clinical judgment to statistical tests). Even if RCTs were at this point the testing standard in clinical research in Britain, the regulator did not officially adopt them as a yardstick. Provided that the regulatory body established its independence from the industry (hence, its financial impartiality), it was possible to submit evidence gathered from different sources and the decision would be made on a case by case basis.[23]

Impartiality in clinical trials is therefore the more socially desirable the bigger the public concern about biases, and this seems to depend entirely on the context in which trials take place. In Germany, for instance, the Drug Law was also revised in the aftermath of the thalidomide catastrophe. Yet this did not bring centralised control over clinical trials, which was considered costly and inefficient. Instead, it was agreed that the Federal Chamber of Physicians (BÄK) Drug Commission and the German Society for Internal Medicine issue guidelines for drug testing that the manufacturers should follow. Unlike Britain or the United States, in Germany therapeutic reformers did not form a coalition with statisticians after the II World War. Arthur Daemmrich's [2004, pp. 53-54] hypothesis is that, as a reaction against the terrible experiments conducted by the Nazi doctors during the war, the German medical profession strongly defended the necessity to treat patients individually, beyond any research protocol reducing them to standardised cases. In consequence, placebos and double blind experiments were often avoided, even if their virtues against biases were known and praised. The BÄK's reputation was based on the defence of patients' rights and not even the thalidomide scandal could shatter it throughout the 1960s and 1970s. While RCTs became more and more widely used, the 1976 Drug Law still granted the medical profession the right to

[23] On the creation and early trials of the MRC see [Cox-Maksimov, 1997]. The regulatory dimensions are discussed in [Abraham, 1995; Ceccoli, 1998].

set testing standards, even if the BÄK had been already accused of pro-industry bias and the socialist party had demanded a "neutral" examination of drugs by a central agency.

The political demands and expectations placed on clinical trials were different in all these countries. However, for regulatory purposes, the testing standard adopted was always justified on the grounds of its purported impartiality, no matter if whether was clinical expert judgment, laboratory tests, or RCTs. Historians and sociologists are probably right in explaining this regulatory concern for impartiality as the result of external public pressure. However, it is open to discussion whether the adoption of a testing standard was always epistemically blind. It may be true that the statistical foundations of RCTs were never well understood by the medical profession as a regulatory yardstick either in the United States or in Britain at the time of their adoption, or even later. However, in both cases, the medical profession, and even the public, seemed to understand quite well in what sense RCTs offered real protection against biases in the conduct and interpretation of medical experiments. RCTs provided proper impartial grounds for regulatory decisions. As I pointed out in the previous sections, randomisation certainly helps in preventing selection bias independently of its statistical grounds. Significance testing was understood less as discretionary interpretation rule than mere clinical expert judgment. On these grounds, the adoption of RCTs as a testing standard for regulatory purposes seems epistemically justified[24].

All in all, the social process that led to the adoption of frequentist RCTs as a regulatory standard may have been interest-driven, but it was not epistemically blind. If we still adhere to the principle that regulatory clinical trials should be independent of the particular interests of the manufacturers, any alternative testing methodology should be at least as impartial as our current RCTs are. However, the situation is today far more complicated than in the 1950s.

As we saw in the previous sections, as patients we may prefer to avoid randomisation, but as consumers of pharmaceutical compounds we may want them to be fairly tested by an independent authority. For the pharmaceutical consumer, the situation is today

[24] Of course, RCTs are not the only means to implement a fair test, but just part of a larger set of tools: see [Evans et al., 2006] for an overview. The interested reader can visit the James Lind Library for a general view of the evolution of fair tests over the world: http://www.jameslindlibrary.org/ See also the Project Impact site: http://www.projectimpact.info/ [both accessed in July 2009]

paradoxical in this respect. On the one hand, tight regulatory standards generate *lags*: it takes more time for a new drug to reach its targeted consumers, with significant economic costs for the producer. On the other hand, consumers are, more than ever, wary of potential fraud in the testing process, if the industry is entirely free to conduct them[25]. In the United States, for-profit private contractors conduct about 75 percent of all clinical trials, in which the pharmaceutical industry invests billions of dollars annually. There is growing evidence of bad testing and reporting practices biasing the results, such as: enrolling patients with a milder disease or healthier than the population who will actually receive the drug, using a dose of a comparable drug that is outside of the standard clinical range, using misleading measurement scales, etc[26]. In this context, randomisation and significance testing alone do not guarantee an unbiased clinical trial: various surveys have found significant degrees of association between private sponsorship and positive conclusions for the experimental drug in published trials[27]. Of course, these results are open to interpretation (it may simply be the case that the industry only funds and publishes trials of products that are considered better than the standard therapy), but caution about bias is advisable.

In sum, frequentist clinical trials are controversial from a methodological and ethical perspective, but have worked reasonably well so far as an impartial regulatory standard. However, there is a clear need for improvement on this front as well: we want regulatory trials to be both more impartial and more efficient (and, in particular, quicker). The *prima facie* strength of the Bayesian approach to clinical trials is that they promise improvement along these three dimensions (methodological, ethical and regulatory). Let us present how they work.

2.1. BAYESIAN TRIALS: A 25 YEARS HISTORY

Let me begin this second part of the paper with a brief summary of the development of the Bayesian approach to clinical trials during the last thirty years. I follow here Deborah Ashby's [2006] review, where she distinguishes three main periods. In the first one, ranging from 1982 to 1986, several experimental designs were launched and some

[25] For an overview of the literature on the *drug lag* and related topics, see [Comanor, 1986]. [Carpenter, 2004] provides an analysis of patients' influence on FDA decisions. The risks of pharmaceutical fraud are discussed in [Krimsky, 2003]
[26] For a quick general overview of these practices see [Jain, 2007].
[27] E.g., [Lexchin, et al., 2003; Yank et al., 2007].

were even implemented [Kadane, 1996]. But the computational power needed to implement more ambitious trials was still lacking. This became gradually available between 1987 and 1991, when the BUGS computer simulation package was created. Then came a period of consolidation (1992-1996), with a regular flow of publications on Bayesian trials and the first hints of regulatory attention to this approach. In the following ten years the variety of ideas and experiences accumulated deserved a first textbook [Spiegelhalter, Abrams & Miles, 2004]. Over the last ten years many phase I and II trials have been conducted following Bayesian principles, since these are allowed by the regulations. Phase III trials for drugs are still rare, due to regulatory restrictions, but they are already accepted by the FDA for medical devices[28]. And a Bayesian meta-analysis has been accepted as evidence in 2003 in the approval of a therapeutic compound[29]. Bayesianism has not yet reached the mainstream medical literature: according to Ashby, it will be the next frontier. But there is growing debate on whether the FDA should approve Bayesian designs for regulatory purposes, and if this occurs, this last boundary will soon be crossed, just as it happened with standard RCTs.

I will first introduce the more elementary concepts in the Bayesian approach to clinical trials, together with an attempt to classify the different methodologies in their design and analysis. The point of this section is to show that Bayesian clinical trials are constitutively diverse and can be tailored to multiple purposes, so no straightforward overall comparison with standard RCTs is possible. In order to illustrate this diversity, in the following two sections I will briefly review two different Bayesian trials. The first one, conducted during the 1980s, exemplifies how very elaborate ethical considerations can be incorporated into the design of a trial through a statistical representation of expert judgment. The second one, designed and conducted at the beginning of this decade, illustrates instead the potential efficiency of Bayesian trials and their impact on the regulatory process.

The following three sections will thus cover the basic items considered in the first part of the paper: epistemic, ethical, and regulatory issues. On these grounds I will provide a final discussion of the relative merits of each approach, frequentist and Bayesian, in the concluding section.

[28] I will not consider here the case of medical devices: see [Campbell, 2005] for a discussion.
[29] See [Berry, 2006, p. 29] for a quick review. The published source is [Hennekens et al., 2004].

2.2 BAYESIAN APPROACHES: A QUICK INTRODUCTION

> The basic paradigm of Bayesian statistics is straightforward. Initial beliefs concerning a parameter of interest, which could be based on objective evidence or subjective judgment or a combination, are expressed as a prior distribution. Evidence from further data is summarized by a likelihood function for the parameter, and the normalized product of the prior and the likelihood form the posterior distribution on the basis of which conclusions should be drawn [Spiegelhalter, Freedman and Parmar, 1994, p. 360][30].

Suppose we are interested in finding out the true mean difference (δ) between the effects of two treatments. The statistic x_m would capture the difference observed in the sample of participants in a comparative trial. The statistic x_m would here be normally distributed as expressed in the following density function:

$$p(x_m) = \phi(x_m | \delta, \sigma^2/m)$$

where m would be the number of observations of the mean response recorded so far in the trial and δ and σ^2/m would stand for the mean and variance of the distribution. This first equation provides the *likelihood function*: it shows the support lent by the trial data to the possible values of the mean difference between treatments.

Our initial beliefs about the true mean difference (δ), excluding all evidence from the trial, could be expressed thus by this density function:

$$p_0(\delta) = \phi(\delta | \delta_0, \sigma^2/n_0)$$

Bayes theorem would allow us to weight our prior by the likelihood function[31], obtaining the posterior distribution of δ:

$$p_m(\delta) \propto p(x_m | \delta) p_0(\delta)$$

$$= \phi\left(\delta \left| \frac{n_0\delta_0 + mx_m}{n_0 + m}, \frac{\sigma^2}{n_0 + m} \right. \right)$$

[30] In this section, I will follow two standard introductions: [Spiegelhalter, Freedman and Parmar, 1994] and [Spiegelhalter, Abrams and Miles, 2004]. Donald Berry has produced very concise overviews of the Bayesian approach to clinical trials: e.g., [Berry, 1993 and 2006].

[31] In the usual expression of Bayes theorem, the product of the prior and the likelihood function is divided by the normalising factor $p(x_m)$.

This is the expression of our beliefs about δ after m observations. The posterior mean would provide a point estimate of the true mean difference (δ) between treatments. The posterior mean ± 1.96 posterior standard deviations would provide a 95% credible interval estimate of δ.

By way of example, Spiegelhalter, Freedman and Parmar [1994] provide the following Bayesian analysis of a conventional RCT. This trial studied the effects of levamisole (LEV) in combination with 5-fluorouracil (5-FU) for patients with resected cancer of the colon or rectum: that is, LEV+5-FU *versus* control. The main outcome measure in this trial was the duration of patients' survival.

Two prior distributions were constructed for the analysis. The first one was a *sceptical* prior formalizing the belief that large treatment differences are not likely. For instance, we may initially believe that the mean difference δ_0 will be 0. The prior should be spread to encompass a range of treatment differences considered plausible by the experts who designed the experiment. The probability of observing a mean difference equal or superior to the minimal clinically worthwhile benefit was set to 0.05 (the type I error α of the original trial). Assuming a value for σ, we can calculate n_0 and specify the sceptical prior distribution $p_0(\delta)$. An enthusiastic prior would concordantly represent the beliefs of those "individuals who are reluctant to stop when results supporting the null hypothesis are observed". They would expect the mean difference δ_0 to be the minimal clinically worthwhile benefit. This second distribution would be spread with the same σ and n_0 than the sceptical prior.

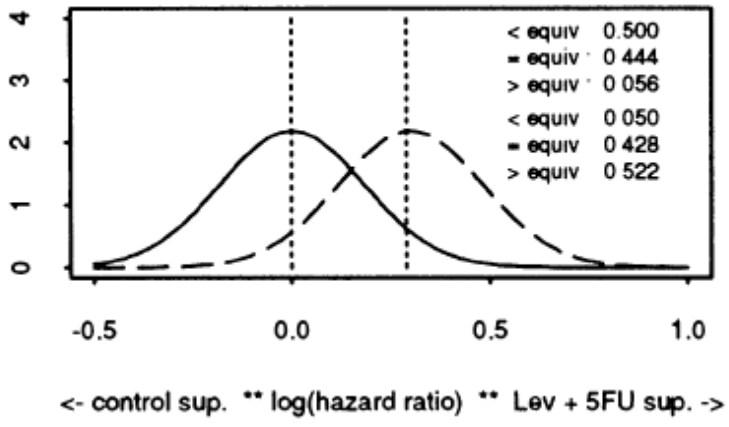

Fig.1 (from [Spiegelhalter, Freedman and Parmar, 1994])

In fig.1 we can see the sceptical (continuous line) and the enthusiast prior (intermittent line), with the probabilities of falling below, within and above the range of equivalence between treatments (dotted vertical lines) in the right hand corner[32].

With the data from the *m* patients gathered in the original trial, we can calculate the observed sample difference x_m and the corresponding likelihood for LEV+5 *versus* control, as shown in fig.2. The probability that LEV+5-FU is an inferior treatment seems low (0.003), though the probability of it being superior is just moderate (0.777)

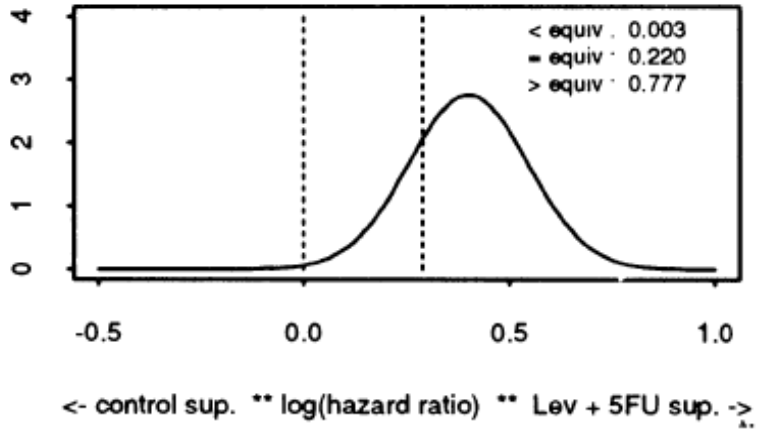

Fig.2 (from [Spiegelhalter, Freedman and Parmar, 1994])

Weighting the priors with the likelihood, we obtain their posterior distributions (fig.3)

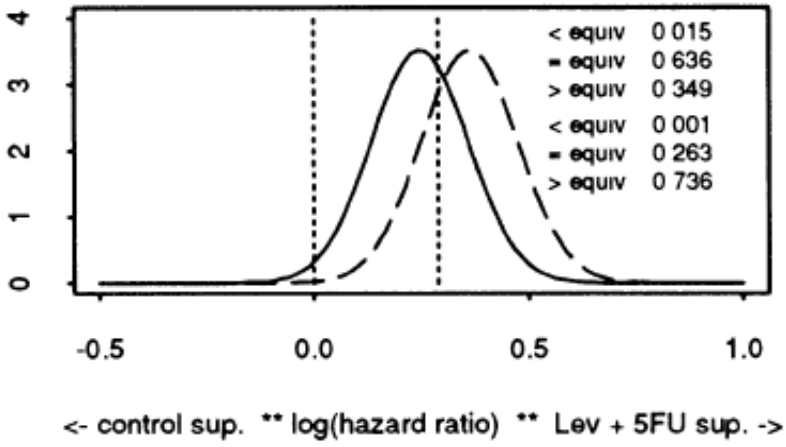

Fig.3 (from [Spiegelhalter, Freedman and Parmar, 1994])

[32] The range of equivalence is the space between the null hypothesis and the minimal clinically important difference, measured as an increase in average survival of a given number of months.

Should anyone holding the sceptical prior (continuous line) accept the efficacy of LEV+5-FU? Even if the posterior mean is now closer to the upper limit of the range of equivalence than the prior mean (fig.1), it is still within this range. The sceptic can reasonably refuse to accept the superiority of LEV+5-FU over the control treatment on the basis of the trial data.

Despite this straightforward illustration, *there is no such thing as a single Bayesian approach*, not only to clinical trials, but generally. But for concreteness, let us just focus here on the general approaches to clinical trials classified by Spiegelhalter, Abrams and Miles in their textbook [2004, pp. 112-13].

One classificatory criterion is the type of prior used in each approach. In the *empirical Bayes* approach, hyperparameters for the distribution of an array of studies assumed exchangeable can be estimated directly from these studies through a meta-analysis. In the *proper Bayes* approach the priors are constructed with either empirical data or subjective opinions (obtained through elicitation methods). *Objective* or *reference* prior distributions are used in the *reference Bayes* approach. These priors summarise a minimal amount of information: for instance, a uniform (e.g., flat) probability distribution over the range of interest; or the sceptical and enthusiastic priors of the previous example.

We can also differentiate these approaches according to their methods of analysis and reporting. In the *empirical Bayes* approach, a frequentist meta-analysis of several studies is reinterpreted, under certain assumptions, as an approximation of a Bayesian estimate. In the *reference Bayes* and *proper Bayes* approaches, the analysis is a direct application of Bayes's theorem. However, in the former, depending on the priors, the posterior distribution would approximate the conclusions of a frequentist likelihood analysis. Spiegelhalter, Abrams and Miles distinguish a fourth Bayesian approach to clinical trials: the *full Bayes* approach, in which decision theory is incorporated into the analysis so that judgments about treatments depend on the maximization of an expected utility function (with subjective probabilities).

Hence, depending on the approach implemented, a Bayesian clinical trial will yield results that will diverge more or less from the conventional RCT. By way of example, we can compare the strength of a *p*-value for or against a given hypothesis with the corresponding *Bayes factor*. This latter is, in its simplest form, the ratio between the

likelihoods of two alternative hypotheses, i.e., the probability of the data, assuming their truth:

$$\text{BF} = p(data \mid H_0) / p(data \mid H_1)$$

The *p*-value is precisely the probability of observing a certain range of values (observed and unobserved), assuming the truth of a hypothesis. The BF is independent of the priors on the hypothesis[33]: it just compares how probable the observed data are assuming the truth of each hypothesis. If H_0 states that there is no difference between treatments regarding a certain parameter ($\theta = 0$) and H_1 encompasses a range of alternative values of θ, the *minimum Bayes Factor* proposed by Steve Goodman takes, among these alternative values, the one that provides the "smallest amount of evidence that can be claimed for the null hypothesis (or the strongest evidence against it) on the basis of the data".

Of all possible $\theta \neq 0$ (H_1), we take the one which makes higher the probability of obtaining the observed data, assuming the truth of H_1 . For this value of θ, the BF will be minimum: a small BF implies that the observed data will be much more probable under H_1 than under H_0, just as a small *p*-value implies that there is a small probability of obtaining data as extreme or more than the one observed if H_0 is true, which is why we should reject it.

[33] Yet, it does depend on the prior distribution within hypothesis. The BF impinges on the prior probabilities through Bayes theorem, which for the comparison between these two hypotheses takes the form:

$$\frac{p(H_0 \mid y)}{p(H_1 \mid y)} = \frac{p(y \mid H_0)}{p(y \mid H_1)} \times \frac{p(H_0)}{p(H_1)}$$

P Value (Z Score)	Minimum Bayes Factor	Decrease in Probability of the Null Hypothesis, %		Strength of Evidence
		From	To No Less Than	
0.10 (1.64)	**0.26** (1/3.8)	75 50 17	44 21 5	Weak
0.05 (1.96)	**0.15** (1/6.8)	75 50 26	31 13 5	Moderate
0.03 (2.17)	**0.095** (1/11)	75 50 33	22 9 5	Moderate
0.01 (2.58)	**0.036** (1/28)	75 50 60	10 3.5 5	Moderate to strong
0.001 (3.28)	**0.005** (1/216)	75 50 92	1 0.5 5	Strong to very strong

Table 1 (from [Goodman 1999b])

Table 1 provides a comparison between one particular minimum BF[34] and the corresponding one-sided *p*-value (for a fixed sample size), for a range of values of the test statistic Z. For the conventional *p*-value threshold, 0.05, the BF is 0.15, meaning that the null hypothesis gets 15% as much support as the best supported alternative value of θ. As Goodman observes, this is a moderate strength of evidence against H_0. If the prior probability of H_0 is 0.75, the impact of the corresponding likelihood will yield a posterior of just 0.44. Only with a very low initial probability (0.26) will we obtain a posterior of 0.05.

This illustration just shows that, for BF of a certain form, a Bayesian analysis can be as demanding as a conventional hypotheses testing or even more so. However, as Spiegelhalter, Abrams and Miles [2004, p. 132] point out, there is no simple monotonic relationship between Bayes factors and *p*-values. If we choose a different form for the BF, they can diverge from the *p*-values. In other words, the traditional frequentist approach to clinical trials and a possible Bayesian alternative will depend on a combination of principled and practical considerations that can justify this choice in case of discrepancy. Since this justification depends entirely on the type of Bayesian approach we choose, it is better to examine a couple of well-articulated examples to see how strong the Bayesian case can be.

[34] Assuming a normal distribution and $H_1 \neq 0$, the minimum Bayes Factor would be:

$$\mathrm{BF}_{\min} = \exp(-z_m^2 / 2)$$

where z_m is the standardised test statistic for H_0. For further details see [Goodman, 1999b].

2.3 THE VERAMAPIL VS NITROPRUSSIDE TRIAL

Our first example is a clinical trial aimed at determining the relative efficacy of two drugs, veramapil and nitroprusside, in controlling hypertension immediately after open-heart surgery. The trial was conducted over 30 months, from September 1984 to March 1987 in Baltimore, at Johns Hopkins Hospital, by a team led by E. Heitmiller and T. Blanck. The statistical advisors were led by J. Kadane at Carnegie Mellon University in Pittsburgh. Both drugs were already available, though verapamil was used for different heart conditions and counted as the experimental treatment in the comparison. A conventional RCT had already been attempted unsuccessfully at John Hopkins in the early 1980s, and in 1984 the idea arose of conducting instead a pilot Bayesian study. The trial implemented for the first time an approach developed by Kadane, N. Sedransk and T. Seidenfeld (hereafter KSS), in the early 1980s. Following the classification outlined above, the KSS approach, as implemented in the trial, would count as a *proper Bayes* approach: priors are elicited from a group of experts to be updated through Bayes theorem. However, Kadane and his coauthors show a clear sympathy for the *full Bayes* approach: decision theory plays a certain role in the conceptual foundations of the KSS methodology (e.g.. [Sedransk, 1996]), even if in this particular trial utility functions were not elicited or postulated to account for any of the choices made.

The main goal of the KSS approach is ethical: it is aimed at improving the allocation of treatments in a trial, so that patients receive a treatment that at least one expert would recommend in view of his personal characteristics. This way, they are protected against treatments that are unanimously considered inferior at the point they enter the trial. The elicited prior for the variable measuring the effect of each treatment probabilistically depends on a set of covariates (diagnostic indicators) and treatment. Depending on the values of these covariates in each patient, a computer will calculate which of the two treatments each expert would recommend for him, according to the expert's updated prior.

Whereas the implementation of the clinical equipoise principle in a frequentist RCT presupposes that the medical community has no statistical grounds to judge one treatment as superior until a significant conclusion is reached, in a Bayesian approach at least some actionable evidence can be attained earlier, depending on one's prior and the data accumulated throughout the trial. The KSS approach uses this evidence in the

following manner: a patient will only receive a treatment if at least one expert judges it *admissible,* given his characteristics, at the point at which he enters the trial. From then on, the patient will never receive a treatment that the panel of experts in the trial agree to consider inferior at that point.

Sedransk [1996] provides an excellent formal analysis of the notion of *admissibility* as implemented in the KSS design. It hinges on the following basic principles (for which Sedransk provides an axiomatic statement as well)[35]:

> P1: The outcome following treatment with any admissible treatment must be scientifically interpretable
>
> P2: Admissibility must be determined based on current information including data already gathered in the course of the clinical trial
>
> P3: A set of K experts is sufficient when the addition of any other expert (i.e., any other relevant scientific opinion) cannot change the admissibility or inadmissibility of any treatment.

For a treatment to be admissible P1 requires that its effects can be traced to clearly defined factors (excluding therefore those "alternative" therapies without clear causal mechanisms to back up the experts' opinion). P2 differentiates the KSS approach from standard RCTs since the evidence accumulated throughout the trial impinges on the definition of an admissible treatment. P3 is also crucial for the design of the study, since the trial will terminate only when the data gathered bring to an agreement the panel of experts whose priors are elicited for the study. The cogency of the results of the trial will therefore depend on the range of opinions represented in the panel. P3 establishes a sufficiency criterion to assess the diversity of this range.

Several admissibility criteria can potentially satisfy this set of principles, of which the simplest one defines an admissible therapy "as a treatment considered superior or equivalent (to the proper comparison treatment(s)) by at least one expert in the panel"[36]. However, this criterion presents no particular difficulty for P1 and P2, but it will only comply with P3 if the views about each treatment in the scientific community are just a

[35] Sedransk actually presents eight principles, but in order to simplify the discussion I will just consider three, those that she deems the "basic premises" for the KSS designs [Sedransk, 1996, p. 109]
[36] More formally, as Kadane puts it, "if at least one (updated) expert would consider it (in the computer) to have lower expected deviation from target than the other treatment" [Kadane, 1994, p. 223].

few and are fully represented in the panel. Otherwise, the admissibility criterion may not secure that patients do not receive an inferior treatment.

A variety of allocation rules based on admissibility criteria are possible within the KSS approach, sometimes departing from standard randomisation. The "statistical price" to pay for the ethical constraint imposed on allocation rules is that the likelihood function should explicitly condition on the patients' characteristics that are considered in the allocation of a treatment. As Kadane and Seidenfeld [1996] show, the likelihood function would then be of the general form:

$$f_\theta(P_j) \propto \prod_j^J f_\theta(O_j|T_j, X_j)$$

Where X_j is the vector of relevant characteristics of jth patient, T_j is the treatment assigned to the jth patient and O_j the corresponding outcome. The past evidence up to and including the jth patient is expressed by $P_j = (O_j, T_j, X_j, O_{j-1}, T_{j-1}, X_{j-1}, \ldots, O_1, T_1, X_1)$. θ is a vector of the parameters that determine the probability of an outcome O_j for a patient j given characteristics X_j and treatment T_j.

This is the likelihood function that will be used in the allocation of treatments during the trial; conditioning on X_j, the set of diagnostic indicators used in the allocation of the treatment, makes explicit all the information on the outcome O_j carried by the assigned treatment. If the computer assigns treatments according to this information alone, all other sources of bias in the allocation will be excluded. Even if randomisation is acceptable in a Bayesian perspective in order to prevent selection biases, the KSS approach achieves this by virtue of its own design[37]. However, the allocation algorithm designed by Sedransk will make use of it in order to balance the independent variables.

Let us now briefly review how the KSS methodology was implemented in the veramapil vs. nitroprusside trial conducted at Johns Hopkins Hospital[38]. Five experts representing a range of medical opinions about the treated condition were identified. Once the criteria of eligibility for the trial were set, the anaesthesiologist in charge of the study independently chose the four most important variables for predicting a prognosis for each of the participant patients. Then the prior opinion of each expert on the outcome

[37] See [Kadane and Seidenfeld, 1990] for the details of this argument and a wonderful discussion of randomisation from a Bayesian perspective.
[38] This summary draws from the papers compiled in [Kadane, 1996, pp. 129-219].

(the effects on arterial pressure) was elicited as a function of these covariates and the treatment administered[39].

The elicitation method designed by Kadane and his coauthors required an hour-long telephone interview. The prior was estimated assuming that the treatment outcome depended on the four predictor variables according to a normal linear model. There were 16 possible combinations of these variables and, therefore, 16 possible patient types. For each of these, each expert's prior would allow us to estimate the effect of each treatment and therefore the expert's preference for either veramapil or nitroprusside.

Once the trial started, whenever a suitable patient was recruited, the values of each of the four predictor variables were measured and entered into a computer, which yielded the mean arterial pressure predicted by each expert for a patient with such values according to each treatment. The computer also implemented an allocation rule as follows: if all the experts predicted a higher (better) mean arterial pressure with the same treatment, this one was assigned to the patient; otherwise, the computer would assign one at random with the constraint of maximising balance among treatments regarding the predicting variables. After a treatment was administered to a patient, the lowest mean arterial pressure recorded was also entered into the computer and the experts' priors were updated. The updated priors were then used to deliver predictions for new patients entering the trial[40].

All in all, 29 patients completed the trial, 17 in the verapamil arm and 12 receiving nitroprusside. Even if the allocation rule made more likely that certain types of patients received one of the treatments more often, no statistically appreciable effect was detected. The results can be summarised in a table showing the treatment each expert would recommended for each type of patient before and after the treatment, using for

[39] E.g.: "For patients on beta blockers and calcium antagonists who have no previous history of hypertension and no wall motion abnormality, what is your median for the average deviation of mean arterial pressure from 80mmHg?" [Kadane, 1996, p. 171]. The methodology of this elicitation procedure was exposed in full in [Kadane et al., 1980].

[40] Due to a "gap in communication" between the medical team and the statistical advisors about how to measure the more beneficial outcome for patients, two different endpoints were used in the trial (each one with its own set of elicited priors): the lowest value of the mean arterial pressure and the average deviation from a target pressure, both over 30 minutes after the patient received the treatment (LADEV). An additional measure was used in the transition between these two. Also, due to a bug in the computed program, the treatments were not assigned according to the original allocation rule. However it was always a function of the patients' characteristics alone and therefore the results were not biased by this change. For a discussion of these complications see Kadane's section on "Operational History and Procedural Feasibility" [Kadane, 1996, pp. 171-176]

this end the priors elicited for LADEV, once updated with the data collected in the trial. This table shows "an overall trend towards preferring verapamil over nitroprusside" [Kadane and Sedransk, 1996, p. 177]. The prior and posterior distributions of each expert for the effects of each treatment in each type of patient are also presented. Kadane and Sedransk do not provide an aggregate of these opinions showing the degree of consensus attained and suggest instead a more general assessment, using standard indexes in cardiology to evaluate the effects of each treatment [Heitmiller et al., 1996].

Given the difficulties that hindered this particular implementation of the KSS approach, it is understandable that the trial yielded no strong conclusion. Actually, this study is not defended on the basis of the statistical strength of its results, but rather for its ethical superiority in terms of the standard of care provided to the participants. Let us then briefly examine this Bayesian methodology in the light of the ethical issues that arise in frequentist trials. It has already been mentioned that in the KSS framework it is possible to incorporate the actual beliefs of the medical community about a treatment: clinical equipoise can therefore be measured rather than merely postulated as in conventional RCTs. On the patients' side, regarding their autonomy, it is open to discussion whether there is real understanding of the informed consent form regarding the allocation procedure. The participants in the verapamil trial had to deal with the following paragraph:

> The drug to be used in your case would be chosen with a recently developed statistical technique which incorporates the opinions of experts in the field concerning which drug is best for you, based on a variety of characteristics of the disease process, such as any history of high blood pressure or abnormal heart movements, rather than on an actual consideration of your case. If these opinions lead to the conclusion that only one of the drugs is allowable for you, that drug will be used. If both are found to be allowable, the assignment will be based on the need for balance in the characteristics of participants receiving each drug [Kadane, 1996, p. 141]

Whether patients can understand this paragraph better than the usual sentence about flipping coins is a purely empirical question. A priori, it does not seem very plausible that they do. The autonomy of the participant in KSS trials seems to be grounded more on their desires than on their beliefs. If patients expect to benefit personally from their participation in a trial, it can be argued a priori that a KSS trial gives them a better

expected utility to do so under very reasonable assumptions [Emrich and Sedransk, 1996]. Notice that this does not amount to straightforward choice of treatment by the patient [Kadane and Seidenfeld, 1996, pp. 118-119], but the KSS allocation rule probably does more to meet the demand for the personal recommendation of the physician than standard randomisation.

This is also relevant to the justice of the allocation procedure. If the distribution of costs and benefits in a KSS trial admits a utilitarian justification from the patient's point of view, *a fortiori* it will be equally justifiable from a contractarian perspective: if the treatment assigned is conditional on just the set of covariates capturing the relevant diagnostic indicators, no patient can claim that the allocation was intended to favour one person over another[41]. Hence, the KSS allocation rule admits a broader justification than randomisation, as far as the principle of justice is concerned.

Lastly, the admissibility rule implemented in the verapamil trial provides a very strong implementation of the principles of beneficence and non-maleficence: patients will not receive a treatment that no expert recommends, and they have a better chance of receiving one they can personally benefit from than with randomisation. This is also relevant for the discussion of the trial-stopping rules, the other ethical contentious issue raised by frequentist trials. Let me quote Kadane again:

> Whether to stop is a different kind of decision in a design of this sort than it is in a classically randomised design. In the latter, there can be agonizing decisions about whether to suspend operations when it is fairly clear which treatment is best (either overall or for a subclass of patients), but the results are not yet "significant". By contrast, in the trial suggested above, patients are protected from clearly bad treatments, so the decision of whether to continue has no ethical component. Rather, it is merely a question of whether the cost of continued data collection is repaid by the information gained. [Kadane, 1994, p. 223]

All in all, the KSS approach seems to comply better with the principles of autonomy, justice, and beneficence than do standard RCTs. However, its scope is somewhat more restricted: as Kadane [1994, p. 222] points out as well, the KSS approach will only offer protection to patients against inferior treatments if the results are gathered at a pace

[41] Assuming, of course, that nobody can decide at what point a patient enters a trial: the later he does, the more accumulated information he will benefit from.

quick enough to update the experts' priors before new patients enter the trial. In the verapamil trial the relevant data were ready for collection from each patient an hour after the surgical procedure. "In slower, more chronic diseases, there might be little or no information to capture at this step, and consequently little or no advantage to patients (or anyone else) in using these ideas" [Kadane, 1994, p. 222]. Not much was said about the advantages of the KSS approach from a regulatory perspective, but I will discuss this further in the final conclusion.

2.4 THE ASTIN TRIAL

Our second example is ASTIN (Acute Stroke Therapy by Inhibition of Neutrophils), a phase II clinical trial conducted between 2000 and 2001 in order to test a neuroprotective therapy to stop or slow the death of brain cells in acute ischemic stroke[42]. Very few treatments are available for this condition, despite the tens of thousands of patients randomised into clinical trials over the last four decades. The ASTIN design was intended to provide a more *efficient* approach to clinical trials, improving the use of scarce patient resources and accelerating the development of promising therapeutic agents. ASTIN was described as follows:

> A Bayesian sequential design with real-time efficacy data capture and continuous reassessment of the dose response allowed double-blind, randomised, adaptive allocation to 1 of 15 doses (dose range, 10 to 120 mg) or placebo and early termination for efficacy or futility. The primary end point was change from baseline to day 90 on the Scandinavian Stroke Scale (SSS), adjusted for baseline SSS. [Krams et al., 2003, p. 2543]

This is an instance of so-called *adaptive designs*: in trials of this sort, the design can be periodically modified depending on the evidence about certain hypotheses provided by the accumulated data. In ASTIN both treatment allocation and stopping rules were adaptive in a sense we will discuss in detail below.

ASTIN was a multi-centre international trial, sponsored by a pharmaceutical company (Pfizer). The trial was designed by Donald Berry and Peter Mueller, from the University of Texas M.D. Anderson Cancer Center. Under the leadership of Berry, over the last decade the centre became an international reference in the conduct of Bayesian clinical

[42] The neutrophil inhibitory factor UK-279,276.

trials[43]. Of the 964 trial protocols registered at M.D. Anderson between 2000 and early 2005, 178 used both a Bayesian design and a Bayesian analysis, namely for monitoring efficacy and toxicity. Some trials implemented Bayesian adaptive randomisation and dose finding techniques and, to a lesser degree, hierarchical models and predictive probabilities were also incorporated. For the last thirty years, Berry has advocated a *full Bayes* approach to clinical trials, grounding his arguments both on statistical and ethical considerations. However, Berry and his team at M.D. Anderson work under a strict regulatory system which requires approval from various internal panels and, sometimes, the FDA and other national bodies. Due to these regulatory constraints, most of trials at M.D. Anderson were phase I/II or phase II, supported by extensive simulations of their operating characteristics showing their degree of equivalence with standard frequentist trials. As Berry [2004, p. 186] puts it,

> At least for the near future they will be used as tools, with justifications following a more or less traditional frequentist course. As time passes and as researchers and regulatory folk become more accustomed to Bayesian ideas, they will be increasingly accepted on their own terms.

The ASTIN trial is no exception in this respect, and its original design was described by their own authors as a "frequentist cake with Bayesian icing" [Berry et al., 2002, p. 154]. This is why the efficiency of these constrained Bayesian designs is so prominently emphasised. Even if the "playing field" is not levelled, Bayesian trials can provide a more efficient solution to one of the main regulatory issues of our time: scientific innovation goes much faster than the development of new therapies and this delay is partly caused by the time constraints imposed by the current regulatory regime of RCTs.

Bayesian phase II trials such as the one we will discuss here can be more efficient in the following ways [Krams et al., 2005, p. 1341]: the participant patients will be treated more effectively thanks to an adaptive allocation procedure that incorporates the available information about the more efficient dosage; their design allows a quicker and more reliable choice of the dose to be used later in the phase III trial; if the regulatory

[43] See Berry's profile in [Couzin, 2004]. For an overview of the trials conducted at M.D. Anderson, see [Biswas et al., 2009].

authority permits, it is possible to make a seamless (and therefore quicker) transition from the dose finding to this confirmatory phase of the trial[44].

The aim of the ASTIN trial was to identify the *minimum dose with satisfactory effects*, defined as the ED_{95}: this dose would deliver 95% of the maximum efficacy, minimizing unacceptable adverse reactions. ASTIN sought a point estimate of ED_{95} with minimal variance. In order to achieve this goal, a standard phase II design may use between three and five doses and placebo. Each dose will be tested on an equal number of patients, usually fixed, independently of their comparative efficacy that will only be revealed at the end of the trial. The patients' reactions will provide the basis to estimate the *dose-response curve*. However, part of the observations may be wasted depending on the adjustment between the true range of efficacy of the drug and the dosage tested in the trial. The separation between the doses tested will constrain the accuracy of the ED_{95} estimate. Ideally, it would be better to test many different doses, but the number of patients that this would require to ground the power of a standard design is prohibitive.

The ASTIN trial tested 15 doses and placebo. In order to learn quickly and make the sample size as small as possible an adaptive treatment allocation rule was implemented. The rule was grounded on a formal decision model that calculated, at each point in the trial, the expected utility of choosing a given dose with a view to minimise the expected variance of the response at the ED_{95}. Once the optimal dose Z_j was chosen, the next patient could receive either placebo (with a fixed probability p_0) or a dose in the neighbourhood of Z_j (the remaining probability 1- p_0 was split uniformly over all of them)[45].

Another adaptive feature of the ASTIN trial was an optimal stopping rule. Once a week, in view of the available data, it had to be decided whether to end the trial abandoning the drug (futility), continue with the dose-finding phase or finish it switching to a phase III trial. A stopping rule grounded on another formal decision model was initially constructed, but the trial implemented a simpler approach, based on bounds of posterior probability, that the authors summarised as follows:

[44] The original design of ASTIN [Berry et al., 2002] envisaged the possibility of this seamless transition between phase II and phase III, but it was not finally implemented. Inoue et al., 2002 provide another sequential design for a seamless phase II/III trial.

[45] This fixed lower bound p_0 for placebo granted that there would be a group of patients (at least 15% of the total) providing a "comparison benchmark" in the study, as expected by the regulator: see [Walton, 1995, p. 352].

> The stopping rule in ASTIN continuously asked the following questions: (1) Does our estimate of the dose–response suggest that there is <10% chance of success for any dose (success was defined as a >3-point recovery over and above placebo as measured by a stroke scale)? If so, then stop for futility. (2) For the best dose, is the response good enough to conclude that there is >90% chance of success? If so, then stop for efficacy and switch to a confirmatory trial, comparing the "best dose" against placebo. [Krams et al., 2005, p. 1343]

The effects of this sort of therapy are measured around 90 days after the stroke, assessing the patient's neurological deficit with a standardised scale. In order to update the dose allocation system before this three-month deadline, a predictive longitudinal model was built to estimate the score for eligible patients in the Scandinavian Stroke Scale (SSS). Once measured, the true day 90 score replaced the estimate. The model was built on the evidence gathered in the Copenhagen Stroke Database[46] and was updated with the periodical responses obtained from patients in ASTIN .

Finally, in ASTIN the probability model for the dose-response curve was a normal dynamic linear model[47]. In the initial week of the trial, the prior estimate was flat, with a placebo effect of 10 points change from the SSS baseline (calculated from the Copenhagen Database). Such prior would not influence much the final results of the analysis and therefore its validity was never a concern for the regulators, who could rely entirely upon the study data [Walton et al., 2005, p. 352]. Updating this model with the study data yielded posterior estimates and 95% posterior credible intervals of the dose-response curve, ED_{95} and the effect over placebo at the ED_{95}.

The successful conduct of this trial depended on a computer system that recorded and processed the information entered by the investigators, ran the software implementing all the statistical models, delivered the dose for each patient to the investigators, and assessed the stopping rules, helping to monitor the progress of the study [Berry et al., 2002, pp. 127-134]. In the actual trial, the computer system was run by a private company independently of the sponsor.

The trial process could therefore be charted as follows[48]:

[46] A compilation of data gathered over two years in a Copenhagen facility from 1351 pharmacologically untreated stroke patients.
[47] See [Berry et al., 2002] for details.
[48] I quote from [Krams et al., 2005, p. 1343].

[1] A patient enters the trial. Baseline data are entered into the system;

[2] Patient is randomised in blinded fashion to placebo or "optimal" dose to learn about research question;

[3] Dose assigned is converted to particular vial numbers, allowing for blinded administration of study drug;

[4] Patient's response data are entered into the system as they progress through the study;

[5] Patient's final outcome is predicted using a longitudinal model (the prediction is substituted by the final response, as soon as it becomes available)

[6] Based on the currently available data, the system updates the "estimate" of the dose response curve and its uncertainty;

[7] Each day the algorithm implements a decision rule and recommends to either:

[8] A0: stop the study because of futility (based on the posterior probability that the treatment has an effect smaller than a minimum clinically relevant size) or

[9] A2: stop dose finding and moves to a large confirmatory study (based on the posterior probability that the treatment has an effect larger than some clinically relevant size);

[10] A1: continue dose finding study (the recommendation of the system is reviewed by the IDMC, which incorporates clinical judgment and factors in safety issues);

[11] The dose allocator chooses a dose from a list of possible doses that will optimise learning about the ED_{95} or some aspect of the dose-response curve. The database used to determine the dose is continually updated as outcome data from patients are gathered.

Before the trial was started ASTIN was simulated under a wide range of assumptions in order to convince both the sponsors and the regulatory authorities of its soundness [Berry et al., 2002, pp. 135-154]. Simulations provided both optimal parameters for the algorithm and the operating characteristics of the design, allowing a comparison with a

standard RCT. The following table compares, for instance, the sample size required in each [Grieve and Krams, 2005, p. 345]

	Power of traditional design		Adaptive design (max. 1000 pts)	
Benefit over placebo	80%	90%	% Stopped for efficacy	Median number of patients
0	—	—	0.02	501
2	2432	3220	0.56	644
3	1080	432	0.90	416
4	608	808	0.95	280

[Table 2]

For a dose-response curve that reached a plateau at 2, 3 or 4 points benefit over placebo, on the left side there is the number of patients needed for a 80% and 90% power. On the right side, for an adaptive design with a maximum of 1000 patients, you can see the percentage of trials that would stop for the same benefit over placebo and the median number of patients required in each. The trial was designed to detect a 3 point benefit over placebo[49].

In the actual trial, 966 patients were randomised, 26% of them to placebo. 40% of the patients were allocated to the top three doses. Quoting from the published results, "UK-279,276 did not produce any statistically significant effect on any of the efficacy variables at any dose or dose category for any of the analysed populations" [Krams et al., 2003, p. 2545]. After 48 weeks, the Independent Data Monitoring Committee that oversaw the trial decided that it could be stopped for futility and no more patients were admitted. The algorithm allowed a conclusion of futility at week 40, so the number of patients recruited might have been smaller. However, the trial protocol required at least 500 assessable patients before stopping. Those already in the trial were monitored for 13 additional weeks without a positive-dose response.

A less conservative protocol could have stopped the trial much earlier: apparently, similar conclusions could have been reached with half of the recruited patients [Walton et al., 2005, p. 356]. However, with sequential stopping rules, a frequentist design could have been effective with fewer patients than those estimated in table 2. As William du

[49] For a more extensive comparative discussion of sample sizes in both approaches, see the contribution of Land and Wieand in [Berry et al., 2002, pp. 169-174] and the rejoinder in pp. 176-180.

Mouchel observed, the inflexibility of standard RCTs is more a consequence of the regulatory framework than of the frequentist approach itself [Walton et al., 2005, p. 354]. The panel discussion of the ASTIN trial organized in 2005 in a FDA-sponsored symposium revealed a general appreciation of the simplicity of implementing, e.g., the stopping rule in a Bayesian approach. Yet it remains an open question to what extent it is necessary to become a fully committed Bayesian in order to benefit from the ASTIN techniques. I will return to this point in the final discussion.

Finally, notice that, from an ethical perspective, the ASTIN trial is at least as defensible as a standard frequentist trial, and perhaps more so. Going again through the principles we examined in the KSS trial, the autonomy of the patient is certainly respected. When adaptive randomisation schemes are implemented in a M.D. Anderson trial, the informed consent form incorporates clauses along these lines:

> If you are ... eligible to take part in the study, you will be randomly assigned (as in the toss of a coin) to one of two treatment groups. Participants in one group will receive [regimen 1]. Participants in the other group will receive [regimen 2]. At first, there will be an equal chance of being assigned to either group. As the study goes along, however, the chance of being assigned to the treatment that worked better so far will increase. [Biswal et al., 2009, p. 214]

Again, this seems no more difficult to understand than standard randomisation techniques in a conventional RCT, and plausibly the patients will be happy that their chances of being assigned to the better treatment are gradually increased. Beneficence and non-maleficence are equally well observed. From the point of view of justice, the stopping rule originally designed for the trial is particularly interesting. This rule allowed maximisation of the value of each stroke patient entering the trial in order to optimise treatment for the overall population and the individual patients [Berry et al., 2002, pp. 119-124]. From a utilitarian perspective, the sacrifices of the trial participants will be minimised and justified by the welfare the tested treatment would bring to this bigger collective [Krams et al., 2005, p. 1343][50]. However, the contractarian argument to justify the distribution of costs and benefits among trial participants would also apply here.

[50] From this perspective, it is really worth considering the procedure to calculate sample sizes developed in [Inoue et al., 2005].

3. CONCLUDING DISCUSSION: FREQUENTIST VS BAYESIAN TRIALS

The examples discussed in the previous section show that, on the one hand, we have highly standardised frequentist RCTs, the design of which evolved under increasing regulatory pressure over the last 50 years. On the other hand, we have a plurality of Bayesian approaches to clinical trials: depending on which principles we want to implement, there is a wide range of designs available and more will certainly come in the future. What would a fair comparison under these circumstances be? Let us examine this question considering the three dimensions discussed in this paper: epistemological, ethical, and regulatory.

Starting with the first one, we saw that *p*-values and confidence intervals are often misinterpreted in the medical literature as if they provided direct probabilities for particular events in clinical trials (§1.1). If this is not just a misunderstanding, but rather the expression of the sort of probability assignment the medical profession is interested in, this is an argument for the Bayesian approach, in which these probabilities can be correctly calculated. The objections against randomisation we examined in section 1.2 do not apply, in principle, to its use in Bayesian trials, since it does not provide any inferential grounds: randomisation can be defended in a Bayesian perspective as a device against allocation biases [Berry and Kadane, 1997]. Even in this respect, there are alternatives to randomisation in a Bayesian approach, like conditioning on the allocation mechanism (and implementing it in a computer in order to assign treatments): the KSS approach provided a nice illustration of this possibility. We saw in section 1.4 that RCTs were mainly adopted in Britain and the United States for the warrant they provided against biases. Bayesian trials can provide such warrants, using randomisation if necessary.

Hence, in principle, Bayesianism is a suitable alternative for the potential epistemic demands of the medical profession. From a pure research perspective, any kind of Bayesian trial provides an excellent tool to conduct experiments to learn about therapies, and in non-regulatory contexts their use is growing fast [Biswas et al., 2009]. The thorny question is what kind of Bayesian approach should be preferred for the design and analysis of clinical trials in a regulatory context, where experiments should prove the efficacy and safety of a compound. It is at present dubious whether there is a purely epistemic response to this issue.

As I briefly mentioned in the discussion of the ASTIN trial, from a purely pragmatic perspective it may seem possible to use the more suitable technique for the goals of each trial, be it frequentist or Bayesian, without paying much attention to coherence (e.g. [Walton et al., 2005, p. 354]). However, the main epistemic argument for Bayesianism is that it makes it possible to carry out the entire design and analysis of a trial within a coherent framework (e.g., [Walton et al., 2005, p. 356]). In this respect, the apex of coherence would be provided by a *full Bayes* approach, in which every decision to be made in a trial could be explicitly formalised. This may not be very attractive for many practitioners: as it was observed in the discussion of the ASTIN design, the utility functions that feature in decision models are often simplistic in order to facilitate computations [Berry et al., 2002, p. 167]. Yet, as Berry often notes, the decisions will be made anyway and the formalisation contributes to clarify our choices and make them more transparent to every stakeholder in the trial. The verapamil and the ASTIN trials are both supported by expected utility calculations that are certainly relevant from the patient's perspective.

The open question here is whether it is possible to incorporate in a unified decision model all the interests at stake. John Whitehead [1993, p. 1410] presented this problem as follows. Three different goals are usually pursued in phase III clinical trials: regulatory agencies (acting on behalf of patients and consumers) want to keep out of the market ineffective and harmful compounds; pharmaceutical companies want to introduce into the market effective and safe compounds; finally, clinicians are interested in acquiring information on the relative characteristics of the experimental and control treatments. This is certainly a simplification, since all parties are interested in all aspects of the trial, but, it shows nonetheless that there is no single decision maker in clinical trials. But from a Bayesian perspective, this multiplicity of agents is difficult to encompass in a unified model[51]. There may be limits to the implementation of a full Bayes approach in a regulatory context.

As Whitehead points out, in standard RCTs all these interests are somehow represented in the different elements of the analysis: small *p*-values express the concerns of the regulator; high-powered trials give a better chance of showing the efficacy of a compound and thus are in the producer's interest; and the estimates of the comparative

[51] This is too technical an issue to discuss here, but it is certainly not neglected by the authors we are considering here: see, for instance, the compilation of essays in [Kadane, Schervish and Seidenfeld, 1999].

difference between treatments will serve the clinician's interests. Of course, as we saw, it is easy to approximate all these aspects of standard RCTs from a Bayesian perspective, even without a full-fledged decision model. But, again, there are too many ways of approximating these characteristics in a Bayesian trial. In the current epistemology of science, this is more a virtue than an obstacle: as the methodological debate on evidence-based medicine illustrates (e.g., [Worrall, 2007]), scientific data serve different practical purposes and it is good to have different approaches to assess them (e.g., [Cartwright, 2006]). However, in a regulatory context, this plurality of standards may complicate the final decision: whether to authorise or not the commercial distribution of a drug. As Robert Temple, an FDA officer put it in an informal debate on the incorporation of Bayesian approaches to regulatory decisions:

> Of course, everybody knows that "p < 0.05" is sort of stupid. Why should it always be the same? Why shouldn't it be adjusted to the situation, to the risks of being wrong in each direction? 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. Maybe we adjust it sometimes when we feel we have to but you simplify the process a little bit so you can get done. I don't want to have to have a symposium for every new trial to decide on an acceptable level of evidence. [Berry et al., 2005, p. 303]

The FDA has been revising their views on acceptable evidence for regulatory purposes over the last decade. Two landmarks in this process are the so-called *Evidence Document* and the *Critical Path Initiative* report[52]: both texts acknowledge that quicker phase III trials drawing on broader data sources are necessary in order to accelerate the approval of new drugs. The current regulatory process is costly for pharmaceutical companies and deprives patients of access to potentially life-saving drugs for years. Complaints about this *drug lag* date back to the late 1960s and early 1970s, when the FDA enforced the current regulatory regime requiring two trials. However, nowadays the Bayesian approach provides a viable alternative to RCTs in order to meet this demand for faster trials and it has been defended precisely along these lines: patients

[52] Food and Drug Administration, *Innovation or stagnation? Challenge and opportunity on the critical path to new medical products* (2004) and *Providing clinical evidence of effectiveness for human drug and biological products. Guidance for industry* (1998), both available at http://www.fda.gov (accessed in July 2009).

want quick access to a better standard of care from the early testing stages just as much as the industry wants faster trials.

In my view, an argument about the way Bayesian trials can help to protect the consumer from incorrect regulatory decisions is still lacking. In particular, it is still undecided which is the best design to cope, within a Bayesian framework, with the growing financial pressure exerted by pharmaceutical companies in the conduct of for-profit clinical trials. The KSS admissibility criteria illustrate the sort of practical issues involved in this process. If an expert decides to recommend a treatment independently of the accruing data, patients will not be protected from inferior treatments; thus it is necessary to incorporate a (fourth) admissibility principle in order to prevent such situations[53]. *Mutatis mutandis*, a similar principle should be applied to the selection of priors in a trial conducted for regulatory purposes, so that pharmaceutical companies do not make abusive use of exaggeratedly optimistic priors. That is, this should occur unless, as happened in the ASTIN trial, we use priors with minimal information in order not to influence the trial data, thus diminishing the potential to exploit the information available before the trial.

I think Steven Goodman is right in pointing out the necessity of a middle ground between the potential flexibility of Bayesian approaches and the necessity of standardised Bayesian procedures that secure good (and quick) regulatory decisions [Berry, 2005, p. 304]. Once these procedures are agreed, their ethical superiority to standard RCTs may not be as outstanding as it currently appears in the examples discussed in the previous section, but this should not be the crucial consideration in our choice of a design for regulatory purposes. Even if it is an imperative to conduct clinical trials with the highest ethical standards, in the current regulatory regime most of them are conducted for the sake of consumer protection. In my view, this latter goal should prevail, as long as our societies deem it necessary to have regulatory agencies overseeing pharmaceutical markets.

4. REFERENCES

[53] The principle goes as follows: "At the outset, no experimental treatment is guaranteed perpetual or even very long-term admissibility" [Sedransk ,1996, p. 77].

[Abraham, 1995] J. Abraham. *Science, Politics and the Pharmaceutical Industry*. N. York: St. Martin's Press, 1995.

[Ashby, 2006] D. Ashby. Bayesian Statistics in Medicine: A 25 Year Review, *Statistics in Medicine* 25: 3589–3631, 2006.

[Basu, 1980] D. Basu. Randomization Analysis of Experimental Data: The Fisher Randomization Test, *Journal of the American Statistical Association* 75: 585-581, 1980.

[Baum et al., 1994] M. Baum, J. Houghton and K. Abrams. Early Stopping Rules--Clinical Perspectives and Ethical Considerations, *Statistics in Medicine* 13: 1459-1472, 1994.

[Beauchamp and Childress, 2001] T. L. Beauchamp and J. F. Childress. *Principles of Biomedical Ethics*. New York: Oxford University Press, 2001.

[Berry, 1993] D. Berry. A Case for Bayesianism in Clinical Trials (with Discussion), *Statistics in Medicine* 12: 1377-1404, 1993.

[Berry, 2004] D. Berry. Bayesian Statistics and the Efficiency and Ethics of Clinical Trials, *Statistical Science* 19: 175-187, 2004.

[Berry, 2005] D. Berry. Clinical Trials: Is the Bayesian Approach Ready for Prime Time? Yes!, *Stroke* 36: 1621-1622, 2005.

[Berry, 2006] D. Berry. Bayesian Statistics, *Medical Decision Making* 26: 429-430, 2006.

[Berry et al., 2002] D. Berry, P. Müller, A. P. Grieve, M. Smith, T. Parke, R. Blazek, et al. Adaptive Bayesian Designs for Dose-Ranging Drug Trials, in C. Gatsonis et. al. (eds.), *Case Studies in Bayesian Statistics Vol. V*. New York, NY: Springer, pp. 99-181, 2002.

[Berry et al., 2005] D. Berry, S. N. Goodman and T. Louis. Floor Discussion, *Clinical Trials* 2: 301-304, 2005.

[Berry and Kadane, 1997] S. M. Berry and J. B. Kadane. Optimal Bayesian Randomization, *Journal of the Royal Statistical Society, Series B: Methodological* 59: 813-819, 1997.

[Biswas et al., 2009] S. Biswas, D. Liu, J. Lee and D. Berry. Bayesian Clinical Trials at the University of Texas M. D. Anderson Cancer Center, *Clinical Trials* 6: 205-216, 2009.

[Buchanan and Miller, 2005] D. Buchanan and F. Miller. Principles of Early Stopping of Randomized Trials for Efficacy: A Critique of Equipoise and an Alternative Nonexploitation Ethical Framework, *Kennedy Institute of Ethics Journal. Je* 15: 161-178, 2005.

[Campbell, 2005] G. Campbell. The Experience in the FDA's Center for Devices and Radiological Health with Bayesian Strategies, *Clinical Trials* 2: 359-363, 2005.

[Cannistra, 2004] S. A. Cannistra. The Ethics of Early Stopping Rules: Who Is Protecting Whom?, *Journal of Clinical Oncology* 22: 1542-1545, 2004.

[Carpenter, 2004] D. Carpenter. The Political Economy of FDA Drug Review: Processing, Politics, and Lessons for Policy, *Health Affairs* 23: 52-63, 2004.

[Carpenter and Moore, 2007] D. Carpenter and C. Moore. Robust Action and the Strategic Use of Ambiguity in a Bureaucratic Cohort: FDA Scientists and the Investigational New Drug Regulations of 1963, in S. Skowronek and M. Glassman (eds.), *Formative Acts: American Politics in the Making*. Philadelphia: University of Pennsylvania Press, pp. 340-362, 2007.

[Cartwright, 2006] N. Cartwright. Well-Ordered Science: Evidence for Use, *Philosophy of Science.* 73: 981-990, 2006.

[Cartwright, 2007] N. Cartwright. Are RCTs the Gold Standard?, *Biosocieties* 2: 11-20, 2007.

[Ceccoli, 1998] S. J. Ceccoli. *The Politics of New Drug Approvals in the United States and Great Britain.* Unpublished Thesis (Ph. D.), Washington University, 1998.

[Chalmers, 2005] I. Chalmers. Statistical Theory Was Not the Reason That Randomization Was Used in the British Medical Research Council's Clinical Trial of Streptomycin for Pulmonary Tuberculosis, in G. Jorland, G. Weisz and A. Opinel (eds.), *Body Counts: Medical Quantification in Historical and Sociological Perspective*. Montreal: McGill-Queen's Press, pp. 309-334, 2005.

[Comanor, 1986] W. Comanor. The Political Economy of the Pharmaceutical Industry, *Journal of Economic Literature* 24: 1178-1217, 1986.

[Council, 1948] Medical Research Council. Streptomycin Treatment of Pulmonary Tuberculosis, *British Medical Journal* 2: 769-782, 1948.

[Couzin, 2004] J. Couzin. The New Math of Clinical Trials, *Science* 303: 784-786, 2004.

[Cowles and Davis, 1982] M. Cowles and C. Davis. On the Origins of the .05 Level of Statistical Significance, *American Psychologist* 37: 553-558, 1982.

[Cox-Maksimov, 1997] D. Cox-Maksimov. *The Making of the Clinical Trial in Britain, 1910-1945: Expertise, the State and the Public.* Cambridge University, Cambridge, 1997.

[Daemmrich, 2004] A. A. Daemmrich. *Pharmacopolitics: Drug Regulation in the United States and Germany*. Chapel Hill: University of North Carolina Press, 2004.

[Edwards, 2007] M. N. Edwards. *Control and the Therapeutic Trial: Rhetoric and Experimentation in Britain, 1918-48*. Amsterdam: Rodopi, 2007.

[Edwards et al., 1998] S. Edwards, R. J. Lilford and J. Hewison. The Ethics of Randomised Controlled Trials from the Perspectives of Patients, the Public, and Healthcare Professionals, *British Medical Journal* 317: 1209-1212, 1998.

[Emrich and Sedransk, 1996] L. Emrich and N. Sedransk. Whether to Participate in a Clinical Trial: The Patient's View, in J. B. Kadane (ed.), *Bayesian Methods and Ethics in Clinical Trial Design*. New York: Wiley, pp. 267-305,1996.

[Epstein, 1996] S. Epstein. *Impure Science. Aids and the Politics of Knowledge*. Berkeley-Los Angeles: University of California Press, 1996.

[Evans et al., 2006] I. Evans, H. Thornton and I. Chalmers. *Testing Treatments. Better Research for Better Healthcare*. London: The British Library, 2006.

[Featherstone and Donovan, 1998] K. Featherstone and J. L. Donovan. Random Allocation or Allocation at Random? Patients' Perspectives of Participation in a Randomised Controlled Trial, *BMJ* 317: 1177-1180, 1998.

[Flory and Emanuel, 2004] J. Flory and E. Emanuel. Interventions to Improve Research Participants' Understanding in Informed Consent for Research: A Systematic Review, *JAMA* 292: 1593-1601, 2004.

[Gillies, 1973] D. Gillies. *An objective theory of probability*. London: Methuen, 1973.

[Gifford, 1986] F. Gifford. The Conflict between Randomized Clinical Trials and the Therapeutic Obligation, *Journal of Medicine and Philosophy. N* 86: 347-366, 1986.

[Gifford, 1995] F. Gifford. Community-Equipoise and the Ethics of Randomized Clinical Trials, *Bioethics. Ap* 9: 127-148, 1995.

[Gigerenzer, 1989] G. Gigerenzer et al.. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge [England]; New York: Cambridge University Press, 1989.

[Goodman, 1999a] S. N. Goodman. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy, *Annals of Internal Medicine* 130: 995-1004, 1999.

[Goodman, 1999b] S. N. Goodman. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor, *Annals of Internal Medicine* 130: 1005-1013, 1999.

[Goodman, 2007] S. N. Goodman. Stopping at Nothing? Some Dilemmas of Data Monitoring in Clinical Trials, *Ann Intern Med* 146: 882-887, 2007.

[Grieve and Krams, 2005] A. Grieve and M. Krams. ASTIN: A Bayesian Adaptive Dose–Response Trial in Acute Stroke, *Clinical Trials* 5: 340-351, 2005.

[Hackshaw, 2009] A. Hackshaw. *A Concise Guide to Clinical Trials*. London: Wiley-Blackwell, 2009.

[Hartnett, 2000] T. Hartnett (ed.). *The Complete Guide to Informed Consent in Clinical Trials*. Springfield (VA): PharmSource Information Services, 2000.

[Heitmiller et al., 1996] E. Heitmiller, J. B. Kadane, N. Sedransk and T. Blanck. Verapamil Versus Nitroprusside: Results of the Clinical Trial II, in J. B. Kadane (ed.), *Bayesian Methods and Ethics in Clinical Trial Design*. New York: Wiley, pp. 211-219,1996.

[Hennekens et al., 2004] C. H. Hennekens, F. M. Sacks, A. Tonkin, J. W. Jukema, R. P. Byington, B. Pitt, et al. Additive Benefits of Pravastatin and Aspirin to Decrease Risks of Cardiovascular Disease: Randomized and Observational Comparisons of Secondary Prevention Trials and Their Meta-Analyses., *Archives of Internal Medicine* 164: 40-44, 2004.

[Inoue et al., 2005] L. Y. T. Inoue, D. Berry and G. Parmigiani. Relationship between Bayesian and Frequentist Sample Size Determination, *The American Statistician* 59: 79-87, 2005.

[Jain, 2007] S. Jain. *Understanding Physician-Pharmaceutical Industry Interactions*. Cambridge [England]; New York: Cambridge University Press, 2007.

[Kadane, 1994] J. B. Kadane. An Application of Robust Bayesian Analysis to a Medical Experiment, *Journal of Statistical Planning and Inference* 40: 221-232, 1994.

[Kadane, 1996] J. B. Kadane. *Bayesian Methods and Ethics in a Clinical Trial Design*. New York: Wiley, 1996.

[Kadane et al., 1999] J. B. Kadane, M. J. Schervish and T. Seidenfeld. *Rethinking the Foundations of Statistics*. Cambridge [England]; New York: Cambridge University Press, 1999.

[Kadane and Sedransk, 1996] J. B. Kadane and N. Sedransk. Verapamil Versus Nitroprusside: Results of the Clinical Trial I, in J. B. Kadane (ed.), *Bayesian Methods and Ethics in Clinical Trial Design*. New York: Wiley, pp. 177-210,1996.

[Kadane and Seidenfeld, 1990] J. B. Kadane and T. Seidenfeld. Randomization in a Bayesian Perspective, *Journal of Statistical Planning and Inference* 25: 329-345, 1990.

[Kadane and Seidenfeld, 1996] J. B. Kadane and T. Seidenfeld. Statistical Issues in the Analysis of Data Gathered in the New Designs, in J. B. Kadane (ed.), *Bayesian Methods and Ethics in Clinical Trial Design*. New York: Wiley, pp. 115-125,1996.

[Krams et al., 2005] M. Krams, K. R. Lees and D. A. Berry. The Past Is the Future: Innovative Designs in Acute Stroke Therapy Trials, *Stroke* 36: 1341-1347, 2005.

[Krams et al., 2003] M. Krams, K. R. Lees, W. Hacke, A. P. Grieve, J.-M. Orgogozo, G. A. Ford, et al. Acute Stroke Therapy by Inhibition of Neutrophils (Astin): An Adaptive Dose-Response Study of Uk-279,276 in Acute Ischemic Stroke.[See Comment], *Stroke* 34: 2543-2548, 2003.

[Krimsky, 2003] S. Krimsky. *Science in the Private Interest: Has the Lure of Profits Corrupted Biomedical Research?* Lanham: Rowman & Littlefield Publishers, 2003.

[Lehmann, 1993] E. L. Lehmann. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?, *Journal of the American Statistical Association* 88: 1242-1249, 1993.

[Levi, 1982] I. Levi. Direct Inference and Randomization, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2: 447-463, 1982.

[Levine, 1988] R. J. Levine. *Ethics and Regulation of Clinical Research*. New Haven-London: Yale University Press, 1988.

[Lexchin et al., 2003] J. Lexchin, L. Bero, B. Djulbegovic and O. Clark. Pharmaceutical Industry Sponsorship and Research Outcome and Quality: Systematic Review, *British Medical Journal* 326: 1167-1170, 2003.

[Macklin, 2004] R. Macklin. *Double Standards in Medical Research in Developing Countries*. Cambridge: Cambridge University Press, 2004.

[Mainland, 1960] D. Mainland. The Use and Misuse of Statistics in Medical Publications, *Clin. Pharmacol. Ther.* 1: 411-422, 1960.

[Marks, 1997] H. M. Marks. *The Progress of Experiment. Science and Therapeutic Reform in the United States, 1900-1990*. N. York: Cambridge University Press, 1997.

[Marks, 2000] H. M. Marks. Trust and Mistrust in the Marketplace: Statistics and Clinical Research, 1945-1960, *History of Science* 38: 343-355, 2000.

[Montori et al., 2005] V. M. Montori, P. J. Devereaux, N. K. J. Adhikari, K. E. A. Burns, C. H. Eggert, M. Briel, et al. Randomized Trials Stopped Early for Benefit: A Systematic Review, *JAMA* 294: 2203-2209, 2005.

[Mueller et al., 2007] P. S. Mueller, V. M. Montori, D. Bassler, B. A. Koenig and G. H. Guyatt. Ethical Issues in Stopping Randomized Trials Early Because of Apparent Benefit, *Annals of Internal Medicine* 146: 878-881, 2007.

[Neyman, 1957] J. Neyman. 'Inductive Behavior' as a Basic Concept of Philosophy of Science, *Review of the International Statistical Institute* 25: 7-22, 1957.

[Papineau, 1994] D. Papineau. The Virtues of Randomization, *British Journal for the Philosophy of Science* 45: 437-450, 1994.

[Piantadosi, 2005] S. Piantadosi. *Clinical Trials. A Methodological Approach*. Hoboken (NJ): Wiley, 2005.

[Pocock, 1983] S. J. Pocock. *Clinical Trials: A Practical Approach*. Chichester [West Sussex]; New York: Wiley, 1983.

[Porter, 1995] T. M. Porter. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J.: Princeton University Press, 1995.

[Seidenfeld, 1979] T. Seidenfeld. *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher*. Dordrecht-London: Reidel, 1979.

[Sedransk, 1996] N. Sedransk. Admissibility of Treatments, in J. B. Kadane (ed.), *Bayesian Methods and Ethics in Clinical Trial Design*. New York: Wiley, pp. 65-113,1996.

[Sismondo, 2009] S. Sismondo. Ghosts in the Machine: Publication Planning in the Medical Sciences, *Social Studies of Science* 39: 171-198, 2009.

[Spiegelhalter et al., 2004] D. J. Spiegelhalter, K. Abrams and J. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: John Wiley, 2004.

[Spiegelhalter et al., 1994] D. J. Spiegelhalter, L. S. Freedman and M. K. B. Parmar. Bayesian Approaches to Randomized Trials, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 157: 357-416, 1994.

[Sterne and Smith, 2001] J. A. C. Sterne and G. D. Smith. Sifting the Evidence—What's Wrong with Significance Tests?, *British Medical Journal* 322: 226–231, 2001.

[Stone, 2007] P. Stone. Why Lotteries Are Just, *The Journal of Political Philosophy* 15: 276-295, 2007.

[Taylor et al., 1984] K. Taylor, R. Margolese and C. Soskolne. Physicians' Reasons for Not Entering Eligible Patients in a Randomized Clinical Trial of Surgery for Breast Cancer, *N. Engl. J. Med.* 310: 1363-1367, 1984.

[Toth, 1998] B. Toth. *Clinical Trials in British Medicine 1858-1948, with Special Reference to the Development of the Randomised Controlled Trial.* University of Bristol, Bristol, 1998.

[Urbach, 1985] P. Urbach. Randomization and the Design of Experiments, *Philosophy of Science. JE* 85: 256-273, 1985.

[Urbach, 1993] P. Urbach. The Value of Randomization and Control in Clinical Trials (with Discussion), *Statistical Science* 12: 1421-1431, 1993.

[Vandenbroucke, 2004] J. P. Vandenbroucke. When Are Observational Studies as Credible as Randomised Trials?, *Lancet* 363: 1728-1731, 2004.

[Walton et al., 2005] M. Walton, R. Simon, F. Rockhold, W. DuMouchel, G. Koch and R. O'Neill. Panel Discussion of Case Study 3, *Clinical Trials* 2: 352-358, 2005.

[Whitehead, 1993] J. Whitehead. The Case for Frequentism in Clinical Trials (with Discussion), *Statistical Science* 12: 1405-1419, 1993.

[Worrall, 2007] J. Worrall. Evidence in Medicine and Evidence-Based Medicine, *Philosophy Compass* 2: 981-1022, 2007.

[Worrall, 2008] J. Worrall. Evidence and Ethics and Medicine, *Perspectives in Biology and Medicine* 51: 418-431, 2008.

[Yank et al., 2007] V. Yank, D. Rennie and L. Bero. Financial Ties and Concordance between Results and Conclusions in Meta-Analyses: Retrospective Cohort Study, *British Medical Journal* 335: 1202-1205, 2007.