Quanticate

MAKE LIFE SIMPLER

Bayesian Study Design: The Pragmatic Solution for Phase II Clinical Development

Introduction

It is a dilemma faced by many pharmaceutical and Biotech companies on a frequent basis: you have a New Chemical Entity (NCE) which you hope will successfully treat a particular disease; the NCE has completed phase I and now attention is focused on designing the first phase IIa study to assess efficacy; and you want to collect robust data to demonstrate the efficacy of the NCE, but budget and resources are tight.

A fully powered clinical trial is expensive, and is unlikely to be appropriate given the lack of information about the likely efficacy profile of the NCE at this early stage of development. A quick, small, cheap 'look and see' study may be easy to run, but without a proper up-front consideration of the study design, may lead to results that are unreliable, difficult to interpret or not statistically robust. An alternative approach, designing such studies in a Bayesian framework, enables clinical development teams to plan studies that are practical, appropriately sized, and with a clear understanding of the statistical robustness of the results and of the chances of success and failure.

This article will give a brief overview of Bayesian methods and introduce some simple graphical tools to assess the properties of particular Bayesian designs. Future articles will extend these ideas to look at flexible study designs with interim analyses, and evaluation of the data at the end of the study within the Bayesian framework.

Traditional Approaches

Traditionally, clinical trials of NCEs have been designed to provide confirmatory evidence of a therapeutic benefit of a particular magnitude over and above that of some reference therapy. Typically, this might be evidence of superior efficacy, but alternatively might be non- inferior or equivalent efficacy but with other advantages such as improved tolerability or more favourable dosing regimen.

A typical starting point is to construct a null and alternative hypothesis as follows:

Null hypothesis: The NCE and the reference therapy have the same effect on the outcome of interest

Alternative hypothesis: The NCE and the reference therapy have different effects on the outcome of interest

Data are gathered during the course of the trial and analysed at the study end to assess the strength of evidence in support of the null hypothesis, typically presented as a p value. If this is sufficiently small, the null hypothesis is rejected in favour of the alternative, and it is concluded that there is a difference between the NCE and reference therapy.

The study includes enough subjects to keep the probability of incorrectly rejecting the null hypothesis when it is in truth correct (the probability of a type 1 error) to a certain level (alpha) and to ensure there is a sufficiently high probability (power) of correctly rejecting the null hypothesis when it is in truth wrong. The sample size is calculated under the assumptions that the magnitude of difference between the NCE compared to the reference is some known, pre-specified value, and the variability of the outcome of interest is also known.

This methodology (often referred to as Frequentist) works well in cases where there is good reason to believe that the assumptions made at the start, about the magnitude of the difference between the NCE and the reference therapy (and also about the variability of the outcome of interest), are correct. This information can often be gleaned, for example, from previous, similarly designed studies. In this case the study may be viewed as confirmatory – providing further evidence that the benefit of the NCE is indeed of the particular desired magnitude. But what if we don't have that sort of information to hand? Does it really make sense to design trials to demonstrate evidence of a pre-specified magnitude of effect when we don't have any information beforehand to suggest that the NCE is likely to be able to show such an effect?

This is typically the case in early phase clinical trials, such as proof of concept (PoC) studies which are usually the first assessment of efficacy of an NCE. Rather than setting up the study to meet a hurdle without any evidence that we are likely to be able to jump it, it is perhaps more reasonable to design these initial studies to assess the probability that the NCE has any benefit at all, and also the probability that the NCE might have the magnitude of benefit that other therapies have, or that the regulators require. These are questions that the classical Frequentist statistical approaches are unable to address, and therefore an alternative framework must be considered.

Quanticate

Bayesian Methods

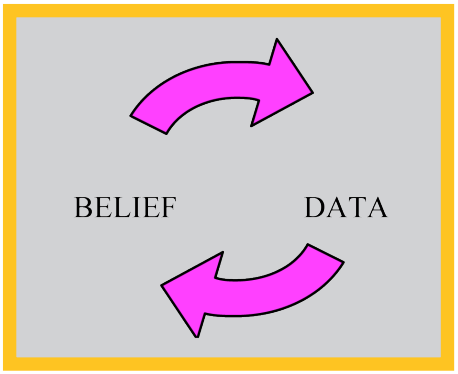
Bayesian methods can be used in clinical trials to answer questions such as 'What is the probability that the NCE has a benefit over the reference therapy?' or 'What is the probability that the magnitude of benefit over the reference therapy is at least a certain amount?'. They also have significant utility in adaptive study designs.

The basic Bayesian approach works in the following way (technical terms in parentheses):

i) I have a belief about the likely magnitude of effect of the NCE, and I am able to express how confident I am that my belief is correct (Prior belief)
ii) I gather some data to explore what the likely magnitude of effect of the NCE might be (Likelihood)
iii) Based on the data I have collected, I update my belief about the likely magnitude of effect of the NCE (Posterior belief).

Once we have our Posterior belief (iii), we can use this as an updated Prior belief (i) and go through the process again, gathering more data (for example by including more subjects in the study, or conducting a new study) and further updating our belief. Figure 1 shows how this can be viewed as a simple cycle:

Figure 1: Bayesian Philosophy - The relationship between data and belief



The first question that must be considered when applying a Bayesian approach is that of the initial belief. From where does my prior belief come? This question is not a trivial one, and can be a particularly difficult question to address for the first study, as there is rarely any reliable data available upon which to construct a prior opinion. However, a simple solution is to start from a position of ignorance. The prior belief becomes 'I do not have any reason to believe that any one value for the magnitude of benefit of the NCE over the reference therapy is any more likely than any other'. In other words, all possible values for the treatment effect are equally plausible.
I then gather some data and see how this belief changes as a result of what I observe in my study. In this case, the posterior belief is entirely driven by the observed data and not by the prior opinion. We call this type of prior belief (that of ignorance) a 'non-informative' prior.

Aside: Understanding Effect Size

So that the examples that follow can be understood in the most general sense, we will be expressing our desired treatment effect in terms of the Effect Size (ES).

The standard effect size is defined as follows:

Effect Size (ES) = Treatment difference/Standard Deviation

For example, a NCE that delivers a difference in treatment means (compared to a reference therapy) of 3.5 points, where the standard deviation of the outcome measure is 10 points, has an ES of 0.35.

Application: Decision Making Rules and Sample Size for a PoC study

Suppose we have a NCE which is about to enter its first efficacy study, and will be compared to placebo. We want to show that the NCE is better than placebo, and our target effect size (based on data from currently marketed agents in our indication of interest) is 0.4. Since this is our first efficacy study, we do not wish to do a large confirmatory study, rather we would like to explore what the efficacy profile might look like, and if there is a reasonable probability that the NCE does provide benefit, then development will proceed to the next stage.

There are two key questions to consider:

I) How many subjects should we use?
2) What level of confidence do we want that the NCE is effective in order to further invest in development?

The two questions are closely linked, and we can use simulations within the Bayesian framework to help us assess the appropriateness of different sample sizes and decision rules. We will begin from a position of prior ignorance, i.e. we do not want to commit ourselves to any particular view on what the likely effect of the NCE may be, and therefore we shall apply a non-informative prior. Table 1 shows the different scenarios that we will explore:

Table 1: Overview of scenarios

	Scenario 1	Scenario 2	Scenario 3
Sample Size	Varied: 20 to 100 per group	Fixed: 50 per group	Fixed: 50 per group
Decision Rule - continue development if...	Posterior probability that the ES is >0 is at least *threshold*	Posterior probability that the ES is >0 is at least *threshold*	Posterior probability that the ES is >0.4 is at least *threshold*
Threshold	Fixed: 90%	Varied: 50% - 90%	Varied: 10% - 90%

Scenario 1: Choosing the Sample Size

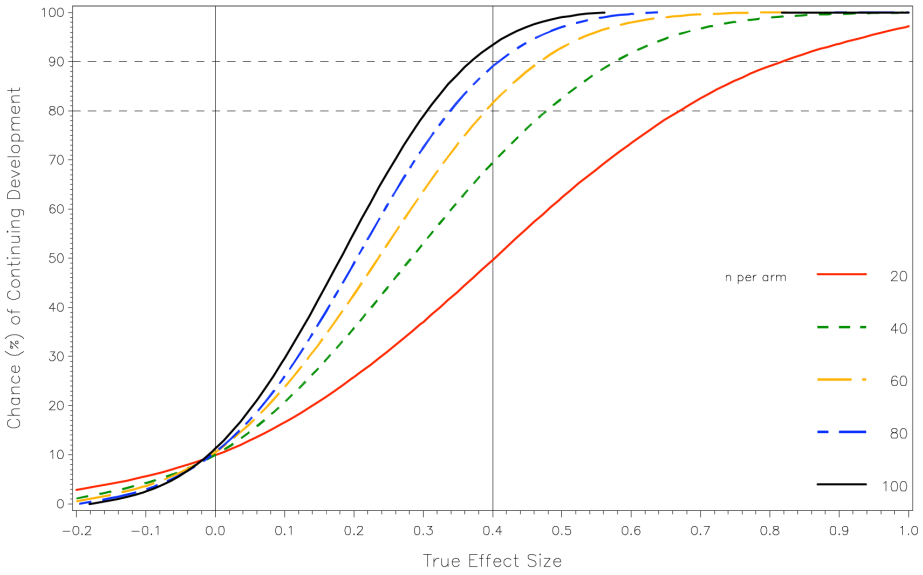
Let's start off by fixing the answer to question 2. Suppose we decide that we will continue development if, at the end of our study, the posterior probability that the NCE is better than placebo is at least 90%. This is equivalent to saying the posterior probability that the true effect size is >= 0 is at least 90%.

We want to choose an appropriate sample size for the study such that, under this decision making-rule, we would choose to continue the development of the NCE if the true actual effect size of the NCE compared to placebo is 0.4 or more. We also want to avoid choosing to continue to develop the NCE if in truth it is no better than placebo (i.e. a false positive outcome).

We can simulate, for different sample sizes, the probability that this decision rule would lead us to make a decision to continue development, under the assumption of various possible values for the actual true effect size. In Figure 2, the x-axis represents these possible true effect sizes (with values greater than 0 indicating a benefit of the NCE), and the y –axis shows the chance of making a 'continue development' decision. The lines on the figure represent different choices of sample sizes (number of subjects per treatment arm).

Figure 2: Simulation results showing the chance of continuing development based on different sample size (assuming a non-informative prior distribution)

Decision Rule: Posterior Probability that (Effect Size > 0) is > 90%



Note: 0.4 is the target effect size. Results are based on 10000 simulations.

From Figure 2 we can assess the benefits of increasing the sample size in the study. With 40 subjects per group, if the NCE really does have an effect size of 0.4, then with this decision rule we will continue development 70% of the time (green short dashed line). If we include 80 subjects per group we will continue development 88% of the time (blue mixed dashed line). If the NCE is no better than placebo (effect size=0) then there is 10% chance we would continue development even though the NCE is not effective (a false positive outcome). If the NCE is worse than placebo (effect size <0) then there is only a small chance of making a 'continue development' decision, and this chance decreases for larger sample sizes.

Scenario 2: Evaluating the Appropriateness of Decision Criteria based on Strength of Evidence for a Drug Effect

In scenario 1, our rule for deciding whether or not to continue development at the end of the study was set up front, such that we would decide to continue development if, at the end of our study, the posterior probability that the NCE is better than placebo is at least 90%. We then looked at how many subjects we needed to include in the study in order to have confidence in the decision we make when applying that rule.

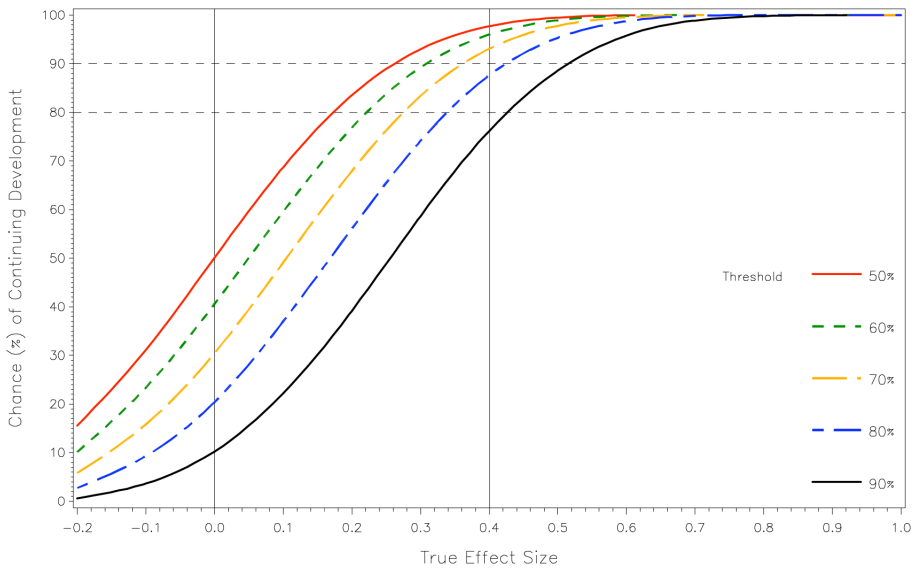
However, we do not have to use 90% as our threshold for continuing development, we could choose a different value, such as 80%, or 95%. In the situation where the budget for the study is tight and we have a limit to the number of subjects we can use, we may need to look at different thresholds for decision making in order to ensure we make sensible decisions given the limited amount of data available.

Suppose we have a sample size which is fixed at 50 per treatment arm. We decide that we will continue to develop the NCE providing that the posterior probability that the NCE works (ES>0) is at least some threshold value, such as 60%, 80%, 95%. It is important to select the threshold before conducting the study, so that we can have confidence that it will lead to sensible decision making when we have the study results. We can use simulation to help us choose what the most appropriate threshold might be.

In Figure 3, as previously, the x-axis represents these possible true effect sizes (with values greater than 0 indicating a benefit of the NCE), and the y –axis shows the chance of making a 'continue development' decision. The lines on the figure represent different choices of threshold.

Figure 3: Simulation Results showing the chance of continuing development for different posterior probability thresholds for success based on a sample size of 50 subjects per treatment arm (assuming a non-informative prior distribution).

Decision Rule: Posterior Probability that (Effect Size > 0) is > threshold



Note: 0.4 is the target effect size. Results are based on 10000 simulations.

From Figure 3 we can see, for example, that if we were to choose 50% as our threshold (red solid line), then if the NCE really does have an effect size of 0.4, we will continue development >95% of the time. However if the NCE is in truth no better than placebo (effect size=0) then there is 50% chance we would continue development even though the NCE is not effective. Clearly this is a high risk of a false positive outcome and we probably want to look at a different decision rule.

If instead we set the threshold at 80% (blue mixed dashed line), we can see that if the NCE really does have an effect size of 0.4, with this decision rule we will continue development ~87% of the time, but if the NCE is no better than placebo (effect size=0) then there is now only a 20% chance we would continue development even though the NCE is not effective.

The clinical development team can then consider whether 20% is still too high a risk of a false positive result, or whether they want to look at alternatives, which may include using a higher threshold but accepting the increased risk of a false negative decision (i.e. failing to develop a NCE that in truth is effective), or else increasing the sample size.

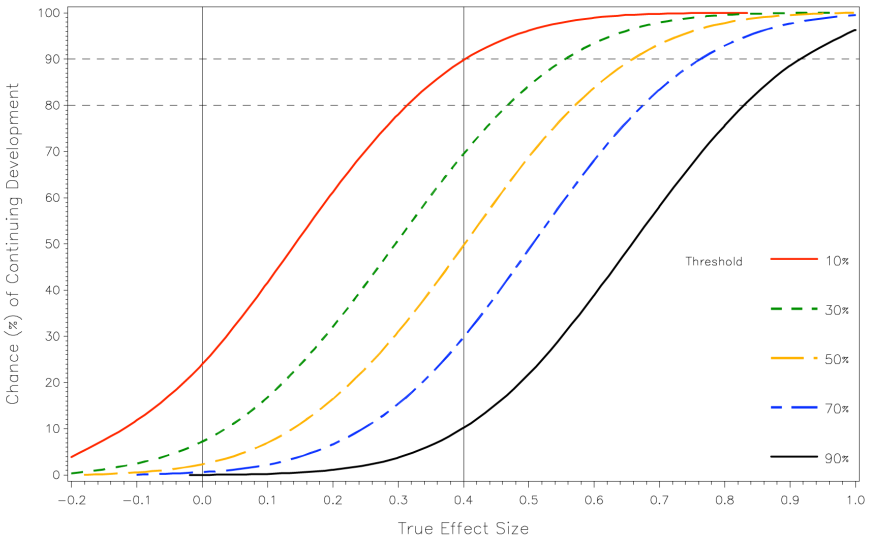
Scenario 3: Evaluating the Appropriateness of Decision Criteria based on Strength of Evidence for the Target Drug Effect

So far, our decision rules have been based on the posterior probability that the NCE performs better than placebo (i.e. that the effect size is >0). It may be that the clinical development team are interested in the posterior probability that the effect size is at least as good as the target of 0.4.

We can simulate this in exactly the same way as in scenario 2, with the same assumed sample size of 50 per treatment arm. The only difference is that in Figure 3 the thresholds for continuing development now relate the posterior probability that the effect size is >0.4, instead of >0.

Figure 4: Simulation Results showing the chance of continuing development for different posterior probability thresholds for success based on a sample size of 50 subjects per treatment arm (assuming a non-informative prior distribution)

Decision Rule: Posterior Probability that (Effect Size > 0.4) is > threshold



Note: 0.4 is the target effect size. Results are based on 10000 simulations.

From Figure 4, we can see that by raising the development hurdle such that we are now interested in the probability of the true effect size being > 0.4 rather than >0, the chances of deciding to continue development are considerably reduced.

For example, if the NCE really does have an effect size of 0.4, by setting our posterior probability (that the ES is >0.4) threshold at 50%, we will continue development only 50% of the time (yellow long dashed line). Even if we lower the threshold to 30% (green short dashed line), we would still only continue development ~70% time if the NCE does in truth have an effect size of 0.4. However if the NCE is in truth no better than placebo (effect size=0) then there is only a small chance of a false positive outcome.

In practice, although these decision rules have instinctive appeal ('We will only continue development if there is a high chance that the NCE works as well as we want it to'), they can lead to poor decision making in these early, relatively small phase IIa studies. Rules such as those in scenarios 1 and 2 are generally more helpful at this stage of development.

Concluding Thoughts

Bayesian methods have particular value in situations where the rigour of traditional Frequentist methods is not appropriate, or when there is uncertainty over what the expected drug effect may be. They enable clinical development teams to design studies of a realistic size with a clear understanding of the likely study performance and the robustness of the results.

The scenarios presented show the utility of the methodology in the simplest of cases, a normally distributed endpoint with a non-informative prior distribution. Some people may feel that there is no such thing as prior ignorance, and that it is not realistic to assume that we start our trial with no idea about the likely magnitude of NCE benefit. However, the use of such 'non-informative' priors opens up the opportunity to design trials and evaluate results within the Bayesian framework, without worrying that our choice of prior might be inappropriately dominating our posterior belief, compared to the data that we observe.

It is easy to apply the methods described in this paper with different priors, for example an optimistic prior that expresses confidence that the NCE will deliver the desired effect, or a pessimistic prior that assumes that the NCE is no better than the reference therapy. Combining the study data with various different priors can provide a valuable assessment of the robustness of study results.

Bayesian approaches mirror the scientific method, and encourage discussion amongst the team at an early stage about the relative merits of choice of sample size and decision rule, with respect to the associated risks of false negative or false positive outcomes. Within the Bayesian framework, it is possible to answer questions of how likely the NCE is to be efficacious, and to evaluate what the most likely effect will be. This information can be used to design future confirmatory trials.

This article has shown how the use of Bayesian methods in a clinical trial setting provides an intuitive framework for evaluating the evidence for drug effects. Future articles will illustrate the benefits of the Bayesian approach in determining decision rules for interim analyses, and also some helpful ways of presenting and interpreting the results from Bayesian analyses at the end of a study.

Should you wish to receive future articles on this subject, please register your interest by e-mailing enquiries@quanticate.com with "Interested in further articles on Bayesian methodology" in the Subject line.

