

Title: **Commentary on FDA guidance on Bayesian statistics in medical device clinical trials**

Authors: **Roger Sewell**
Senior Consultant at Cambridge Consultants, Cambridge, UK
MA (Cantab) in Mathematics
BM BCh (Oxon)

DM (Oxon)

Elisabeth Crowe
Senior Consultant at Cambridge Consultants, Cambridge, UK
BSc (Dun) in Physics
MSc in Medical Physics with Information Technology

A shorter version of this article is due to appear in the March/April 2010 issue of The Regulatory Affairs Journal: Devices Vol 18 Issue 2.

1 Introduction

On 5th February 2010 the FDA issued its revised guidance document “Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials” (“the Guidance”, [1]). This article provides commentary on the Guidance and describes the practical implications for a medical device sponsor considering taking advantage of this useful methodology. For an initial introduction to Bayesian Inference see [1, 2].

2 Why do we believe that Bayesian inference is the right way to proceed?

We believe this Guidance to be an improvement on the draft version issued in 2006. It makes clearer what the FDA will expect of a company using this technique in the evaluation of its devices, and gives sponsors more confidence that this technique may be acceptable to the FDA.

We are very pleased to see the FDA moving in this direction because we believe Bayesian Inference to be the right way of analysing clinical data in order to give the right answer to the right question. Here the “right question” is “What is the probability that this device is safe and effective, given the data that we have observed?”

The Bayesian method and its advantages are outlined in the Guidance. The main advantages over traditional frequentist methods are:

- Trials may be able to be designed to be shorter in length or smaller in size due to the superior use of the information than is possible with frequentist methods, reducing their cost. The principal reason for this is that data collection can stop as soon as the probability that the device is satisfactory given the data has

increased beyond the necessary (usually 0.95) level, and the data can be analysed continuously without making any difference to the size of effect that needs to be observed in each analysis.

- The method is adaptable and can allow sponsors to make changes in the trial without starting afresh. For example, suppose a group of patients have been exposed to a particular device, and roughly one tenth have experienced an adverse effect. Examining the data, investigators realise that it is Rhesus negative patients who are suffering the adverse effect. The Bayesian approach permits retrospective analyses on subgroups (conditional on suitable priors for connections between the subgroups being given) and permits correct inference to be made on them.
- The Bayesian paradigm provides correct calculation of the probability that the device is safe and effective given the data.

The Bayesian method can be proved to be optimal for a mathematically posed problem. Suppose we know beforehand which values of a parameter θ are likely and which not, i.e. we know a prior distribution $P(\theta)$. Suppose we also know, for any true value of θ , the probability distribution $P(D|\theta)$ of the data D we have observed. Then the Bayesian posterior distribution

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

is the distribution $Q_D(\theta)$ that gives the maximum possible value of the apparent Shannon information

$$J(\theta; Q_D) = \int P(\theta, D) \log \frac{Q_D(\theta)}{P(\theta)} d(\theta, D) = E \log \frac{Q_D(\theta)}{P(\theta)},$$

and any other distribution giving the same value is $P(\theta|D)$ -almost-surely equal to $P(\theta|D)$.

- Some problems are specified to some extent by the data provided. For example, it may be that when testing a multi-dose inhaler an investigator is uncertain whether the doses will be distributed normally, log-normally, or according to one of the Student, log-Student, or skew-Student distributions. In such a situation, Bayes provably chooses optimally between the range of mathematical formulations offered by the investigator and then solves that or those formulation(s) optimally, despite the fact that these distribution families do not all have the same number of parameters. What Bayes will not do is suggest the distribution families to consider.
- Similarly, the Bayesian approach can help one to look for outliers in the data, for example points that are too far deviated to be plausibly normally distributed. However, we would caution it is often appropriate first to ask whether the data is indeed normally distributed in the first place – in our experience Student distributions are far more common in real life.

The Guidance recommends that investigators do not switch from a frequentist to a Bayesian method; we would agree and would further recommend that sponsors consider using Bayesian methods from the outset for all trials.

3 Planning to use Bayesian statistics for trials

If a device sponsor is interested in using this statistical approach for a trial, we would recommend the following steps.

1. Employ an expert. Bayesian methods can be complex, especially when choosing the prior information, analysing available data and formulating mathematical equations that fully define the model.
2. Examine the options for prior data carefully.
3. Consider re-analysing data from a previous trial using Bayesian methods – to give a concrete example from which to build understanding of the method and its outputs, and of the necessary differences in trial design. In particular it is worth reflecting on whether the trial might have been completed successfully earlier and at less cost.
4. Seek guidance on appropriate software tools and how to customise them for the particular trial in question. The Guidance cites WinBUGS and its associated packages, BRUGS and OpenBUGS. Other libraries and toolkits, e.g. for Matlab®, are also available for constructing software to deal efficiently with unusual problems.
5. Sponsors should ensure they have a confident Bayesian statistician available to help them discuss the proposal for the trial with the FDA and to answer submission questions. The use of Bayesian methods is still new and controversial. A well-founded approach based on solid experience will be more likely to succeed.

As with every aspect of clinical trial design, early planning is more likely to result in an efficient and effective trial that is also acceptable to the FDA. The possible effects on your trial brought about by using Bayesian methods are:

- There are many opportunities to use interesting features in a design which would not be available in a typical frequentist design. For example:
 - Multiple reanalyses without penalty can save unnecessary duration and cost;
 - Adapting the operation (e.g. dosing) of the device as a result of data collected to date may reduce the risk of failure due to wrong dosage or eliminate the need for multiple dose-ranging studies in advance.

The sooner the FDA is informed of such plans, the more probable it is that they may find them acceptable.

- Simulations of the proposed trial are appropriate to provide estimates of cost and duration, and probabilities that various branches may end up being taken.
- The need to meet with the FDA and discuss the preferred plans and prior distributions at the start of the trial is just as important as for frequentist trials.

The FDA do admit the possibility of making changes to the trial protocol mid-trial and the validity of Bayesian analyses in this situation, but this course cannot be relied on as a substitute for making appropriate plans in advance.

4 Related issues

The Guidance makes it clear that sponsors will be required to submit the software used for the analysis. It is clear that the software used must perform as expected and be designed using a thorough approach that may be similar to that recommended for software forming part of a medical device. We would recommend the development includes at minimum issue and review of:

- Software requirements
- Software design specification
- Software test specification
- Software test results

for each iteration of the software development. This should help to avoid software errors that could affect the analysis, give an inaccurate outcome and be very costly in terms of FDA credibility and time taken to resolve the errors.

For pharmaceutical trials all the same issues apply, in that we believe that the Bayesian approach is superior to frequentist methods in allowing the right questions to be asked and answered efficiently; again it would represent the least burdensome approach, and may shorten the time to approval.

A further area in which a Bayesian approach may be very valuable, and an area in which the FDA is beginning to define its approach to regulation, is that of biomarkers. It is rare to find a single biomarker that can predict the likelihood of a particular outcome, be that cancer recurrence or prognosis, or treatment effectiveness. Thus multi-marker panels of data are being explored as diagnostic tools by many companies. Often the answer to the question “What set of biomarkers can give useful information about an outcome” is data led and can be answered via a Bayesian analysis of the data collected. We hope the FDA will be accepting Bayesian techniques in this growing field as they work with sponsors seeking to obtain qualification of biomarker panels.

5 Type 1 error rate

Whilst we commend the publication of the Guidance and are excited about the possibilities it raises for improved analysis of clinical trial data, we believe it contains some unresolved issues.

In particular we would like to draw attention to sections 4.8 and 7.1-7.2 of the Guidance. While we feel that a small step forward has been made in the admission that true Bayesian methods take no account of the type 1 error rate (page 29 paragraph 3), the FDA nonetheless has failed to truly take the step from the frequentist paradigm to the Bayesian paradigm by still wanting to regulate the type 1 error rate.

We have, of course, no objection to calculating the various parameters listed in section 4.8. These include the type 1 error rate (probability of approving a device that is just outside the desired parameter range), the type 2 error rate (probability of not approving a device that has precisely nominal characteristics), the probability of not approving a device that is just inside the desired parameter range, the expected number of measurements/samples needed, the distribution of the duration and cost of the trial, the prior probability that the device has parameters in the desired range, etc. Many of these pieces of information are needed for making decisions that the trial is worth funding. Our disagreement with the FDA comes when they want to base decisions on whether a Bayesian analysis is admissible on whether its type 1 error rate is small enough.

The interest in type 1 error rate comes entirely from the frequentist paradigm, where the question being asked is “If this device has parameters just outside the desired region, what is the probability of the trial approving the device?”. The question we should be asking is “What, in the light of the data, is the probability that the device has parameters inside the desired region (i.e. is safe and effective)?”, which is the nearest answerable question to “Is the device safe and effective?”. i.e. we should be examining the device under test, not examining the trial.

In the following example we illustrate how the Bayesian approach helps to minimise the burden on the sponsor to the extent appropriate to the strength of the prior information being used. The extreme case is where for some reason a justifiable prior distribution leads to no experiments being needed in the trial at all.

Suppose we have developed, tested, and approved a drug delivery device to deliver drug A in aqueous solution, a drug with very narrow therapeutic range, with doses log-normally distributed and within 1% of the label claim in 99% of doses. Suppose now that all are agreed in advance that drug B, which has a very wide therapeutic range, will be safely and efficaciously delivered even if up to 5% of the doses delivered are outside the range $2/3$ to $3/2$ of the label claim. For simplicity we assume that the label claim dose is the same for both drugs. (We realise this example is unlikely to be the case for real world drugs, but it allows us to convey the ideas).

We now consider three sets of prior information (assumptions) one might choose to adopt – the choice being agreed with the FDA at the start of the trial:

1. The first set (model 1) says that *a priori* we consider that a solution of drug B will definitely behave exactly the same as a solution of drug A, and that the mean and standard deviation of the log dose will be the same for both drugs.
2. The second set (model 2) says that dose is again log-normally distributed, the standard deviation of log-dose is almost certainly the same as when using drug A, but *a priori* the median dose of drug B delivered by the device is larger than

the dose of drug A delivered by an unknown factor α , log-normally distributed with median 1.0 and standard deviation of log dose 3 nepers. In other words one thinks that there is probability of around 0.32 that the median delivered dose will be too high or too low by a factor of at least $e^3 \approx 20$.

3. The third set (model 3) says that not only is α distributed as in model 2, but also we believe that the standard deviation of log dose with drug B is greater than that for drug A by a factor β that has probability 0.8 of being more than 1.0 and could be as high as 10 with probability 0.1. (For the technically interested we are here assuming that $1/\sigma^2$ has mean 40000 nepers⁻² and is Gamma distributed with shape 0.5.)

Under all three models we will assume that the Bayesian algorithm is set to continue making measurements until either the posterior probability of the device being approvable or the posterior probability of the device not being approvable has reached 0.95.

Under model 1 *no measurements are required* as dosing accuracy of a solution of drug A has already been demonstrated to a much greater accuracy than is needed for drug B. But the type 1 error rate for this trial of zero measurements is clearly 1.0, much higher than the usually desired 0.05 – i.e. if in fact drug B is not adequately accurately dosed by the device, such a trial will definitely not detect it.

Under model 2, one is almost certain to need exactly one measurement to show whether the dose of drug B delivered is safe and effective or not. The type 1 error rate, however, is then about 0.93 (far greater than the desired 0.05), and it comes from the extremely unlikely possibility that the standard deviation is actually 50 times larger than expected – but the type 1 error rate takes no account of the unlikely nature of the parameter values that cause it.

Under model 3, we are likely to need around six measurements on average to show whether the dose of drug B delivered is safe and effective or not – but the type 1 error rate is still 0.53, much higher than wanted by the FDA. Figure 1 and Figure 2 illustrate operation under the conditions specified by model 3. Figure 1 shows how a device is approved or not by a Bayesian designed trial, and Figure 2 shows how the number of measurements needed depends on how close to the borderline of acceptability the device is.

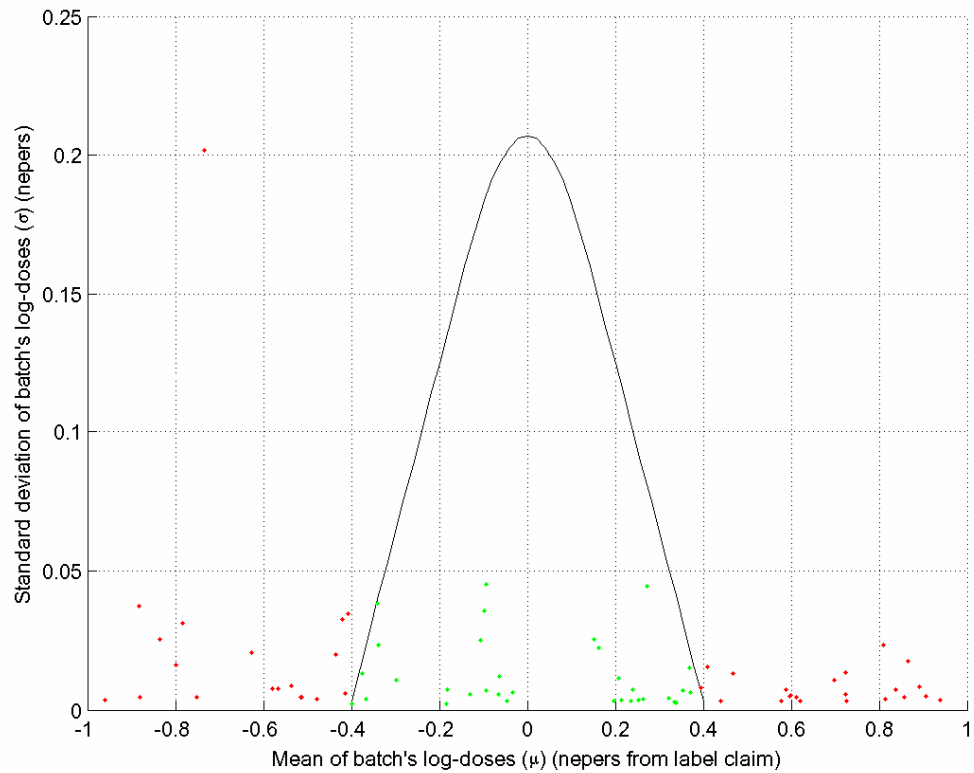


Figure 1: A Bayesian inference algorithm operating on devices drawn from the alleged prior of model 3. The black curve indicates the region of acceptability of delivered dose. The green dots indicate devices passed by the algorithm, and the red dots indicate devices failed by the algorithm.

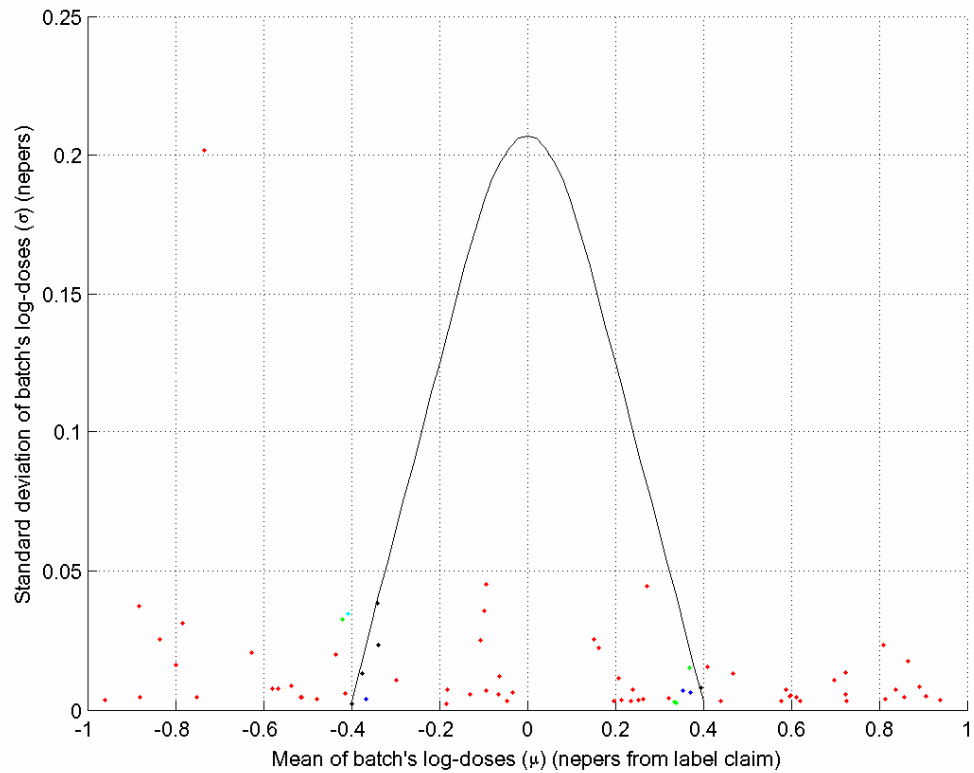


Figure 2: The number of measurements needed for a Bayesian inference algorithm in the regime of model 3 to make a decision on whether the delivered dose is acceptable or not. Red indicates just one measurement was made, green two, blue three, cyan four, magenta five, yellow six, and black seven or more. It is clear that the number of measurements required (average 5.6) depends on how near the delivered dose is to the boundary of acceptability – just as it should do.

On the other hand what happens when the type 1 error rate is being measured is shown in Figure 3; very unusual devices only are tested (in that they are all improbably located just on the bad side of the boundary of acceptability), and the Bayesian inference is “confused” by the fact that it is rating such devices as very unlikely to be seen – as it has been told to by the given prior. Whatever one may think is the appropriate prior to describe devices likely to be seen, it certainly isn’t the distribution evidenced by the points of Figure 3.

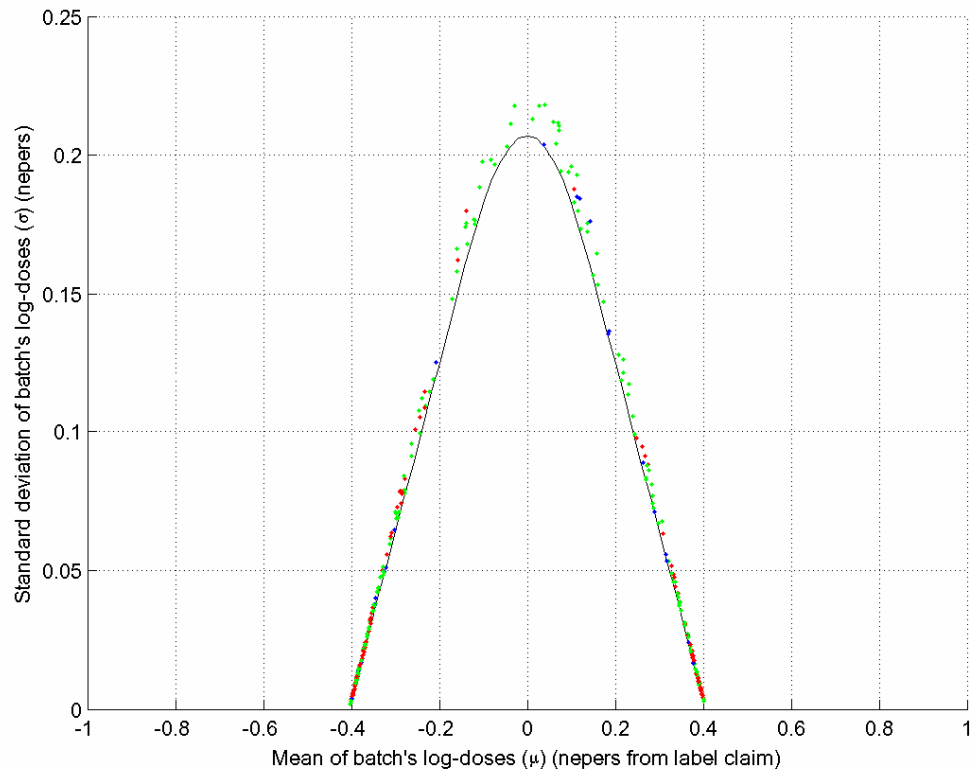


Figure 3: The same Bayesian inference algorithm as in Figure 1 being tested to measure its type 1 error rate. Note that the devices being tested are very unlikely under the alleged prior distribution, and all close to the borderline of acceptability and on the bad side of it. Many are passed (green), giving a high type 1 error rate of 0.53 – but this fails to take into account that seeing such devices is very unlikely in the first place.

What do we need to do in the situation modelled above to make the type 1 error rate below 0.05? There are a number of options listed in the Guidance, but essentially we need to either commit ourselves to making more measurements than are necessary, or discount or ignore the information we already have, or permit devices to fail that really should pass, any of which, of course, leads to our trial costing more – or in the last case even failing our device when it should have passed.

An example of forcing the type 1 error rate down by discounting the information we already have is the following fourth model:

4. Model 4 says that the standard deviation of log-dose is not β times bigger than with model 3, but 50β times larger. This is saying “We are 80% sure than when we put drug B in this device the device will suddenly become 50 to 500 times less accurate” – which might be reasonable if drug B were detergent or other surfactant, but usually will not be.

Under devices still drawn from model 3, but with the prior of model 4 to hold the type 1 error rate down, we are likely to need on average around 12 measurements per good device passed (ignoring measurements on failing devices) to show whether the dose of drug B delivered is safe and effective or not, and the type 1 error rate is now 0.08 (still not quite down to 0.05). Figure 4 shows the pass/fail performance under these conditions.

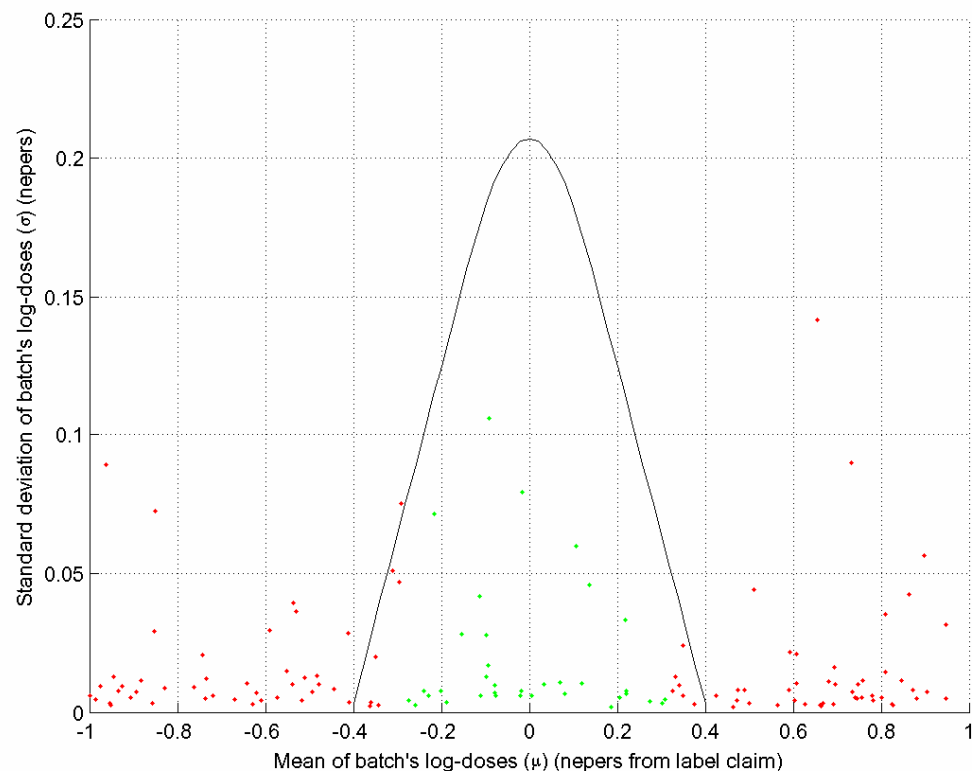


Figure 4: Pattern of pass/fails using Bayes testing devices from the prior of model 3, but using the prior of model 4 for inference.

As can be seen, in addition to the larger number of measurements needed, we are now also failing about 25% of the good devices.

The failures can be overcome by reducing the acceptable probability of failing a good device in the inference design, but this then raises the number of measurements needed further, from an average of 5.6 under the prior of model 3 to an average of more than 24 for each good device passed (ignoring the tests on devices that fail), and also pushes the type 1 error rate back up slightly to 0.09.

We hope it is now clear that the truly least burdensome approach is to use Bayes properly, without attempting to control type 1 error rate.

In our view a desirable approach for the FDA to take to Bayesian trials consists of the following steps. We base these around the example of a disposable drug delivery device delivering only a single dose of a drug.

1) The FDA specifies criteria of acceptability for the probability distribution of the dose

First, uniformly for all similar devices for different sponsors for the same drug and route, the FDA specifies acceptable limits on the centiles of the probability distribution of the dose. For example, for a drug delivery device that delivers just one dose of a drug, the acceptability specification might be that the probability of the dose being outside the range $\left[\frac{10}{11}, \frac{11}{10}\right]$ times the label claim be less than 0.05, and the probability of the dose being outside the range $\left[\frac{10}{12}, \frac{12}{10}\right]$ times the label claim be less than 0.01, and that trials for approval must show that the posterior probability of both these restrictions being met is at least 0.95.

2) Sponsor defines structure of a suggested statistical model

Next, the sponsor defines the structure of a statistical model for the device to be tested. For example, for a drug delivery device delivering only a single dose per device, the parameters might be the those of a log-Student distribution for the dose x , namely the median dose μ delivered, the scale s of the distribution $\left(E \frac{1}{\left(\log \frac{x}{\mu}\right)^2}\right)$, and the shape

m of the distribution. The sponsor delivers this to the FDA as a variable definition document, which optionally may also include suggestions for the priors on these parameters. Here the log-Student distribution is

$$P(x | \mu, s, m) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(m + \frac{1}{2}\right)}{\Gamma(m)} \frac{\left(\frac{m}{s}\right)^m}{x \left(\frac{m}{s} + \frac{1}{2} \left(\log \frac{x}{\mu}\right)^2\right)^{m + \frac{1}{2}}}$$

3) The FDA specifies priors on the parameters

The FDA then states a proper joint prior on the parameters (in the example μ, r, m), or gives reasons why they think that the proposed model structure is inappropriate (for example the sponsor might have made an unjustified assumption of normality).

It is expected that such a prior would either be “broadly agnostic” or based on previous data. Thus it would be expected that:

- Narrow peaks on values that lead to the dose distribution being just unacceptable would have to be justified by previous history of such devices having that property (i.e. we do not expect to be given the distribution of the points in Figure 3);
- Different sponsors submitting similar devices would be given the same broadly agnostic prior after any appropriate shifting and scaling unless there was evidence suggesting that one sponsor has a much better past history with similar devices than another.

Discussion or negotiation might then take place if required because of the sponsor objecting to the FDA’s choice of priors. In general, however, this will not be in the sponsor’s interest unless the FDA has contravened one of the bulleted points above.

Once agreement has been reached at this point, it is then binding on both parties for the later analysis of the trial unless both parties mutually agree to change it. If the manufacturer later needs to add further variables, the FDA retains the right to set the prior distribution on those further variables conditional on those already agreed under similar restrictions to the bulleted points above.

4) The sponsor then designs and executes a trial

The sponsor then designs a trial to collect data according to a protocol. This protocol may, for example, permit reanalysis of the data after every single collected data point, or permit analysis of subgroups (though in the latter case the agreement with the FDA would need to include prior information on the relations between subgroups). The sponsor is free to conduct simulations of the trial before actually executing it (and would be well advised to do so, in order to evaluate for example the likely cost of the trial). There should be neither obligation nor prohibition to calculate type 1 error rate or the prior probability of the device being found to be acceptable.

5) Analysis of the trial

Analysis of the data collected is then done using standard Bayesian inference and if necessary MCMC techniques as already discussed in the Guidance, using the prior already agreed with the FDA. The posterior probability that the dose meets all the acceptability criteria is calculated, and if it is above the specified lower limit (in the example 0.95) the device is approved. In particular, the calculation of type 1 error rate plays no part in the approval decision.

Benefits

The FDA is charged with safeguarding the interests of the public, while burdening the industry as little as possible consistent with the safeguarding duty. The above

suggestion provides the FDA with the opportunity to state a consistent and suitably sceptical prior from which consistent inference may be made. The requirement to reach agreement on the prior before data is collected gives the sponsor confidence that his trial, appropriately conducted, will result in approval, if indeed his trial shows the posterior probability that the device is acceptable is sufficiently high. Since it is possible to design Bayesian trials that make the probability of not accepting a device which is in truth acceptable arbitrarily low, and expend money on continuing the trial only when it is necessary to do so, this minimises the burden on the sponsor as is required by law.

6 Summary

The guidance is well prepared and appears to mark a significant step forward towards using the power of Bayesian techniques to facilitate medical device trials. There is one outstanding important unresolved issue. The complexity and cost of using these techniques is not insignificant, but could be easily outweighed by the gains in terms of smoother and more cost effective clinical trials.

7 References

1. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, 5th February 2010, <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm>
2. Charlish P., “Bayesian analysis and medical device regulation”. The Regulatory Affairs Journal: Devices 14(2) 23rd May 2006.