Comment on Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2006D-0191)

Michael Brimacombe, Ph.D.
Associate Professor,
Department of Preventive Medicine and Community Health,
New Jersey Medical School – UMDNJ & School of Public Health - UMDNJ

Overview:

The use of Bayesian methods is a useful addition to the acceptable set of statistical approaches that can be applied in the setting of clinical trials. The current document is an introductory level overview of Bayesian methods, written from a very "Bayesian as a separate methodology" perspective. In terms of providing guidance for actual application of Bayesian methods in clinical trials, it is seriously lacking in focus, details and balanced perspective. I disagree with making smaller sample size the selling point of any approach to clinical trials or the potential of using the sequential "updating" aspect of Bayes theorem (long known in other fields) to justify "peeking" at the data. Smaller sample sizes may one day be viewed as a serious weakness if side effects in long-term follow-up occur and initial small sample sizes and thus their Bayesian justification become a target of criticism.

Bayesian methods (in my view, and there are many approaches to Bayesian statistics) are an extension of likelihood methods. Given a likelihood function for the set of parameters in a real-world problem we can (1) take aspects of the likelihood function itself (Mode or maximum likelihood estimate, derivative or "score function", difference on a log scale between two parameter values or "likelihood ratio") as pivotal quantities,

find their frequentist sampling distributions and obtain p-values and confidence intervals OR (2) calibrate the likelihood function into a probability density (posterior distribution) on the parameter space via the use of prior distributions and then average out the parameters we don't need, using summaries of the resulting marginal posteriors on some scale (odds ratios, central intervals about the mode etc..). Bayesian methods correspond to (2).

In large samples the central limit theorem dominates either approach and similar answers are typically obtained (the M.L.E. and posterior mode are often very close). With a caveat that the integration (which is really summation) will induce a central limit theorem like effect or "shrinkage" of the variation of the remaining marginal posterior of interest, thus Bayes methods tend to give narrower "confidence" intervals. The number of parameters in the model can greatly affect this shrinkage effect and there is always opportunity for "over-fitting" and increasing the potential for differences between cases and control groups, simply by over-parameterizing the underlying model, a subtle difficulty in the application and interpretation of Bayesian approaches.

Many of the difficulties existing in the application of Bayesian methods are subtle and difficult. The current document needs to be augmented to help statisticians employing this approach, some of whom may have only M.Sc. level training in smaller companies, to deal with problems they will almost certainly encounter. I list some of these difficulties below.

Specific Comments:

1. The choice of prior is non-trivial as it is used to calibrate the likelihood function and related information in the model-data combination. In this document, there should be a table, giving a set of typical standard priors for typical contexts arising in clinical trials (for eg. Bioequivalence). If you were developing a new bioassay, would you not give careful advice on proper calibration? Why should a new statistical approach be exempt?
2. A section of "similarities" should be added to the differences listed. Frequentist methods will often agree with Bayesian approaches and it may be very informative to know when they do or do not.
3. It should be noted that "prior" may not always reflect existing belief, rather it may for example, reflect a desire to be non-informative where the likelihood is most informative (Jeffrey's prior). There should also be some mention regarding typical improper priors as well as the selection of priors when there are many parameters, some of which are not expected to be independent of each other.
4. The meaning of evidence alters when Bayesian approaches are employed. In a frequentist setting the standard proof by contradiction scientific approach is used: assume the null hypothesis true and see if the observed sample contradicts this assumption. Bayes methods do not really do this; they provide more detailed measures of acceptability by placing a probability distribution directly on all the points in the parameter space. It is a question of taste how to interpret the weight given to the null value. Bayes factors, posterior odds ratios or just a plot of the marginal posterior with tail areas or credible interval are all possibilities (with

their supporters and detractors). Which approach to use should be standardized by the FDA.

5. As mentioned above, over-fitting and shrinkage effects can be used to obtain narrower credible intervals in many settings. This obviously can be abused. Guidance should be given, in the context of statistical models typically used in standard clinical trials, so that initial analysis of the data be conducted in the simplest model possible (Occam's Razor). The goal being to minimize over-fitting.
6. Clinical trials at times involve toxicity and other applications involving nonlinear models. Here it is rare that "prior" means prior belief. Typically issues of reparameterization must be examined, with priors used to simply obtain a useful posterior. The set of priors to be used is often quite limited. It is important to note that not all prior – likelihood combinations will yield a useful posterior.
7. Hierarchical models (nested structures in the design of the study, for eg. in multi-site trials) are a challenge. Bayesian methods can be viewed as naturally hierarchical and therefore useful in this setting. But not all prior – likelihood combinations yield useful posteriors. Having statisticians fish about for a "mathematically useful" prior in these settings verges on being scientifically unsound (The chosen prior may be the only prior yielding significant results). What is guiding the calibration of the likelihood in these settings? Feasibility? That is not an argument for serious application in real-world clinical trials. There is no serious discussion of this. Again, standardization of prior selection (which I

am viewing as likelihood calibration) is essential for comparison and integration of studies conducted in different labs or repeated over extended time periods.

8. Frequentist analysis should be requested where possible, as a reference point to interpret Bayesian results. As noted above, there is much overlap due to the common use of likelihood. Sample sizes should be compared, and the typically smaller resulting Bayesian sample sizes justified and discussed in light of shrinkage effects. Sample size calculations should again reflect, on a standard basis, narrowly defined models (See 5 above). If Bayesian sample sizes are for example 20% less than frequentist calculations, it may indicate the model is unstable.
9. Non-linearity affects the accuracy of MCMC methods. The geometry of the likelihood function affects the convergence rates of these methods and they often require initial runs (burn-in) to be useful. A formal reporting mechanism should be set up regarding the reporting of the computing approach taken in developing the Bayesian analysis. This should not be left open to the individual lab or statistician.
10. As some statisticians have pointed out, Bayesian methods, as they involve many assumptions and nuances, are rather hopeful in their application to the real world. They should be applied primarily in settings where the science is mature and the calibration of the likelihood via subjective priors agreed upon and standardized to some point across labs. In younger or new areas of application Bayesian methods may not be suitable as there is insufficient information regarding priors. If this is

the case, improper priors (making the likelihood essentially the posterior) may be the most appropriate. Guidance should be given.

Summary

To conclude, there are useful extensions in Bayesian methods for clinical trials, or any area of science, (for eg. the predictive density and measures of evidence). There is also a substantial practical overlap with standard likelihood based methods and I would emphasize the overlap rather than the differences with frequentist likelihood methods.

The current document, written at an introductory level, is insufficient as a guidebook for the application of Bayesian methods and their interpretation. More detail should be given to limitations and challenges in their application (modern Bayesian methods and the computing techniques motivating them are still very new). Acceptable calibrations (priors) should be listed, at a minimum, for standard clinical trial settings.

The guide as currently written sells the approach (do you still need to do this?), but leaves the details (and problems) to the user. This is not acceptable. Given the need for comparability in clinical trial settings, it seems greater standardization is necessary. A new statistical method, as part of the scientific process of evaluation, should be required to better explain its strengths and weaknesses and give guidance for standardized application.