

Floor discussion

Donald Berry, Steven N Goodman, Thomas A Louis, Robert Temple

Dr Wakim: Is there a book on introduction to Bayesian methods, specifically for clinical trials?

Dr Berry: Yes. One that has recently been published and is really quite nice, is by Spiegelhalter, Abrams and Myles, on clinical research and Bayesian statistics. If you want something that is really elementary, – you know, no integral signs, motivation from the beginning – there is a wonderful little book, written by somebody named Berry – that is titled, “Statistics: a Bayesian perspective.” It was written for university freshmen for “Stat 101” but is very accessible to MDs.

Dr Wakim: My second question is for Dr Goodman, about confidence intervals. You talked about the problems with interpretation of P -values, but doesn't use of confidence intervals solve that problem?

Dr Goodman: That is an excellent question. Another way to frame my example of the small trial and a big trial, would be to note that the confidence intervals were very different in the two settings and that those differences would predict that the evidence was different. But the confidence interval doesn't answer the evidential question. It is very difficult to know how to combine the size of the effect and the precision of the effect into an evidential measure that one can use in an inferential framework. Non-Bayesians often stress the value of looking at confidence intervals, and they are correct, but it is hard to know how to use that information in a formal inferential way. Confidence intervals do get us away from P -values, but the fact that most people just look to see whether the null hypothesis is included in the confidence interval shows us that we cannot stop and declare them the solution.

Dr Berry: Can I just add to that? Confidence intervals, being study-specific, are also not useful for making decisions. If there is other information out there, it is not incorporated in the confidence interval.

Dr Louis: From my point of view it is far better to compute the posterior probability that A is better than B or, if you like, posterior probability that A is better by some threshold value. It is hard enough to

psychologically process what is going on when you are have symmetric confidence intervals, let alone when you have skewed confidence intervals. So in my view, probability-based computations are better to work with.

Dr Temple: Let me just press this a little bit. One of the things you said, Steve (and we have all learned that this is true) is that just because the P -value is less than 0.05, it doesn't mean you can say anything about the probability that there is a true treatment effect.

Dr Goodman: I want to amend that. It doesn't mean that you can't say anything, but that you don't know exactly what to say.

Dr Temple: Okay, but what I want to ask about is the common case in drug development where you don't have too much of a validated view about how likely you are to succeed. We are not usually willing to concede much in the way of prior probabilities for a new class of drugs. So, you then run the study and you show the confidence intervals for the various possible outcomes. How close is this to a Bayesian analysis when there isn't a strong prior? We all recognize that sometimes you do want to incorporate prior information (actually, sometimes I would like to incorporate negative information, like when 10 sepsis trials in a row have failed to show any benefit I am not sure whether the one that does ought to be considered a true positive – I might have a very low prior in that circumstance). But in a lot of cases I don't have substantial prior information. Would it be wrong and simple-minded to think that the two approaches get a lot closer together when there isn't an informative prior?

Dr Goodman: It is true that when you have no background knowledge the answers in the two situations start to get closer. But the reason I emphasized the strength of evidence issue is that measuring the strength of evidence properly doesn't have to depend on your willingness to set a prior probability. It is very important to get the proper strength of evidence measure. A P -value of 0.05 or 0.01 does not appropriately represent the strength of evidence against the null. But it is true that, properly done, the answers from the two methods

when there is little or no prior information will come much closer than they might when you have a lot of prior evidence. Still, the weight of the evidence and how you quantify it will be fairly different.

Dr Siegel: My question is for Dr Berry and it has to do with your discussion about stopping trials early when the results look really good. *The New England Journal of Medicine* published a review of data safety monitoring boards recently in which they gave two examples of interim analyses where the results did look very good. The *P*-value was well below 0.05, but the prespecified boundaries hadn't been met. So, the data safety monitoring board decided to continue the trial even though the results really looked terrific. When they continued the trial, in both cases it turned out there was no difference between drug and control. So, my question for you is how can we be really sure we wouldn't prematurely reach conclusions in a few cases that a drug is effective when it really isn't?

Dr Berry: This is related to Steve's argument about the evidentiary impact of *P*-values. A *P*-value of less than 0.05 is not as strong as one would like it to be and frequently turns around. I suppose what I am going to say is associated with some *P*-value, that is, if the evidence is sufficiently strong to stop the trial in terms of posterior probabilities or predictive probabilities, or whatever, that would correspond to a *P*-value that is rather low.

Dr Woodworth: As a Bayesian, I am concerned about having to use Type I error as a design criterion in designing a Bayesian study. The Type I error, it seems to me, is roughly equivalent to doing a Bayesian pre-posterior analysis with no mass below the null hypothesis, for example, a non-inferiority trial with no possibility of more than negligible inferiority.

I would like to ask whether you see the possibility of the FDA accepting a different form of justification of the operating characteristics of the trial. For example, to use full-blown pre-posterior analysis based on prior and loss function but to demonstrate its insensitivity or robustness to a variety of priors and losses.

Dr Berry: I will let the FDA answer the FDA question but I agree with you about the noninferiority issue. In the second case study I plan to give an alternative to the standard approach. With respect to Type I error, it is so ingrained in the regulatory process and in medical circles generally, I don't see it going away as a criterion. Bob [Temple], do you?

Dr Temple: Anything is possible, but first we have to understand in a tangible way the alternative ways of expressing things. For a start, I would like to see

both analyses submitted together so we can understand better what the properties are. I don't want to speak for everybody at FDA, but there is massive inexperience with Bayesian approaches to evaluating trials of investigational drugs and we are nervous about priors. We have seen many things that were "true" turn out not to be true. So it seems difficult to judge how much credit to give for prior evidence. I gave a long list of cases where it might be easy but the harder cases are when you don't have many studies or much experience. How much weight do you give, for example, to pharmacologic properties? Some believe that if you understand (or think you understand) the pharmacology, the burden on subsequent evidence ought to be much less. That doesn't seem totally unreasonable, yet I have many examples of things that everybody knew were going to be true that turned out not to be true.

Dr Berry: A problem is that it is limiting to have to do both because there are many Bayesian procedures that don't really permit you to calculate a Type I error. To bend in your direction we have to limit the kinds of things we do to those where we can actually simulate the trials and calculate Type I errors.

Dr Temple: What I had in mind is designing the usual frequentist trial, and then simultaneously look at it as if you were doing it in a Bayesian way. The results are the results. Show us the difference in analyses, with and without priors, because most of us really don't understand these analyses yet and we are therefore nervous about using them to make important decisions.

Dr Goodman: In my first case example I am going to show both, the traditional Type I error and then a formal pre-posterior analysis that uses the prior. At this stage it is very important that both sides speak the same language to the extent possible, and type-1 error is part of that language. If we don't find common ground, where we can recognize when we are talking about the same thing, we will be talking by each other rather than to each other. The Bayesians are going to be talking about likelihoods, Bayes factors or posterior probabilities, and the users of the traditional methods are going to be talking about error rates and confidence intervals. So the reason to cite Type I error rates right now is so everyone can be clear that we are working towards the same goal, trying to get things right. I don't have any problem putting a traditional Type I error rate along side the Bayesian measures and then describing in the body of the protocol exactly how to look at both types of numbers.

Dr Thompson: My question is for Don Berry. In your presentation, in the five or so examples you gave, you referred to "the predictive probability of

getting statistical significance at the end of the trial". I wanted to know what you mean by statistical significance in the Bayesian context.

Dr Berry: Yes, it sounds strange coming from a Bayesian, doesn't it? It is calculating the probability of " $P < 0.05$ " at the end of the trial. So, it is a Bayesian calculation of a frequentist measure.

Dr Thompson: Why are you interested in that?

Dr Berry: Because I am pragmatic.

Dr Thompson: Because of submission to the FDA? Is that it?

Dr Berry: Not necessarily, but the eventual FDA assessment of the trial could be the driver. Consider two cases. In the first, a trial is designed from a frequentist perspective. During the course of the trial I may want to know the probability that the trial will be a success, based on the current results. That is inherently a Bayesian notion applied to a frequentist set-up. Perhaps the FDA will approve only if the trial results are statistically significant. In the second case the trial is designed to be Bayesian but the FDA has stated in advance that they will use a noninformative prior distribution and require a 5% posterior probability of superiority for approval. This is essentially the same as requiring a (one-sided) P -value of 5%. The company's prior distribution may be informative, and the predictive probabilities may be calculated accordingly. But it makes sense from a decision perspective to calculate predictive probabilities of someone else's probabilities.

Ms. Collina: I have two questions and maybe if they can't get answered right now, these issues will be touched on later. One is to Don [Berry] and the other is to Bob [Temple], with FDA.

Some wonderful and compelling arguments were made about the Bayesian approach, but we haven't really heard very specifically what we lose. I know it is hard for you, Don, but what do we lose? We have heard the benefits, and they are compelling, but I am still not clear based on this presentation, to what extent we are losing something that is valuable in evaluating safety, in really answering a question. Quicker is not always better.

The second question really goes to the public policy question, which is, why straddle? We've heard a rationale and some compelling arguments for the Bayesian approach, and trials using this approach are being done – why can't the FDA formalize its use? The FDA response is, we have just started to look at it and we are nervous. Is it scientific nervousness? Is it inertia around, "this is how we have always done it?" What is the hesitation?

Dr Berry: I will let somebody else answer the second one. I can answer the first one. This is

something we worry about. For example, when I did that seamless Phase 2/3 trial I showed the null hypothesis and the alternative hypotheses and you might expect that you would gain something in one of those and lose something in the other but what we showed is that in fact you gain in both. Generally we lose very little. We always evaluate what we lose. In case study 2, for example, when modeling the relationship between the 12-month and the 24-month results you may say, well, just introducing the model has the potential for loss because suppose the model is not right; suppose there is no relationship at all. Well, we evaluated that. In that example we lost essentially nothing. But it is something that always concerns us.

Dr Temple: You know, this was supposed to be a nontechnical session; Don went through all of his examples but I can't tell whether I believe them. We all need to see the details, including people who can understand the mathematics well enough to know whether they buy that we can use these approaches without making more mistakes in drug approvals.

There seems to be an underlying assumption that evaluating new drugs is going to somehow get easier if you use these methods; I don't think that is true at all. In a number of areas of drug developments my priors would be extremely negative because there have been many failures of a particular thing in the past.

Let me give you an example. We must have had eight different drugs studied to treat septic shock and all of them failed until one, done by Lilly, finally showed benefit. The trial showed a statistically significant benefit with a conventional P -value analysis that was reasonably robust. But what would my prior be when there are eight straight failures? What should the strength of evidence be to overcome such a negative prior? It probably wouldn't be less than we ask for now; it might be way more. This type of implication needs to be explored.

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.

My point is that all of these things need to be well enough understood so we can actually implement procedures that won't drive everybody

crazy, that won't involve constant arguments about the strength or nature of each assumption every time. This is a very hard thing to do and does not facilitate development because there are too many debates – you have one person who is a little more conservative; another person who is a little more aggressive and you don't necessarily come easily to consensus. I think we need to understand the properties of the approaches we take to evaluation of study results, what the risks are in both directions of being too stringent, over-stringent perhaps, or being too lenient. Until we all understand these approaches thoroughly, which will require looking at a lot of examples, it is going to be hard to adopt them just because Don can show some examples of where he thinks they produced sensible conclusions.

Dr Louis: Just a quick comment about Bob's first point: there is absolutely nothing in this morning's presentations that implied that a Bayesian approach is always faster and/or cheaper. I will say it is always better but sometimes better really means longer. In fact, I think the device group has several examples where the emerging data were very different from the priors; in such cases it actually takes a bigger trial to overcome the prior or, if you would like, convince the community. It is a big mistake for people to enter into this, let's say in the FDA or other places, saying Bayesian approaches will always make things faster or less expensive I think, properly done, they will always make things better because they do bring in all the appropriate uncertainties but, as in Don's example, occasionally they push for a bigger trial than you would expect.

Dr Goodman: I am going to echo what Bob said: we have to look at why the FDA developed the procedures it developed in reaction to what. It was very, very important to have clear procedures that everybody could agree on. Now, the problem was, and is, that having clear and consistent procedures

often produces a certain automaticity that sometimes doesn't make sense. We have to figure out whether there are methodologies that can reduce that automaticity, that facilitate the incorporation of expert opinion in a way that is meaningful, so we don't make nonsensical decisions.

But that is very, very difficult in a public policy position where you have to be both transparent and to have methods that are not *ad hoc*. What you lose by adopting a new, improved approach is some of the automaticity; what you gain is some reasonableness. But you can't give away the store. Part of this public discussion will be how we back away from over-mechanized inference, while not allowing the same forces that forced FDA to produce the clear procedures they have now to take advantage of new-found flexibility to produce incorrect decisions. We need to figure out the middle ground; where we can and should bend the procedures; and what the procedures are that we collectively – companies, academics, regulators and policy-makers – can all be comfortable with.

Participants

Donald Berry: Frank T. McGraw Chair for Cancer Research, Professor and Chair, Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center

Steven N Goodman: Associate Professor, Departments of Oncology, Pediatrics, Epidemiology and Biostatistics, Johns Hopkins School of Medicine and Public Health

Thomas A Louis: Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Robert Temple: Director, Office of Medical Policy, Center for Drug Evaluation and Research, FDA