Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 561–581

https://doi.org/10.29220/CSAM.2017.24.6.561
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Bayesian methods in clinical trials with applications to medical devices

Gregory Campbell[1,a]

[a]GCStat Consulting LLC, USA

### Abstract

Bayesian statistics can play a key role in the design and analysis of clinical trials and this has been demonstrated for medical device trials. By 1995 Bayesian statistics had been well developed and the revolution in computing powers and Markov chain Monte Carlo development made calculation of posterior distributions within computational reach. The Food and Drug Administration (FDA) initiative of Bayesian statistics in medical device clinical trials, which began almost 20 years ago, is reviewed in detail along with some of the key decisions that were made along the way. Both Bayesian hierarchical modeling using data from previous studies and Bayesian adaptive designs, usually with a non-informative prior, are discussed. The leveraging of prior study data has been accomplished through Bayesian hierarchical modeling. An enormous advantage of Bayesian adaptive designs is achieved when it is accompanied by modeling of the primary endpoint to produce the predictive posterior distribution. Simulations are crucial to providing the operating characteristics of the Bayesian design, especially for a complex adaptive design. The 2010 FDA Bayesian guidance for medical device trials addressed both approaches as well as exchangeability, Type I error, and sample size. Treatment response adaptive randomization using the famous extracorporeal membrane oxygenation example is discussed. An interesting real example of a Bayesian analysis using a failed trial with an interesting subgroup as prior information is presented. The implications of the likelihood principle are considered. A recent exciting area using Bayesian hierarchical modeling has been the pediatric extrapolation using adult data in clinical trials. Historical control information from previous trials is an underused area that lends itself easily to Bayesian methods. The future including recent trends, decision theoretic trials, Bayesian benefit-risk, virtual patients, and the appalling lack of penetration of Bayesian clinical trials in the medical literature are discussed.

Keywords: prior, hierarchical modeling, adaptive designs, simulation, exchangeability, treatment response adaptive randomization, subgroup analysis, predictive posterior distribution

## 1. Introduction

The scope of the paper is not a complete review of all the work that has been done on Bayesian methods in clinical trials. There are two publications about a decade ago that help to set that stage, with extensive reviews of Bayesian methods, one in biostatistics by (Ashby, 2006) and the other in biopharmaceuticals by (Grieve, 2007). In particular, with little personal experience on toxicity monitoring (Phase I drug trials), dose finding (Phase II drug trials or on pharmacokinetic or pharmacodynamic studies for drugs), or postmarket safety, any review of these topics will not be attempted here. Instead, a review of the methods and applications of Bayesian methods for clinical trials is provided, with an emphasis on medical device applications.

[1] GCStat Consulting LLC, 14605 Sandy Ridge Road, Silver Spring, MD 20905, USA. E-mail: GCStat@verizon.net

Published 30 November 2017 / journal homepage: http://csam.or.kr

### 1.1. Bayes Theorem

Bayes Theorem was posthumously published by an English minister (Bayes, 1763). For sets $A$ and $B$, it often written as:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}.$$

Reverend Thomas Bayes was interested in conditional updating, of how to incorporate information into the updating of a probability.

Consider a simple diagnostic example of a two-by-two table with entries $a, b, c, d$ with the rows as Disease $(a, b)$ and Non-disease $(c, d)$ and the columns as Test Positive $(a, c)$ and Test Negative $(b, d)$. Then sensitivity (SENS) is $a/(a + b)$ and specificity (SPEC) is $d/(c + d)$. If the table is generated by natural sampling (as opposed to stratified sampling), then the positive predictive value (PPV) $= a/(a + c)$ and negative predictive value (NPV) is $d/(b + d)$. If the sampling is stratified so that for example the number of diseased and non-diseased are fixed in advance but the prevalence $p$ is known or can be estimated separately, then PPV $= p\,\mathrm{SENS}/(p\,\mathrm{SENS} + (1 - p)(1 - \mathrm{SPEC}))$ and NPV $= (1 - p)\mathrm{SPEC}/((1 - p)\mathrm{SPEC} + p(1 - \mathrm{SENS}))$.

## 2. A brief history of Bayesian statistics in the twentieth century

There is nothing controversial about Bayes Theorem itself or its conditional probability application. The controversy erupts when the following is done. Let $A$ denote data generated in the current experiment (trial) and $B$ denote the parameter of interest that you want to investigate. Then $P(B)$ denotes the prior probability of $B$ and Bayes Theorem provides a way to update the prior using the likelihood function to calculate the distribution of the data given the quantity or parameter of interest. The numerator is then the product of the prior and the likelihood function. The denominator is ignored since it is just a constant after the integration and we are left with the proportionality statement below. Then the posterior distribution $q$ of the parameter $\theta$ given data $D$ is proportional to:

$$q(\theta|D) \propto L(\theta|D)\pi(\theta),$$

where $L(\theta|D)$ is the likelihood function and $\pi(\theta)$ is the prior distribution and $\propto$ is the proportionality symbol to indicate that the denominator, which is the integration over $\theta$ of the numerator, can be ignored since it does not depend on $\theta$.

The controversy in some people's minds occurs because now the unknown state of nature $\theta$ is now viewed as a random quantity and technically not a fixed number. This is a subjective way to view reality; moreover, since in most cases two knowledgeable people could have very different priors for the unknown state of nature and hence likely different posteriors and then arrive at possibly different decisions. This can create a difficulty in science reporting since the data is expected to speak for itself, so the tradition has been to use the likelihood only and not allow any subjectivity into the scientific process. However, the more global way in which science works is that each study is integrated into the knowledge base that already exists. The result is that science does not usually turn on a dime based on a single publication but rather integrates that knowledge into what is already known. The reporting of a study only in terms of its likelihood is that it is not in quantitative relation to all the other work that had preceded it. The big advantage with the Bayesian approach is that one could produce posterior distributions that addresses the quantitative integration of the current data with the prior information. This enables one to calculate, for example, the probability that the alternative hypothesis was true of the null hypothesis or that the parameter $\theta$ of interest is greater than 0 or 5 or whatever.

Since there is uncertainty in the parameters as they are viewed as random variables, now probability statements can be made about where a particular value is captured in a (credible) interval. The explanation of a frequentist confidence interval is much more convoluted and in fact most people (not statisticians) interpret confidence interval as if they were credible intervals.

A longstanding problem for the Bayesians had been how to calculate the posterior for a given prior and data represented by the likelihood. This is relatively simple if for a binomial problem with a beta distribution as a prior, since the update is again a beta distribution and it is just a matter of updating the parameters of the prior. So the beta distribution is called a conjugate prior for the binomial and correspondingly the Dirichlet distribution for the multinomial. For the normal data, the conjugate prior is normal and so is the posterior. However, if the prior is not so well behaved so there is no known conjugate prior, the posterior distribution cannot generally be calculated. So arguments that ensued in the statistical community were either philosophical or theoretical since little real data analysis could be done for many real-world problems. I suspect that it was the subjective nature of the Bayesian approach that generated the most heat in the debates that ensued between Bayesians and the non-Bayesian frequentists. The philosophical issue was also whether statisticians were willing to agree to placing distributions on parameters that were unknown to them as opposed to thinking of them as fixed constants.

One approach to the inability to calculate posteriors for many applied problems was to elicit subjective priors from established experts in the relevant field of expertise and then force that information into the form of a conjugate prior. This elicitation effort requires skill by the facilitator and some understanding of statistics by the experts, as the information for each was converted to one or more parameters of a probability distribution and then averaged over all the collected experts to obtain as a prior, a probability distribution from the experts. Chaloner *et al.* (1993) presents a graphical approach for prior elicitation. The generated prior may not correspond to the prior for any of the participating experts and may not be generalizable to other panels of experts. Prior elicitation is both time- and resource-intensive.

My exposure to Bayesian ideas began in graduate school. A course in statistical decision theory with loss functions and utilities emphasized that Bayes rules are admissible. The book by Box and Tiao (1973) had recently been published with a focus of on Bayesian inference based on estimation rather than hypothesis testing. Ferguson (1973) had introduced the concept of the Dirichlet process prior that allowed priors to be placed on the space of distributions that allows for a mean any cumulative distribution (see MacEachern (2016) for an introduction) and then my doctoral thesis with Prof. Myles Hollander as an advisor considered rank order estimation using Dirichlet process priors (Campbell and Hollander, 1978). As a faculty member in the Department of Statistics at Purdue University, I then applied this to a famous rank-order optimal stopping problem called at the time the secretary problem, a hiring problem with a fixed number of candidates but at the end of each successive interview the decision has to be made by the hiring official to hire the individual or not. Using prior information, the optimal Bayesian strategy to hire the best candidate performs much better than no prior (Campbell, 1982). Another application was to prediction intervals (Campbell and Hollander, 1982).

## 2.1. Bayesian statistics in the late 1990s

There had been considerable attention devoted to Bayesian statistics for clinical trials by the late 1990s. Jerome Cornfield and Laurence Freedman at National Institutes of Health (NIH) and Max Parmar and David Spiegelhalter at UK's Medical Research Council had been especially instrumental in encouraging the use of Bayesian methods in clinical trials, with a handful of applied successes. There were several important papers on the use of Bayesian statistics for clinical trial by the 1990s, including Simon (1991), Ashby and Machin (1993), Berry (1993), Parmar *et al.* (1994), and Spiegelhalter *et al.* (1994). Monitoring clinical trials using predictive Bayesian methods was of particular interest (Carlin and Sargent, 1996; Fayers *et al.*, 1997; Freedman *et al.*, 1994; Greenhouse and Wasserman, 1995; Grieve *et al.*, 1991; Parmar *et al.*, 1996). Discussions of the $p$-value fallacy and of Bayes factors have been especially noteworthy (Goodman, 1999a, 1999b).

In the latter part of the twentieth century, the computing revolution generated enormous computing power to solve computationally intense problems and this combined with an algorithmic revolution of Markov Chain Monte Carlo (MCMC) had enormous implications for Bayesian computing. MCMC can be accomplished using the Metropolis-Hastings algorithm with, as a special case, the Gibbs sampler, which was introduced by Geman and Geman (1984). Thus by early 1990s Gibbs sampling was a well-accepted and handy way to calculate posterior distributions for Bayesians. This allowed for the calculation of the posterior distribution for any prior so that no longer were priors restricted to only the class of conjugate priors. First editions of books on Bayesian data analysis in 1995 and 1997 by Gelman *et al.* (2013) and Carlin and Louis (2009), respectively, guided the way. A minor technical problem emerged in that the MCMC chain may not always converge or might do so very slowly. The result was that no longer were Bayesians wed to the simple toy problems that started with only conjugate priors but now any prior could be used reflecting a great versatility that did not exist before.

What was the state of Bayesian applications in a regulatory environment in 1998, the date that the International Conference on Harmonisation E-9 Statistical Principles for Clinical Trials (Food and Drug Administration, 1998) was finalized? At that time there were no examples of Bayesian medical product clinical trials submitted to regulatory agencies in the United States, Europe, or Japan. However E-9 did define Bayesian approaches and frequentist methods and, further, allowed that while the guidelines largely refers to frequentist methods, the use of Bayesian approaches may be considered as well.

## 3. The Food and Drug Administration Bayesian initiative for medical devices

Around 1997, The Food and Drug Administration (FDA) embarked on an effort to consider the use of Bayesian statistics for submissions to the agency by medical device companies. This effort led by me had the full support of Bruce Burlington, the Director of FDA's Center for Devices and Radiological Health (CDRH).

The thinking regarding the application of Bayesian ideas to the clinical trials for medical devices for regulatory submission is as follows: There is often a great deal of prior information for a medical device. This is so in part because often the mechanism of action is well-known, physical as opposed to pharmacokinetic/pharmacodynamics, and local as opposed to systemic. Moreover, the nature of device development is that devices evolve whereas pharmaceutical drugs once discovered remain virtually unchanged. As a consequence, a device changes over time from model to model in small and sometimes not so small steps whereas in contrast a new molecular entity is basically unchanging. Further, as mentioned earlier, the revolution in computing hardware and MCMC had taken place, enabling the posterior calculation for any Bayesian prior. Of particular interest to the device industry was the possibility of bringing good technology to market sooner or with less current data by leveraging prior information. This was of great appeal to the device industry and so the FDA Bayesian initiative was well received by the innovative device industry.

Berry (1997) was asked by FDA to write a brief white paper on Bayesian statistics in medical device clinical trials. The result was a 78-page document entitled "Using a Bayesian approach in medical device development" (Berry, 1997). FDA also hosted internal short courses and seminars for statisticians as well as physicians and engineer reviewers by experts including Don Berry, Frank Harrell, Jay Kadane, Tom Louis, Steve Goodman, and Don Rubin.

The use of subjective priors was considered for the initiative, especially the skeptical and enthusiastic priors of Spiegelhalter *et al.* (1994). While this approach can be very helpful if the prior is not too skeptical, a true skeptic can wipe out any advantage of the Bayesian approach. It was not that FDA and a medical device company could not come to an agreement about a particular subjective prior but the concern was that, at an FDA Advisory Committee, there could be one or more members of the panel who would be very skeptical of the choice of prior and would have chosen a much more skeptical prior. An ensuing debate about subjective priors at the Advisory Committee was thought to not be helpful for FDA or the company. Further, a complete skeptic could reduce any advantage of prior information to zero, negating the entire approach. While Berger and Berry (1988) points out that all approaches are subjective, but the Bayesian approach is at least explicit in its subjectivity whereas the frequentist is "silently subjective", the decision was made to avoid subjective priors and rely only on priors developed from data from previous clinical studies. An added concern relates to the inherent difficulty for an FDA reviewer to evaluate a subjective prior without using information gleaned from other proprietary submissions. In particular, it could be very difficult to evaluate the validity of a company's subjective prior without unconsciously or subconsciously taking into account information derived from other companies' proprietary information.

There were several early lessons learned at FDA in the initiative. For any device there is often a lot of prior information for device studies and the mechanism of action of the device is often well understood. Further, using (good) prior information can often get to the same decision faster without any lowering of the scientific standards, without "lowering the bar". However, for a Bayesian submission that uses prior information, everything must be prospectively designed and pre-specified including the prior information. Further, it is not good science to switch the approach from frequentist to Bayesian or vice versa. For a Bayesian design, a company needs to consult early with CDRH, ensuring that the prior information identified in advance and mutually agreed upon. The planning frequently requires extensive simulations under different scenarios and, furthermore, it is most helpful if a company has a solid Bayesian statistician as an employee or consultant to help guide the statistical aspects of the submission.

## 3.1. Educational outreach efforts

Appreciated early was the importance of educational outreach efforts. In 1998 FDA and Health Industry Manufacturers Association (HIMA; later renamed Advanced Medical Technology Association (AdvaMed)) co-sponsored the workshop "Bayesian Methods in Medical Devices Clinical Trials". Statisticians and clinicians gave presentations, there were three case studies by medical device manufacturers and breakout sessions and a roundtable discussion. In the 2004, the conference "Can Bayesian approaches to studying new treatments improve regulatory decision-making?" was held at the NIH co-sponsored by FDA's three human product centers (CDRH, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER)) and the Department of Biostatistics at Johns Hopkins University. The August, 2005, issue of volume 2 of the journal Clinical Trials is devoted to this conference and contains articles by Bayesians (Berry, 2005; Goodman, 2005; Louis, 2005), and FDA officials (Alderson, 2005; Campbell, 2005; Temple, 2005; Woodcock, 2005) as well as 3 case studies, one of which concerned a medical device (Lipscomb *et al.*, 2005), and panel discussions.

### 3.2. A simple hierarchical Bayesian model

Hired by CDRH on an InterAgency Agreement, Malec (2001) considered the following clinical study of size 500 patients for a hypothetical coronary stent trial. The primary endpoint is proportion $p$ of patients who have target vessel failure (TVF), which is to be minimized. The research question was what would be the influence using Bayesian hierarchical modeling of prior information in the form of a previous study of size 250 on the same population with the identical endpoint that had 50 TVFs out of 250 for a proportion of $p_0 = 0.20$. The challenge is to estimate accurately the proportion of patients in the current study of 500 who have TVF. It is assumed that the current study is on the same patient population and follows the same clinical protocol so that it is reasonable to suppose the two studies are combinable to some extent. In the current study the observed proportion of failures is $p_t = x/500$ where $x$ is the number of failures. The methodology is to use Bayesian hierarchical model to provide an estimate of TVF that uses the prior information in combination with the current data to improve the estimate. Let $p_B$ ($B$ for Bayesian) denote the mean estimate of the posterior distribution. One approach is to build the following Bayesian hierarchical model, where $n_i$ denotes the size of study $i$ and $m_i$ the number of TVFs observed for $i = 0, t$:

$$\begin{aligned} m_i &\sim \text{Binomial}(n_i);\\ \mu_i &= \log\left\{\frac{p_i}{1-p_i}\right\};\\ \mu_r, \mu_t &\sim N\left(\nu, \frac{1}{\gamma}\right);\\ \nu &\sim N\left(\omega, \frac{1}{\tau}\right);\\ \gamma &\sim \text{Gamma}(\alpha,\beta). \end{aligned}$$

Note that the modeling is on the logit transformation $\mu$ of the proportions $p_0$ and $p_t$. Its two means are unknowns and are assumed to be from a normal distribution with parameters $\omega$ and $1/\gamma$. Non-informative normal and gamma priors are then placed on these latter two parameters. Using values of $\tau = 0$ and $\alpha$ and $\beta = 10^{-10}$ produces a proper prior distribution, as reported in Malec (2001).

Suppose $p_t = 0.170$. If we accepted the prior study as equivalent evidence or complete poolability, the weighted average is

$$\frac{(0.17)500 + (0.20)250}{750} = 0.18,$$

with (non-Bayesian) confidence interval of (0.153, 0.207). In contrast, the 95% *de novo* confidence interval, which borrows no information from the previous trial, is (0.137, 0.203), centered at 0.170. The Bayesian approach allows for some but not complete borrowing. The Bayesian mean estimate is $p_B = 0.177$. Note that this value is intermediate between no borrowing 0.170 and complete borrowing 0.180. The 95% credible posterior probability interval is (0.147, 0.207). This interval is 10% shorter than the *de novo* one but larger than the one with complete borrowing. The advantage of the hierarchical model is that it takes into account the study-to-study variability to appropriately inflate the variance. An advantage of the Bayesian approach is that one can use it not just to estimate the mean but any function of the entire posterior distribution and so it is straightforward to calculate quantities such as $P(p_B > 0.20)$, $P(p_B > 0.23)$, and $P(p_B < 0.17)$, for example, or to generate other credible intervals.

The effective sample size as defined by Malec (2001) can be expressed as $n_t$ times the ratio of the variances without any borrowing to the one with borrowing. (An alternative way to think of this is as the sample size that would correspond to the Bayesian credible interval with no prior if the distribution is approximately symmetric.) In this case, the amount of borrowing is slightly larger than 150 from the previous study of size 250.

Figure 1 in Malec (2001) provides some insight into how the hierarchical model works. If the proportion is near 0.20 there is a lot of borrowing but very rarely as much as complete poolability, almost up to the entire additional 250. If the proportion is less than 0.08 or greater than 0.34, then there is no borrowing but if the proportion is say 0.12 (or 0.28), just far enough away, there is negative borrowing, meaning the model needs to overcome the incorrect prior and can result in an effective sample size that is less than 400 whereas the sample size with no borrowing is 500. This is especially notable since the tendency might be for the company to choose as a previous study with a low TVR, hoping to move the observed rate downward from the current observed rate (in this case 0.20) but if this is not based on sound scientific thinking, it can backfire. From a regulatory perspective this has a lot of appeal. The result is that companies who wanted to go down the Bayesian pathway had to be quite confident in the prior information that was chosen for the prior distribution.

## 3.3. The challenge and early decisions

The challenge was how to do Bayesian statistics in a regulatory environment for medical products generally and for medical devices in particular.

One early decision was to restrict to prior information directly based on quantitative, rather than subjective, information from data from previous clinical studies. The argument is then about the previous study and its data quality and not about some subjective opinion. Further, companies need legal access to the data. Legal access could be a study done by the same company overseas or on a very similar device or they could get permission to get legal access to another company's data. Reports of studies published in the literature generally are insufficient for such prior information since patient-level data is usually required. The description of an approved PreMarket Approval (PMA) application from a competitor's Summary of Safety and Effectiveness could not be used since, although it is public, it is the property of that company.

For a Bayesian submission using prior information, FDA and the company need reach an agreement on the validity of any prior quantitative information. This would entail a discussion by clinicians and engineers from the agency and the company. An important question is whether all the appropriate previous data are being used. This is to avoid a company "cherry picking" the data, selecting only advantageous prior data that would pull the current data in a favorable direction while ignoring other less favorable studies. Furthermore, the quality of the prior information needs to be high since a prior that is not in alignment with the current study could prove costly to the company.

What is different for both the company and the FDA with a Bayesian approach as opposed to a frequentist one is that there are different decision rules for clinical study success. This means no longer relying on $p$-values or confidence intervals but rather on some aspect specified in advance associated with the posterior distribution. A Bayesian approach needs a different decision rule, based on the posterior distribution, either a posterior probability exceeding some predetermined value or a credible interval that behaves in a predetermined fashion. This is important since companies (sponsors) deserve to understand what constitutes success in a clinical trial. For most frequentist trials this has been couched as a statistically significant result, a small $p$-value, for the primary effectiveness endpoint using hypothesis testing. In the Bayesian setting this could mean achieving a pre-specified probability of superiority based on the posteriori distribution or the probability that some difference exceeds a pre-specified value with a pre-specified probability.

One could make the argument that the evidentiary criterion is the similar but that with the use of prior information, one can get to the same decision (conclusion) oftentimes with fewer patients and hence faster. While there has been resistance in some quarters to using good prior information, a good question is: Why ignore it if it is available? Should a company with such information be treated the same way as a company without it? Why should the burden be the same for two companies, one of which has good information and the other none? (It may help to think about prior information in terms of equivalent number of patients in a previous study.) Is it fair to penalize a firm and not let them use their legal prior information that could be a competitive advantage? The catch is that while many companies might initially claim that they have very good prior information from previous studies, it is quite another matter to understand that the company is usually making a very expensive bet on that assertion and in some cases that realization has given some companies pause to carefully question the strength of such data.

The early focus at CDRH was on borrowing strength from previous clinical trial data to conclude something about the current study. For FDA the focus is not on a class of medical devices but on a particular device, a particular submission. Note that this is different from Bayesian meta-analysis. For meta-analysis the effort is to make a conclusion about the totality of all the studies whereas in this hierarchical model application, the effort is to borrow strength, to leverage information, from other studies in order to make a conclusion only about the current study of interest. Bayesian hierarchical modeling is not unrelated to random effects models and shrinkage estimation that can be frequentist.

There were a number of myths or misperceptions that needed to be addressed. There was the impression that FDA submissions for medical devices clinical trials were dominated by Bayesian ones but in fact only 5% to 10% of the CDRH submissions have been Bayesian over a twenty year history. Thus, the FDA statistical staff is not mostly Bayesian nor do the FDA Bayesians handle only Bayesian submissions. Further, there was an impression that a Bayesian approach would be easier and constitute a lowering of the regulatory bar. What is true is that if a standard statistical analysis and a Bayesian analysis were to always yield the same basic conclusion, there would be no reason to consider a different approach. Often in the Bayesian approach there is prior information that is ignored in the frequentist approach. Another false impression is that FDA would compel a company to do a Bayesian submission even if they did not want to. It is true that a Bayesian approach would in many cases be a tradeoff between what could be a lighter clinical burden but a heavier statistical/computational one with the Bayesian approach but that choice was the company's to make.

There was a recognized need to bring the industry and FDA review staff up to speed. CDRH at FDA offered 3 or 4 times within a period of 12 years the course "Bayesian statistics: what the non-statistician needs to know", a four-week course (2 lectures per week) meeting one morning a week featuring presentation by Bayesians in CDRH's Division of Biostatistics. The course was designed for non-statisticians who were reviewers and was attended by physicians (medical officers), engineers and other scientists.

### 3.4. General hierarchical Bayes borrowing strength

The basic general hierarchical model for combining data across studies and hence synthesizing clinical evidence consists of several levels in the hierarchy. Exchangeability is a fundamental concept underlying statistical inference and of particular importance in Bayesian trials. Units (patients or trials) are considered exchangeable if the probability of observing any particular set of outcomes on those units is invariant to any re-ordering of the units (Food and Drug Administration, 2010).

For the lowest level 1, patients ($y$) are exchangeable within each study but not from study to study:

$$y_j|\theta_j, f \sim P(y_j|\theta_j, f).$$

At level 2, it is assumed that studies are exchangeable within patient populations:

$$\theta_j|f \sim P(\theta_j|f).$$

At the highest level, level 3 there is a prior $p(f)$, which is usually non-informative. Note that the choice of a non-informative prior distribution is sometimes difficult since there may be more than one way to parameterize the problem and then place what looks like the obvious non-informative prior on the parameter of interest. There is not usually only one non-informative prior.

### 3.5. Adaptive Bayesian methods and modeling

The other big idea besides Bayesian hierarchical modeling is adaptive Bayesian designs, to use accumulating information from the trial, usually with a non-informative prior distribution, to make pre-planned changes to the trial. As defined in the FDA guidance on adaptive designs for medical devices (Food and Drug Administration, 2016a), an "adaptive design for a medical device clinical study is defined as a clinical study design that allows for prospectively planned modifications based on accumulating study data without undermining the study's integrity and validity." For complicated Bayesian adaptive designs, it is also always the case that there is no analytical way to derive the operating characteristics of the design. An important component, then, is to perform simulations to understand the operating characteristics of the design under all reasonable and maybe some (slightly) unreasonable scenarios. If that were the end of the story, then for large trials where the only adaptation is to stop early, one could use the large-sample asymptotics that have been so well developed for group sequential trials. One real advantage comes if there are more complicated adaptations than merely interim analyses to stop early for success or futility. These complications could include changing the randomization ratio, dropping an arm if there are at least 3 arms, dropping a subset, etc. Another very real advantage comes from modeling to use intermediate endpoints to predict the primary effectiveness (efficacy) endpoint. The basic functional form of the prediction model would be specified but the accumulating data would be used to constantly improve the parameter estimation of the model. The result would be a model to predict the primary endpoint for patients who have not reached the final endpoint and also an estimate of how good the model is in its prediction. The result would be the ability using the predictive posterior probability to fast-forward to what is likely or expected to happen for patients who have not reached the final endpoint, including future patients who have not even been enrolled in the trial yet. And the bonus would be that there would be a distribution for the eventual success of the trial so one could calculate how likely it is to occur.

One example is the use of piece-wise exponential models to model the success (survival) over time for a time-to-event study. The parameters of each of the exponential pieces are estimated as data accumulates. This could allow, based on the predictive posterior probability distribution, a halt in the recruitment but continuation of the trial for already enrolled patients. This strategy can save monetary and clinical resources since unnecessary patients that are not needed are not enrolled and the confidence that the eventual success is highly likely.

An example of an adaptive Bayesian model is as follows: Build a theoretical model to predict the 2-year primary endpoint using intermediate endpoints, using a non-informative prior. Recruit 400 patients initially for a trial ($i = 1$). Then at stage $i$, update estimates of predictive model parameters and decide to either: (1) stop early for success (posterior probability of superiority $> a_i$); (2) stop recruiting but continue the trial (predictive post. prob. $> b_i$); (3) stop for futility (pred. post. prob. $< c_i$); otherwise continue to recruit 50 additional patients for $I = 2, 3, 4$. Simulations are performed to determine $a$'s, $b$'s, $c$'s and desired operating characteristics (Type I error and power). This approach by itself merely generates in an arduous manner what could easily be achieved by appealing to a group sequential rule such as O'Brien-Fleming or, more generally, the alpha spending of Lan and DeMets. The big advantage of the Bayesian approach with a non-informative prior is that the accumulating data is being used not only to stop early for success or futility but also to model to the primary outcome variable using intermediate endpoints. The accumulating data is used to estimate the parameters of the model to predict the primary endpoint. The predictive probability that is generated is then used to decide to stop recruiting additional patients in to the trial since the predictive probability is indicating that when the current patients all reach their primary endpoints that the resulting posterior probability is likely to lead to success. Note that the sample size is not known at the beginning of the study. The book by Berry *et al.* (2011) on Bayesian adaptive design provides more detail on Bayesian adaptive designs.

### 3.6. Type I error and priors

What about the probability of Type I error for hierarchical Bayesian designs and for Bayesian adaptive designs? For the use of previous information in the form of previous studies in hierarchical modeling, the Type I error is inflated compared to the situation of no prior information. However, if the prior data makes the null hypothesis more unlikely, it may be no surprise that the Type I error probability calculated under the unlikely null hypothesis is inflated. Then extremely stringent Type I error probability control does not make as much sense since there is already evidence that the null is not true. Some new thinking is needed about the tight control of Type I error in the situation of hierarchical borrowing. For Bayesian adaptive and Bayesian hierarchical designs, it is crucial to understand the operating characteristics of the design, in particular the Type I error probability and the statistical power. This is usually accomplished by simulation under many possible scenarios since there is rarely a closed-form formula for the operating characteristics of most Bayesian designs.

### 3.7. Sample size

Another issue concerns the sample size for a Bayesian trial that uses hierarchical modeling or Bayesian adaptation. It is clear for the latter where it is possible to perform sample size reassessment or other preplanned adaptations so as to possibly increase the sample size under certain pre-specified circumstances. See Grieve *et al.* (1991), Dmitrienko and Wang (2006), and Saville *et al.* (2014). For hierarchical Bayesian modeling, since the amount of borrowing is unknown under a hierarchical model or with a Bayesian power prior approach, one needs to plan the study in an adaptive fashion. If it is decided that 650 patients are needed to make the inference and the size of the previous information is 250 patients, then one approach would be to plan the study for at least 400 new patients or realistically for 500 or 550 and then during the study assess how much borrowing is happening and curtail the study when the effective sample size is 650.

### 3.8. Food and Drug Administration guidance for medical device clinical trials

A draft of the FDA Guidance on the Use of Bayesian Statistics in Medical Device Clinical Trials was released for public comment in May of 2006, followed by a public meeting on it in July of the same year. It was finalized an in 2010 (Food and Drug Administration, 2010). It describes the two main types of Bayesian submissions, Bayesian hierarchical modeling with data from previous studies and

Bayesian adaptive trial usually with a non-informative prior where it is the accumulating data in the trial that is used to make preplanned changes to the trial.

### 3.9. Center for Devices and Radiological Health review and successes

Pennello and Thompson (2007) provide experience concerning the review of Bayesian medical device submissions to FDA. Bayesian submissions generally require more work for the statistical reviewer. That has been especially true when no validated statistical software for Bayesian analysis existed. The review of Bayesian submissions requires an FDA statistical staff that can capably handle such challenges. The review at the design stage, the investigational device exemption (IDE) stage, is perhaps even more challenging since it involves reviewing the simulations the company has provided to understand the operating characteristics of the design.

As far as the early success of the FDA Bayesian initiative, see Campbell (2005), Irony and Simon (2006), and Bonangelino *et al.* (2011). Since the initiative began, at least 25 PreMarket Approval applications have been approved by FDA where the primary design and analysis has been Bayesian. There has been at least one PreMarket Notification (called a 510(k)) that has also been approved along with a great many IDEs that allow companies to commence with clinical trials. For a complete list of publicly available submissions see Campbell (2011, 2013). A report on implantable medical devices is Pibouleau and Chevret (2011).

### 3.10. Bayesian software

The primary software package for Bayesian analysis is Bayesian inference Using Gibbs Sampling (BUGS).. However, BUGS or its Window's version WinBUGS is often used but it is not a commercial (non-validated) product (Spiegelhalter *et al.*, 2003). Since WinBUGS is not a commercial product, there has been no incentive for its developers at UK's Medical Research Council to validate the software. A concern by FDA is that any software package used in a regulatory submission produces results that are reliable. When a software product is not validated, this poses a challenge for FDA reviewers. The lack of such validated software has required in many instances that the statistical reviewer needs to repeat the Bayesian analysis using a different software package. Other worries are whether the convergence depends on the seed that is used and whether the Markov chain has converged. For simulations at the design stage, careful checking of the computer program may be necessary. CDRH at FDA entered into a Cooperative Research and Development Agreement with Cytel Inc., for “Software for Bayesian Clinical Trials” which was completed in 2013 (Food and Drug Administration, 2017). More recently there are validated software packages for Bayesian analysis, including Bayesian capabilities within SAS/STAT (Stokes *et al.*, 2015). There is a critical need for software in the planning of a Bayesian design. Often the amount of simulation is much greater at the design stage than at the analysis stage in order to understand and calibrate the design so that it has the desired operating characteristics.

### 3.11. Power priors

A quandary has arisen if the hierarchical model relies on only one prior study. If there is only one previous trial, then the number of studies to estimate the study-to-study variability is only two, the previous study and the current one, leading to a highly unstable estimate of the variance. And even three studies do not lead to a very good estimate of the study-to-study variance. In such circumstances, this is reflected in the lack of robustness of the results to the exact form of the non-informative prior. At the very least the analysis would need to study the sensitivity associated with the selection of the non-informative prior.

Instead of a Bayesian hierarchical model, another exciting development was the introduction of the conditional power prior by Ibrahim and Chen (2000). For a parameter $\theta$ and data $D_0$ from a previous study, the conditional power prior is given by:

$$\pi(\theta|D_0, \alpha_0) \propto \pi_0(\theta)L(\theta|D_0)^{\alpha_0},$$

where $\pi_0(\theta)$ is the prior and $L(\theta|D_0)$ is the likelihood based on the prior data $D$. The parameter $\alpha_0$ is the exponent for the likelihood based on the prior data and indicates a range of borrowing from none ($\alpha_0 = 0$) to all ($\alpha_0 = 1$), where the latter is complete patient-level exchangeability (complete pooling) and if $\alpha_0$ is zero, then there is no borrowing and no prior information is used. So a power prior with a fixed $\alpha_0$ between 0 and 1 allows for some but not complete borrowing. Then the posterior distribution $q$ is:

$$q(\theta|D_0, D, \alpha_0) \propto \pi_0(\theta)L(\theta|D_0)^{\alpha_0}L(\theta|D).$$

The Bayesian approach now is to place prior on $\alpha_0$ and use the current data expressed through its likelihood $L(\theta|D)$ to calculate the posterior for $\alpha_0$. See Bae *et al.* (2008) for an application of power priors to one-way ANOVA.

An effort to improve upon the power prior is the commensurate prior approach of Hobbs *et al.* (2011, 2012). Commensurate power priors introduce a parameter $\tau$ that is a measure of how commensurate the current data are with the historical data. So the conditional prior is of the form $\pi^{\text{CPP}}(\theta, \alpha_0, \tau|D_0)$. For a location parameter $\theta$, the commensurate parameter $\tau$ can be a precision parameter estimated from the data from the historical studies and the current one. So place a (non-informative) prior on $\alpha_0$ and another on $\tau$. Murray *et al.* (2014) have extended this theory to time-to-event data.

## 3.12. Bayes and historical controls

The use of historical controls from previous trials has been the "low hanging fruit" for the application of Bayesian methods (Viele *et al.*, 2014). Of note is that using historical controls in a hierarchical Bayesian model can give rise to adaptive randomization (Hobbs *et al.*, 2013). They are not without some risk since one could in theory "cherry pick" the choice of which historical controls to use and which not. The other problem is that even without "cherry picking" one could be drawn to a particular trial or trials that have a favorable historical control for the current use. It is good advice for the FDA and the company to consider carefully all high-quality prior studies that are thought to be exchangeable.

The Bayesian experience has been most helpful for adaptive designs generally. See Campbell (2013) for a discussion of the similarities and differences of Bayesian and adaptive designs. The Bayesian experience helped considerably in the development of an FDA guidance document on Adaptive Designs for Medical Device Clinical Investigations (Food and Drug Administration, 2016a). CDRH has seen about 250 adaptive submissions from 2007 to 2013 according to a manuscript by Yang *et al.* (2016), which reports adaptive design practice at CDRH from January 2007–May 2013, of which about 30% have been Bayesian.

## 3.13. Treatment response adaptive randomization

Adaptive randomization is an idea that has a long history. Randomized play-the-winner rule for clinical trials were introduced by Wei and Durham (1978) and applied in a randomized clinical trial for a medical device, extracorporeal membrane oxygenation (ECMO), for the treatment of persistent pulmonary hypertension of newborn babies. In this trial, an urn is initialized with one black ball and one red, the black ball (B) corresponding to the conventional medical therapy and the red ball (R) to ECMO. If the patient is a success another ball of the same color is added to the urn and if a failure (death) a ball of the opposite color is added. In the University of Michigan trial, the babies received therapy in the following order: RBRRRRRRRRRR, at which point the trial was stopped in favor of ECMO. Adaptive trials such as ECMO are difficult for frequentists to analyze since the sample space may not be fixed at the beginning of the trial and the stopping rule may not be well established in advance. The trial and its many possible statistical analyses are reported in Ware (1989). The challenge for frequentists is trying to describe the sample space for an adaptive trial, although there have been some interesting large sample theory developments (Hu and Rosenberger, 2006). Ethics play a large role in many adaptive trials and certainly in most pediatric clinical trials. Earlier more general Bayesian discussion of ethics includes publications by Ashby and Machin (1993), Royall (1991), Kadane (1996), and Berry (2004). Treatment response adaptive trials can be more ethical and consequently can lead to great participation by investigator and patients if there is an effort to randomize patients to maximize more patients to the better therapy based on the accumulating data in the trial.

More recent work on Bayesian trial monitoring that keeps track of the probability of success and of futility by Data Monitoring Committees, even if the trial is not Bayesian, can be found by Dmitrienko and Wang (2006) and Saville *et al.* (2014).

### 3.14. The likelihood principle

A basic tenet of Bayesian statistics is the likelihood principle. It states simply that “all the information about a parameter $\theta$ in a trial is expressed in its likelihood”. See Berger and Wolpert (1988) and Berry (1987) for further discussion of the principle. This might make sense in a study such as ECMO where it may not be altogether clear about how the trial was conducted in terms of stopping rules. However, the probability of a null hypothesis for a trial with multiple possible looks for stopping for success or futility can be shown to depend on not just the likelihood but on the chance of making different decisions. In the context of making decisions, looks are not free (as simulations of Type I error demonstrate). It may help to distinguish between information about the parameter versus the decisions that need to be made based on the data. The information does not depend on anything but the likelihood but making decisions for pre-specified statistical plans can be affected by, for example, the number of looks or the number of primary endpoints.

### 3.15. Bayesian and pediatric extrapolation

In the FDA pediatric guidance on extrapolation for medical device studies (Food and Drug Administration, 2016b), the appendix advocates the use of Bayesian hierarchical modeling in the following way: At the top level of the hierarchy, adult studies and the pediatric studies are assumed exchangeable, an assumption that requires medical confirmation. At the second level, each of the adult studies is assumed exchangeable with each other and the pediatric ones with each other as well. Finally, at the lowest level of the hierarchy, patients in each study are assumed to be exchangeable with each other. A prior, usually a non-informative one, is placed at the top of the hierarchy. A more recent publication is Gamalo-Siebers *et al.* (2017). Bayesian methods can also be utilized in trials for rare diseases where prior information and adaptive techniques can be brought to bear. An additional complication is that there may be a very limited study population, which can have Bayesian implications (Cheng and Berry, 2003).

### 3.16. Bayesian subset (subgroup)

Dixon and Simon (1992) and Simon (2002) have considered Bayesian analysis of subgroups. This is similar to a random effects model for multiple centers or shrinkage estimation. This can help to address the well-known regression to the mean. If particular subgroups are of interest in a clinical trial, it is possible in advance to build a hierarchical model that identifies these important subgroups. These can then “borrow” from each other or “gain strength”. However, what about a failed trial with a promising subgroup? It is certainly inappropriate to “cherry pick” and then to ignore all data except the subgroup. Under what extraordinary circumstances would it be possible to rely on the data from this single study as prior information? What if anything can be learned from a failed study that has a very impressive subgroup? Is it real or spurious? The worry is picking up a possibly spurious subset (subgroup) that cannot be confirmed in a later study. It is fundamental question about regression to the mean and ultimately about reproducibility.

### 3.17. An example of a failed trial with an interesting subgroup

The following is an instructive example. The pivotal Acute Myocardial Infarction with Hyperoxemic Therapy (AMIHOT) clinical study for the TherOx® Downstream Aqueous Oxygen System in treating post acute myocardial infarction (AMI) patients had not been a success in terms of its three primary endpoints but there was an interesting subgroup that was not planned for in the original analysis; that is to say that although there were subgroups identified in advance in AMIHOT, no effort was planned for a confirmatory analysis of them. In this case the interesting subset was anterior AMI patients revascularized within 6 hours of AMI onset. So a second study AMIHOT-II was planned to investigate only this subgroup. Only one of the three primary effectiveness endpoints, infarct sized based on imaging, was selected for the second study (this selection was not taken into consideration in what follows). The idea was to use the data from AMIHOT in a Bayesian analysis of this new trial. However, it would not be good science to use as a prior only the data in the subgroup of interest. Instead the company built a Bayesian hierarchical model that used all the data in AMIHOT in a hierarchy that allowed for exchangeability of this subgroup and three other subgroups in AMIHOT and then, at a higher level of the hierarchy, study exchangeability between AMIHOT and AMIHOT-II. There were two endpoints, one an effectiveness endpoint, infarct sized based on imaging, for a superiority claim and one a safety endpoint, the 30-day MACE rate with a non-inferiority delta of 6%. The study was powered so that there was insufficient power for AMIHOT-II if it stood on its own, but sufficient power if some strength was borrowed from the AMIHOT data. The Bayesian design and analysis of the study was described in Stone *et al.* (2009). The posterior probability of superiority for the effectiveness endpoint was 96.9% and 99.5% for the non-inferiority safety endpoint, both exceeding the pre-specified success thresholds of 95% (Food and Drug Administration, 2009); hence the study was a success. A rough calculation indicated that the amount of borrowing from AMIHOT was about 3 patients out of the subgroup of 49 in a trial of 289, a calculation that is similar to a value conveyed to me by the Bayesian statistician for the company, John Boscardin (personal communication). The 2009 meeting of the FDA Circulatory Systems panel of the Medical Device Advisory Committee recommended to FDA that the submission was non-approvable and it was subsequently turned down by FDA.

### 3.18. Non-inferiority and Bayes

Non-inferiority is a uniquely regulatory concept. Generally scientists do not set out to show their product is non-inferior by some fixed amount before a study begins. Instead of superiority as the alternative hypothesis, the null hypothesis is inferiority by at least a pre-specified amount d, the non-inferiority margin, and the alternative is non-inferiority by more than d. In a two-arm non-inferiority trial of an experimental product to a known active control, a Bayesian approach seems most natural, incorporating prior information explicitly from earlier studies of the active control to a placebo control. Interesting Bayesian papers on non-inferiority are Chen *et al.* (2011) using historical controls for medical devices and Gamalo *et al.* (2014) for non-infective drug products.

### 3.19. Adaptive Designs Accelerating Promising Trials into Treatments

A recent educational effort to encourage the use of Bayesian methods has been Adaptive Designs Accelerating Promising Trials into Treatments (ADAPT-IT) (Meurer *et al.*, 2012). The idea was to work with groups who were preparing NIH grants and FDA submissions for confirmatory trials to consider using Bayesian methods and to study the perceptions of key stakeholders involved in the process (Meurer *et al.*, 2016). Of the five projects, two were cooling medical devices. (Interestingly, another cooling trial using Bayesian methods is Pedroza *et al.* (2016).)

## 4. The future

### 4.1. Trends

One trend has been a shift from Bayesian hierarchical models to Bayesian adaptive ones. The reasons are many. One is that often a careful examination of exchangeability results in concluding that the new study is not exchangeable since it is more likely to be better in some sense than its predecessors. A second is that there remains some apparent discomfort in some quarters associated with the inflation of the Type I error using prior data from prior studies.

There is currently a more thoughtful approach to the mindless use of $p$-values to assess statistical significance. The American Statistical Association (ASA) recently issued a statement on statistical significance and $p$-values (Wasserman and Lazar, 2016) and a number of Bayesians have heartily joined the discussion.

There has been some interest in the pharmaceutical drugs concerning the development of seamless Phase II-III trial for drugs, namely, that if planned carefully at the outset it is possible to view data from Phases II and III in drug development as combinable (Inoue *et al.*, 2002). This same idea can be expanded to medical device trials, planning at the outset to combine pilot and pivotal trials.

Freedman and Spiegelhalter (1992) and Müller *et al.* (2017) considered using Bayesian decision theoretic methods in clinical trials, using statistical decision theory for trial design and analysis, deciding when to curtail a study, when the loss of enrolling more patients is larger than that of stopping (for either success or failure). An early example with an application to emergency medicine trials was Lewis and Berry (1994). There is now renewed interest in this approach with a focus on benefit-risk decisions. Waddingham *et al.* (2015) has considered using Bayesian methods in benefit-risk assessment. For FDA this requires quantitative (in non-economic public health) measures of benefit as well as risk. Often in premarket submissions this is a balance between safety and effectiveness or, for diagnostic ones, sensitivity versus specificity.

A recent article by Haddad *et al.* (2017) as part of a project on modeling and simulation by the Medical Device Innovation Consortium uses engineering modeling for medical device performance to generate “virtual patients” that could be combined in conjunction with clinical trial data using the power prior.

Regulatory authorities besides FDA medical device center appear to be more receptive to Bayesian methods than ever, as noted recently by Campbell (2017).

The Bayesian initiative is just one of several innovations launched by FDA for medical devices. See Campbell and Yue (2016) for other statistical innovations for medical devices pioneered by FDA.

### 4.2. Bayesian reporting in the medical literature

Why are there so many Bayesian articles in the statistical literature and so few in the medical journals? Partly this is a reporting challenge of what should be in a report for a Bayesian trial, which have been addressed by Hughes (1993) and, for adaptive trials, by Sung *et al.* (2005). There is now a huge disconnect between what has been developed in Bayesian theory and what is used in clinical trial practice. This gap is now quite apparent between Bayesian methods for clinical trials developed in theoretical publications in statistical journals and their use or lack of it in applications as reported in medical journals. Chevret (2012) reports the use of Bayesian adaptive designs in the statistical and medical literature. Berry (1993), Kadane (1995), Spiegelhalter *et al.* (2004), and Gönen (2009) have all written that the time is ripe for Bayesian methods. While that sentiment has now largely been embraced by the statistical community and its journals, it has not been so for the most biomedical journals. Campbell (2011) provides a small list of medical journal articles for Bayesian medical device trials. He reports that many such trials are often not reported as such in the NIH website clinicaltrials.gov that tracks most clinical trials. Some biomedical journal editors have insisted on frequentist analysis for trials that are designed and analyze as Bayesian, according to Don Berry (personal communication, September 12, 2017). It is quite clear that Bayesian methods have an image problem in biomedical science. In the face of such obstacles it is important to celebrate the Bayesian successes and, further, for statisticians to continue to lead the way.

## References

Alderson NE (2005). Editorial, *Clinical Trials*, **2**, 271–272.

Ashby D (2006). Bayesian statistics in medicine: a 25 year review, *Statistics in Medicine*, **25**, 3589–3631.

Ashby D and Machin D (1993). Stopping rules, interim analyses and data monitoring committees, *British Journal of Cancer*, **68**, 1047–1050.

Bae RN, Kang YH, Hong MH, and Kim SW (2008). Multiple comparison for the one-way ANOVA with the power prior, *Communications for Statistical Applications and Methods*, **15**, 13–26.

Bayes T (1763). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, **53**, 370–418.

Berger JO and Berry DA (1988). Statistical analysis and the illusion of objectivity, *American Scientist*, **76**, 159–165.

Berger JO and Wolpert RL (1988). *The Likelihood Principle* (2nd ed), Institute of Mathematical Statistics, Hayward.

Berry DA (1987). Interim analysis in clinical trials: the role of the likelihood principle, *The American Statistician*, **41**, 117–122.

Berry DA (1993). A case for Bayesianism in clinical trials, *Statistics in Medicine*, **12**, 1377–1393.

Berry DA (1997). Using a Bayesian approach in medical device development, Technical Paper, Retrieved October, 2017, from: http://ftp.isds.duke.edu/WorkingPapers/97-21.ps

Berry DA (2004). Bayesian statistics and the efficiency and ethics of clinical trials, *Statistical Science*, **19**, 175–187.

Berry DA (2005). Introduction Bayesian methods III: uses and interpretation of Bayesian tools in design and analysis, *Clinical Trials*, **2**, 295–300.

Berry SM, Carlin BP, Lee JJ, and Müller P (2011). *Bayesian Adaptive Methods for Clinical Trials*, CRC Press, Boca Raton.

Bonangelino P, Irony T, Liang S, *et al.* (2011). Bayesian approaches in medical device clinical trials: a discussion with examples in the regulatory setting, *Journal of Biopharmaceutical Statistics*, **21**, 938–953.

Box GEP and Tiao GC (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

Campbell G (1982). The maximum of a sequence with prior information, *Communications in Statistics. Part C: Sequential Analysis*, **1**, 177–191.

Campbell G (2005). The experience in the FDA's Center for Devices and Radiological Health with Bayesian strategies, *Clinical Trials*, **2**, 359–363.

Campbell G (2011). Bayesian statistics in medical devices: innovation sparked by the FDA, *Journal of Biopharmaceutical Statistics*, **21**, 871–887.

Campbell G (2013). Similarities and differences of Bayesian designs and adaptive designs for medical devices: a regulatory view, *Statistics in Biopharmaceutical Research*, **5**, 356–368.

Campbell G (2017). Regulatory acceptance of Bayesian statistics. In E Lesaffre, G Baio, and B Boulanger (Eds), *Bayesian Statistics Applied to Pharmaceutical Research*, CRC Press, Boca Raton.

Campbell G and Hollander M (1978). Rank order estimation with the Dirichlet prior, *Annals of Statistics*, **6**, 142–153.

Campbell G and Hollander M (1982). Prediction intervals with a Dirichlet-process prior distribution, *The Canadian Journal of Statistics*, **10**, 103–111.

Campbell G and Yue LQ (2016). Statistical innovations in the medical device world sparked by the FDA, *Journal of Biopharmaceutical Statistics*, **26**, 3–16.

Carlin BP and Louis TA (2009). *Bayesian Methods for Data Analysis* (3rd ed), Chapman and Hall/CRC, Boca Raton.

Carlin BP and Sargent DJ (1996). Robust Bayesian approaches for clinical trial monitoring, *Statistics in Medicine*, **15**, 1093–1106.

Chaloner K, Church T, Louis TA, and Matts JP (1993). Graphical elicitation of a prior distribution for a clinical trial, *The Statistician*, **42**, 341–353.

Chen MH, Ibrahim JG, Lam P, Yu A, and Zhang Y (2011). Bayesian design of noninferiority trials for medical devices using historical data, *Biometrics*, **67**, 1163–1170.

Cheng Y, Su F, and Berry DA (2003). Choosing sample size for a clinical trial using decision analysis, *Biometrika*, **90**, 923–936.

Chevret S (2012). Bayesian adaptive clinical trials: a dream for statisticians only?, *Statistics in Medicine*, **31**, 1002–1013.

Dixon DO and Simon R (1992). Bayesian subset analysis in a colorectal cancer clinical trial, *Statistics in Medicine*, **11**, 13–22.

Dmitrienko A and Wang MD (2006). Bayesian predictive approach to interim monitoring in clinical trials, *Statistics in Medicine*, **25**, 2178–2195.

Fayers PM, Ashby D, and Parmar MK (1997). Tutorial in biostatistics Bayesian data monitoring in clinical trials, *Statistics in Medicine*, **16**, 1413–1430.

Ferguson TS (1973). Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**, 209–230.

Food and Drug Administration (1998). Guidance for industry: E9 statistical principles for clinical trials, Retrieved October, 2017, from: https://www.fda.gov/downloads/drugs/guidancecompliancer egulatoryinformation/guidances/ucm073137.pdf

Food and Drug Administration (2009). FDA panel presentation for Therox P080005, Retrieved October, 2017, from: https://www.fda.gov/ohrms/dockets/ac/09/slides/2009-4419s1-01.pdf

Food and Drug Administration (2010). The use of Bayesian statistics in medical device clinical trials: guidance for industry and Food and Drug Administration staff, Retrieved October, 2017, from: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/

Food and Drug Administration (2016a). Adaptive designs for medical device clinical studies: guidance for Industry and Food and Drug Administration staff, Retrieved October, 2017, from: https://www.fda.gov/downloads/medicaldevices/

Food and Drug Administration (2016b). Leveraging existing clinical data for extrapolation to pediatric uses of medical devices: guidance for Industry and Food and Drug Administration staff, Retrieved October, 2017, from: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulatio nandGuidance/GuidanceDocuments/UCM444591

Food and Drug Administration (2017). FDA CRADAs, Retrieved October, 2017, from: https://www.fda.gov/scienceresearch/collaborativeopportunities/

Freedman LS and Spiegelhalter DJ (1992). Application of Bayesian statistics to decision making during a clinical trial, *Statistics in Medicine*, **11**, 23–35.

Freedman LS, Spiegelhalter DJ, and Parmar MK (1994). The what, why and how of Bayesian clinical trials monitoring monitoring, *Statistics in Medicine*, **13**, 1371–1383.

Gamalo MA, Tiwari RC, and LaVange LM (2014). Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products, *Pharmaceutical Statistics*, **13**, 25–40.

Gamalo-Siebers M, Savic J, Basu C, *et al.* (2017). Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation, *Pharmaceutical Statistics*, **16**, 232–249.

Geman S and Geman D (1984). Stochastic relaxation, Gibbs distribution and Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB (2013). *Bayesian Data Analysis* (3rd ed), Chapman & Hall/CRC, Boca Raton.

Gönen M (2009). Bayesian clinical trials: no more excuses, *Clinical Trials*, **6**, 203–204.

Goodman SN (1999a). Toward evidence-based medical statistics, I: the *P* value fallacy, *Annals of Internal Medicine*, **130**, 995–1004.

Goodman SN (1999b). Toward evidence-based medical statistics, II: the Bayes factor, *Annals of Internal Medicine*, **130**, 1005–1013.

Goodman SN (2005). Introduction Bayesian methods I: measuring the strength of evidence, *Clinical Trials*, **2**, 282–290.

Greenhouse JB and Wasserman L (1995). Robust Bayesian methods for monitoring clinical trials, *Statistics in Medicine*, **14**, 1379–1391.

Grieve AP, Choi SC, and Pepple PA (1991). Predictive probability in clinical trials, *Biometrics*, **47**, 323–330.

Grieve AP (2007). 25 years of Bayesian methods in the pharmaceutical industry: a personal, statistical bummel, *Pharmaceutical Statistics*, **6**, 261–281.

Haddad T, Himes A, Thompson L, Irony T, Nair R, and MDIC Computer Modeling and Simulation Working Group Participants (2017). Incorporation of stochastic engineering models as prior information in Bayesian medical device trials, *Journal of Biopharmaceutical Statistics*, **10**, 1–15.

Hobbs BP, Carlin BP, Mandrekar SJ, and Sargent DJ (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials, *Biometrics*, **67**, 1047–1056.

Hobbs BP, Carlin BP, and Sargent DJ (2013). Adaptive adjustment of the randomization ratio using historical control data, *Clinical Trials*, **10**, 430–440.

Hobbs BP, Sargent DJ, and Carlin BP (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models, *Bayesian Analysis*, **7**, 639–674.

Hu F and Rosenberger W (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*, Wiley, Hoboken.

Hughes MD (1993). Reporting Bayesian analyses of clinical trials, *Statistics in Medicine*, **12**, 1651–1663.

Ibrahim JG and Chen MH (2000). Power prior distributions for regression models, *Statistical Science*, **15**, 46–60.

Inoue LYT, Thall PF, and Berry DA (2002). Seamlessly expanding a randomized phase II trial to phase III, *Biometrics*, **58**, 823-831.

Irony T and Simon R (2006). Application of Bayesian methods to medical device trials, in *Clinical Evaluation of Medical Devices, Principles and Case Studies* (2nd ed), Humana Press, New York, 99–116.

Kadane JB (1995). Prime time for Bayes, *Controlled Clinical Trials*, **16**, 313–318.

Kadane JB (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*, Wiley, New York.

Lewis RJ and Berry DA (1994). Group sequential clinical trials: a classical evaluation of Bayesian decision-theoretic designs, *Journal of the American Statistical Association*, **89**, 1528–1534.

Lipscomb B, Ma G, and Berry DA (2005). Bayesian predictions of final outcomes: Regulatory approval of a spinal implant, *Clinical Trials*, **2**, 325–333.

Louis TA (2005). Introduction Bayesian methods II: fundamental concepts, *Clinical Trials*, **2**, 291–294.

MacEachern SN (2016). Nonparametric Bayesian methods: a gentle introduction and overview, *Communications for Statistical Applications and Methods*, **23**, 445–466.

Malec D (2001). A closer look at combining data among a small number of binomial experiments, *Statistics in Medicine*, **20**, 1811–1824.

Meurer WJ, Lewis RJ, Tagle D, *et al.* (2012). An overview of the adaptive designs accelerating promising trials into treatments (ADAPT-IT) project, *Annals of Emergency Medicine*, **60**, 451–457.

Meurer WJ, Legocki L, Mawocha S, *et al.* (2016). Attitudes and opinions regarding confirmatory adaptive clinical trials: a mixed methods analysis from the Adaptive Designs Accelerating Promising Trials into Treatments (ADAPT-IT) project, *Trials*, **17**, 373.

Müller P, Xu Y, and Thall P (2017). Clinical trial design as a decision problem, *Applied Stochastic Models in Business and Industry*, **33**, 296–301.

Murray TA, Hobbs BP, Lystig TC, and Carlin BP (2014). Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data, *Biometrics*, **70**, 185–191.

Parmar MKB, Spiegelhalter DJ, and Freedman LS (1994). The CHART trials: Bayesian design and monitoring in practice, *Statistics in Medicine*, **13**, 1297–1312.

Parmar MKB, Ungerleider RS, and Simon R (1996). Assessing whether to perform a confirmatory randomized clinical trial, *Journal of the National Cancer Institute*, **88**, 1645–1651.

Pedroza C, Tyson JE, Das A, Laptook A, Bell EF, and Shankaran S (2016). Advantages of Bayesian monitoring methods in deciding whether and when to stop a clinical trial: an example of a neonatal cooling trial, *Trials*, **17**, 335.

Pennello G and Thompson L (2007). Experience with reviewing Bayesian medical device trials, *Journal of Biopharmaceutical Statistics*, **18**, 81–115.

Pibouleau L and Chevret S (2011). Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices, *Journal of Clinical Epidemiology*, **64**, 270–279.

Royall RM (1991). Ethics and statistics in randomized clinical trials (with discussion), *Statistical Science*, **6**, 52–88.

Saville BR, Connor JT, Ayers GD, and Alvarez J (2014). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials, *Clinical Trials*, **11**, 485–493.

Simon R (1991). A decade of progress in statistical methodology for clinical trials, *Statistics in Medicine*, **10**, 1789–1817.

Simon R (2002). Bayesian subset analysis: application to studying treatment-by-gender interactions, *Statistics in Medicine*, **21**, 2909–2916.

Spiegelhalter DJ, Abrams KR, and Myles JP (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, Chichester.

Spiegelhalter DJ, Freedman LS, and Parmar MKB (1994). Bayesian approaches to randomised trials, *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **157**, 357–416.

Spiegelhalter D, Thomas A, Best N, and Lunn D (2003). WinBUGS version 1.4.1 user manual, Retrieved October, 2017, from: http://www.mrc-bsu.cam.ac.uk/wpcontent/uploads/manual14.pdf

Stokes M, Chen F, and Gunes F (2015). An introduction to Bayesian analysis with SAS/STAT software: paper SAS1775-2015, Retrieved October, 2017, from: http://support.sas.com/resources/papers/proceedings15/SAS1775-2015.pdf

Stone GW, Martin JL, de Boer MJ, *et al.* (2009). Effect of supersaturated oxygen delivery on infarct size after percutaneous coronary intervention in acute myocardial infarction, *Circulation: Cardiovascular Interventions*, **2**, 366–375.

Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, and Tomlinson GA (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study, *Journal of Clinical Epidemiology*, **58**, 261–268.

Temple R (2005). How FDA currently makes decisions on clinical studies, *Clinical Trials*, **2**, 276–281.

Viele K, Berry S, Neuenschwander B, *et al.* (2014). Use of historical control data for assessing treatment effects in clinical trials, *Pharmaceutical Statistics*, **13**, 41–54.

Waddingham E, Mt-Isa S, Nixon R, and Ashby D (2015). A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment, *Biometrical Journal*, **58**, 28–42.

Ware JH (1989). Investigating therapies of potentially great benefit: ECMO (with discussion), *Statistical Science*, **4**, 298–340.

Wasserman RL and Lazar NA (2016). The ASA's statement on $p$-values: context, process and purpose, *The American Statistician*, **70**, 129–133.

Wei LJ and Durham S (1978). The randomized play-the-winner rule in medical trials, *Journal of the American Statistical Association*, **73**, 830–843.

Woodcock J (2005). FDA introductory comments: clinical study design and evaluation issues, *Clini-*

*cal Trials*, **2**, 273–275.

Yang X, Thompson L, Chu J, *et al.* (2016). Adaptive design practice at the Center for Devices and Radiological Health (CDRH), January 2007 to May 2013. *Therapeutic Innovation and Regulatory Science*, **50**, 710–717.

*Received November 1, 2017; Revised November 16, 2017; Accepted November 16, 2017*