VIEWPOINT

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1736

Published online 21 January 2016 in Wiley Online Library

Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development

Andrew P. Grieve*

The use of Bayesian approaches in the regulated world of pharmaceutical drug development has not been without its difficulties or its critics. The recent Food and Drug Administration regulatory guidance on the use of Bayesian approaches in device submissions has mandated an investigation into the operating characteristics of Bayesian approaches and has suggested how to make adjustments in order that the proposed approaches are in a sense calibrated. In this paper, I present examples of frequentist calibration of Bayesian procedures and argue that we need not necessarily aim for perfect calibration but should be allowed to use procedures, which are well-calibrated, a position supported by the guidance. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: bayesian statistics; drug development; calibration; operating characteristics; simulation; planning

1. INTRODUCTION

I have been a pharmaceutical statistician for 40 years, and for 35 of them, I have advocated a greater use of Bayesian methods in all aspects of drug development. I work for a company providing services to the pharmaceutical industry, having previously worked for large pharmaceutical companies, and therefore, when addressing the issues of using Bayesian methods in drug development, I look at them through pharmaceutical eyes as well as Bayesian eyes. In the early 1980s, when I began investigating and implementing the use of Bayesian methods in the chemical–pharmaceutical industry, the practical use of Bayesian methods was almost unheard of. This was not because Bayesian ideas were new. Thomas Bayes' original paper appeared posthumously in 1763 [1]; Laplace in 1774 was using posterior distributions in practical applications [2]; in 1875/1876, Helmert and Lüroth developed the t-distribution as the posterior distribution for a population mean [3–6] followed quickly by Edgeworth with uninformative prior distributions on the mean and variance [7]; in 1898, when studying uncertainty in estimating the correlation coefficient, Karl Pearson used a Bayesian approach [8]; Gosset (Student) developed the sampling distribution of the correlation coefficient in 1908 to simplify the calculations required to determine the appropriate posterior distribution [9]. As far as I am aware, the earliest reference to an adaptive clinical trial design was a paper by William Thompson in 1933 [10] based on Bayesian ideas. Despite this early work, there is no evidence to suggest that in the field of clinical trial methodology Bayesian ideas caught on a notable exception being Cornfield's [11,12] work on a Bayesian approach to sequential clinical trials. Things have, however, changed. Both in the United States [13] and Europe [14], there has been significant activity in bringing to the attention of practicing statisticians the advantages of a Bayesian perspective in clinical trials and more generally in medical research [15] and pharmaceutical R&D [16].

In the context of drug development, Bayesian ideas are now formally acceptable to the major international regulatory authorities: 'The use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust' (ICH, Guideline on Statistical Principles in Clinical Trials - E9 [17]). While I found this endorsement somewhat lukewarm, it did at least open the door to the use of Bayesian techniques in drug trials intended to provide evidence to support marketing authorisation. The subsequent Food and Drug Administration (FDA) guidance on using Bayesian methods in clinical trials of medical devices gave greater support but with some caveats [18] of which more later.

If the regulatory environment continues to become more Bayesian-friendly, what other obstacles could prevent their use? Whilst in the past, there were five major obstacles: philosophy, trust, conservatism, tools and expertise, the latter two are less of an issue these days with the development of MCMC methodology and the increasing number of undergraduate and post-graduate courses in statistics and/or biostatistics that cover Bayesian methodology. In this paper, I want to look at an aspect of the remaining obstacles, namely, the requirement to control, in a frequentist sense, the false positive rate of Bayesian procedures, sometimes thought of as calibration. To this end, in this paper, I will look at three major areas of application: monitoring clinical trials, the use of historical information and Bayesian adaptive randomisation designs. In Section 2, I look at the guidelines that have been developed covering the design and reporting of a different study types carried out in biomedical research. These include specific guidelines covering Bayesian approaches developed both by academic groups as well as regulatory authorities. I introduce the idea of calibrated Bayesian procedures, an idea that is missing from the academic guidelines, but central to the regulatory guideline. In Sections 3– 5, I cover three specific applications of Bayesian methods and look at the implications of control of the

Innovation Centre, Icon PLC, Globeside Business Park, Marlow, Buckinghamshire, SL7 1HZ, UK

**Correspondence to: Innovation Centre, Icon PLC, Globeside Business Park, Marlow, Buckinghamshire, SL7 1HZ, UK.*
E-mail: andrew.grieve@ICONPLC.com

false positive rate of these procedures. The applications include Bayesian monitoring of group sequential designs (Section 3), the utilisation of historical control information to enhance clinical trial designs by forming proper prior distributions (Section 4) and Bayesian adaptive randomisation designs (Section 5). In Section 6, I look at the design, analysis and reporting of simulation experiments to control the false positive rate of Bayesian procedures, covering the use of fractional factorial designs and the choice of simulation sample size. I conclude with a discussion section.

2. REPORTING OF BAYESIAN ANALYSES OF CLINICAL TRIALS

In 1996, the Consolidated Standards of Reporting Trials statement was first published [19], combining the independent recommendations of Standardised Reporting of Trials [20] and the Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the biomedical literature [21]. Since then, guidelines have been developed for reporting observational studies in epidemiology (STROBE) [22], genetic association studies (STREGA) [23], systematic reviews and meta-analysis (PRISMA) [24], diagnostic studies (STARD) [25], health economic evaluations (CHEERS) [26], qualitative research (COREQ) [27] and its synthesis (ENTREQ) [28], public health improvement studies (SQUIRE) [29], case reports (CARE) [30], the development of protocols (SPIRIT) [31] and the reporting of statistical analyses and methods (SAMPL) [32]. These guidelines provide important means to improve the quality of the design, conduct and analysis and reporting of biomedical research. With the exception of SAMPL [32], the guidelines do not refer to Bayesian designs, analysis or reporting. In the following sections, we look at the literature on Bayesian analyses and reporting.

2.1. Academic Reporting of Biomedical Research

Four academic groups have proposed guidelines for the reporting of Bayesian analyses: ROBUST [33], BAYESWATCH [34] and BaSiS [35] and SAMPL [32]. All four guidelines cover topics that might be expected including the following: specification and justification of the prior distribution; specification of the statistical model; what software is used; details of the convergence criteria if MCMC techniques are used; how aspects of the posterior inferences are to be summarised; and what sensitivity analyses are to be conducted. One aspect of sensitivity analysis that is rarely considered is to provide readers with the ability to change the prior distribution, and therefore to construct their own posterior distributions. Historically, such Bayesian communication has not been a simple process (Lehmann and Goodman [36]). Two approaches have been proposed. The first is to utilise a family, or community, of priors, which should include a likelihood analysis, essentially a Bayesian analysis of all parameters using a uniform prior distribution, an optimistic prior distribution and a pessimistic prior distribution. This idea was introduced by Hildreth [37] over 50 years ago and brought up to date by Spiegelhalter, Freedman and Parmar [14]. Alternatively, the basic model can be embedded in a hierarchical structure and the sensitivity of posterior conclusions to changes in parameters within the hierarchy explored. This approach is related to Draper's continuous model expansion accounting for uncertainty in model choice [38]. Draper commented 'it is preferable to perform model expansion continuously' in which the alternative is discrete model uncertainty. I am unconvinced that this is generally true as it is often of interest to treat null models discretely rather than as part of a continuum. Grieve [39,40] has used the latter approach to allow readers to input a subjective assessment of competing models in the context of crossover designs. Missing from these reporting guidelines is any idea of presenting the operating characteristics of the chosen Bayesian procedure. This requirement is the subject of the next section.

2.2. Regulatory Bayesian Guideline

In February 2010, the Centre for Devices and Radiological Health (CDRH) of the FDA issued Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials [18]. The guidance has its origins in CDRH's considerable experience in the use of Bayesian approaches in clinical trials of medical devices. The guideline introduces Bayesian ideas in comparison with standard approaches and looks at the benefits and potential pitfalls of their use. It defines prior distributions, likelihoods, posterior distributions and predictive distributions; it covers the more complex issues of exchangeability and the likelihood principle; it describes the planning and analysis of Bayesian clinical trials in some detail.

The section covering technical details is of particular interest as it presents the information to be provided in the trial protocol covering specific issues relating to the Bayesian aspects of the trial. These include the prior information to be used, the criterion for success, the justification of the sample size, operating characteristics (power and type I error), the prior probability of the study claim, the effective sample size and the programme code. When considering the operating characteristics and the type I error, it is worth noting a remark of LeBlond that '…just because Bayesian methods do not inherently rely on the hypothetical repeated trial p-value concept, this does not mean that this metric is not available when Bayesian methods are used. Modern computer simulation allows us to evaluate the Type 1 error rate of any statistical decision-making approach, including Bayesian approaches'. [41]

Whilst most regulatory guidances strongly emphasise strict control of the type I error of procedures, that is not the case in this Bayesian guidance. For example, it recognises the importance of assessing the operating characteristics of Bayesian approaches but states that they 'strive for reasonable control of the type I error', leaving open what should be understood by 'reasonable'. There is a recognition that if we are to use prior information 'that it may be appropriate to control the type I error at a less stringent level than when no prior information is used', a remark we will return to later. Of course the guidance reserves the right on a case-by-case basis to judge the amount by which such control can be relaxed and the degree of discounting of prior information.

This theme is taken up in an extensive section of the guidance dealing with aspects of the simulation of operating characteristics. It includes the remark that if 'the FDA considers the type I error rate of a Bayesian experimental design to be too large; we recommend you modify the design or the model to reduce that rate'. Determination of 'too large' is specific to a submission because 'some sources of type I error inflation .... may be more acceptable than others'. Suggested approaches to decreasing the type I error include increasing the probability criterion for a successful trial; adjusting interim analyses, where appropriate; discounting prior information; increasing the maximum sample size; changing the study termination criterion; and changing the method for determining the type I error. We will take up these themes in subsequent sections when

considering interim monitoring of clinical trials and the use of historical information.

2.3. Calibration of Bayesian Procedures

There are at least three types of calibration in statistics: assay calibration (Hunter and Lamboy [42]), forecasting calibration (Bickel *et al.* [43]), and more recently, the frequentist calibration of Bayesian procedures (Rubin [44]). Rubin argues that frequentist calculations are useful both in understanding and validating Bayesian statements. The former has to do with communicating to lay consumers of Bayesian statistics, Bayesian solutions, which are difficult to understand. The latter is useful 'for making Bayesian statements, scientific', by which he means making them empirically testable, and it is this idea, which is of interest in our context. Rubin defines a Bayesian as being calibrated if the probability statement that he/she makes have the coverage that is being asserted and argues that frequentist investigation of procedures, which are going to be recommended for routine use, are justifiable, if not, absolutely essential. Little in a series of papers has also championed this approach [45–47] There are numerous examples in the literature of the frequentist properties of Bayesian solutions being superior to traditional approaches. Grieve [48] investigated a Bayesian approach to the problem of comparing two normal populations in which we wish to show that both the locations and variances of the two populations are equivalent. Generalising the two one-sided *t*-test approach used in bioequivalence, Bauer and Bauer [49] developed a multi-test procedure based on two pairs of one-sided tests in which equivalence in terms of location and variability is established if all four tests lead to a rejection of their respective null hypotheses. Grieve's [48] approach was to calculate the posterior probability that the differences between means and the ratio of the variances of the two populations lie within the rectangle defined by the null hypotheses proposed by Bauer and Bauer [49]. Based on simulations, Grieve [48] showed that the Bayesian approach was 'well-calibrated', in fact, he was able to show that the results were 'uniformly better than and Bauer and Bauer's corrected test procedure and are generally better than .... the one they recommend'.

3. MONITORING OF CLINICAL TRIALS

A year before the results of the Medical Research Council trial of Streptomycin [50] appeared, the trial that introduced Fisher's ideas of randomisation into modern medical research, Abraham Wald published his book on sequential analysis, which he had developed during his wartime work on military ordnance manufacture [51]. Predating Wald's work, the earliest sequential test procedure in which, in contrast to statistical tests developed in agriculture, the number of observations is not fixed in advance – goes as far back as Dodge and Romig [52] who developed a double sampling procedure. The advantage of these schemes, recognised by Dodge and Romig and by Bartky [53], is that on average, they require fewer observations than traditional single sampling schemes.

The development of similar, sequential ideas in medicine in the 1950s is due to Peter Armitage in the UK and Irwin Bross in the US [54–58]. In 1969, Armitage *et al.* [59] investigated issues associated with multiple tests of accumulating data showing that the use of repeated significance tests at level α increases the overall probability of a type I error above α. Their work led directly to the development of group sequential designs and stopping rules [60–64] and the α–spending function approach of Lan and DeMets [65]. We noted earlier that Cornfield had addressed issues in sequential trials from a Bayesian perspective in the 1960s. From 1983 onwards, there has been a stream of authors who have addressed monitoring of clinical trials from a Bayesian perspective [66–78]. In the next section, we look at one of these approaches in depth.

3.1. A General Structure of Bayesian Monitoring

We follow the development of Grossman *et al.* [78]. That is, we assume that a clinical trial is being run to compare two treatments (A and B) in which T interim analyses are planned after each block of n/T of patients per group and that a maximum of n patients in each group will be available. For each block of patients, an estimate $d_i (i = 1, .., T)$ is available where the $d_i \sim N(\delta T \sigma_\delta^2 / n)$ are independently distributed from each other; δ is the expected treatment effect; $\sigma_\delta^2 = 2\sigma^2$ and σ^2 is known. Further, we assume there is prior information available about δ, equivalent to *fn* patients per group with prior mean δ_0. Under these assumptions at the end of the tth block, the posterior distribution for δ is

$$p(\delta \mid d_1, \ldots, d_t) \sim N\left(d_t^*, \sigma_\delta^2 / n_t\right)$$

where

$$d_t^* = \frac{nt\left(d_1 + d_1 + \ldots + d_t\right)/T + fn\delta_0}{\frac{nt}{T} + fn} = \frac{tD_t + Tf\delta_0}{t + fT}$$

and

$$n_t = \frac{nt}{T} + fn$$

At each interim, the posterior probability that δ is greater than a pre-specified value δ_C is determined, and the decision to the study is made if this probability is greater than a minimum value $1 - \psi_t$, that is

$$P(\delta > \delta_C | D_t) = 1 - \mathbf{\Phi}\left[\frac{\delta_c(tD_t) + Tf\delta_0/(t + fT)}{\sigma\delta/\sqrt{nt/T + fn}}\right] > 1 - \Psi_t \quad (1)$$

implying that

$$\psi_t > \mathbf{\Phi}\left[\frac{\delta_C - \left(tD_t + Tf\delta_0\right)/(t + fT)}{\sigma_\delta/\sqrt{nt/T + fn}}\right]$$

which in turn requires that

$$D_t > \frac{\sqrt{T}\sqrt{t + fT}Z_{\psi_t}\sigma_\delta}{t\sqrt{n}} + \delta_C\frac{t + fT}{t} - \frac{Tf\delta_0}{t} \quad (2)$$

For the moment, we assume that the prior information is real, and fixed, then if we are interested in controlling the experiment wise type I error, we have a number of options to consider. First, we could leave δ_C fixed and choose to have different probabilities, ψ_i, at each interim. Second, we could choose a common value for ψ and adjust the common δ_C. Third, we could adjust the common probability ψ. Finally, keeping ψ fixed, we could choose a different cutoff δ_{C_i} at each interim i $(i = 1, \ldots T)$. We investigate each of these in turn.

3.2. Example 1: $T = t = 1, \delta_C = 0, \psi_t = \psi$ (for all t).

This example corresponds to a single analysis, in other words, no interims. Under these assumptions, the stopping rule (2) becomes

$$D > -\frac{\sqrt{1+f}Z_\psi \sigma_\delta}{\sqrt{n}} - f\delta_0 \quad (3)$$

and we can investigate the frequency properties of the rule by noting that under the null hypothesis $D \sim N\left(0, \sigma_\delta^2/n\right)$ implying that

$$Prob\left[D > -\frac{\sqrt{1+f}Z_\psi \sigma_\delta}{\sqrt{n}} - f\delta_0\right] = 1 - \Phi\left(-Z_\psi\sqrt{1+f} - \frac{\sqrt{f}\sqrt{nf}\delta_0}{\sigma_\delta}\right) \quad (4)$$

To control this probability at the $1-\alpha/2$ level requires that

$$Z_{1-\psi} = \frac{Z_{1-\alpha/2} + \sqrt{f}Z_0}{\sqrt{1+f}} \quad (5)$$

where $Z_0 = \sqrt{fn}\delta_0/\sigma_\delta$ is the prior standardised effect size. Figure 1 **A**) displays contours of ψ necessary to control the one-sided type I error at 2.5% for varying f and Z_0. If the standardised effect size is large then ψ must be considerably reduced to control the type I error. In contrast, for small Z_0 and large f, the nominal level ψ may be relaxed. This is intuitively correct because in this case, the prior distribution is providing a significant penalty towards zero.

Another example of a single analysis is given by $T = t = 1$, $\psi_t = \psi$ (for all t) for which a similar process leads to a requirement that the standardised Bayesian decision criterion $Z_C = \delta_C\sqrt{n}/\sigma_\delta$ satisfies the following relationship

$$Z_C = \frac{\sqrt{f}Z_0 - \left(\sqrt{1+f} - 1\right)Z_{0.975}}{1+f}$$

in order to control the type I error. The contours of Z_C necessary to control the one-sided type I error at 2.5% for varying f and Z_0 can also be displayed. Whichever approach is used, we are effectively discounting the prior information. To see this, consider again Case 1 and substitute (5) into (3) to give

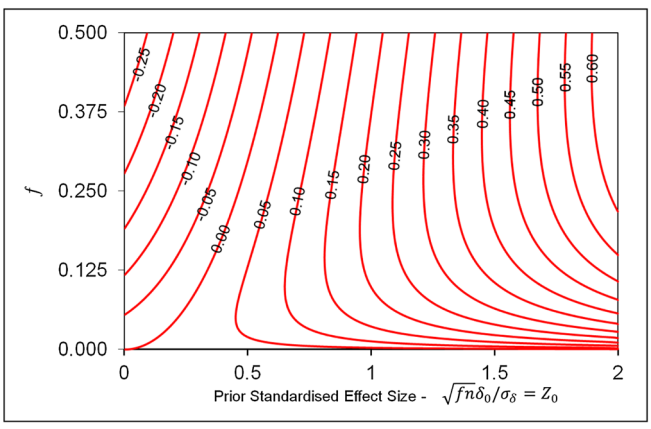

Figure 1. Bayesian decision rules giving a one-sided type I error of 2.5%

$$D > \frac{\sigma_\delta Z_{0.975}}{\sqrt{n}}$$

This is the standard, frequentist decision rule. In other words, in these 2 cases, requiring strict control of the type I error results in 100% discounting of the prior information. A similar result is shown in an appendix to Pennello and Thompson [79]. This result is important in the context of the remark in the FDA's Bayesian guidance that 'it may be appropriate to control the type I error at a less stringent level than when no prior information is used'. If when using prior information in a single analysis we require absolute control of the type I error, and such procedures we might term 'perfectly-calibrated', then we must throw away any prior information. I would argue that the FDA's remark is recognition of this phenomenon and an endorsement of a less strict control of type I error, and such procedures we will term 'well-calibrated'.

3.3. Example 2: $\psi_t = 0025, \delta_C = 0, \delta_0 = 0$ (Sceptical Prior Distribution).

Freedman, Spiegelhalter and Parmar [14] describe the formal construction of a sceptical prior distribution in the following way. Consider a normal distribution, centred at zero and with a small probability, $.\gamma$, of achieving the alternative hypothesis δ_A. Then, if the prior is as previously defined, but with $\delta_0 = 0$, the following relationship will hold

$$\delta_A = -\frac{\sigma_\delta Z_{1-\gamma}}{\sqrt{fn}}$$

If the trial has been designed with size α and power $1-\beta$ to detect the alternative hypothesis δ_A, from which the sample size can be determined from

$$n = \frac{\sigma_\delta^2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\delta_A^2}$$

then substitution gives

$$f = \left(\frac{Z_\gamma}{Z_{1-\alpha/2} + Z_{1-\beta}}\right)^2$$

In what follows, we simply assume that the sceptical prior distribution is $\delta \sim N\left(0, \sigma_\delta^2/(fn)\right)$. Setting $\psi_t = 0.025$, $\delta_C = 0$, $\delta_0 = 0$ in (1) gives

$$Prob\ (\delta > 0 \mid D_t) = \Phi\left(\frac{\sqrt{nt}D_t}{\sigma_\delta\sqrt{T}}\sqrt{\frac{t}{t+fT}}\right) > 0.975$$

which is equivalent to increasing the standard, unadjusted critical region by a factor

$$\sqrt{\frac{t+fT}{t}}$$

and in this context, Grossman *et al.* [78] refer to f as the 'handicap'. The frequentist properties of this 'handicapping' are not as simply derivable as they were in the previous two cases. For T=2 – a single interim – the frequentist type I error can be calculated using a bivariate normal probability function, for example, the SAS function PROBNRM. For T >3, Grossman *et al.* [78] use simulation to determine the handicap f that controls the two-sided type I error at 5% and 1% and to get sufficient accuracy simulated

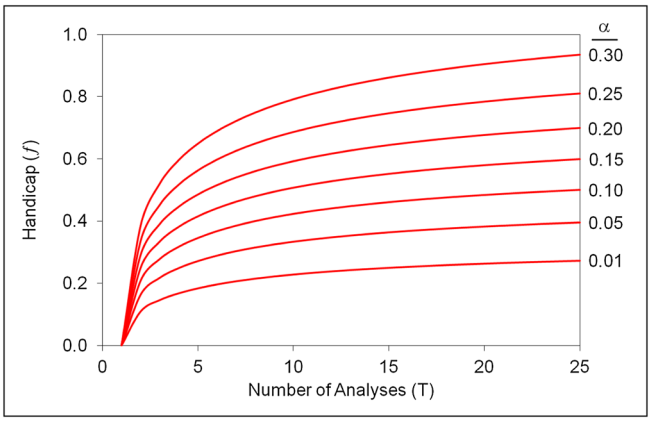

Figure 2. Handicaps (f) to control the two-sided α for up to 25 analyses.

20 000 000 trials. As an alternative, a combination of the algorithm derived by Armitage *et al.* [59] and a Newton–Raphson algorithm can be used. SAS PROC IML has subroutines, which perform the appropriate integrations or a modified version of the FORTRAN programme described by Reboussin *et al.* [80] can be used, the approach used here. Figure 2 displays the handicaps required to control the two-sided type I error at 1,5(5) 30% for up to 25 interim analyses.

The extensive results in this figure support the conclusions of Grossman *et al.* [78]. For example, the shallow slope as T increases for T > 5, independent of the type I error, supports their conclusion that this approach is robust in the sense if it were planned to conduct 4 or 5 interims, there is little lost in terms of the type I and type II errors if 8–10 are actually carried out.

4. USE OF HISTORICAL INFORMATION

At a meeting around the turn of the millennium hearing, Greg Campbell of FDA's CDRH division commented that there was a concern that the use of 'subjective priors may allow unscrupulous companies and/or statisticians to attempt to pull the wool over the regulators' eyes.' I think this is very unlikely, and if it were that easy, we probably need a different set of regulators! Nonetheless, the 2010 FDA Bayesian Guideline is clear: '…Bayesian methods are usually less controversial when the prior information is based on empirical evidence such as data from clinical trials. However, Bayesian methods can be controversial when the prior information is based mainly on personal opinion (often derived by elicitation from "experts")' [18]. The strength of the regulatory message has been clearly received by researchers, even subjective Bayesians: 'To present a Bayesian analysis in which the company's own prior beliefs are used to augment the trial data will, in general, not be acceptable to an external agency' (O'Hagan and Stevens [81]). Even if it is accepted that sponsors in general have no 'unscrupulous' hidden agendas in the use of Bayesian methods nonetheless, as already noted, Bayesian statisticians need to calibrate their methods, and I cannot imagine that a method which it can be shown unduly inflates the false positive rates would be acceptable.

Secondly, I agree with Professor Stephen Senn that 'There can be few areas where the discipline of statistics is conducted with greater discipline' [82] than in the pharmaceutical industry. Part of this pharmaceutical discipline is documentation, a fundamental creed of pharmaceutical statistics, and it will be no different with Bayesian methods. The prior distributions will need to be specified in the protocol, as will utility functions if required. They will need to be justified. They will not be able to be changed. In my mind, it is unlikely that undocumented, subjective prior distributions will be allowed.

Here are two examples of the use of historical information as priors in clinical trials, one theoretical and one real.

4.1. Example 1 (Cont.): $T = t = 1, = \delta_C = 0, \psi_t = \psi$.

This example had no interims and supposing that we fixed the Bayesian decision rule at $\alpha/2$ which would correspond, using an uninformative prior, to a one-sided test of significance at the $\alpha/2$ level, then from (4), the probability of a positive trial under the null hypothesis is

$$1 - \Phi\left(-Z_{\alpha/2}\sqrt{1+f} - Z_0\sqrt{f}\right) \tag{6}$$

where Z_0 is the prior standardised effect size. For a given f and Z_0, (6) can be used to calculate the effective type I error. We noted before that if Z_0 is small and f is large, the nominal level ψ could be relaxed and still control the type I error at the required level. This was a consequence in such cases of the prior distribution providing a significant penalty towards the null. One way of seeing this is to ask, for a given value of f, what value of Z_0 when substituted in (6) returns a value of $\alpha/2$. This value for the prior standardised effect size provides a break-even point, if the sponsor's chosen standardised prior distribution is smaller than the break-even, the resulting test is conservative. In contrast, if the sponsor's chosen standardised prior is larger than the break-even, the resulting test is anti-conservative. Table I illustrates the size of these break-even points for prior distributions with information-content up to 25% of the trial itself and for nominal one-sided type I error rates of 0.005, 0.025, 0.050 and 0.100.

4.2. Example 4: A Bayesian Adaptive Dose-Ranging Clinical Trial.

Some of the details of this example have been changed to preserve client confidentiality, but the essential features have been retained. This study was run as a multi-centre, randomised, Bayesian, adaptive design to evaluate the efficacy, safety and tolerability of doses of a new drug in a post-operative condition. The primary efficacy objective of the study was to identify the lowest dose of the drug that exceeded an increase in mean response of 0.8 units compared with control. The primary efficacy objective of the study was to identify the lowest dose of the drug that exceeded an increase in mean response of 0.8 units compared to control. A dose of the drug was to be declared successful if

Table I. Standardised prior mean break-even points for a range of prior sample size fraction (f) and type I error ($\alpha/2$).

		Prior sample size fraction (f).					
		0.01	0.05	0.1	0.15	0.2	0.25
Type I	0.005	0.128	0.284	0.398	0.481	0.55	0.608
error	0.025	0.098	0.216	0.303	0.366	0.418	0.463
($\alpha/2$)	0.05	0.082	0.182	0.254	0.307	0.351	0.388
	0.1	0.064	0.142	0.198	0.24	0.274	0.303

the posterior probability that the mean response of that dose compared to control was greater than 0.8 units was greater than 30%.. The cut-off of 30% was chosen to control the type I error of the complete null scenario, that is, when the mean response in all dose arms was the same as the control arm, at 5%. Based on approximately 3600 control-treated patients, a prior distribution for the mean response in the control arm was developed with a discounting of the information down to 40 patients. The study was designed to have a maximum of 200 completing patients with an interim analysis after 20 patients per treatment arm to allow the study to terminate for futility or to allocate the remaining patients to one of three doses depending on pre-specified decision criteria.

A series of simulation scenarios were run to assess the type I error and the power not only for the planned adaptive design but also for a fixed design.

Figure 3 displays two of the scenarios. Scenario A is the complete null scenario in which the mean responses in the control and treatment arms were all the same as the historical control mean (a value of 2 units). Scenario B is also a null scenario in the sense that none of the treatment arms have a mean response exceeding an increase of 0.8 units. In the case of Scenario A, the empirical type I error of the adaptive design was 4.7%, confirming the choice of 30% for the posterior probability cut-off to control the type I error. In contrast, the empirical type I error for Scenario B was > 95%. Why? The expected posterior mean of the mean response in the control arm is 2.2 units, and therefore, both of the highest doses are likely to meet the effectiveness criterion of 0.8 units. This arises because the low prior control mean reduces the posterior control mean, and therefore overestimates the dose effect.

The issues raised in this section are of increasing interest in a drug development context. Neuenschwander [83] have argued that historic control information is always important and relevant in the prospective design of clinical trials but also has a role to play in the analysis of the trials, and their approach has been applied to develop a prior for a proof-of-concept trial in chronic obstructive pulmonary disease [84]. Viele *et al.* [85] review methods for the borrowing of historical information including methods for checking prior/data compatibility. More recently, Albers and Lee [86] have investigated an approach to calibrating the prior distribution for a normal model with conjugate prior by adjusting the hyper-parameters.

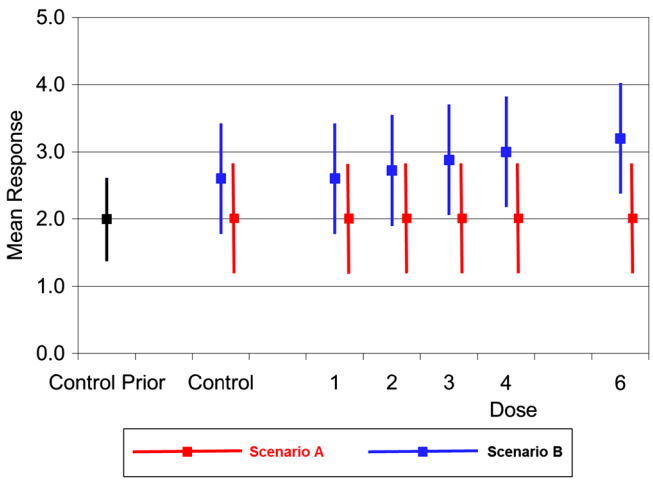

Figure 3. Null SimulationScenarios: a Bayesian Adaptive Dose Selection Design.

5. BAYESIAN ADAPTIVE RANDOMISATION DESIGNS

The ASTIN trial [87], designed to determine the ED_{95} of NIF, a glycoprotein, originally derived from the canine hookworm, in the treatment of acute stroke, entailed a change in the allocation ratio between the 16 arms of the study using a complex Bayesian approach involving a smoothing algorithm, a longitudinal model to predict the long-term outcome from a series of short-term outcomes and a predictive approach to choosing the next dose. Such complexity is not a requirement of adaptation involving allocation changes in a Bayesian framework. Thall and Wathen [88] introduce a Bayesian approach with its roots in Thompson's original proposal [10].

Suppose we are designing a trial to compare the response rate π_A of a control with the response rate π_B of an active treatment. From a Bayesian perspective, the posterior distribution of the parameters π_A and π_B contains all the relevant information to make inferences about the relative efficacy of the treatments. For example, the posterior mean of $\pi_A - \pi_B$ provides a point estimate of the absolute rate reduction of the active compared with placebo; the posterior mean of π_A/π_B plays the same role for the relative risk, and any other summary measure of treatment effectiveness can be similarly treated. Corresponding to each measure, a credible interval, so-called to differentiate it from a confidence interval, provides a measure of our certainty, or uncertainty, in the estimate. Alternatively, probabilities of direct interest can be used to, for example, determine the posterior probability that the active response rate is greater than that of control $P(\pi_A < \pi_B|\mathbf{X})$, and it is this latter quantity that forms the basis of the approach suggested by Thall and Wathen [88]. Suppose, at a given point during the course of the trial, that n_A patients have been randomised to control of whom r_A respond and that n_B patients have been randomised to active of whom r_B respond. Further, if the prior distributions for the response probabilities π_A and π_B are Beta $(\alpha_A\beta_A)$ and Beta $(\alpha_B\beta_B)$, respectively, then their posterior distribution $p(\pi_A, \pi_A|\boldsymbol{X})$ is

$$\frac{\pi_A^{r_A+\alpha_A-1}\,(1-\pi_A)^{n_A-r_A+\beta_A-1}\,\pi_B^{r_B+\alpha_B-1}\,(1-\pi_B)^{n_B-r_B+\beta_B-1}}{B\,(r_A+\alpha_A, n_A-r_A+\beta_A)\,B\,(r_B+\alpha_B, n_B-r_B+\beta_B)}$$

The posterior probability that the response rate on active is greater than on control given by $P(\pi_A < \pi_b|\mathbf{X}) = \int_0^1\int_0^{\pi_A} p(\pi_A, \pi_B|\mathbf{X})d\pi_B\, d\pi_A$. Using the relationship between the incomplete beta function and the binomial distribution, this can be written as the cumulative hypergeometric distribution

$$\sum_{s=max(r_B+\alpha_B-r_A-\alpha_A,0)}^{r_B+\alpha_B-1} \frac{\binom{r_A+r_B+\alpha_A+\alpha_B-1}{s}\binom{n_A+n_B-r_A-r_B+\beta_A+\beta_B-1}{n_B+\alpha_B+\beta_B-1-s}}{\binom{n_A+n_B+\alpha_A+\alpha_B+\beta_A+\beta_B-2}{n_A+\alpha_A+\beta_A-1}}$$

(Altham [89]). In the uniform case (α_A=α_B=β_A=β_B= 1), this probability was first described by Liebermeister [90] (see also Seneta [91], Ineichen [92] and Seneta *et al.* [93]) and in German statistical circles as an alternative to Fisher's exact test. Thompson [94] also considered the uniform case and showed that

$P(\pi_A < \pi_B|\mathbf{X})$ can be written as

$$\sum_{k=0}^{min(b-1,W-w)} \frac{\binom{W}{w+k}\binom{B}{b-1-k}}{\binom{W+B}{w+b-1}}$$

where $W = n_A + 1,\ B = n_B + 1,\ w = n_A - r_A + 1$ and b= $n_B - r_B$ +1. This is the probability under sampling without replacement from a mixture of W white balls and B black balls that we will get w white balls before b black balls. This interpretation allowed Thompson to develop a randomisation 'machine' based on $P\pi_A < \pi_B|X$. He made a box in the form of an isosceles triangle in which he put $n_A + 1$ white and $n_B + 1$ black balls. The balls were then shuffled and allowed to line up along the hypotenuse. If $n_A - r_A + 1$ white balls were found before $n_B - r_B + 1$ black balls, the next subject was allocated to treatment A, otherwise to treatment B. Figure 4 illustrates four cases giving rise to randomisation to treatments B, B, A and B, respectively. Thompson used his 'machine' to simulate the conduct of his Bayesian adaptive design.

Early on in such a trial, there will be little information about the parameters that we are interested in, a consequence is that $P(\pi_A < \pi_B|\mathbf{X})$ is likely to be highly variable, and therefore, there is a considerable risk that it will unbalance the samples in favour of the inferior treatment. To overcome this, Thall and Whalen [88] have proposed the following simple modification in which the probability of randomising a patient to active, p_{ran} is

$$p_{ran} = \frac{P(\pi_A < \pi_B|\mathbf{X})^C}{P(\pi_A < \pi_B|\mathbf{X})^C + [1 - P(\pi_A < \pi_B|\mathbf{X})]^C}$$

where C is arbitrary.

Figure 5 shows the impact of choice of C on the randomisation probability p_{ran}. The choice $C = 0$ corresponds to traditional randomisation with no change in the randomisation probability during the course of the trial. The choice $C = 1$ corresponds to the randomisation probability being $P(\pi_A < \pi_B|\mathbf{X})$, which, as we remarked, is unstable. For small values of C, there is little change from straightforward randomisation, and therefore, it would be sensible to start the study with a small value of C and increase it as $P(\pi_A < \pi_B|\mathbf{X})$ becomes more stable.

With this in mind, Thall and Wathen propose using $C = n/(2N)$ in which n is the current sample size and N is the trial's maximum sample size. Figure 6 illustrates the difference between using this approach and a value $C = 1$. Based on the design considered by Thall and Wathen, a maximum of 200 patients were to be allocated to either control or active. 10^5 adaptive clinical trials were simulated using $\pi_A = 0.25$ and $\pi_B = 0.25\ (0.05)\ 0.45$, and the adaptive probabilities were calculated for each simulated trial. The figure displays the mean randomisation probability to B as a function of the patient number as a function of C and π_B. There is an apparent advantage of the choice $C = 1$ because it reacts more quickly to the accumulating data than does the choice $C = n/(2N)$, and the rate of reaction increases as the response rate of the active treatment increases. For example, for the case $\pi_B = 0.45$, the randomization probability rising rapidly to, on average 0.80 by patient number 40 for $C = 1$ with the corresponding randomisation probability being approximately 0.55 for $C = n/(2N)$. On the negative side, the randomisation probabilities for $C = 1$ are much more variable. Figure 7 shows the standard deviations of the randomisation probabilities in the case $\pi_A = 0.25$, $\pi_B = 0.45$ again based on 10^5 simulations. At patient number 40, the standard deviation of the probabilities is approximately 0.23 $C = 1$ whilst for $C = n/(2N)$. It is less than a quarter of this value. The use of $C = n/(2N)$ is recommended because it provides a substantial imbalance when imbalance is appropriate, has little risk of unbalancing in the wrong direction and maintains virtually identical power and average overall

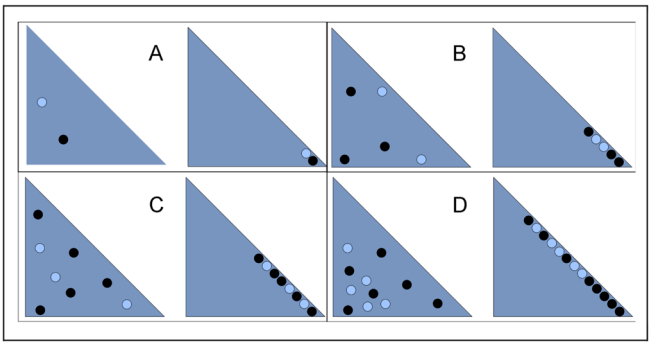

Figure 4. Determination of posterior probabilities using Thompson's machine. A) rA=0, nA=0, rB=0, nB=0, W=1, B=1, w=1, b=1; B) rA=0, nA=1, rB=2, nB=2, W=2, B=3, w=2, b=1; C) rA=1, nA=2, rB=1, nB=4, W=3, B=5, w=2, b=4; D) rA=2, nA=4, rB=4, nB=7, W=5, B=8, w=3, b=4

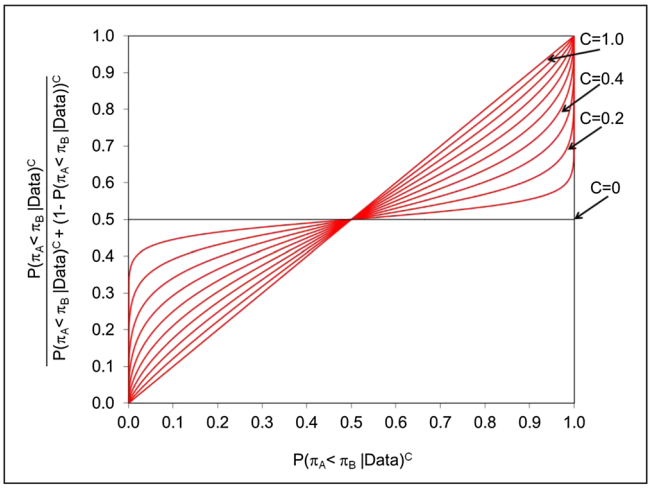

Figure 5. Thall and Wathen transformation of posterior probabilities.

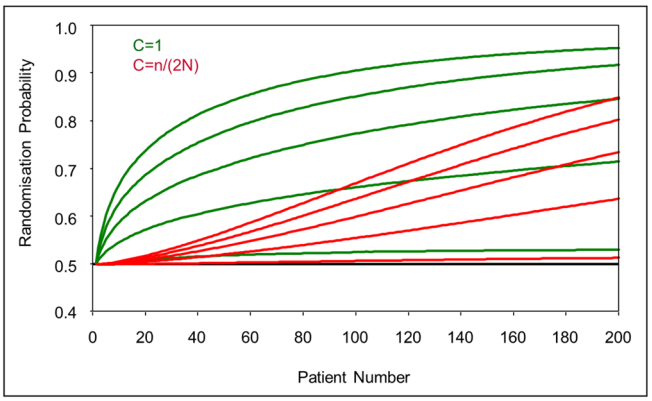

Figure 6. Randomisation probabilities (10^5 simulations) $\pi_A = 0.25, \pi_B = 0.25(0.05)0.45$.

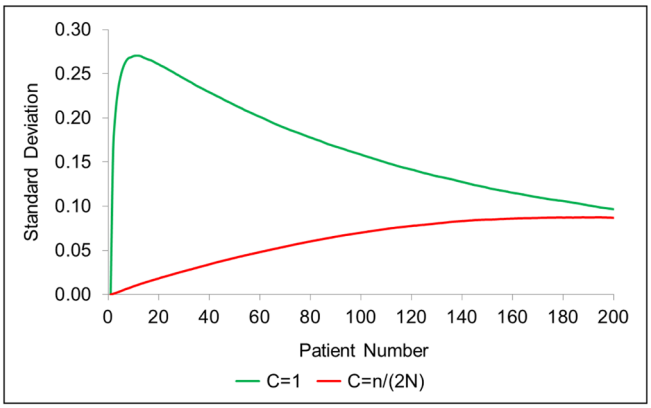

Figure 7. Standard deviation (SD) of randomisation probabilities (10^5 simulations), $\pi_A = 0.25, \pi_B = 0.45$.

sample size compared with conventional randomisation and is relatively stable.

Thall and Wathen [88] chose the following stopping rule:

- If $P(\pi_A < \pi_B|\mathbf{X})$ >0.99 stop and "choose" B
- If $P(\pi_A < \pi_B|\mathbf{X})$ <0.01 stop and "choose" A (futility)

We should expect that such a design is likely to have an elevated type I error because the study essentially has 200 interim analyses, one after every patient. In order to investigate the operating characteristics of the aforementioned stopping rule, we again resort to simulation. An added complication is that the control rate is a nuisance factor. Figure 8(A) displays the outcome of a simulation experiment for a range of null hypotheses $\pi_A = \pi_B \in (0.05, 0.95)$ based on 10^6 simulations and shows the estimated type I error as a function of the control rate. The estimated type error rate range from approximately 0.015 to 0.08 – anti-conservative to conservative.

There are different possibilities for adjusting the type I errors. The first would be to reduce the number of interims. For example, we could consider only calculating the posterior probabilities $P(\pi_A < \pi_B|\mathbf{X})$ after every 10 or 20 patients rather than after every patient. What we investigate here is adjusting the cut-off value 0.99 in the stopping rule. This inevitably will be a function of the nuisance parameter, the control rate. Figure 8(B) illustrates this relationship showing the critical value necessary to control the type I error at 0.025 one-sided as a function of the control rate. The maximum value of this critical value, 0.99747, occurs approximately when the control rate is 50% and is relatively constant when the control rate is in the range 0.15 to 0.8. If we fix the critical value at this maximum for all control rates and re-calculate the type I error, we get the results shown in Figure 8(C). The figure shows that when the control rate is again in the range 0.15 to 0.8, the type rate error is between 0.02 and 0.025. Of course, this value is only the maximum value for the case of 200 patients; if we use another total number of patients, this critical value needs to be recalculated. Figure 8(D) illustrates how the maximum value increases as a function of the maximum sample size. The slope of the relationship is steep up to a sample size of 50 and then reduces, but still stays positive.

Examples of the use of this type of design have been reported by Maki *et al.* [95] and Giles *et al.* [96] and whilst they have an intuitive appeal, their use is not without criticism. We have

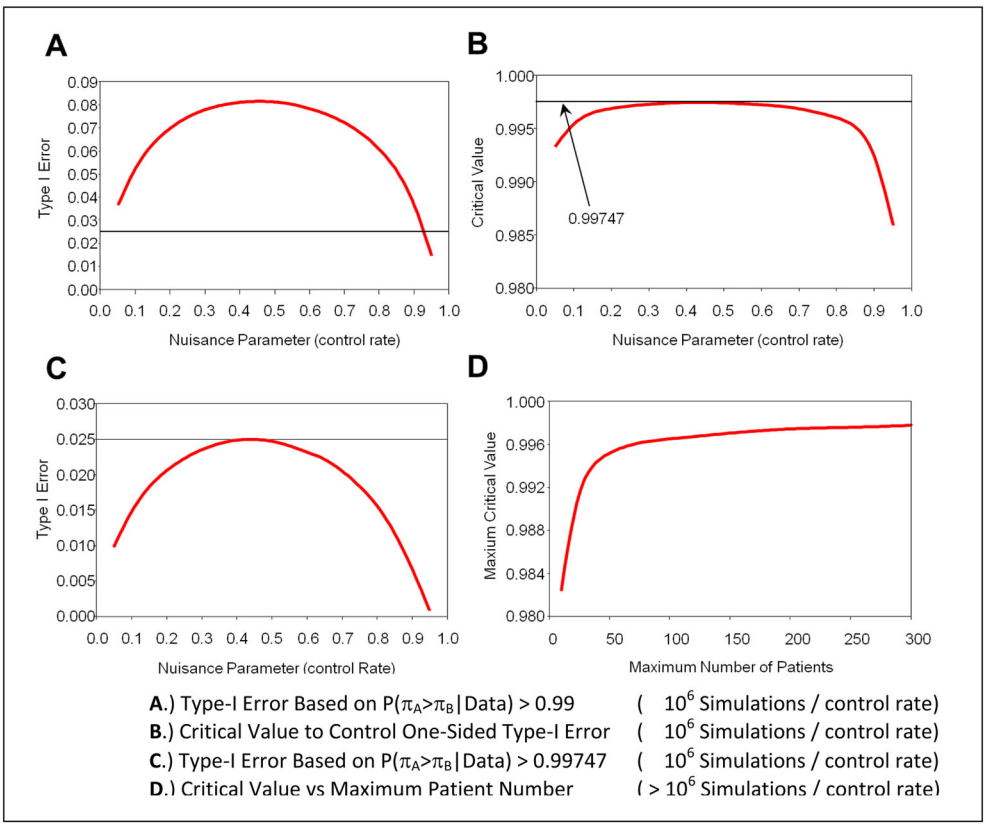

Figure 8. A.) Type I error based on $P\pi_A > \pi_B|\text{Data}) > 0.99$ (10^6 Simulations / control rate) **B.)** Critical value to control one-sided type I error (10^6 Simulations / control rate) **C.)** Type I error based on $P(\pi_A > \pi_B|\text{Data}) > 0.99747$ (10^6 Simulations / control rate) **D.)** Critical value versus maximum patient number (> 10^6 Simulations / control rate)

concentrated this investigation on the type I error as there is an asymmetry in the handling of type I and type II errors in drug regulation and a design that is calibrated tends to mean a design that controls the type I error. There are, of course, other aspects that are of interest.

The proportion of patients who are allocated to the more effective treatment is important because these designs were introduced from an ethical perspective to maximise this metric; the overall response rate across all patients in the trial being a second, and the overall 'power' of the design being a third. Korn and Freidlin [97] investigate aspects of the Thall and Wathen design and conclude adaptive randomisation 'is inferior to 1:1 randomization in terms of acquiring information for the general clinical community and offers modest-to-no bene?ts to the patients in the trial, even assuming the best-case scenario of an immediate binary outcome'. In commenting on their results, Berry [98] has argued that that the greatest benefits of adaptive randomisation are likely to accrue for trials with more than two arms. I agree with this conclusion. Similar to the group sequential case in Section 3 when T=1, greater complexity gives more scope for Bayesian designs. More recently, the original authors have revisited these designs [99] and whilst they give some support for their use in early phase trials, such as phase I-II dose-finding, it is not unconditional and they recommend a number of strategies to mitigate some of the issues with their use. However, this support does not carry over to later-phase trials, indeed their final conclusion is for 'RCTs where treatment comparison is the primary scientific goal, it appears that in most cases designs with fixed randomization probabilities and group sequential decision rules are preferable to AR scientifically, ethically, and logistically'.

6. PLANNING AND CONDUCTING SIMULATION STUDIES

The operating characteristics of the ASTIN study, referred to previously, were determined by simulation. In planning these simulations, the statistical team identified a large number (>2000) simulation scenarios, which were of interest. The amount of time necessary to run these simulations would have delayed the start of the trial. In these circumstances, it is important to consider more efficient ways to conduct the simulations. In reality, the majority of the individual scenarios that were considered to be of interest could be expressed in terms of six individual factors, four of which had two levels, and two had four levels. This factorial structure opens the possibility of utilising experimental design to efficiently explore the design of the simulation study.

6.1. Experimental Design in Planning Simulation Experiments

Giovagnoli and Zagoraiou [100] consider the design of 'virtual experiments', in our terminology simulation experiments and conclude that their design needs to be 'efficient so as to gather information in the best possible way'. They point out that in simulation experiments, we are able to experiment on a far wider design space than in real experiments. In particular in simulation experiments, it is possible to arbitrarily increase the numbers of factors and to increase the number of levels of each factor to ensure those that are clearly of interest to investigate. Statisticians designing such multi-factorial experiments need to remember the basic tenet of good experimental design, which is not to investigate one factor at a time because if we do that, we are likely to hide interactions between factors. If we contemplate investigating a large number of parameters, then, as Holford *et al.* [101] remark in this situation 'it is even more critical, than in an actual experiment, to capitalise on ideas from the statistical sub-field of experimental design with factorial experiments'; in particular, we should consider the use of fractional factorial designs.

This is the approach taken in the case of ASTIN. A fractional factorial design, a $1/4$ replicate of the basic $2^4 4^2$ factorial experiment, was used. Such a design results in the aliasing of factors and their interactions. The design used resulted in a pattern of aliasing such that no main effects were aliased with any 2-factor or 3-factor interactions. Second, no 2-factor interactions were aliased with each other. These aliasing structures are important in practice when considering how to set up such a design we need to be aware of the importance of individual factors. For example, if one particular factor or the interaction of factors is of particular importance then we can allocate the codes to the factors in such a way to ensure that these factors and interactions are not aliased.

6.2. Analysis and Reporting of Simulation Experiments

Once we have accepted the principle that simulation investigations are experiments, we can utilise standard analytic armoury associated with such designs to analyse and report the results. For designs with replication standard analysis of variance is used to analyse each metric separately. For example in ASTIN, separate analysis of variance were carried out for bias, defined as the difference between the estimated effect compared with placebo at the ED_{95} compared with the true effect; the power of the study; the accuracy of the estimated ED_{95}; the accuracy of the estimate of variability; and the proportion of patients allocated to placebo. For non-replicated designs, there is an issue that models are saturated, which leaves no degrees of freedom to estimate the error and the appropriate standard deviation τ against which to test the estimated contrasts, or effects. In such cases available, techniques rely on the principle of 'effect sparsity', which assumes that we can expect only a small number of contrasts to be non-zero. Three approaches have been proposed. Daniel [102] suggested the use of a half-normal plot of the estimated contrasts associated with the contrasts that define the fractional factorial design. The 'effect sparsity' principle suggests that non-zero contrasts will appear as outliers on the plot. The subjective nature of the interpretation of such plots has been criticised, and in response to Box and Meyer [103], developed a more formal approach based on the determination of the posterior probability that each individual contrast is non-zero, by modelling the individual contrasts as a sample from a scale-contaminated normal distribution. The prior information depends on two parameters, the probability that an individual contrast is effective (α) and an inflation factor (k) applied to the standard deviation caused by a non-zero effect. Box and Meyer suggest values of α =0.2 and k=10 from an empirical analysis of published examples. The posterior probabilities are plotted as a bar graph with guidelines to aid in interpretation. A third approach is due to Lenth [104], which first determines an estimate of τ and uses this to re-estimate τ based on a trimmed median procedure. The error estimate, called a pseudo-standard error, uses the assumption that approximately 20% of the effects will be active, and the rest of the effect estimates are zero-mean jointly normal random variables. Simulation results are often reported in aggregate form, and whilst this can answer many questions of interest, it cannot answer them all. Grieve and Krams [87] displayed the posterior distribution of the ED_{95} derived from

the data collected in the ASTIN trial, a distribution that is bimodal. The outcome is consistent with a flat dose response curve and because in such circumstances the dose-curve response curve will consist of a series of random bumps, some of which will disappear through smoothing. After the event, it was possible to identify individual clinical trial simulations within the fractional factorial experiment, which exhibited a similar posterior to that seen in ASTIN. If we concentrate solely on aggregated properties, we may miss important characteristics, which are of more than just academic interest in the running of a clinical trial.

6.3. Proof by Simulation

Posch *et al.* [105] have investigated adaptive designs with treatment selection at an interim and sample size re-estimation. Their interest was in controlling the family wise error rate in a strong sense, meaning under all possible configurations of hypotheses Their approach was to compute the type I error rate for a traditional test-statistic for a pre-specified adaptation rule, with the computation being based on simulation, and they were interested in understanding the limitations of such a simulation-based approach. One of their conclusions was that researchers have to be careful with the assumptions behind the simulations. In particular, they raise the point that the choice of seed for simulation has an impact on the estimated type I error. Their argument is as follows. Because Monte Carlo estimates of the type I error rate are not exact, in the sense that they are subject to random error, the choice of a particular seed for a random number generator can impact on the estimate of the type I error rate estimate. A consequence of which is that if researchers use a strategy of searching for a seed that minimises the estimated type I error rate, this will inevitably lead to an underestimation of type I error rate. As an example, they consider a case in which the type I error rate is estimated in a simulation experiment by the percentage of significant results among 10^4 simulated Randomized clinical trial (RCTs). Suppose the true type I error rate is 0.026, the chance of observing an empirical type I error rate less than 0.025 is

$$p = \sum_{i=0}^{0.025N-1} \binom{N}{i} 0.026^i 0.974^{N-i}$$

so that the average run length (ARL) that is required to find a seed giving an empirical type I error less than 0.025 is ARL=1/p. For $N = 10000$ simulations $p = 0.256$ and ARL = 4 whilst for $N = 100000$ simulations $p = 0.022$ and ARL = 45. This raises the possibility that researchers could practically search for an advantageous seed giving an acceptable empirical type I error. The fallacy in this argument is in accepting that a simulation sample size of 10000 or even 100000 is sufficiently large for the purpose. If the purpose is to discriminate between type I error rates of 0.025 and 0.026 then these sample sizes are inadequate because such sample sizes are sufficient only to deliver powers of 0.1 and 0.5, respectively, to discriminate between type I errors of 0.025 and 0.026. If we wish to have 80% power to detect a difference between an empirical type I error of 0.026 and the null hypothesis of 0.025, a simulation sample size of over 190000 is needed and for 90% power, 260 000 is needed. These sample sizes correspond to ARLs of 380 and 1600. Figure 9 shows the relationship between the ARL and simulation sample size to discriminate between 0.025 and 0.026 type I errors.

Concentrating on comparisons of 2.5% and 2.6% suggests that Posch and colleagues are aiming at having procedures that

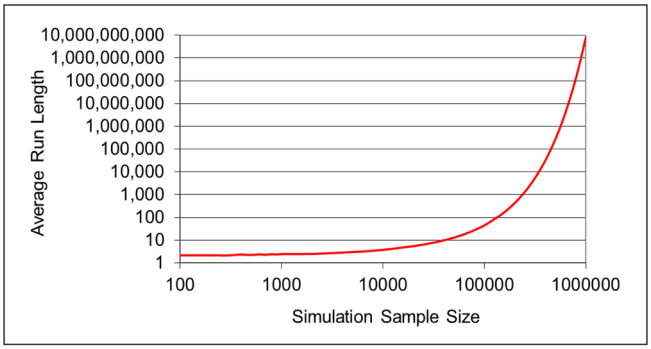

Figure 9. Average run length to find a 'Good Seed' .

we previously termed perfectly-calibrated, in which case perhaps they should have been requiring the ability to discriminate between 2.5% and 2.51%, which would require a simulation sample size of approximately 10^7. These sample sizes are very large, but can be reduced by the use of techniques such as anitithetic variates (Hammersley and Morton [106]).

The view that control of the type I error is unprovable by simulation was re-iterated in a panel discussion on adaptive designs at the Third Annual FDA/DIA Statistics Forum held in Washington in 2009 (Wang and Bretz [107]). The panellists agreed that it is impossible to determine the α level of a confirmatory trial based only on the empirical type I error rate from a simulation. The reasons for this were the following: first, to simulate any statistical procedure, a statistical model is chosen, which may not be true in practice; second, in relying on simulations as proof, it is assumed that amongst the configurations studied are those that give rise to a maximal type I error rate; third, the simulations are reproducible; finally for adaptive designs, the adaptation rules need to be completely pre-specified, and this does not allow for changes arising from the practicalities of running clinical trials. The panellists were positive about the use of simulation to evaluate the robustness of theoretical arguments concerning the control of the type I error. A referee commented that criticality of model choice applies even to analytical derivation of type I error rates. The second reason may not be unique to simulations either. As models and tests become more complicated, they may include additional parameters whose values are not clear under a null hypothesis.

7. DISCUSSION

In this paper, we have looked at a number of examples in which frequentist calibration of Bayesian procedures has been undertaken. Perhaps the most important question that needs to be addressed is the following: should calibration of Bayesian procedures be perfect? Or is being 'well-calibrated' sufficient? My view has always been that it is the latter that we should be seeking.

To illustrate this last point, take as an example the ASTIN trial. In that trial, a termination rule, based on pre-specified bounds of posterior probability, was used to recommend stopping recruitment for either futility or efficacy. The prime purpose of the trial was to identify the ED_{95}, and the stopping rules were based on the mean response at the dose closest to the estimate of the ED_{95}. If the lower 80% credible interval for the difference to placebo was greater than 2 points on the Scandinavian Stroke Scale, the study would stop for efficacy. In contrast, if the upper 80% credible interval for the difference to placebo on the scale

was less than 1 point, the decision would be to stop the study for futility. The cut-offs of 2 points and 1 point for efficacy and futility, respectively, were chosen after discussions with the clinical members of the team, these were not statistically driven. Having established those cut-offs, the study simulations were conducted to assess the 'type I' of the procedure. For the case of a flat dose-response curve, simulations showed that the adaptive design would wrongly conclude a clinically worthwhile effect at the ED_{95}. In the event, these simulations showed that there was a 2% one-sided chance of falsely identifying a clinically worthwhile difference, and this was deemed acceptable. What would have been the course of action if the empirical type I error had been above 2.5%? My preference would have been to leave the cut-offs unaltered and to have changed the information requirement, in this case by modulating the posterior probability associated with the credible interval. This is consistent with the FDA's Bayesian guidance in which one option to better control the type I error is to increase the probability criterion for a successful trial.

The distinction between perfectly-calibrated and well-calibrated procedures is in a sense the distinction between a position which says that the importance of controlling the type I error is paramount and of greater importance than minimising the type II error. More recently, there have been suggestions that the Neyman–Pearson approach, which minimises the type II error for a fixed type I error, is not the most appropriate approach in regulated scientific enterprises [108–110]

Acknowledgements

Most of this work was presented during the Design and Analysis of Experiments programmes held at the Isaac Newton Institute for Mathematical Sciences in Cambridge in the autumn of 2011 and the summer of 2015. I am grateful for the comments of anonymous reviewers which led to a measurable improvement.

REFERENCES

[1] Bayes T. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 1763; **53**:370–418.
[2] Laplace PS. Mémoire sur la probabilité des causes par les évenemens. *Mémoires de mathématique et de physique presentés a l'Académie royale des sciences, par divers savans, & lûs dans ses assemblées* 1774; **6**:621–656.
[3] Helmert FR. Über die Bestimmung des wahrscheinlichen Fehlers aus einer endlichen Anzahl wahrer Beobachtungsfehler. *Zeitschrift für Mathematik und Physik* 1875; **20**:300–3.
[4] Helmert FR. Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und uber einige damit in Zusammenhang stehende Fragen. *Zeitschrift für Mathematik und Physik* 1876; **21**:192–218.
[5] Helmert FR. Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit Astronomische. *Nachtrichten* 1876; **88**:113–32.
[6] Lüroth J. Vergleichung von zwei Werten des wahrscheinlichen Fehlers. *Astronomische Nachtrichten* 1876; **87**:209–20.
[7] Edgeworth FY. The method of least squares. *The London, Edinburgh, and Dublin Philosophical Magazine Series* 1883; **5** (16):360–375.
[8] Pearson K, Filon LNG. Mathematical contributions to the theory of evolution. IV On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London Series A* 1898; **191**:229–311.
[9] Student. Probable error of a correlation coefficient. *Biometrika* 1908; **7**:302–310.
[10] Thompson W. On the likelihood that an unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 1933; **25**:285–294.
[11] Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* 1966; **20**:19–23.
[12] Cornfield J. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *Journal of the American Statistical Association* 1966; **61**:577–594.
[13] Berry DA. A case for Bayesianism in clinical trials. *Statistics in Medicine* 1993; **4**:1377–1393.
[14] Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society Series A* 1994; **157**:357–416.
[15] Ashby D. Bayesian statistics in medicine: a 25 year review. *Statistics in Medicine* 2006; **25**:3589—3631.
[16] Ashby D, Grieve AP. 25 years of Bayesian methods in the pharmaceutical industry: a personal, statistical bummel. *Pharmaceutical statistics* 2007; **6**:261–281.
[17] International Conference on Harmonisation. E9:Statistical Principles for Clinical Trials, 1996. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf (Accessed July 2014).
[18] FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. —Draft Guidance for Industry and FDA Staff, 2010. http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm071121.pdf (Accessed July 2014).
[19] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Journal of the American Medical Association* 1996; **276**:637–639.
[20] Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. *Journal of the American Medical Association* 1994; **272**:1926–1931.
[21] Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature: a position paper. *Annals of Internal Medicine* 1994; **121**:894–895.
[22] Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Medicine* 2007; **4**(1625):1654.
[23] Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, Williamson RE, Zou GY, Hutchings K, Johnson Y, Tait V, Wiens M, Golding J, van Duijn C, McLaughlin J, Paterson A, Wells G, Fortier I, Freedman M, Zecevic M, King R, Infante-Rivard C, Stewart A, Birkett N. Strengthening the Reporting of Genetic Association studies (STREGA)–an extension of the STROBE statement. *European Journal of Clinical Investigation* 2009; **39**:247–266.
[24] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 2009; **151**:264–269.
[25] Bossuyt PM, Reitsma JB, Bruns DE, Constantine A, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D. de Vet HCW, and Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine* 2003; **138**:W1–12.
[26] Husereau D, Drummond M, Petrou S, Carswell S, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E. Consolidated health economic evaluation reporting standards (CHEERS) statement. *BMC Medicine* 2013; **11**:80.
[27] Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* 2007; **19**:349–357.
[28] Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research, ENTREQ. *BMC Medical Research Methodology* 2012; **12**:181.
[29] Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney S. Publication guidelines for improvement studies in health care: evolution of the SQUIRE project. *Annals of Internal Medicine* 2008; **149**:670–676.

[30] Gagnier JJ, Gunver KG, Alyman DG, Moher D, Sox HR. The CARE guidelines: consensus-based clinical case reporting guideline Development. *Headache: The Journal of Head and Face Pain* 2013; **53**:1541–1547.
[31] Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin J, Doré C, Parulekar W, Summerskill W, Groves T, Schulz K, Sox H, Rockhold FW, Rennie D, Moher D. SPIRIT 2013 Statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine* 2013; **31**(158):200–207.
[32] Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines, Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors; 2013.
[33] Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology* 2005; **58**:261–268.
[34] Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technology Assessment* 2000; **4**:1–130.
[35] The BaSiS Group. September 13 2001-last update, Bayesian standards in science (BaSiS), http://lib.stat.cmu.edu/bayesworkshop/2001/BaSis.html. (Accessed July 2014).
[36] Lehmann HP, Goodman SN. Bayesian Communication A Clinically Significant Paradigm for Electronic Publication. *Journal of the American Medical Informatics Association 7* 2000; **3**:254–266.
[37] Hildreth C. Bayesian statisticians and remote clients. *Econometrica: Journal of the Econometric Society* 1963; **31**:422–438.
[38] Draper D. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B (Methodological)* 1995; **57**:45–97.
[39] Grieve AP. A Bayesian analysis of the twoperiod crossover design for clinical trials. *Biometrics* 1985; **41**:979–990.
[40] Grieve AP. Extension of a Bayesian approach to the analysis of the twoperiod crossover to include baselines. *Statistics in Medicine* 1994; **13**:905–929.
[41] LeBlond D. FDA Bayesian statistics guidance for medical device trials – application to process validation. *Journal of Vlaidation Technology* 2010; **16**:24–33.
[42] Hunter WG, Lamboy WF. A Bayesian analysis of the linear calibration problem.". *Technometrics* 1981; **23**:323–328.
[43] Bickel JE, Floehr E, Kim SD. Comparing NWS PoP forecasts to third-party providers. *Monthly Weather Review* 2011; **139**:3304–3321.
[44] Rubin DR. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics* 1984; **12**:1151–1172.
[45] Little R. Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician* 2006; **60**:213–223.
[46] Little R. Calibrated Bayes for, statistics in general, and missing data in particular. *Statistical Science* 2011; **26**:162–174.
[47] Little RJ. *Calibrated Bayes, an alternative inferential paradigm for official statistics*, Vol. 28, 2012, pp. 309–334.
[48] Grieve AP. Joint equivalence of means and variances of two populations. *Journal of Biopharmaceutical Statistics* 1998; **8**:377–390.
[49] Bauer P, Bauer MM. Testing equivalence simultaneously for location and dispersion of two normally distributed populations. *Biometrical Journal* 1994; **6**:643–660.
[50] Medical Research Council Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment for pulmonary tuberculosis. *British Medical Journal* 1948; **ii**:769–82.
[51] Wald A. *Sequential Analysis*. John Wiley: New York, 1947.
[52] Dodge HF, Romig HG. A method of sampling inspection. *The Bell System Technical Journal* 1929; **7**:613–631, Geller NL.
[53] Bartky W. Multiple sampling with constant probability. *Annals of Mathematical Statistics* 1943; **14**:363–377.
[54] Armitage P. Sequential tests in prophylactic and therapeutic trials. *Quarterly Journal of Medicine* 1954; **23**:255–274.
[55] Armitage P. Sequential methods in clinical trials. *American Journal of Public Health* 1958; **48**:1395–1402.
[56] Bross I. Sequential medical plans. *Biometrics* 1952; **8**:188–205.
[57] Bross I. Sequential clinical trials. *Journal of Chronic Diseases* 1958; **8**:349–365.
[58] Armitage P. *Sequential Medical Trials, 2nd Edition*. Blackwell Scientific: Oxford, 1975.
[59] Armitage P, McPherson CK. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A* 1969; **132**:235–244.
[60] Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* 1971; **44**:793–797.
[61] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
[62] Pocock SJ. Interim Analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1982; **38**:153–162.
[63] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* 1976; **34**:585–612.
[64] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
[65] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
[66] Berry DA. Interim analysis in clinical trials: classical vs. Bayesian approaches. *Statistics in Medicine* 1985; **4**:521–526.
[67] Berry DA. Interim analysis in clinical trials: the role of the likelihood principle. *The American Statistician* 1987; **41**:117–122.
[68] Freedman LS, Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring. *Controlled Clinical Trials* 1989; **10**:357–367.
[69] Lewis RJ, Berry DA. A classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Assoication* 1994; **89**:1528–1534.
[70] Freedman LS, Spiegelhalter DJ, Parmar MKB. The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine* 1994; **13**:1371–1383.
[71] Fayers PM, Ashby D, Parmar MKB. Bayesian data monitoring in clinical trials. *Statistics in Medicine* 1997; **16**:1413–1430.
[72] Cronin KA, Freedman LS, Lieberman R, Weiss HL, Beenken SW, Kelloff GJ. Bayesian monitoring of Phase II trials in cancer chemoprevention. *Journal of Clinical Pharmacology* 1999; **8**:705–711.
[73] Parmar MKB, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *The Lancet* 2001; **358**:375–381.
[74] Ashby D, Tan S-B. Where's the utility in Bayesian data-monitoring of clinical trials. *Clinical Trials* 2005; **2**:197–208.
[75] Goodman SN. Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine* 2007; **46**:882–888.
[76] Emerson SS, Kittelson JM, Gillen DL. Bayesian evaluation of group sequential trial designs. *Statistics in Medicine* 2007; **26**:1431–1449.
[77] Gsponer T, Gerber F, Bornkamp B, Ohlssen D, Vandemeulebroecke M, Schmidli H. A practical guide to Bayesian group sequential designs. *Pharmaceutical statistics* 2014; **13**:71–80.
[78] Grossmann J, Parmar MK, Spiegelhalter DJ, Freedman LS. A unified method for monitoring and analysing controlled trials. *Statistics in Medicine* 1994; **13**:1815–1826.
[79] Pennello G, Thompson L. Experience with reviewing Bayesian medical device trials. *Journal of Biopharmaceutical Statistics* 2007; **18**:81–115.
[80] Reboussin DM, DeMets DL, Kim KM, Lan KKG. Computations for Group Sequential Boundaries Using the Lan-DeMets Spending Function Method. *Controlled Clinical Trials* 2000; **21**:190–207.
[81] O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making* 2001; **21**:219–230.
[82] Senn S. Consensus and controversy in pharmaceutical statistics (with discussion). *Journal of the Royal Statistical Society: Series D (The Statistician)* 2000; **49**:135–176.
[83] Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clinical Trials* 2010; **7**:5–18.
[84] Di Scala L, Kerman J, Neuenschwander B. Collection, synthesis, and interpretation of evidence: a proof-of-concept study in COPD. *Statistics in Medicine* 2013; **32**:1621–1634.
[85] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S, Micalle S, Roychoudhury S, Thompson L. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* 2014; **13**:41–54.
[86] Alber SA, Lee JJ. *Calibrating the Prior Distribution for a Normal Model with Conjugate Prior*, 2015. To appear.

[87] Grieve AP, Krams M. ASTIN: a Bayesian adaptive dose–response trial in acute stroke. *Clinical Trials* 2005; **2**:340–351.
[88] Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cance* 2007; **43**:859–866.
[89] Altham PME. Exact Bayesian analysis Of a 2x2 contingency table, and Fisher's "Exact" significance test. *Journal Of The Royal Statistical Society Series B* 1969; **31**:261–269.
[90] Liebermeister C. Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Sammlung klinischer Vorträge* 1877; **110**:935–962.
[91] Seneta E, Carl Liebermeister's Hypergeometric Tails. *Historia Mathematica* 1994; **21**:453–462.
[92] Ineichen R. Der "Vierfeldertest" von Carl Liebermeister (Bemerkungen zur Entwicklung der medizinischen Statistik im 19. Jahrhundert). *Historia Mathematica* 1994; **21**:28–38.
[93] Seneta E, Seif FJ, Liebermeister H, Dietz K. Carl Liebermeister (1833-1901): a pioneer of the investigation and treatment of fever and the developer of a statistical test. *Journal of Medical Biography* 2004; **12**:215–221.
[94] Thompson W. On the theory of apportionment. *American Journal of Mathematics* 1935; **57**:450–456.
[95] Maki RG, Wathen JK, Patel SR, Priebat DA, Okuno SH, Samuels B, Fanucchi M, Harmon DCX, Schuetze SM, Reinke D, Thall PF, Benjamin RS. Baker LH, and Hensley ML Randomized phase II study of gemcitabine and docetaxel compared with gemcitabine alone in patients with metastatic soft tissue sarcomas: results of sarcoma alliance for research through collaboration study 002. *Journal of Clinical Oncology* 2007; **25**:2755–2763.
[96] Giles FJ, Kantarjian HM, Cortes JE, Garcia-Manero G, Verstovsek S, Faderl S, Thomas DA, Ferrajoli A, O'Brien S, Wathen JK, Xiao LC, Berry DA, Estey EH. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *Journal of Clinical Oncology* 2003; **21**:1722–1727.
[97] Korn EL, Freidlin B. Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology* 2011; **29**:771–776.
[98] Berry DA. Adaptive clinical trials: the promise and the caution. *Journal of Clinical Oncology* 2011; **29**:606–609.
[99] Thall PF, Fox PS, Wathen JK, Some Caveats for Outcome Adaptive Randomization in Clinical Trials. In *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects*, Sverdlov O (ed.) CRC Press: Boca Raton; 2015; pp. 287–305. in press.
[100] Giovagnoli A, Zagoraiou M. Simulation of clinical trials: a review with emphasis on the design issues. *Statistica* 2012; **72**:63–80.
[101] Holford NHG, Hale M, KO HC, Steimer J-L, Sheiner LB, Peck CC. Simulation in drug development, October 2014. Good Practices, 1999. Available at http://bts.ucsf.edu/cdds/research/sddgpreport.phpAccessed.
[102] Daniel C. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* 1959; **1**:311–341.
[103] Box GEP, Meyer RD. *An analysis for unreplicated fractional factorials.Technometrics* 1986; **28**:11–18.
[104] Lenth RV. Quick and easy analysis of unreplicated factorials. *Technometrics* 1989; **31**:469–473.
[105] Posch M, Maurer W, Bretz F. Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical statistics* 2011; **10**:96–104.
[106] Hammersley JM, Morton KW. A new Monte Carlo technique: antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society* 1956; **52**:449–475.
[107] Wang S-J, Bretz F. From adaptive design to modern protocol design for drug development: part I. Editorial and summary of adaptive designs session at the third FDA/DIA statistics forum. *Drug Information Journal* 2010; **44**:325–331.
[108] Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE* 2012; **e32734**:7.
[109] Pericchi LR and Pereira CAB. Changing the paradigm of fixed significance levels: Testing Hypothesis by Minimizing Sum of Errors Type I and Type II. *Brazilian Journal of Probability and Statistics* 2015, to appear.
[110] Grieve AP. How to test hypotheses if you must. *Pharmaceutical Statistics* 2015; **14**:139–150.