Applied Stochastic Models in Business and Industry

Special Issue Paper

Received 2 July 2016, Revised 15 November 2016, Accepted 19 November 2016 Published online 13 January 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2222

Clinical trial design as a decision problem

Peter Müller[a*†], Yanxun Xu[b] and Peter F. Thall[c]

The intent of this discussion is to highlight opportunities and limitations of utility-based and decision theoretic arguments in clinical trial design. The discussion is based on a specific case study, but the arguments and principles remain valid in general. The example concerns the design of a randomized clinical trial to compare a gel sealant versus standard care for resolving air leaks after pulmonary resection. The design follows a principled approach to optimal decision making, including a probability model for the unknown distributions of time to resolution of air leaks under the two treatment arms and an explicit utility function that quantifies clinical preferences for alternative outcomes. As is typical for any real application, the final implementation includes some compromises from the initial principled setup. In particular, we use the formal decision problem only for the final decision, but use reasonable ad hoc decision boundaries for making interim group sequential decisions that stop the trial early. Beyond the discussion of the particular study, we review more general considerations of using a decision theoretic approach for clinical trial design and summarize some of the reasons why such approaches are not commonly used.

Keywords: Bayesian decision problem; Bayes rule; nonparametric Bayes; optimal design; sequential stopping

1. Introduction

We discuss opportunities and practical limitations of approaching clinical trial design as a formal decision problem. Using a case study as a running example keeps the argument focused and specific. We review a study that was set up to compare a hydrogel sealant (Progel) against standard care for patients who develop air leaks after pulmonary resection. The main features of the design are the elicitation of a utility function that quantifies clinical preferences for time to resolve the air leaks and a nonparametric Bayesian prior for the distributions of the resolution time under the two treatment arms. A Bayesian nonparametric (BNP) model is a prior for an unknown probability measure that is not restricted to a specific parametric family. Both features are important. The utility function is only meaningful if the probability model allows learning about detailed features of the event time distribution, and the nonparametric model is only needed when the decision hinges on such details. In the upcoming discussion, we focus mainly on the features of the decision problem. A complete discussion of the design and the trial appears in [1], including extensive simulations to evaluate the design's operating characteristics (OCs) under alternative scenarios.

The use of decision theoretic approaches in Bayesian clinical trial design is rare. Commonly used methods use Bayesian inference to compute posterior probabilities of clinically meaningful events or inference summaries for key parameters but then use these summaries for reasonable, but ad hoc designs. See, for example, [2] or [3] for a recent review. Beyond clinical trial design, Bayesian decision theoretic approaches are not widely used for optimal design in general. The review by [4] discusses commonly used Bayesian approaches to optimal design, focusing mainly on design problems related to learning about unknown parameters but clearly recognizing the importance of more general problems. To date, this paper remains one of the most cited and most comprehensive reviews of Bayesian optimal design. In their final discussion, Chaloner and Verdinelli write 'It is clearly helpful, in the design process, to carefully consider the reason the experiment is being done and to consider what utility should be used. [...] it would also be interesting to see alternatives constructed and explored in future research'. In the following discussion, we review one such construction which is in the spirit of what Kathryn Chaloner might have wished to see.

[a] *Department of Mathematics, University of Texas at Austin, USA*
[b] *Department of Applied Mathematics and Statistics, Johns Hopkins University, USA*
[c] *Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, USA*
**Correspondence to: Peter Müller, Department of Mathematics, University of Texas at Austin, USA*
[†] *E-mail: pmueller@math.utexas.edu*

The discussion is restricted to the particular decision problem of clinical trial design and even further focused by following a particular example. The arguments are meant to highlight the benefits of following a formal decision theoretic setup and also to indicate the practical limitations and challenges that are involved.

2. Decision problem

2.1. Framework

We set up the design problem as a Bayesian decision problem, following a formal, principled approach. See, for example, [5] or [6] for a description of the general framework. Briefly summarized, the ingredients of a Bayesian decision problem are a probability model for all relevant unknown quantities, including data, future data, and unknown parameters of interest and a utility function that formalizes relative preferences of a decision maker under hypothetical data and assumed parameters. The optimal action, known as the Bayes rule, is the action that maximizes utility, in expectation over all unknown variables and conditional on all known variables. To be specific, let y denote the data. Often, some data may already be observed at the time of decision making. In that case, we partition the data vector into (y, y_0) denoting the observed data by y_0 (but we will simply use y when all data is observed). Let θ denote unknown parameters, and let d denote the decision. The Bayesian probability model usually is given as a sampling model for the data, $p(y, y_0 \mid \theta, d)$ and a prior model, $p(\theta)$. Including d here in the conditioning set is a slight abuse of notation and only indicates that the sampling model can be indexed by d. In general, the prior also could be indexed by d, but this usually is not performed. Finally, let $u(d, \theta, y)$ denote the utility, defined as a function of a possible action d, assumed parameters θ, and hypothetical data y. The utility function represents the decision maker's relative preferences for actions under an assumed truth and data. Let A denote the action set of all possible decisions under consideration. The *Bayes rule* is

$$d^\star = \arg\max_{d \in A} \int u(d, \theta, y)\, p(\theta, y \mid y_0, d)\, dy\, d\theta. \tag{1}$$

Letting $U(d \mid y_0) = \int u(d, \theta, y)\, p(\theta, y \mid y_0, d)\, dy\, d\theta$, we can write $d^\star = \arg\max U(d \mid y_0)$. Here, the conditioning bar in $U(\cdot)$ indicates the conditioning on y_0 in the expectation. One can argue, from first principles, that a rational decision maker should act as if he or she were maximizing expected utility $U(d \mid y_0)$ [7].

An important detail in the setup is the statement of the action set A. Usually, the set of possible actions is highly restricted, to avoid unintuitive, unreasonable, or impractical decisions. A good choice of action set avoids awkward solutions $d^\star$. Mathematically, a choice of a probability model and a utility function implies an optimal decision $d^\star$. But the mapping is very indirect through the integration and maximization in (1), and technical details in the choice of the probability model and utility function could lead to unintended solutions if care is not taken to restrict A suitably. We will give examples of this later on.

2.2. Terminal decision

The design of the Progel study involves two different types of decisions. After each patient cohort is treated, and their outcomes are observed, we decide whether to continue accrual (Sequential Stopping Decision). Let $a_c \in \{0, 1\}$ denote the continuation decision, with $a_c = 0$ indicating continuation and $a_c = 1$ for early stopping. We index cohorts by $c = 1, \ldots, C$, with a_1 denoting the continuation decision after the first cohort, etc. The study includes a maximum sample size. That is, we restrict $a_C = 1$. Upon stopping, we decide whether or not to report Progel as superior (*Terminal Decision*). Let $d \in \{0, 1\}$ denote the terminal decision, with $d = 0$ for recommending standard care versus $d = 1$ for Progel. In summary, the action set is $a_c \in \{0, 1\}$ and $d \in \{0, 1\}$.

Ideally, all decisions should be made as Bayes rule with respect to the same underlying utility function. However, this is where the need to construct a practicable implementation that actually can be used to conduct clinical trial parts with the principled approach. For the Progel trial, only the terminal decision d is implemented as a Bayes rule. In contrast, the stopping decisions are based on a reasonable group sequential decision boundary. Details of the latter are discussed in the next section. We first discuss the terminal decision and explain the terminal decision graphically, in Figure 1.

Posterior inference. Index $j = 1$ for Progel and $j = 0$ for standard care, with sample sizes n_1 and n_0. Let y_{ji} denote the time until resolution of air leaks for patient $i = 1, \cdots, n_j$ in treatment arm j. Let $G_j(y)$ denote the distribution of time to resolution of air leaks, $j = 0, 1$. That is, $y_{ji} \sim G_j$ independently for all $i = 1, \cdots, n_j$. We will use a BNP model for G_j. That is, G_j itself is the unknown parameter, for which the BNP model defines a prior $p(G_j)$. Figure 1(a) shows hypothetical posterior means for G_0 and G_1, conditional on all patients up to a particular time during the trial. For some patients, actual event times are recorded, while for others the time until resolution of air leaks is censored at the current time. Note that G_1 has a point mass at $y = 0$, corresponding to a nonzero probability of immediate resolution of air leaks, before day 1.

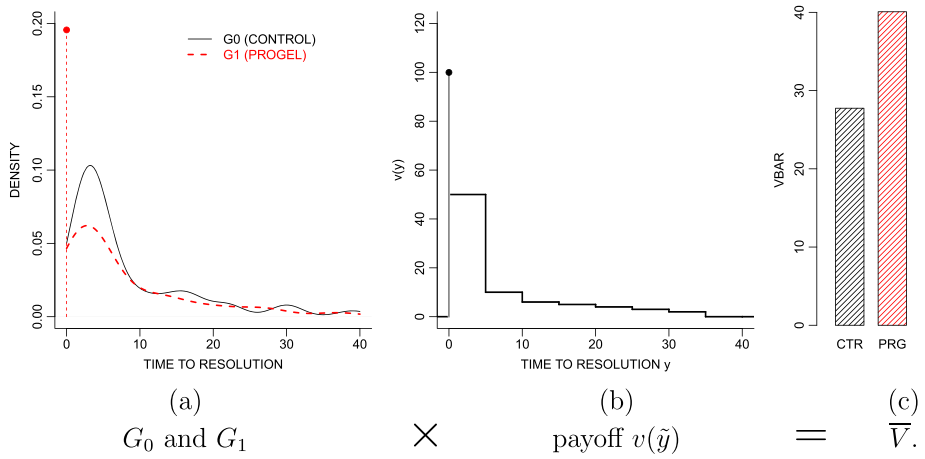

Figure 1. The random resolution times under each treatment (left panel, a) are weighted with elicited payoffs (center panel, b), giving the weighted average payoff $\overline{V}_0$ and $\overline{V}_1$ (right panel, c) for the two treatments. Inference in the BNP model includes averaging with respect to the uncertainty on the resolution time distributions, G_0 and G_1. [Colour figure can be viewed at wileyonlinelibrary.com]

Utility. Let $\tilde{y}$ denote the time until resolution of air leaks for a future patient who is assigned the final recommended treatment. Figure 1(b) shows the elicited payoffs $v(\tilde{y})$ that are used to construct a utility function. For the Progel study, the utility function compares average clinical payoff $v(\tilde{y})$ of resolution times under the two treatment arms. That is, v formalizes clinical desirability for different resolution times, with high values for quick resolution times $\tilde{y} < 5$. Let $\overline{V}_j = \int v(\tilde{y}) dG_j(\tilde{y})$ denote the average payoff under G_j. We then define the utility function

$$u(G_0, G_1, d) = \begin{cases} I(\overline{V}_1 > \overline{V}_0 + 18) \text{ if} d = 1 \\ I(\overline{V}_1 \leq \overline{V}_0 + 18) \text{ if} d = 0. \end{cases}$$

In words, the utility function is an indicator of average clinical payoff under Progel being more than 18 units superior to the average under control. The offset of 18 units in the comparison relates to the minimum clinically meaningful difference (on the scale of v).

Expected utility and Bayes rule. Let y denote all currently observed data, and let $\overline{G}_j = E(G_j \mid y)$ denote the current posterior mean of G_j. Expected utility is computed as

$$U(d = 1 \mid y) = \int u(G_0, G_1, d)\, dp(G_0, G_1 \mid y, d) = p(\overline{V}_1 > \overline{V}_0 + 18 \mid y) \tag{2}$$

and similarly $U(d = 0 \mid y) = 1 - p(\overline{V}_1 > \overline{V}_0 + 18 \mid y)$. To decide between the two treatments, we compare $U(d = 0 \mid y)$ versus $U(d = 1 \mid y)$ as in (1). Panel (c) shows, for these assumed G_0 and G_1 and this utility function, that Progel is the treatment with higher expected utility. In this case, we would report Progel as the recommended treatment.

There is an important point about (2). At the time of the terminal decision, we have already made the earlier sequential stopping decisions. Therefore, the only argument of the expected utility function at this moment is the terminal decision d.

In the actual application, the utility function was elicited from the principal investigator of the study. The highly nonlinear relative preferences for early resolution times in Figure 1(b) reflect the clinical importance of a quick resolution. Leaks that persist beyond the immediate postoperative period of 5 days may result in longer chest tube drainage, greater postoperative pain, increased risk of infection, empyema, thromboemboli, and increased length of hospitalization. For these reasons, the utility function includes a strong preference for early resolution of air leaks.

The probability model underlying inference on G_j in (2), that is, that defines $p(G_j \mid y)$, will be discussed in more detail in the next sections. Briefly, we use a dependent Dirichlet process (DDP) model. This is a BNP model for related random probability measures, in our case G_0 and G_1. We develop the model for y_{ji} on a log scale, that is, for $y_{ji} = \log(T_{ji} + 1)$, where T_{ji} is resolution time in days. In particular, $y_{ji} = 0$ corresponds to immediate resolution. In the following discussion, little would be lost if the DDP model were replaced by a mixture of k, say $k = 3$, normals, that is, assuming $G_j = \sum_{h=1}^{3} w_h N(\mu_{jh}, \sigma^2)$, with common weights w_h for G_0 and G_1, and an order constraint $\mu_{0h} \geq \mu_{1h}$ on the normal location parameters. One more elaboration of the model adds a point mass at $y = 0$, to allow for immediate resolution of air leaks. This is given by $G_j = \pi_j \delta_0 + (1 - \pi_j) \sum_{h=1}^{3} w_h N(\mu_{jh}, \sigma^2)$, where δ_0 denote the point mass 1 at $y_{ji} = 0$. Later, we will

introduce the model actually used, which involves additional generalizations. Still, little would be lost if the actual model were replaced by the simple zero-enriched mixture of normals given previously.

2.3. Sequential stopping decision

The second, more challenging decision is the sequential stopping decision, a_c. After each cohort of patients, up to a maximum sample size, at cohort C, we decide whether or not to stop early for either futility or efficacy. That is, we stop early to recommend standard of care or stop early to recommend Progel. In the actual trial, we used $C = 3$ cohorts.

For a principled solution, one would set this up as a sequential decision problem. Upon stopping, we again would use (2) to make the terminal decision. But the sequential stopping decision must be made first. For the optimal decision in the next to last cohort, $c = C - 1$, before the final cohort is enrolled, we would consider the following expected utility calculation. Let y_0 denote the current data, and let y denote the future data for the possible last cohort. As before, we use $\tilde{y}$ for the response of a future patient assigned the final recommended treatment. The utility of stopping early, given the current data, is

$$U_a(a_{C-1} = 1 \mid y_0) \equiv \int \max_j \left\{ U(d = j \mid y_0, y) \right\} \, dp(y \mid y_0). \tag{3}$$

We use the subindex a on $U_a(\cdot)$ to distinguish this expected utility from $U(d \mid y)$ in (2). Both are expected utilities but include expectations and nested optimization to a different extent. While $U(d \mid y)$ conditions on the trial already being stopped and only averages over G_j and $\tilde{y}$, in (3), we average over the posterior predictive distribution, $p(y \mid y_0)$ for the last cohort (outside integration), substitute the optimal next stage decision (inside maximization), and finally integrate over the posterior distribution on G_j and over $\tilde{y}$ in the nested evaluation of $U(d \mid y_0, y)$. This alternating sequence of integration and optimization is typical for sequential decision problems. For earlier stage continuation decisions, a_c, $c = C - 2, \ldots, 1$, similar expressions are required but with additional levels of integration and maximization. The Bayes rule is then

$$a_c^\star = \arg \max_{h \in \{0,1\}} U_a(a_c = h \mid y_0).$$

In general, the solution $a_c^\star$ is computationally prohibitive, because of the exponential explosion of possible cases and histories that one needs to keep track of in the alternating sequence of optimizations and expectations. In practice, approximate solution strategies are used. See, for example, Berger ([8], Chapter 7) for a general discussion.

In the context of clinical trial design, it is a common practice to replace maximization of (3) by decision boundaries, often specified in terms of clinically meaningful events. This is best explained with an example, again using the Progel trial. Again denoting the current data by y_0, define

$$\eta(y_0) = p(\overline{V}_1 > \overline{V}_0 + 18 \mid y_0)$$

as the posterior probability of the average payoff with Progel being at least 18 points better than under standard care, recalling the definition (2). Instead of the computationally prohibitive solution to (3), we use the following decision boundaries for $c = 1, \ldots, C - 1$:

$$a_c^\star = \begin{cases} 1 \ \text{if} \eta(y_0) < 0.05 \text{ or } \eta(y_0) > 0.90 \\ 0 \ \text{if} 0.05 \leq \eta(y_0) \leq 0.90. \end{cases}$$

These boundaries stop in the face of overwhelming evidence of either futility ($\eta < 0.05$) or efficacy ($\eta > 0.90$). Note that the definition of $\eta(y_0)$ coincides with the expected utility in (2). Thus, the terminal decision can be characterized in terms of η, as $d^\star = 1$ if $\eta(y_0) > 0.5$ and $d^\star = 0$ otherwise.

The specific values for $p_L = 0.05$ and $p_U = 0.90$ were determined by simulating the trial design for each of several pairs, say $p_L = 0.01, 0.05, 0.10$ and $p_U = 0.90, 0.95, 0.99$, and choosing the cut-offs to obtain a design with desirable OCs, which include sample size distributions, and nominally correct and incorrect decision probabilities, such as conventional Type I and Type II error rates. For the general use of such decision boundaries in clinical trial design, see also [9] or [10].

3. Probability model

We deliberately postponed our discussion of the underlying probability model until after discussion of the decision problem. This highlights the separate nature of the probability model for statistical inference and the decision problem on top of it. The two are linked by (2), when we evaluate expected utility by averaging with respect to posterior distributions in the inference model. Ideally, one could describe the nature of the decision and state the utility function without reference

to specific details of the probability model. We only need to assume that there is a well-defined probability model. For reference and completeness, we briefly describe the underlying model for the Progel trial.

Let δ_x denote a point mass at x and let $N(\mu, \sigma^2)$ denote a normal distribution with moments (μ, σ^2) and let $x \sim N^+(m, s^2)$ denote a truncated normal with $x \geq m$. We assume

$$G_j = \pi_j \delta_0 + (1 - \pi_j) \sum_h w_h N(x_{jh}, \sigma^2),$$

with a beta prior on π_j, subject to $\pi_1 > \pi_0$, an infinite version of a Dirichlet prior on the weights w_h, independent $x_{1h} \sim N(\mu_1, \sigma_1^2)$ and conditionally on x_{1h}

$$p(x_{0h} \mid x_{1h}) = \begin{cases} \delta_{x_{1h}} & \text{with probability } \kappa \\ N^+(x_{1h}, \tau^2) & \text{with probability } 1 - \kappa. \end{cases}$$

That is, G_j is modeled as a mixture of a point mass at 0 (no air leak) and a mixture of normals. The models for G_0 $(j = 0)$ and G_1 $(j = 1)$ are linked by assuming $\pi_0 \leq \pi_1$ and $x_{0h} \geq x_{1h}$. The model on G_j marginally is known as Dirichlet process mixture model (Ferguson; [11]). The joint model on (G_0, G_1) is a version of the DDP [12].

4. Other decisions

The Progel study involved the sequential stopping and terminal decision only. We chose this study as the running example of the discussion exactly because of this focused setup and the fact that the decisions are easily described. Other studies may involve many other types of decisions, more complex outcomes, and thus more structured probability models.

A common decision relates to adaptive treatment allocation. For each patient, or patient cohort, we might want to compute different probabilities of assigning the competing treatment arms. In practice, one rarely would use a full decision theoretic implementation for choosing this treatment assignment probability. One common solution is to first consider the (current) posterior probability of any given arm being the optimal arm, say π_t is the probability of treatment t being optimal. Here, optimal could refer to the criterion used in the terminal decision, such as overall survival, progression free survival, etc. A common design is to assign treatment t with probability proportional to π_t. See, for example, [13] for a summary, or [14] for an earlier reference.

Other, more complicated, decisions could arise. For example, an investigator might be interested in inference on subgroups of patients who benefit significantly better or worse from the investigated treatments. This is known as subgroup analysis. See, [15] for a recent review of Bayesian approaches. [16] discuss a setup of the subgroup analysis problem as a formal decision problem. In general, this and many other decisions that are made in the course of a clinical trial often are too complex and involved to be easily approached as formal decision problems.

5. Conclusion

Using a particular clinical study as an example, we have discussed some features of a clinical trial design as a decision problem. In particular, we introduced the setup including a formal description of decisions as elements in an action space, a probability model on all unknown quantities including parameters and data, and a utility function that quantifies relative preferences for alternative actions under assumed parameters and hypothetical data. In that framework, the Bayes rule is the optimal action. In principle, this framework includes all decisions that may have to be taken when one develops a clinical trial protocol. In particular, this should include sequential stopping decisions, treatment allocation, and terminal decisions at the conclusion of the trial.

There are many practical limitations to this setup. To start, the optimal sequential decision is computationally intractable. For this and other practical reasons, it often is convenient to instead use reasonable decision boundaries for sequential stopping rules. We showed this process in the context of the illustrative example.

Given these practical limitations and compromises, the case study still is an example of an actual clinical trial protocol that is closer to a formal decision theoretic framework than many. There are many other reasons why most studies, including trials that use Bayesian designs, do not use decision theoretic approaches. An important one is the difficulty of eliciting suitable utility functions. However, we would argue that the need to state a utility function is a feature, not a problem. Even without stating a utility function, investigators make choices and judgments. The main difference is that the statement of a utility function keeps these choices clearly stated, and facilitates discussion and critique. Examples of elicited utilities for various two-dimensional outcomes in phase I–II clinical trials are given in Yuan, Nguyen, and Thall ([3], Chapters 6, 8, 11, 13, and 14). A general methodology for eliciting utilities for randomized trials with categorical outcomes is given by Murray, Thall, and Yuan [17].

A more serious issue is the nature of the Bayes rule $d^\star$ as an implicit solution of the optimization problem in (2). Mathematically, a given probability model and utility function imply $d^\star$. However, often technical details of the choice of probability model and utility function might eventually imply clinically unreasonable decisions. This is perhaps another reason why investigators are reluctant to use this approach. Of course, one could argue that in this case the utility function fails to reflect clinical utilities (or similarly for the probability model). In practice, when implementing utility-based designs, we have found it quite easy to elicit utilities that accurately reflect investigator's goals. Research physicians typically are very happy to provide their utilities and refine them during the process of learning the impact of a given utility on design properties (OCs). When an investigator sees that the consequence of a given initial set of numerical utilities is a design with undesirable properties, they readily alter their utilities to obtain a design with desirable OCs. That is, in practice, we always show the investigators the consequences of their numerical utilities in terms of design properties, and utility elicitation becomes an iterative process.

Finally, clinical studies involve many stakeholders, with diverse goals and utility functions, leaving it unclear whose utility function should drive the design.

References

1. Xu Y, Müller P, Thall P, Mehran R. A Bayesian nonparametric utility-based comparison of time to resolution of air leaks. *Bayesian Analysis* 2016. DOI:10.1214/16-BA1016.
2. Yin G. *Clinical Trial Design*. John Wiley & Sons: Hoboken, New Jersey, 2012.
3. Yuan Y, Nguyen H, Thall P. *Bayesian Designs for Phase I–II Clinical Trials*, Chapman & Hall/CRC Biostatistics Series: CRC Press, Boca Raton, Florida, 2016.
4. Chaloner K, Verdinelli I. Bayesian experimental design: a review. *Statistical Science* 1995; **10**:273–304.
5. Parmigiani G, Inoue L. *Decision Theory: Principles and Approaches*. John Wiley & Sons: Hoboken, New Jersey, 2009.
6. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley & Sons: Hoboken, New Jersey, 2003.
7. Robert CP. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* 2nd ed. Springer-Verlag: New York, New York, 2007.
8. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York, New York, 1985.
9. Müller P, Berry DA, Grieve AP, Smith M, Krams M. Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference* 2007; **137**(10):3140 –3150.
10. Rossell D, Müller P, and Rosner G. Screening designs for drug development. *Biostatistics* 2007; **8**:595–608.
11. Ferguson TS. Prior distribution on the spaces of probability measures. *The Annals of Statistics* 1974; **2**:615–629.
12. MacEachern S. Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria, VA, 1999; 50–55.
13. Thall P, Wathen J. Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer* 2007; **43**(5):859–66.
14. Berry D, Eick S. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in Medicine* 1995; **14**(3):231–46.
15. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 2011; **8**:129–143.
16. Müller P, Sivaganesan S, Laud P. A Bayes rule for subgroup reporting. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, Chen MH, Dey DK, Mueller P, Sun D, Ye K (eds), Springer: New York, New York 2010; 277 –284.
17. Murray T, Thall P, Yuan Y. Utility-based designs for randomized comparative trials with discrete outcomes. *Statistics in Medicine* 2016; **35**(24):4285–4305.