UNIVERSITY of York

Discussion Papers in Economics

No. 19/01

Cost-Effective Clinical Trial Design: Application of a Bayesian Sequential Stopping Rule to the ProFHER Pragmatic Trial

M. Forster, S. Brealey, S. Chick, A. Keding, B. Corbacho, A. Alban, A. Rangan

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Cost-Effective Clinical Trial Design: Application of a Bayesian Sequential Stopping Rule to the ProFHER Pragmatic Trial*

M. Forster,† S. Brealey,‡ S. Chick,§ A. Keding,‡ B. Corbacho,‡
A. Alban,§ A. Rangan¶

February 7, 2019

Abstract

We investigate value-based clinical trial design by applying a Bayesian decision-theoretic model of a sequential experiment to data from the ProFHER pragmatic trial. In the first applied analysis of its kind to use research cost data, we show that the model's stopping policy would have stopped the trial early, saving about 5% of the research budget (approximately £73,000). A bootstrap analysis based on generating resampled paths from the trial data suggests that the trial's expected sample size could have been reduced by approximately 40%, saving an expected 15% of the budget, with 93% of resampled paths making a decision consistent with the result of the trial itself. Results show how substantial benefits to trial cost stewardship may be achieved by accounting for research costs in defining the trial's stopping policy and active monitoring of trial data as it accumulates.

JEL codes: I10, C44, C61
Keywords: Bayesian sequential experimentation; Randomised clinical trials; Health technology assessment

This is a proof of concept paper which has been issued for discussion. It is intended to illustrate how the model presented may be populated with data from a health technology assessment. Results may change following feedback. It is not intended that the paper represents a comment on the health technologies themselves. The ProFHER trial was funded by the NIHR Technology Assessment programme (project number ref 06/404/502). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

*We thank Helen Handoll for comments on a preliminary draft of the paper. We also thank Matthew Dore and Andrea Palfreeman for preparation of the data used in the ProFHER trial itself. Alban and Chick acknowledge the support of the European Union through the MSCA-ESA-ITN project (676129). Forster acknowledges funding from the Research Infrastructure Support Fund of the Department of Economics and Related Studies, University of York.

†Corresponding author. Department of Economics and Related Studies, University of York, York YO10 5DD, United Kingdom. mf8@york.ac.uk

‡York Trials Unit, Department of Health Sciences, University of York, York YO10 5DD, United Kingdom.

§Technology and Operations Management Area, INSEAD, Fontainebleau, France 77300.

¶Department of Health Sciences, University of York, Faculty of Medical Sciences & NDORMS, University of Oxford and James Cook University Hospital, Middlesbrough.

1 Introduction

There is growing interest in the use of sequential clinical trials to assess the effectiveness and cost-effectiveness of health technologies (Pallmann et al., 2018; Cui et al., 2017; Yin et al., 2017; Wason et al., 2016; Bhatt and Mehta, 2016). Sequential trials offer the opportunity to stop early and bring the better treatment to patients sooner. Interest is supported by the idea of promoting 'value-based' clinical research: with DiMasi et al. (2016) estimating that the out-of-pocket cost of bringing one new compound to market is in the region of $1600 million, commissioners are seeking new ways to make clinical trials more efficient, without compromising the quality of their recommendations.

The development of new, value-based, methodologies for health technology assessment (HTA) raises important and, to a large extent, unexplored, questions about the role they should play in future HTA protocols. For example, in the area of drug development, the current system of evaluation has its roots in demonstrating quality, safety (U.S. Congress, 1938) and efficacy (U.S. Congress, 1962). More recently, countries including Australia, the United Kingdom, France and Canada have introduced the so-called 'fourth hurdle' of cost-effectiveness into approval processes (Taylor et al., 2004). This historical context is important for assessing the potential that new methodologies offer: notwithstanding their economic gains, careful assessment must be made of whether their recommendations will be acceptable to patients, the medical profession and wider society.

This paper takes the Bayesian decision-theoretic model of a sequential and value-based experiment proposed by Chick et al. (2017) and applies it to the ProFHER trial, a pragmatic, multicentre, randomised controlled trial conducted in the United Kingdom which compared surgery and sling immobilisation for the treatment of displaced proximal humeral fracture (Handoll et al., 2015; Rangan et al., 2015; Corbacho et al., 2016). The model is sequential because it assumes that trial data may be monitored as the trial progresses, permitting early stopping or late running as a function of the data as it is observed; it is value-based because its solution is a policy which maximises the expected net benefit of a technology adoption decision, accounting for the costs of carrying out the trial and the costs incurred in switching technologies. The model also provides a value-based rule for optimal trial design – run a sequential trial, run a trial with a fixed sample size, or run no trial at all – as a function of the prior mean for incremental net monetary benefit (INMB).

Although Chick et al. applied their model using a series of simulations based on published papers, these lacked detailed and accurate information on a number of variables and parameters. In this paper, we use the ProFHER trial's outcome, treatment and research cost data to calibrate their model much more accurately than has heretofore been possible. Chick et al.'s model assumed that, prior to the start of the trial, all patients were being treated with the 'standard' treatment and all patients would switch to the 'new' treatment if it were found to be superior. We use the work of Alban et al. (2018) to extend the model to account for the pragmatic nature of the ProFHER trial and the fact that, prior to its commissioning, clinical practice for treatment of proximal humeral fracture in the United Kingdom was mixed (that is, some patients were managed surgically, others were managed non-surgically).[1] We use sequential summary statistics based on the trial's outcome and treatment cost data to chart a path for the posterior mean for INMB. Comparing this path with the stopping policy, we establish when the trial would have stopped and carry out bootstrap and Monte Carlo analyses to assess the

[1] Without the Alban et al. (2018) extension, it would not be possible for us to solve the model.

policy's operating characteristics.

Section 2 describes the ProFHER trial and its financial profile, section 3 outlines the version of the model that we apply in section 4. Section 5 discusses our results.

2 The ProFHER trial

The ProFHER trial was a pragmatic, multicentre, randomised controlled trial conducted in the UK National Health Service (NHS) which investigated the use of surgery versus nonsurgical intervention (sling) to treat patients with a displaced proximal humeral fracture. The trial randomised 250 patients aged 16 years and older who presented to orthopedic departments in NHS hospitals between September 2008 and April 2011 to either: (1) surgical treatment, which consisted of fracture fixation with plate and screws to preserve the humeral head, or humeral head replacement, followed by active rehabilitation, or (2) non-surgical treatment, which consisted of sling immobilisation for the injured arm for as long as was thought necessary, followed by active rehabilitation. Follow-up took place after two years, which was later extended to five years. Results of the clinical and economic evaluations are reported in Rangan et al. (2015), Handoll et al. (2015) and Corbacho et al. (2016).

The trial's results suggested that there was no significant difference between surgical intervention and sling in terms of the primary outcome measure, the Oxford Shoulder Score (OSS). Similar results were obtained for a range of other outcomes.[2] The economic evaluation consisted of a cost-utility analysis taking the NHS perspective, using the EQ-5D-3L questionnaire to measure quality-adjusted life years (QALYs). After two years' follow-up it found that, on average, patients randomised to surgery incurred greater costs, and slightly lower QALYs, than patients randomised to sling: surgical intervention for one patient cost an estimated £1,758 more than sling (the 95% confidence interval was (£1,126, £2,389)) and yielded an estimated 0.0101 fewer QALYs (the 95% confidence interval was (-0.13, 0.11)). A five year follow up, designed to check for potentially late-appearing complications or differential improvement/deterioration in function, found the main results unchanged (Handoll et al., 2017).

The ProFHER trial was funded by the National Institute for Health Research, with a total budget of £1,485,585. Figure 1 shows the cumulative research costs incurred over the lifetime of the project (left axis, dashed blue line), together with the cumulative estimate of INMB at one year (right axis, continuous red line, measured in blocks of ten patient pairs at a time[3]), the outcome measure that is the focus of this paper. INMB is calculated as the difference between the point estimate of the net monetary benefit of surgery minus the point estimate of the net monetary benefit of sling.[4] Hence positive values of INMB suggest that surgery is superior, from the cost-effectiveness perspective, and negative values suggest that sling is superior.

We used best judgement to classify costs according to whether they were fixed – i.e. incurred independently of whether patients were being recruited to the trial – and variable. We

[2]The physical component score, the mental component score, complications related to surgery or fractures, the need for secondary surgery to the shoulder, new shoulder-related therapy and mortality.

[3]The path for INMB is drawn assuming that outcomes were observed in blocks of ten patients allocated to sling, plus ten patients allocated to surgery, which is not what happened in practice because of randomisation. We discuss this assumption further in section 4.

[4]For each technology, net monetary benefit is equal to the point estimate of the quality of life measure at one year, multiplied by a maximum willingness to pay of £20,000 for one quality-adjusted life (in line with guidance from NICE 2013), minus the point estimate of the treatment cost of the technology.

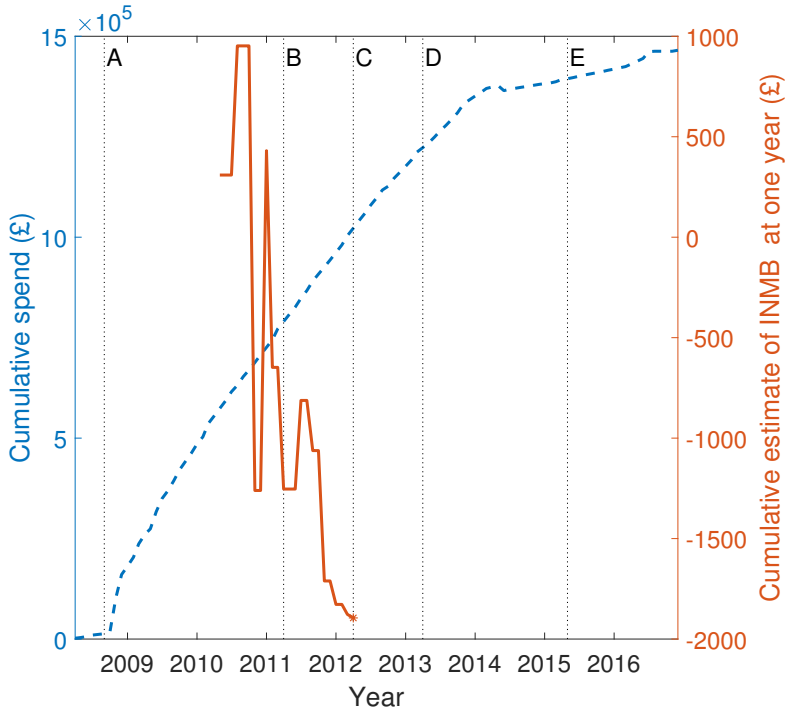


Figure 1: The ProFHER trial's cost profile (left axis) and cumulative estimate of INMB at one year's follow up, plotted in blocks of ten patient pairs (right axis). Letters denote the following: A – recruitment starts; B – recruitment finishes; C – one year follow-up finishes; D – two year follow-up finishes; E – publication of Handoll et al. (2015) and Rangan et al. (2015)

estimated that costs of approximately £161,000 were incurred prior to the recruitment of the first patients in September 2008 (labelled as 'A' in Figure 1). During the recruitment phase (which finished in April 2011, labelled 'B') and the two-year follow-up phase (which finished in April 2013, labelled 'D'), further costs of approximately £1,020,000 were incurred. The main results in Handoll et al. (2015) and Rangan et al. (2015) were reported two years later ('E') and the project concluded at the end of December 2016. Approximately £289,000 of costs were incurred post follow-up, which included covering the tasks of data preparation, cleaning, analysis and report writing. The total spend was therefore approximately £1,470,000. The costs plotted in Figure 1 are those relating to the research project budget itself. Treatment costs were not charged to the ProFHER trial's budget, rather they were funded as part of normal commissioning arrangements within the NHS. For the purposes of this work, we assume that treatment costs would have been the same with or without the trial, on average, across the hospitals participating in the trial.

Blinding meant that the path of INMB shown in Figure 1 was not available to the investigators as the trial progressed. The path shows that, although the estimate of INMB from the first blocks of patients favoured surgery, INMB changed to favour sling in late 2010 and remained that way for the rest of the follow-up period, resulting in an estimate of INMB equal to approximately -£1,800 at the end of follow-up (point 'C').

3 Model of a value-based and sequential clinical trial

Chick et al. (2017) model a two-armed sequential clinical trial in which patients are randomised, in a pairwise and sequential manner, to a new health technology, N, and a control (or standard) health technology, S. The outcomes and costs of treatment for each patient are recorded after a defined follow-up period of $\Delta \geq 0$ units of time. To reflect beliefs concerning the relative cost-effectiveness of the technologies before starting the trial, the model places a prior distribution on the INMB of N compared with S. The objective of the model is to obtain a policy to halt sequential recruitment to the trial. This policy maximises the expected net benefit of carrying out the trial and then recommending one of the two technologies on cost-effectiveness grounds, accounting for any costs incurred in switching technologies, as well as the size of the total population that is expected to benefit from the technology adoption decision.

Denote the effectiveness of the two treatments by the random variables $E_{\mathrm{N}} \in \mathbb{R}$ (for the new technology) and $E_{\mathrm{S}} \in \mathbb{R}$ (for the standard) and define the patient-level treatment costs of using the technologies by the random variables $C_{\mathrm{N}} \in \mathbb{R}_{\geq 0}$ and $C_{\mathrm{S}} \in \mathbb{R}_{\geq 0}$. INMB for pairwise allocation i is the difference between the net benefits of N and S, defined for each technology as the monetary benefit minus the cost, or:

$$X_i = \lambda(E_{\mathrm{N},i} - E_{\mathrm{S},i}) - (C_{\mathrm{N},i} - C_{\mathrm{S},i}), \tag{1}$$

where $\lambda \in \mathbb{R}_{>0}$ is the monetary value of one unit of effectiveness. It is assumed that $X_i \,|\, W \sim \mathcal{N}(W, \sigma_X^2)$, $i = 1, 2, \ldots, Q_{\text{max}}$, where Q_{max} is the maximum number of pairwise allocations that can be made in the trial, σ_X^2 is the variance of INMB in the population (the 'sampling variance'), assumed known for the purposes of this paper, and W is the unknown expected value of INMB in the population (the 'sampling mean'). Beliefs about W are assumed to have a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution. $n_0 = \sigma_X^2/\sigma_0^2$ is its so-called effective sample size.

Assuming a fixed rate of recruitment to the trial, we may express the delay in terms of time, Δ, or pairwise allocations, $\tau \geq 0$. The sequential trial comprises three distinct stages:

1. during Stage I, patients are randomised to the two technologies but no cost or outcome data are observed;

2. during Stage II, cost and outcome data are being observed and there is the option to recruit more patients to the trial or to stop recruitment to the trial;

3. during Stage III, recruitment no longer takes place and outcomes and treatment costs of patients in the 'pipeline' – those who have been treated but who have not yet been followed up for the requisite period of time – are observed.

During Stages II and III, observations are used to update the prior/posterior distribution of W sequentially. If, during Stage II, it is decided not to randomise a further pair of patients, Stage II finishes and the trial moves to Stage III. Once all pipeline patients are followed-up, the adoption recommendation is made for P patients who are expected to benefit from the adoption decision, accounting for any cost incurred in switching from one technology to the other.

Define the adoption decision $\mathcal{D} \in \{\mathrm{S}, \mathrm{N}\}$. Alban et al. (2018) account for the pragmatic nature of the ProFHER trial in which, prior to the start of the trial and owing to the absence of definitive clinical guidance, some patients were treated with surgery and some were treated with sling (implying that fewer than 100% of patients requiring treatment would be affected by

the adoption decision upon the trial's conclusion). We incorporate this extension and define p_{N} as the proportion of patients who are treated with the new technology N (surgery, in the context of the ProFHER trial) prior to the start of the trial, so that the proportion $1-p_{\text{N}}$ receive standard treatment S (sling).

The model assumes a risk-neutral social planner who wishes to maximise expected net benefit, defined as the sum of all patient health benefits, minus the costs of running the trial and switching future patients to one of the two treatments. A valid policy π is defined as a rule which takes the information available in the trial and maps it to an action, which is whether to stop the trial during Stage II or to sample another pair of patients and, once all of the data have been observed, choose the best technology. The information comprises the posterior mean and its effective sample size. The expected reward from carrying out the trial is therefore:

$$\begin{aligned} &V(\pi;\mu_0,n_0) = \\ &-c_{\text{cap}} + \mathbb{E}_\pi\left[-Tc + [\mathbf{1}_{\mathcal{D}_{T+\tau}=\text{N}}(P(1-p_{\text{N}})W - I_{\text{N}}) + \mathbf{1}_{\mathcal{D}_{T+\tau}=\text{S}}(Pp_{\text{N}}(-W) - I_{\text{S}})] \,|\, \mu_0, n_0\right]. \end{aligned} \tag{2}$$

$c_{\text{cap}} \geq 0$ are the fixed costs of the trial, $\mathbf{1}_F = 1$ if F is true and zero otherwise and I_{N} and I_{S} are the costs of switching patients to N and S, respectively. $\mathbb{E}_\pi$ is the expectation induced by policy π and $T \in \{0, 1, \ldots, Q_{\text{max}}\}$ is the number of pairwise allocations made at the time of stopping. A policy π which maximises V may be obtained by using dynamic programming: the discrete time problem is approximated by a continuous time one and the partial differential equation and boundary conditions which result solved. We call a policy π^* which solves the problem an 'Optimal Bayes Sequential' policy.

There are two scenarios in which it is not optimal to enter Stage II: (1) the expected reward of entering Stage II is less than that of running a trial with a fixed number of pairwise allocations in the range $(0, \tau]$. In this scenario, the Optimal Bayes Sequential policy selects the same sample size as a trial designed to maximise the difference between the expected value of sample information and the cost of sampling (we call such a design the 'Optimal Bayes One Stage' design); (2) the value of the prior mean favours one of the two technologies so strongly that the cost of conducting a trial outweighs the benefits. In this scenario, the Optimal Bayes Sequential policy is to run no trial and base the adoption decision on the value of the prior mean alone.

Figure 2 presents the stopping policy for the problem in (sample size $\times$ prior/posterior mean) space. Stages I to III of the sequential trial, as described above, are shown. Before starting the trial, the prior mean μ_0 is compared with the ranges defined on the vertical axis by the letters A, B, C and D. These are defined according to a comparison of the expected value of running the sequential trial, running a trial with $T < \tau$ and running no trial. If μ_0 lies between points C and D, it is optimal to run a sequential trial. For values of μ_0 lying between A and C or D and B, it is optimal to run a trial with $T < \tau$. If the prior mean lies above A or below B, no trial should be run and the adoption decision should be based on the value of the prior mean alone: above A, prior information is strong enough to favour immediate adoption of surgery; below B, it is strong enough to favour immediate adoption of sling. Chick et al. (2017, section 3) provide further discussion.

If it is optimal to run a sequential trial, recruitment of patients takes place during Stage I but no outcomes are observed. At the start of Stage 2, outcomes and treatment costs for the first pairwise allocation are observed and used to update the prior mean. Outcomes then arrive sequentially, the posterior mean is updated and, as long as the posterior mean lies between the upper and lower stopping boundaries, it is optimal to continue to recruit. As soon as the

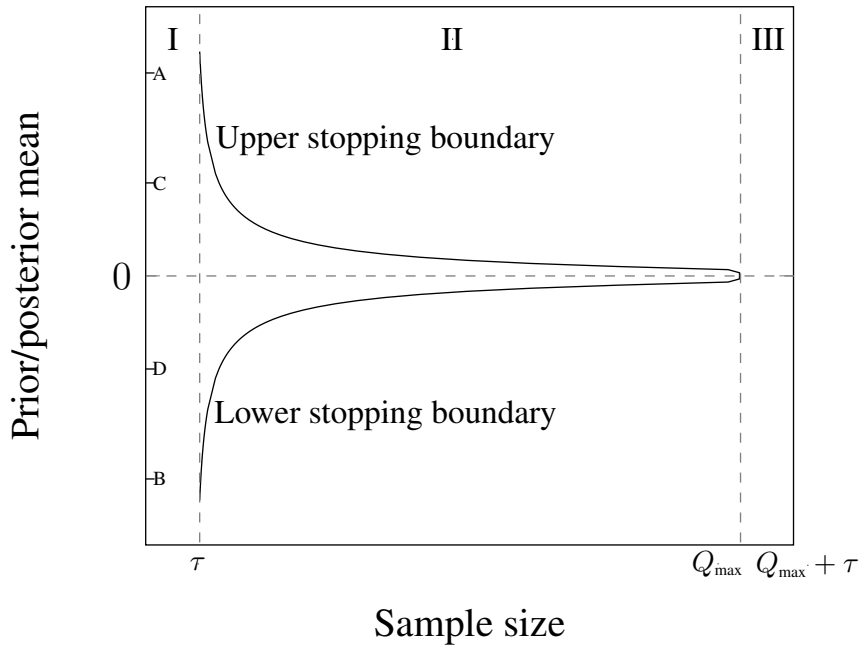


Figure 2: Stopping policy for the Optimal Bayes Sequential problem. Stage III is shown assuming that stopping takes place at the maximum sample size of $Q_{\max}$ pairwise allocations.

posterior mean hits or crosses one of the two boundaries, it is optimal to halt recruitment and move to Stage III.

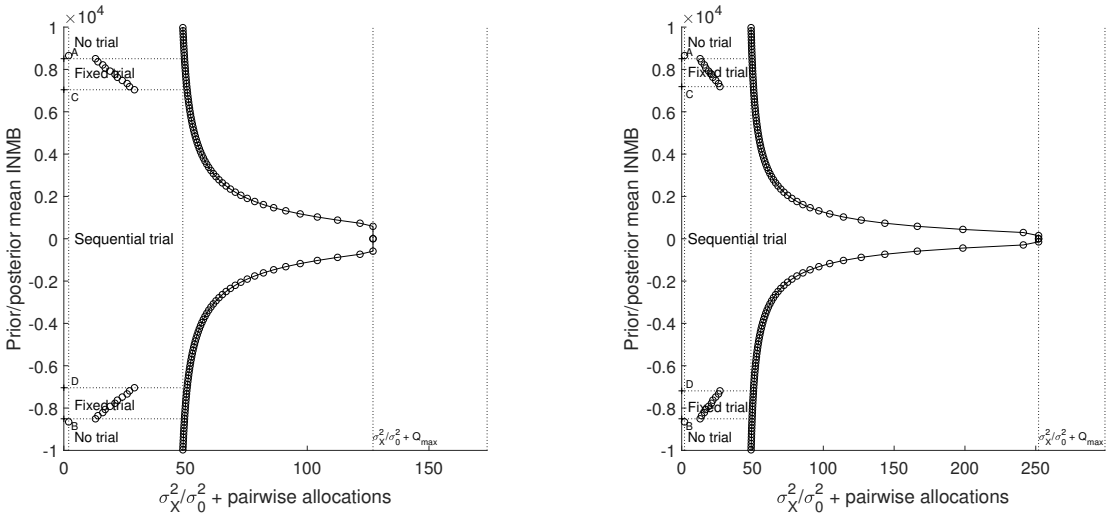
4 Application to the ProFHER trial

The parameter values used for our application to the ProFHER trial, together with their sources, are reported in Table 1. Details of the calculations used to obtain the parameter values are provided in Appendix A. We run two versions of the model. The baseline version assumes that the maximum number of pairwise allocations that could be made, $Q_{\max}$, is equal to 125, so that the sequential trial could recruit a maximum of 250 patients, equal to the actual sample size chosen for the trial. A second version doubles $Q_{\max}$ to 250 pairs of patients. The baseline version ensures that the maximum sample size of the sequential trial cannot exceed that of the ProFHER trial; the latter version offers the opportunity for learning more about cost-effectiveness before deciding to conclude Stage II.

We assume that the delay, Δ, in observing the EQ-5D outcome measure and treatment costs is equal to one year. We assume that the rate of recruitment to the sequential trial is equal to the average rate of recruitment in the ProFHER trial itself (47 pairwise allocations). With follow-up at one year, this implies that the delay, τ, in observing outcomes and treatment costs, is also equal to 47 pairwise allocations. We assume a near non-informative prior, with $\mu_0 = 0$ and n_0 equal to 2 pairwise allocations, representing the lack of evidence for effectiveness and cost-effectiveness at the start of the trial. Finally, the trial costs plotted in Figure 1 are broken down into best estimates of the fixed costs incurred prior to starting recruitment to the trial, costs incurred between the start of recruitment and the end of follow-up (we assume a 50% split between fixed and variable costs during this period), and post-follow-up. This leads to an

Parameter	Definition	Value	Source
p_{N}	Proportion treated with surgery at the start of the trial	0.39	Handoll et al. (2015)
	Fracture incidence rate	7,000 patients per annum	Assumption
	Time horizon for post-decision population	6 years	Assumption
P	Post-decision population to benefit	42,000 patients	Defined from above parameters
σ_X	Sampling standard deviation for INMB	£4,400	ProFHER trial
n_0	Effective sample size for prior distribution	2 pairwise allocations	Assumption
Δ	Delay for observing EQ-5D endpoint (in years)	1 year	Assumption
	Estimated annual rate of recruitment to trial	47 pairwise allocations	ProFHER trial
τ	Delay for observing EQ-5D endpoint (in pairwise allocations)	47 pairwise allocations	Annual rate of recruitment
	Time horizon of trial	32 months	Handoll et al. (2015)
Q_{max}	Total sample size of trial	125 pairwise allocations	Handoll et al. (2015)
I_{N}	Cost of switching $P(1-p_{\text{N}})$ patients to surgery	0	Personal communication
I_{S}	Cost of switching Pp_{N} patients to sling	0	Personal communication
	Estimated spend on fixed costs prior to starting trial	£161,000	ProFHER trial's accounts
	Estimated spend on fixed costs during trial	£510,000	ProFHER trial's accounts
	Estimated spend on variable costs	£510,000	ProFHER trial's accounts
	Estimated spend on fixed costs post follow-up	£289,000	ProFHER trial's accounts
c_{cap}	Total spend on fixed costs	£960,000	ProFHER trial's accounts
	Total spend	£1,470,000	ProFHER trial's accounts
c	Estimated cost per pairwise allocation	£4,080	ProFHER trial's accounts
λ	Monetary value of one QALY	£20,000	NICE (2013)

Table 1: Parameter values



(a) $Q_{\max} = 125$ pairwise allocations (b) $Q_{\max} = 250$ pairwise allocations

Figure 3: Stopping boundaries showing the maximum length of Stages II and III, together with optimal Stage I decisions for two versions of the model ($Q_{\max} = 125$ and $Q_{\max} = 250$)

estimate of the marginal cost per pairwise allocation equal to $c =$ £4,080.

Figure 3 shows the stopping boundaries for the two versions of the model in ($n_0 +$ number of pairwise allocations) $\times$ posterior mean space. The optimal Stage I decisions (run a sequential trial, run a trial with a fixed sample size, run no trial), delineated by the letters A–D as they were in Figure 2, are shown, together with circles showing the sample sizes when $T < \tau$.

4.1 When would the sequential version of the ProFHER trial have stopped?

Figure 4 plots the two Stage II stopping boundaries from Figures 3a and 3b on the same diagram, together with the path of the posterior mean for INMB from the ProFHER trial (in bold black, markers: '$\circ$'). This is drawn using the summary data for effectiveness and treatment costs from the trial, arranged in blocks of ten pairwise allocations and reported in Table 4.[5] The four other paths in Figure 4 are described in section 4.2. Figure 4 shows that, although doubling the maximum sample size increases the maximum length of Stage II, it has little impact on the stopping boundaries.

The first point on the path for the posterior mean, at the start of Stage II and for an effective sample size of 49 pairwise allocations (equal to $n_0 = 2$ plus the delay of 47 pairwise allocations), is equal to the prior mean ($\mu_0 = 0$). Figure 4 shows that, when $Q_{\max} = 125$, Stage II would have concluded after 107 patient pairs had been recruited, with a posterior mean equal to -£1,110. This corresponds to the first point at which the posterior mean lies outside the red stopping boundary. Follow-up of the 47 patient pairs in the pipeline is shown by the remaining

[5]The trial summary data in Table 4 is the same as the data which are used to obtain the path for the cumulative measure of INMB that is plotted in Figure 1, the only difference being that the plot in Figure 1 ca be thought of as being 'frequentist', in that it omits the prior mean and effective sample size. As noted in section 2, patients were not observed in blocks of (ten allocated to surgery + ten allocated to sling) at a time owing to randomisation. We used our best judgement to create such blocks, based on the sequence in which outcomes and treatment cost data was observed, in order to illustrate the path for the posterior mean in as simple a manner as possible.

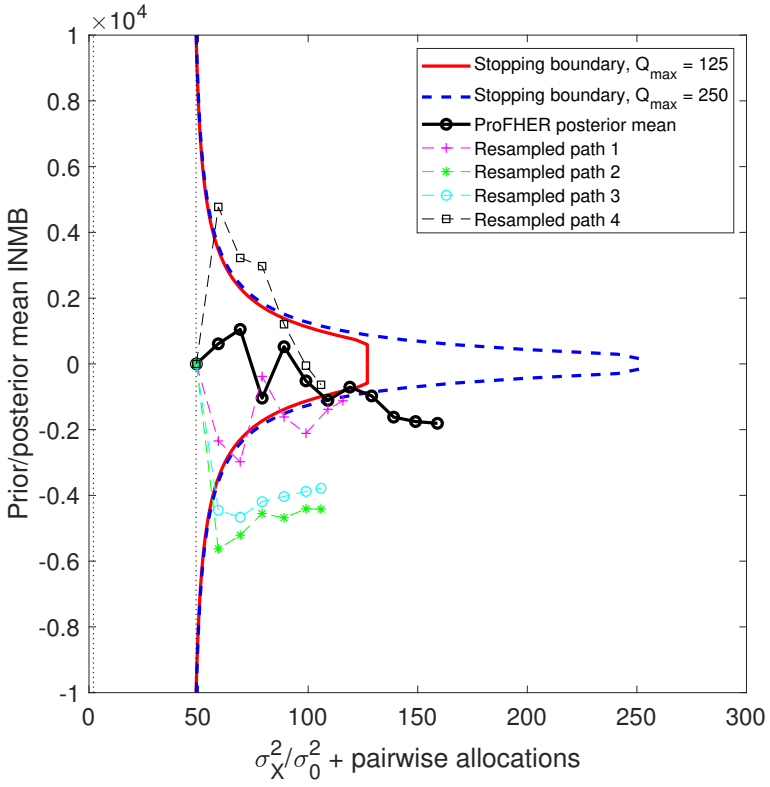


Figure 4: Stopping boundaries for $Q_{\max} = 125$ (red, continuous line) and $Q_{\max} = 250$ (blue, dashed line) pairwise allocations, together with the path for the posterior mean (black, bold line, marker: '○') and four resampled paths (dashed lines, markers: '□', '+', '○' and '*') from the bootstrap

circles on the path and would have led to a decision being made in favour of sling, based on a posterior mean for expected INMB equal to approximately -£1,810. Hence the sequential trial would have stopped earlier, with no change in the adoption decision and little change in the estimate of cost-effectiveness, saving 18 patient pairs and approximately $18 \times$ £4,080 = £73,000 (around 5% of the total cost of the trial).

When $Q_{\max} = 250$, the decision would have been no different, since the first point at which the posterior mean lies outside the blue (dashed line) stopping boundary is the same as that for the case of $Q_{\max} = 125$.[6] Hence, for this application, the stopping decision for stage II, together with the final decision in the trial (favour sling), are relatively insensitive to whether the maximum sample size of the sequential trial is set to the sample size of the ProFHER trial itself, or to double that size.

4.2 Bootstrap analysis

Section 4.1 explored the performance of the sequential trial when applied to the ProFHER trial using the specific sequence of data from the trial itself, arranged in blocks of ten patient pairs. We now analyse the performance of the Optimal Bayes Sequential policy in the average, by taking random draws from that sequence of data using a bootstrap analysis.

We resampled at random, and with replacement, from the data in Table 4. For each iteration

[6] Figure 4 appears to show the point lying precisely on the boundary, which would imply indifference between stopping and continuing Stage II. However, it lies slightly outside the boundary.

	Final decision		
	Sling	Surgery	Total
$Q_{\max} = 125$			
First crossing lower boundary (Sling)	0.814	0.022	0.836
First crossing upper boundary (Surgery)	0.110	0.054	0.164
Total	0.924	0.076	1.000
$Q_{\max} = 250$			
First crossing lower boundary (Sling)	0.824	0.019	0.843
First crossing upper boundary (Surgery)	0.103	0.054	0.157
Total	0.927	0.073	1.000

Table 2: Contingency tables for first crossing and final adoption recommendation decision from the bootstrap analysis

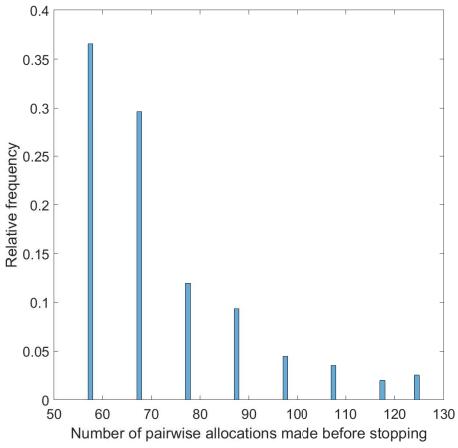
$Q_{\max} = 125$	
Average INMB at stopping (SD; min, max)	-£2,076 (£1,408; -£6,424, £4,753)
Average saving (as % of total spend)	£216,240 (15%)
Average sample size (SD; min, max)	72 (17; 57, 125)
$Q_{\max} = 250$	
Average INMB at stopping (SD; min, max)	-£2,093 (£1,396; -£6,300, £4,465)
Average saving (as % of total spend)	£208,080 (14%)
Average sample size (SD; min, max)	74 (22; 57, 250)

Table 3: Summary statistics for the 100,000 resampled paths used for the bootstrap analysis

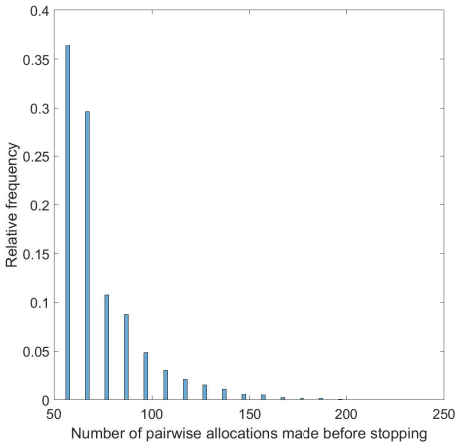
of the bootstrap, we used the sequence of outcome, cost and sample size data, together with the prior mean and prior effective sample size, to create a cumulative path for the posterior mean. For each of the resulting resampled paths for the posterior mean, we established the point at which Stage II would have concluded, by comparing the path for the posterior mean with the stopping boundaries. Four such paths are shown in Figure 4. The pink (markers: '+'), blue (markers: '$\circ$') and green (markers: '*') paths all cross the lower stopping boundary first, before pipeline patients are followed up and the final adoption decision recommends sling. The black dashed path (markers: '$\square$') crosses the upper boundary first, but by the time the pipeline patients are followed up, the posterior mean for INMB is negative and so this path, too, results in a recommendation of sling.

We obtained 100,000 resampled paths for each version of the model ($Q_{\max} = 125$ and $Q_{\max} = 250$) and performance characteristics are summarised in Tables 2 and 3. Table 2 shows that, when $Q_{\max} = 125$, the stopping rule makes the same reccommendation as made in the ProFHER trial (sling) for 92% of the resampled paths. 11% of paths cross the upper boundary first but make a recommendation of sling upon the trial's conclusion; 5% cross the upper boundary first and recommend surgery. 81% of paths cross the lower boundary first and conclude by recommending sling; 2% cross the lower boundary first and conclude by recommending surgery. The lower half of Table 2 shows that these results change only marginally when $Q_{\max}$ is doubled to 250 pairwise allocations.

Table 3 shows that, when $Q_{\max} = 125$, the average of the estimates of the posterior mean for INMB upon the trial's conclusion is -£2,076. The average sample size is 72 pairwise allocations, which is about 58% of the actual sample size of the trial. This represents an average saving of 53 $\times$ £4,080 = £216,240. The lower half of Table 3 shows that, as expected, the average sample sizes and the posterior mean for INMB upon the trial's conclusion change little when the maximum number of pairwise allocations is doubled. The average saving falls

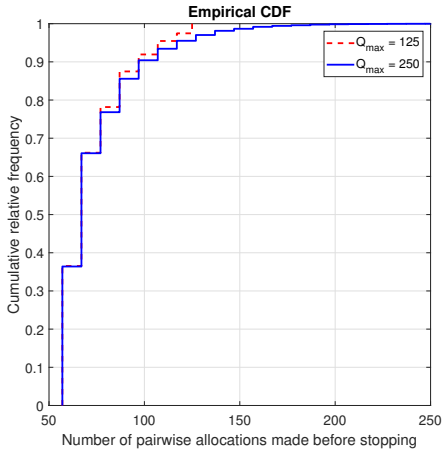


(a) $Q_{\max} = 125$

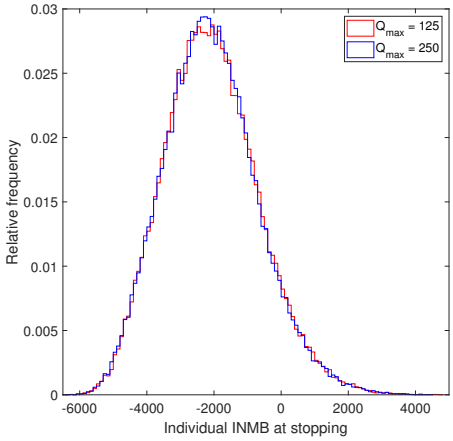


(b) $Q_{\max} = 250$

Figure 5: Relative frequency histogram for the sample size



(a) Empirical CDFs for the stopping time



(b) Histograms for posterior mean INMB

Figure 6: Further graphical analyses from the bootstrap

slightly, to £208,080, owing to the slightly increased average sample size (74 rather than 72 pairwise allocations) and very little gain in the proportion of decisions favouring sling.

Finally, Figures 5 and 6 present some graphical summaries of the bootstrap analysis, which demonstrate some qualitative differences not observed in the numerical results of Tables 2 and 3. Figure 5b shows that setting the maximum sample size to $Q_{\max} = 250$ reduces the height of the bar of the histogram that is seen at $Q_{\max} = 125$ in Figure 5a: by allowing the trial to run for longer, more learning about W takes place. Having a larger maximum sample size also reduces the probability of stopping throughout Stage II, as shown by the empirical CDFs in Figure 6a. However, this makes little difference to the overall performance of the two scenarios, as already seen. Finally, Figure 6b shows that the histograms for the posterior mean for INMB at adoption are almost identical and right-skewed.

4.3 Sensitivity analysis using Monte Carlo simulation

The bootstrap analysis in section 4.2 assumed a single prior mean and a fixed sampling mean. In this section we use Monte Carlo simulation to explore how the sequential model performs when the prior mean is varied over a range of values with sampling means drawn from the resulting prior distributions. We investigate performance characteristics including the average sample size, average reward and the probability of correctly selecting the better technology. We do this for three trial designs: the Optimal Bayes Sequential design with $Q_{\text{max}} = 125$ and $Q_{\text{max}} = 250$ and the design that was used in the ProFHER trial itself (a fixed sample size design with $Q_{\text{max}} = 125$, which we call the 'Fixed' design).

We took a range of values of the prior mean between a lower limit of $-$£10,000 and an upper limit of £10,000. For each value, we took the value of n_0 used for the bootstrap and made 15,000 random draws of W from the resulting distribution. For each draw, we sampled at random from the sampling distribution and used these draws to generate a path for the posterior mean. For each path we calculated the stopping time of the trial and the adoption decision.[7] Results are presented in Figures 7a to 7d.

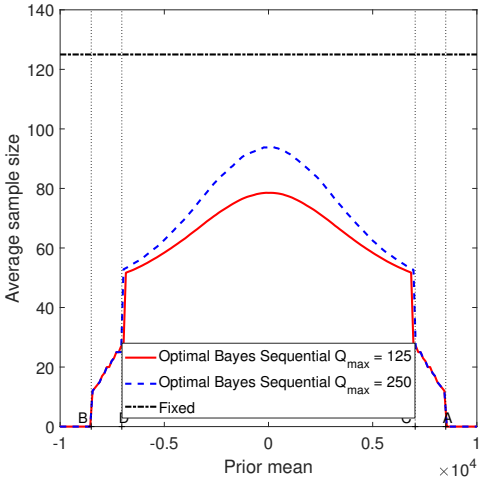
Figure 7a presents the average sample size of the Optimal Bayes Sequential design as a function of the prior mean (red, continuous line for $Q_{\text{max}} = 125$ pairwise allocations; blue, dashed line for $Q_{\text{max}} = 250$) and the Fixed design (black, dash-dot line). The letters A–D correspond to those marked in Figure 2. Figure 7a shows that, for both sequential designs, the average sample size falls the further the prior mean is from zero, reflecting the lower expected sample size that comes when the prior mean favours one of the two technologies. Doubling Q_{max} increases the average sample size.

Figure 7b presents the proportion of times that each design makes the correct technology adoption decision. Doubling the maximum sample size of the sequential design increases the proportion of correct decisions by about one percentage point, reflecting the value of continuing to learn about W. Performance of all three designs is worst in the region of $\mu_0 = 0$, although the correct selection is still made in approximately 96–97% of the replications.

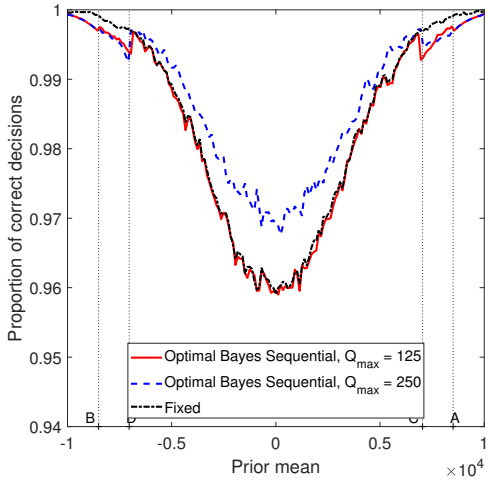
Figure 7c plots the estimate of the 'net gain' of the sequential designs over the Fixed design. Net gain is defined as the difference between the average reward of the Optimal Bayes Sequential and the Fixed design, accounting for both the reward accruing at the point of adoption and the cost of the trial. The sequential designs outperform the Fixed design by between £200,000-£300,000 because both save costs owing to early stopping (Figure 7a), while making a similar proportion of correct decisions (Figure 7b).

Finally, Figure 7d compares the proportion of correct decisions when $Q_{\text{max}} = 125$ from Figure 7b (red, continuous line) with the proportion of correct decisions from what we term a 'frequentist' approach to the MC simulation (green, dash-dot line). For the latter approach, n_0 is set equal to zero, so that the sampling mean is no longer a draw from a prior distribution but is equal to the prior mean for all 15,000 replications. When $\mu_0 = W = 0$, the probability of selecting surgery is equal to one half owing to the fact that the stopping boundaries are symmetric, but it increases the further the W lies from 0.

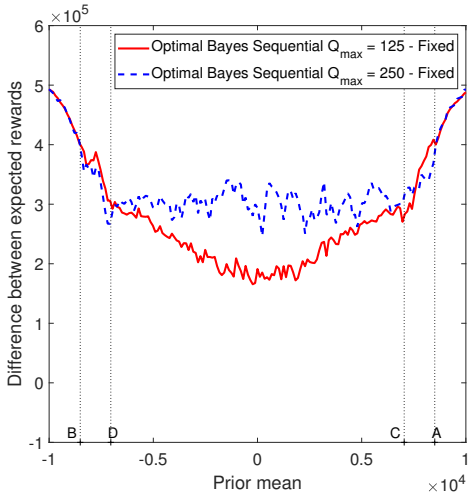
[7] In this subsection, a correct decision is defined according to the value of the draw for W: if $W > 0$ and the adoption recommendation is surgery, this is defined as a correct decision. Similarly, if $W < 0$ and the adoption recommendation is sling, this is defined as a correct decision.



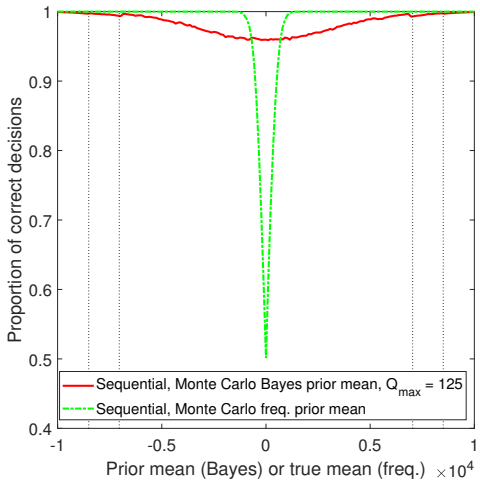
(a) Expected sample sizes



(b) Probability of making the correct selection



(c) Difference in rewards



(d) Probability of making correct selection: comparison of Bayesian and frequentist approaches

Figure 7: Operating characteristics for the Monte Carlo simulation

5 Discussion

With growing interest in value based health care and the use of adaptive and sequential clinical trial designs to assess the effectiveness and cost-effectiveness of new health technologies, there is a need to explore how new approaches to conducting health technology assessments perform, from both economic and more traditional statistical perspectives. In this paper, we have taken a recently published Bayesian decision-theoretic model of a sequential and value-based experiment, together with an extension proposed by Alban et al. (2018), and applied it to the ProFHER trial. We have presented a value-based rule for optimal trial design – run no trial, run a trial with a fixed sample size, run a sequential trial – and assessed performance criteria. Using the data as it accumulated from the trial itself, results show that, had the model's stopping policy been used for the sequence of data observed in the trial, it could have stopped

early, saving about 5% of the research budget (approximately £73,000). A bootstrap analysis based on averaging over resampled paths drawn from the trial's data suggests that the sample size would have been reduced by approximately 40%, saving 15% of the budget, with 93% of resampled paths making a decision consistent with that of the trial itself.

Analysis of the distribution of stopping times shows a skewed distribution, with approximately 35% of resampled paths stopping at 57 pairwise allocations. This means that these 'trials' would have concluded with a sample size of just under half of the patients who were recruited to the ProFHER trial itself. Given that the ProFHER trial has seen mixed uptake of its results in the years following publication of Handoll et al. (2015) and Rangan et al. (2015) (Jefferson et al. (2017) provide an assessment of the impact of the findings of the ProFHER trial on surgeons' clinical practice), it is important to consider whether surgeons would consider such a small sample size credible enough to change practice. One way to address this matter would be to adapt the reward functions in the model to make the probability of uptake of the recommended treatment a function of the strength of the results of the trial – namely, the value of the posterior mean and the effective sample size at the time of stopping – or to impose a lower bound on the stopping time of the model, above which it is thought that the results would be acceptable to health care practitioners. Pursuing these options would involve trading-off some of the expected savings of the trial with the expected benefit of an increased probability that the trial influences practice.

There are many directions for future research. Firstly, the rewards of the adoption decision are a function both of the size of the population to benefit and the cost of switching technologies. The former is hard to estimate pre-trial, when there exists a large degree of uncertainty over the number patients who would meet the trial's inclusion criteria (and who therefore would benefit from the adoption decision in the post-trial period). There is also uncertainty over the time horizon for which the adoption decision will be implemented. The latter requires an assessment of the cost of switching technologies. Secondly, there is the question of defining the fixed and variable costs of the trial itself: the stage II stopping boundaries are a function of the variable costs alone; the decision about which trial design is optimal accounts also for the fixed costs. Finally, the model that we apply assumes that the sampling variance is known, which is unlikely to be the case in practice. This requires that either an estimate of the sampling variance be used at the start of the trial (perhaps with a sensitivity analysis to investigate how the Stage II stopping boundaries changes as the sampling variance changes) or that the methods of Chick et al. (2017, Section 4) are used to obtain bespoke boundaries for the case of unknown variance.

A Parameter values and data from the ProFHER trial

Table 4 presents the point estimates of quality of life and costs, arranged in blocks of ten patient pairs, from the ProFHER trial. These data are used to plot the path for the point estimate of INMB in Figure 1 (ignoring block 0, which refers to the prior mean and its effective sample size). They are also the point estimates used to carry out the bootstrap analysis of section 4.2.

The parameter values used for the analysis in section 4 are sourced and calculated as follows:

1. Estimate of the proportion treated with sling at the start of the trial, p_N: taken from Handoll et al. (2015, page 104). Of 313 non-consenters in the ProFHER trial, 66 were assigned to surgery, 105 to sling, 118 were classified as 'uncertain' and data were missing for the remaining 24. Assume that non-consenters were representative of the overall patient population for which the ProFHER trial was designed and that patients with missing data for treatment, together with those classified as 'uncertain', do not systematically differ from the study population either. Then it is estimated that $p_N = 0.39$ (= 66/171).

	EQ-5D-3L		Treatment costs		Number of observations			
Block	$\bar{E}_{\text{N}}$	$\bar{E}_{\text{S}}$	$\bar{C}_{\text{N}}$	$\bar{C}_{\text{S}}$	$n_{\bar{E}_{\text{N}}}$	$n_{\bar{E}_{\text{S}}}$	$n_{\bar{C}_{\text{N}}}$	$n_{\bar{C}_{\text{S}}}$
0 (Prior)	0	0	0	0	2	2	2	2
1	0.74	0.67	3166	2102	10	10	4	6
2	0.85	0.69	1855	47	10	9	6	6
3	0.66	0.84	2464	120	9	9	5	7
4	0.77	0.44	2191	1150	10	9	5	7
5	0.66	0.73	3921	32	10	10	4	5
6	0.68	0.78	2854	582	10	9	4	8
7	0.78	0.61	2549	1223	10	9	6	3
8	0.60	0.71	3081	3028	10	10	7	5
9	0.52	0.73	2689	20	10	10	5	5
10	0.75	0.81	2434	821	9	10	9	6
11	0.62	0.65	1918	27	10	9	5	3

Table 4: Point estimates of EQ-5D and cost data from the ProFHER trial, together with the number of observations used to obtain each point estimate, arranged in blocks of ten patient pairs. Also included is a row for the prior mean (block = 0), together with the effective sample size of the prior Figure 4.

2. Estimated of switching costs: from personal communications it was believed that these costs would be minor and they are assumed to be equal to zero.

3. Estimate of the sampling variance, σ_X^2: from the 95% confidence interval for INMB at one year provided by the ProFHER trial's data. The point estimate of INMB at one year was -£1601.66 and the upper limit of the 95% confidence interval was -£458.06, based on approximately 60 pairwise allocations. Using a critical value of $t = 2$, $\sigma_X \approx \sqrt{60} \times$ (-£458.66 + £1601.66)/2 $\approx$ £4,400.

4. Estimate of P, the post-decision population to benefit: there appears to be no reliable information on the annual incidence rate of fractures meeting the inclusion criteria for the ProFHER trial. We therefore estimated P using information from a number of sources. Corbacho et al. (2016, page 7) report that there were 3,519 first listed consultant episodes for patients with fractures of the proximal humerus which involved an operation during 2011–12.They assume that 80% of these were displaced fractures involving the surgical neck. They make the conservative assumption that 50% of these cases may change from surgical intervention to non-surgical intervention as a result of the ProFHER trial and calculate a £2.5m saving to NHS England (i.e. 3,519 x 0.8 x 0.5 = 1,408 patients x ΔC= £1,758 = £2.5m). Treatment using sling is classified as an outpatient appointment in the UK, and there are no data on the number of sling administrations that took place during 2011–12. Given that Corbacho et al. (2016) estimate that there were 2,815 (= 0.8 x 3,519) cases of fractures of the proximal humerus involving the surgical neck during 2011-12, we use p_{N} from point 1 above to estimate that 4,403 (2,815 x (1 / (0.39) - 1)) patients would have been treated with sling. We therefore estimate an annual incidence rate of 2,815 + 4,403 $\approx$ 7,000 patients who may be treated either with surgery or sling. We combine this with a total duration for implementing the decision resulting from the trial which is equal to 6 years, so P = 6 $\times$ 7,000 = 42,000.

5. Estimate of c, the marginal cost per pairwise allocation, is calculated using the financial records from the trial (those used to produce Figure 1). Approximately £161,000 was spent prior to recruiting the first patients. This is classified as the fixed set-up cost of the trial. An estimated 50% of the £1,020,000 of costs incurred between the start of patient recruitment and the finish of follow-up is taken to be the variable cost of the trial, giving an estimate of the marginal cost of adding one pairwise allocation to be £510,000/125 = £4,080. The remaining 50% is taken to be a cost (such as overheads) which would have been incurred during the recruitment phase even if no patients were being recruited. Finally, costs of £289,000 are incurred post follow-up. This gives a total spend of £1,470,000.

References

Alban, A., Chick, S. E., and Forster, M. (2018). Update on a Bayesian decision-theoretic approach to value-based clinical trial design. In Rabe, M., Juan, A., Mustafee, N., Skoogh, A., Jain, S., and Johansson, B., editors, *Proceedings of the 2018 Winter Simulation Conference*, page to appear, Piscataway, NJ. IEEE, Inc. 2, 5, 14

Bhatt, D. L. and Mehta, C. (2016). Adaptive designs for clinical trials. *The New England Journal of Medicine*, 375:65–74. 2

Chick, S. E., Forster, M., and Pertile, P. (2017). A Bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society, Series B*, 79(5):1439–1462. 2, 5, 6, 15

Corbacho, B., Duarte, A., Keding, A., et al. (2016). Cost effectiveness of surgical *versus* non-surgical treatment of adults with displaced fractures of the proximal humerus. *Bone and Joint Journal*, 92-B(2):152–159. 2, 3, 16

Cui, L., Zhang, L., and Yang, B. (2017). Optimal adaptive group sequential design with flexible timing of sample size determination. *Contemporary Clinical Trials*, 63:8–12. 2

DiMasi, J., Grabowski, H., and Hansen, R. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47:20–33. 2

Handoll, H., Brealey, S., Rangan, A., et al. (2015). The ProFHER (proximal fracure of the humerus: Evaluation by randomisation) trial - a pragmatic multicentre randomised controlled trial evaluating the clinical effectiveness and cost-effectivesness of surgical compared with non-surgical treatment for proximal fracture of the humerus in adults. *Health Technology Assessment*, 19:1–280. 2, 3, 4, 8, 15

Handoll, H. H., Keding, A., Corbacho, B., Brealey, S. D., Hewitt, C., and Rangan, A. (2017). Five-year follow-up results of the PROFHER trial comparing operative and non-operative treatment of adults with a displaced fracture of the proximal humerus. *Bone and Joint Journal*, 99-B:383–392. 3

Jefferson, L., Brealey, S., Handoll, H., Kelding, A., Kottram, L., Sbizzera, I., and Rangan, A. (2017). Impact of the PROFHER trial findings on surgeons' clinical practice. an online questionnaire survey. *Bone and Joint Research*, 6:590–599. 15

NICE (2013). Guide to the methods of technology appraisal. `https://www.nice.org.uk/process/pmg9/chapter/foreword`, Accessed 2 Feb 2019, London. 3, 8

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16. 2

Rangan, A., Handoll, H., Brealey, S., et al. (2015). Surgical vs nonsurgical treatment of adults with displaced fratures of the proximal humerus: the PROFHER randomized clinical trial. *Journal of the American Medical Association*, 313(10):1037–1047. 2, 3, 4, 15

Taylor, R. S., Drummond, M. F., Salkeld, G., and Sullivan, D. S. (2004). Inclusion of cost effectiveness in licensing requirements of new drugs: the fourth hurdle. *British Medical Journal*, 329:972. 2

U.S. Congress (1938). Act of June 25, 1938 (Federal Food, Drug, and Cosmetic Act), Public Law 75-717, 52 STAT 1040. 2

U.S. Congress (1962). Act of October 10, 1962 (Drug Amendments Act of 1962), Public Law 87-781, 76 STAT 780. 2

Wason, J., Magirr, D., Law, M., and Jaki, T. (2016). Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 25(2):716–727. 2

Yin, G., Lam, C., and Shi, H. (2017). Bayesian randomized clinical trials: from fixed to adaptive design. *Contemporary Clinical Trials*, 59:77–86. 2