Subjective Bayesian Methods in the Design and Analysis of Clinical Trials

Cameron Williams

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

May 2022

Acknowledgements

I would firstly like to thank my supervisors, Dr Kevin Wilson, Dr Nina Wilson, and Prof John Matthews. Their advice, guidance, and feedback over the last years has been invaluable, and I couldn't have asked for a better supervisory team.

I am also extremely appreciative of all of the experts who gave their time to take part in the elicitations. Without their generosity, much of the work in this thesis could not have been completed at all.

I would also like to thank Newcastle University's School of Mathematics, Statistics and Physics. Their support, financial and otherwise, has provided me this opportunity which otherwise would not have been possible.

I am very grateful to the friends I have made during my PhD, in particular my fellow members of PhD Office 7, and those who took part in slightly too long lunch breaks and after work badminton games. Without them, the last few years would not have been nearly as enjoyable as they were.

I would also like to thank the Elliott family for allowing me to flee to their Northumbrian pit village during the COVID-19 lockdowns in 2020. The alternative (remaining locked down in a small mouldy flat, without the company of two friendly dogs) would not have been particularly nice.

I would also like to thank my Australian friends and family. Although we have seen each other far too infrequently, I have appreciated their support and friendship over the years.

Last, but not at all least, I would like to thank my partner, Lucy, for all of her support, encouragement, and baked goods.

Abstract

Assurance provides a Bayesian alternative to commonly used frequentist sample size calculation methods. As part of sample size calculations, an estimate of a treatment's effect size or a test's accuracy is typically required. When using Bayesian methods, these unknown quantities can be represented with a prior distribution, rather than using a single point estimate, allowing for more nuanced information about the unknown quantity to be incorporated into the sample size calculation.

In this thesis, we first review common sample size calculation methods and elicitation techniques. We consider the problem of aggregating expert prior beliefs to form a single prior distribution, to be used in sample size calculations. Common methods of prior distribution aggregation include mathematical methods, which use a mathematical rule to combine priors, and behavioural methods, which provide experts with a framework to assist them in creating an aggregate prior during a group discussion.

Though not a recent development, assurance is not commonly used in practice. We provide a case study of a diagnostic study, investigating a novel diagnostic test for Motor Neurone Disease, for which prior distributions are elicited and aggregated across experts, and sample size calculations are conducted using both frequentist and assurance methods.

As a result of the requirements involved in using each method of aggregation, few comparisons between behavioural and mathematical aggregation methods exist. In order to make comparisons, we structured a series of elicitations as part of the case study. We demonstrate how any method of aggregation outperforms individual experts, and that the Sheffield Elicitation Framework and Classical Method perform best out of the aggregation methods compared. We also demonstrate that all of the considered aggregation methods perform better than a randomly selected individual expert.

In order to explore the behaviour of assurance, we provide a number of simulation studies comparing assurance and power calculations. We investigate the sensitivity of power and assurance to changes in input parameters, the effect of misrepresenting an effect size, and the effect of using different prior distributions in the design and analysis stages of assurance calculations. We consider these behaviours for both Normal and binomial observations.

We use the resulting aggregated prior distributions for assurance and power calculations, to determine appropriate sample sizes within the case study and more generally. We compare assurance calculations with different priors, analysis methods and target values to further demonstrate differences between assurance and power, and their properties. We demonstrate how the choice of model and prior distribution can have a large impact on the final results of a sample size calculation.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Background

An important step in the design of many experiments, including clinical trials and diagnostic accuracy studies, is the calculation of a required sample size. This number informs researchers of the size of the sample they need to collect in order to make scientifically valid conclusions. However, this is often challenging in practice.

In order to calculate a required sample size, most standard approaches require a value for the effect size of interest. Of course, if the effect size of interest was known accurately then there is little need for the experiment in the first place. As such, this effect size must be estimated, or otherwise specified, for the sample size calculations. Often, this can be done by selecting the minimum effect size the researchers deem practically significant, or based on previous studies or expert knowledge.

If the effect size used in the sample size calculations does not correspond closely with what is observed in the experiment, then the results of the analysis may not be sound. One method which may provide a more robust sample size calculation is the use of Bayesian assurance. Instead of a single point estimate for the effect size, a prior distribution is placed on the effect. This means that a range of possible effect sizes are accounted for within the assurance calculation.

While there has been much work developing and applying Bayesian assurance as a method for sample size calculation, there has been less literature investigating the effect of elicited prior distributions, whether from individual experts or aggregated priors. In this thesis, through simulation and a case study, we investigate the effects of prior distributions on assurance calculations, and make comparisons to commonly used statistical power calculations.

In the wider elicitation literature, there have been few attempts to compare behavioural and mathematical approaches of expert judgment aggregation. While many mathematical aggregation methods have been compared, we are not aware of any comparisons between mathematical and behavioural aggregation methods. Furthermore, there are few comparisons between Bayesian and opinion pooling methods of aggregation. In this thesis, we present novel comparisons between expert judgement aggregation methods for prior distributions, some of which have been published in Williams et al. (2021).

1.2 Bayesian Inference

Frequentist statistics typically makes inferences based on hypothesis tests and confidence intervals, and utilises only observations in parameter estimation. Parameters are assumed to have some fixed value to be inferred, and uncertainty is confined to the observations. As a result, testing involves the probability of the observations producing a test statistic at least as extreme than the one observed, assuming a null hypothesis is true.

Bayesian inference, by contrast, considers inferences from an alternative point of view. Instead of assuming parameters take fixed values to be estimated, they are regarded as uncertain quantities to be given probability distributions. In our context, this allows for a probability distribution to be placed on the possible effect sizes.

In order to do this, Bayesian inference utilises Bayes theorem which, for observations X and a parameter θ.

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)} \tag{1.1}$$

where $P(\theta \mid X)$ is the posterior distribution, $P(X \mid \theta)$ is the likelihood, $P(X)$ is a normalising constant, and $P(\theta)$ is the prior distribution.

The prior distribution represents the state of knowledge about θ before incorporating the observations, the likelihood represents the probability of observing the data conditional on θ and the posterior distribution represents the state of knowledge about θ having observed the data.

The likelihood is determined by the type of data being collected. In the cases considered within this thesis, we focus on binomial and normal observations. The form of the prior distributions is typically chosen by the statistician performing the analysis.

There is a view that the requirement for the statistician to choose a prior distributions brings bias and subjectivity to an analysis. One way that this is often addressed is through uninformative priors. A common choice for an uninformative prior is a flat, or uniform, prior distribution. The probability density function for this type of prior is

$$f(\theta) = \frac{1}{b-a} \tag{1.2}$$

where a and b are the minimum and maximum of the range for the parameter θ. In the case of parameters with an infinite support, a flat distribution can still be used. Such a prior distribution, which does not integrate to one, is known as an improper prior distribution.

One common criticism of the use of uniform prior distributions as uninformative priors is that they do not always provide an impartial or unbiased representation of the parameter. For example, a prior distribution which is uniform will not be uniform when transformed, such as on a log or odds scale. The choice of whether the parameter should be uniformly distributed, or the log or odds-transformed parameter should have a uniform distribution, then influences the resulting posterior distribution.

Additionally, if we consider a parameter bounded between zero and one, for example, a uniform distribution would suggest each parameter value is equally likely. However, if interest lay in whether the parameter was greater than 0.01, this uniform prior would provide a 99% probability of this being the case. This would then suggest there is strong prior evidence in favour of a noteworthy result, despite the prior choice aiming to be uninformative.

The subjectivity of a prior distribution can be instead viewed as a benefit of Bayesian statistics (Goldstein, 2006). If there is information known prior to the analysis, it can be included within the prior distribution. For example, this could be in the form of results from previous experiments or, as outlined in Chapter 3, as elicited information from experts in a relevant field.

An important consideration is whose beliefs or knowledge is being represented by the prior distribution. Given a single expert, it makes sense that the prior represents their beliefs. In the case of a group of experts, however, an aggregated prior distribution may not be representative of any single expert's personal views. In such a case, the aggregated prior distribution may be thought to represent the views of a decision maker, who considers information provided by the group. The prior does not necessarily have to belong to a real individual. The Sheffield Elicitation Framework provides a method of elicitation, outlined in Chapter 3, which aims to elicit the opinion of an imaginary third party who has taken into account the evidence and opinions provided by a group of experts (O'Hagan, 2019).

1.2.1 Computation of Posterior distributions

In many of the analyses presented in this thesis, the prior distribution can be chosen to be conjugate to the likelihood. Conjugacy means that the posterior distribution is of the same form as the prior distribution, allowing for a tractable analytical solution. For binomial observations, modelled with a binomial likelihood, the conjugate prior is a Beta distribution. The parameters of the Beta posterior distribution can be calculated analytically, given the prior distribution parameters and observed data.

However, when aggregating prior distributions together mathematically, the resulting distribution is often a pool of multiple distributions. Such distributions may no longer take the form of a simple parametric distribution, and instead require computational methods in order to obtain a posterior distribution (Gelman et al., 2013). One common set of algorithms which are used to do this are Markov chain Monte Carlo, or MCMC, algorithms.

MCMC methods construct a Markov chain, which has the posterior distribution as its stationary distribution. When the Markov chain has reached this stationary distribution, the MCMC is said to have converged to the posterior distribution. It is not known with certainty, however, when the algorithm has reached this convergence. As such, a number of diagnostic methods can be employed. The simplest methods may involve checking trace plots to ensure there is no longer any drift in the sampled values, or checks for autocorrelation (Roy, 2020). There are many additional statistics which can also be calculated to check for convergence (Cowles and Carlin, 1996; Plummer et al., 2006; Smith, 2007).

This thesis will utilise both the JAGS and Stan programs, through the 'rjags' package (Plummer, 2016) and 'stan' package (Carpenter et al., 2017), in order to obtain posterior samples when the prior distribution is not conjugate to the likelihood function.

JAGS, similar to the program BUGS (Lunn et al., 2000), obtains samples from the posterior distribution using Gibbs sampling. Gibbs samplers work by taking draws of each parameter from their full conditional distributions conditional on the most recent draws of all the other parameters in the model.

For example, if we consider a model with data, $\underline{y}$, and three parameters, θ_1, θ_2, and θ_3, the m'th draw of the posterior for each will be given by

$$\theta_1^{(m)} \sim p(\theta_1 \mid \theta_2^{(m-1)}, \theta_3^{(m-1)}, \underline{y}) \tag{1.3}$$

$$\theta_2^{(m)} \sim p(\theta_2 \mid \theta_1^{(m)}, \theta_3^{(m-1)}, \underline{y}) \tag{1.4}$$

$$\theta_3^{(m)} \sim p(\theta_3 \mid \theta_1^{(m)}, \theta_2^{(m)}, \underline{y}) \tag{1.5}$$

By running the sampler for a large number of m iterations, a sample from the posterior is obtained. Gibbs samplers requires initial starting values, when $m = 1$. It may take some time for the sampler to converge to the posterior distribution from these initial values, and so initial draws are often discarded from the posterior samples. By removing these initial values, and keeping only those which have converged to the posterior distribution, the sample should only contain draws from the posterior distribution.

Stan uses a Hamiltonian Monte Carlo (HMC) algorithm in order to obtain posterior samples. HMC uses an additional momentum parameter to more efficiently sample the parameter space (Neal, 2011). This efficiency comes at the cost of increased computational complexity, and as such, both Stan and JAGS have particular advantages and disadvantages for different classes of models.

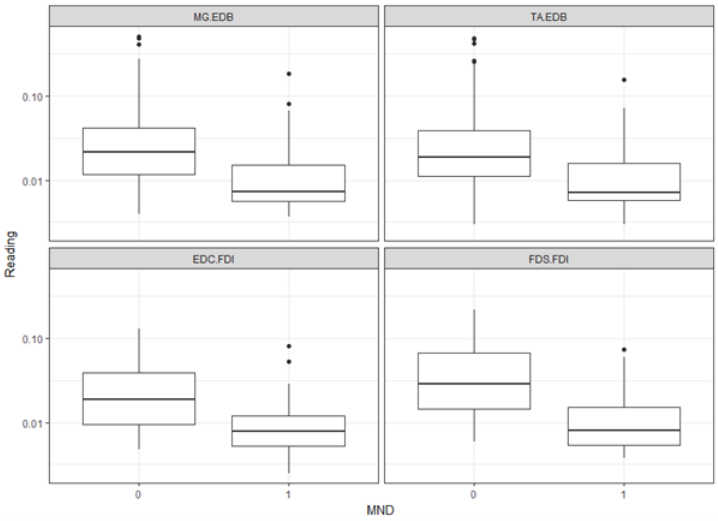


Figure 1.1: BIMC readings from patients with and without MND, log scaled. Patients known to have MND are labelled 1, and those known not to have MND are labelled 0.

1.3 Clinical trial case study

Throughout this thesis, we utilise a diagnostic accuracy study as a case study for the methods discussed. This study had been designed and submitted for review, with feedback suggesting the designers should consider the inclusion of Bayesian methods.

The study focuses on the diagnosis of Motor Neurone Disease (MND). Presently, MND is often diagnosed using the Awaji criteria (de Carvalho et al., 2008; Costa et al., 2012), which involves looking for degeneration of both Upper Motor Neurones (UMN) and Lower Motor Neurones (LMN). While evidence of UMN degeneration is often detected using electromyography, the LMN are usually assessed physically by a clinician. This assessment involves physical manipulation of limbs, and is susceptible to varying ability and experience of clinicians, and difficulty in detecting early-stage or low levels of neurone degeneration.

A new diagnostic test has been developed, and is aimed to replace the clinician's assessment of the LMN. The Beta-band Intermuscular Cohesion (BIMC) test takes electrical readings from pairs of muscles to assess the neurones. This allows for a more standardised test, which may be able to detect deterioration before physical symptoms become severe. It is proposed that BIMC could be used within the Awaji criteria to improve diagnoses.

Figure 1.1 provides initial results from a previous laboratory study. BIMC readings between pairs of muscles, two pairs from the leg (MG.EDB and TA.EDB) and two pairs from the arm (EDC.FDI and FDS.FDI), are provided from a group of patients known to have, or not have, MND from a previous diagnosis. As the plot shows, those with MND tend to have lower readings from the BIMC test.

While the BIMC test is designed to be used alongside further criteria for a diagnosis,

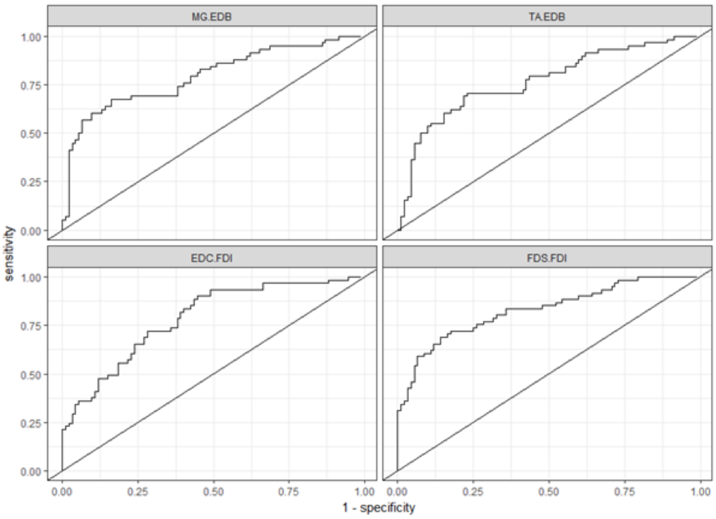


Figure 1.2: BIMC receiver operating characteristic curve

Table 1.1: BIMC Readings

BIMC Reading	Area under curve	PPV	NPV
MG.EDB	0.782	0.931	0.315
TA.EDB	0.759	0.931	0.239
EDC.FDI	0.775	0.967	0.25
FDS.FDI	0.820	0.902	0.391

we also consider how it may perform alone. Figure 1.2 provides receiver operating characteristic (ROC) curves for each muscle pair measured. These curves show the sensitivity and specificity achievable by the BIMC test for different cutoffs. The further the curves are from the diagonal, the better the test is performing. This figure suggests that each of the muscle pairs seems to perform similarly.

Table 1.1 provides further details. The area under an ROC curve can be used to measure the performance of a test, with the range of possible values of the area between 0.5 and 1. As it shows, using the test on the FDS.FDI muscle combination performs best.

The table also considers positive predictive values (PPV) and negative predictive values (NPV). The PPV is the proportion of positive results which are true positives, and the NPV the proportion of negative results which are true negatives. The PPV and NPV values in Table 1.1 have been found by choosing a cutoff which maximises the minimum value of the sensitivity and specificity. It appears from these results that the BIMC test may have some reasonable diagnostic ability, which may further be improved by being used within the Awaji criteria.

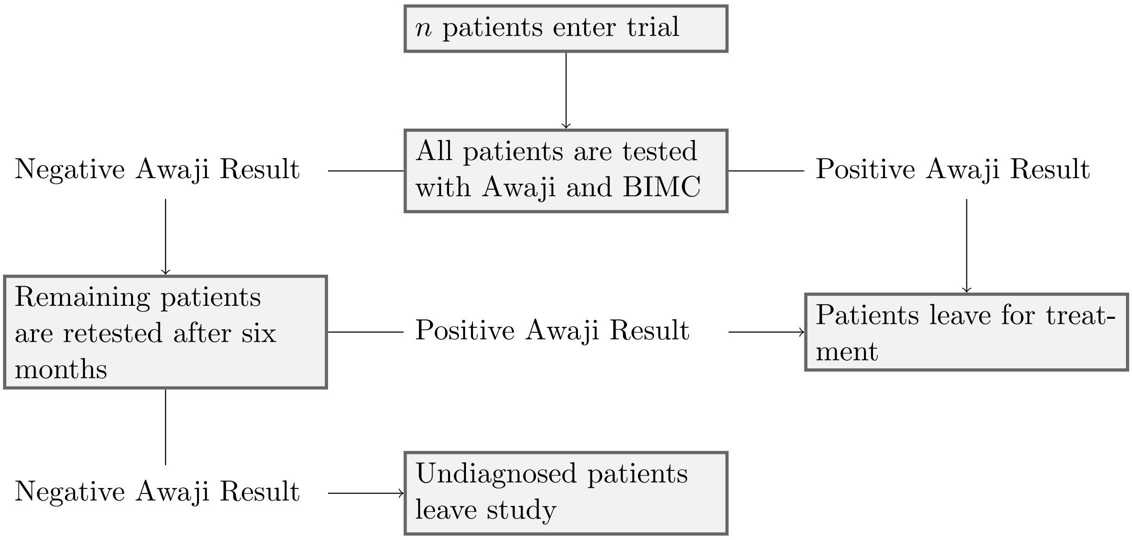

n patients enter trial
All patients are tested with Awaji and BIMC
Negative Awaji Result
Positive Awaji Result
Remaining patients are retested after six months
Positive Awaji Result
Patients leave for treatment
Negative Awaji Result
Undiagnosed patients leave study

Figure 1.3: Format of BIMC study.

1.3.1 Trial Design

The trial aims to compare current Awaji criteria diagnosis with an Awaji criteria diagnosis including a BIMC test. The trial designers anticipate that the inclusion of the BIMC test as part of the Awaji criteria will allow for MND to be diagnosed earlier in patients than use of the Awaji criteria as it is currently used.

Figure 1.3 outlines the design of the study. Patients are initially tested using both the Awaji criteria and the BIMC test. Those who receive a positive diagnosis from the Awaji criteria will leave the study to receive treatment at this point. The results from the BIMC test will not affect the treatment of patients during this study.

After a further six months, those patients who remain in the study will be tested under the Awaji criteria again. Those who receive a positive Awaji criteria diagnosis will be progressed to treatment for the disease. Due to the speed at which MND usually progresses, the researchers anticipate that any of the initial patients with MND will be diagnosed by this stage. As such, for the purpose of this study, the Awaji criteria diagnosis after six months is considered the reference standard against which BIMC will be compared. The Awaji criteria can henceforth be referred to as the reference test (RT), while the BIMC test alongside the Awaji criteria can be referred to as the experimental test (ET).

It is the aim of this study to investigate whether the BIMC test will be able to positively identify patients with MND at the initial time point, who would otherwise have been diagnosed at the later time point after six months. This earlier diagnosis has many positive implications for the treatment of patients with MND and recruitment of patients to future trials into new treatments for MND.

The results from the trial were to be analysed using McNemar's test. We will also present Bayesian alternatives, to allow for a fully Bayesian design and analysis of the study. In order to calculate a sample size based on McNemar's test, estimates of the number of patients receiving each combination of RT and ET results were required. The Bayesian approaches are based around the proportion of patients who are diagnosed six months earlier using BIMC than they would otherwise be with the Awaji criteria alone. The model parameters which are to be elicited have been chosen so they can be used in both sample size calculations.

McNemar's test is outlined in Chapter 2, and the Bayesian model is as follows.

Of the n patients initially recruited for the trial, n_1 will receive a positive Awaji diagnosis at the first time point. We model n_1 as coming from a binomial distribution, with probability η,

$$n_1 \mid \eta \sim Binomial(n, \eta) \tag{1.6}$$

Of the remaining $n - n_1$ patients, n_2 will receive a positive Awaji diagnosis at the second time point. Again, we model n_2 as coming from a binomial distribution, with probability μ,

$$n_2 \mid \mu \sim Binomial(n - n_1, \mu) \tag{1.7}$$

The remaining $n - n_1 - n_2$ patients will then receive a negative Awaji diagnosis at both time points in the trial, and thus not be diagnosed with MND during the course of the trial.

We consider the total number of patients receiving a positive BIMC test, b, as belonging to one of three groups. The first group, b_1, is those who also received a positive Awaji test at the first time point, the second, b_2, those who received a positive Awaji test at the second time point, and third, b_3, those who did not receive a positive Awaji test at all. These three groups are modelled as binomial distributions, with probabilities θ_1, θ_2, and θ_3 respectively,

$$b_1 \mid \theta_1 \sim Binomial(n_1, \theta_1) \tag{1.8}$$
$$b_2 \mid \theta_2 \sim Binomial(n_2, \theta_2) \tag{1.9}$$
$$b_3 \mid \theta_3 \sim Binomial(n - n_1 - n_2, \theta_3) \tag{1.10}$$

We can alternatively express each count in terms of n, by combining the probabilities above appropriately. That is,

Table 1.2: Model Terms

Parameter	Definition
η	P(positive RT at the first time point)
μ	P(positive RT at the second time point| negative RT at the first time point)
θ_1	P(positive ET result at first time point| positive RT at the first time point)
θ_2	P(positive ET result at first time point| positive RT at the second time point)
θ_3	P(positive ET result at first time point| negative RT for both time points)

Table 1.3: Total patients with each test result

Equation	Expected number of...
n	Total sample size
$n\eta$	Total patients with a positive RT at the first time point
$n(1-\eta)\mu$	Total patients with a positive RT at the second time point
$n(\eta+(1-\eta)\mu)$	Total patients with a positive RT
$n\eta\theta_1$	Total patients with a positive ET with a positive RT at the first time point
$n(1-\eta)\mu\theta_2$	Total patients with a positive ET with a positive RT at the second time point
$n(1-\eta)(1-\mu)\theta_3$	Total patients with a positive ET with negative RT at both time points
$n(\eta\theta_1+(1-\eta)\mu\theta_2+(1-\eta)(1-\mu)\theta_3)$	Total patients with a positive ET

$$\begin{aligned}
n_1 \mid \eta &\sim Binomial(n, \eta) \tag{1.11}\\
n_2 \mid \eta, \mu &\sim Binomial(n, [1-\eta]\mu) \tag{1.12}\\
b_1 \mid \eta, \theta_1 &\sim Binomial(n, \eta\theta_1) \tag{1.13}\\
b_2 \mid \eta, \mu, \theta_2 &\sim Binomial(n, [1-\eta]\mu\theta_2) \tag{1.14}\\
b_3 \mid \eta, \mu, \theta_3 &\sim Binomial(n, [1-\eta][1-\mu]\theta_3) \tag{1.15}
\end{aligned}$$

The parameters of the model are summarised in Table 1.2. Prior distributions were then elicited for each of these parameters. While each parameter may be modelled to have a Beta prior distribution to ensure conjugacy, we also considered a number of other distributions to allow for flexibility in the elicitations.

Based on Table 1.2, the expected number of patients within each group can also be calculated. The results are presented in Table 1.3. These quantities were used during the elicitations to present the experts' information back to them, as a check of their values.

The Bayesian analysis has as its subject the parameter θ_2, as this represents the improvement from the BIMC test, i.e. the proportion of patients who would be diagnosed six months earlier.

There are a number of ways that this trial could be analysed from a frequentist perspective. We focus on McNemar's test, as it is the intended analysis method chosen by the trial designers.

In order to calculate a sample size for a McNemar's test, the expected proportions of patients with a positive reference test and negative experimental test result, and negative reference test and positive experimental test result, are required. These values can be inferred from the elicited parameters.

The proportion of patients with a positive Awaji criteria diagnosis at the first time point, but negative BIMC test result, is given by $\eta(1-\theta_1)$. The proportion of patients with a negative Awaji criteria diagnosis at the first time point, but a positive BIMC test result, is given by $(1-\eta)\mu\theta_2+(1-\eta)(1-\mu)\theta_3$. This contains patients from two groups, namely those with a positive BIMC test and a positive Awaji diagnosis after six months, and those with a positive BIMC test and no positive Awaji diagnosis.

1.4 Thesis Outline

The remainder of this thesis is comprised of six chapters.

Chapter 2 reviews methods of sample size calculation. In particular, we focus on the minimum sample sizes required for a chosen statistical power or Bayesian assurance. We review commonly used approaches to determining the inputs to power calculations, and discuss issues which can arise from misspecification, where the chosen inputs do not correspond with the observed effect sizes in the trial. We also review the inputs to assurance calculations, in the form of prior distributions. The prior distributions chosen for the design and analysis stages both affect the assurance. We also consider how the required level of power or assurance is chosen, and some other additional considerations.

Chapter 3 reviews elicitation methodologies, and a number of prior aggregation methods. We outline the elicitation process and popular elicitation techniques, and the cognitive biases which drive their use. We then review elicitation aggregation methods for combining multiple expert judgments into a single prior distribution. We focus on common mathematical aggregation methods, which use a predefined mathematical rule to aggregate distributions, and a behavioural aggregation method, which provides a framework to allow the group of experts to form a consensus view among themselves. We also implement and discuss two elicitations for the aforementioned BIMC case study, which were designed to allow comparisons between prior aggregation methods.

Chapter 4 compares prior aggregation methods using the BIMC case study. We first consider results from the experts from two rounds of elicitation, comparing their performance against each other. We then use cross-validation to compare aggregation methods against each other and the individual experts, across the two groups of experts separately and combined. We finally present the aggregated model parameters required for assurance calculations for the BIMC trial, for each aggregation method considered.

Chapter 5 investigates assurance using simulations. We first present simulations demonstrating how power and assurance change as sample sizes or inputs vary. We then compare assurance and power under similar parameter inputs, to discuss differences under comparable scenarios. We also simulate assurances where the prior used for the design and analysis stages differ, and discuss how this affects the assurance calculations. Simulations considering inputs which do not correspond to observed effect sizes, and inputs from previous trials, are also presented. Finally, we compare aggregation methods at different stages of an assurance calculation.

Chapter 6 investigates the calculation of sample sizes using power and assurance with the aggregated prior distributions from Chapter 4. We use Minimal Clinically Important Differences elicited from experts alongside the aggregated distributions from chapter three as a basis for power and assurance calculations, incorporating both in a number of different ways. We also consider an additional sceptical prior, and both a Bayesian and Frequentist analysis, for assurance calculations and present the resulting sample sizes. Finally, we consider a maximum number of feasibly recruitable patients, and some possible assurance, power, and detectable effect size values which could be achieved for a study of that size.

Finally, we present conclusions and identify future work which could be completed in this area.

Part II

Sample Size Calculations

Chapter 2

Sample Size Calculations

2.1 Introduction

In this chapter we review sample size calculations, using statistical power and Bayesian assurance. Sample size calculations are a necessary step in the design of experiments, and are commonly required when designing and seeking approval for clinical trials.

We begin by outlining the importance of sample size calculations. We then review statistical power, and how it is used to determine sample sizes. We discuss multiple options for determining the required parameters for the calculations, and discuss the implications of misspecifying these inputs. We then discuss Bayesian assurance, which takes the form of the probability of success for a trial. We also consider how prior distributions can be used in the calculations, for both the design and analysis.

Finally, we discuss how a level of power or assurance can be chosen for a trial, before discussing other considerations such as dropout rates.

2.2 Sample Size Calculations

When planning and designing any experiment, the size of the sample to be collected is an important consideration. The larger the sample size, the more accurate and robust the results are likely to be. With this larger sample size, however, comes greater requirements for the experiment. These are typically increased financial costs or additional difficulty in recruiting subjects.

Sample sizes can be chosen to minimise these costs (Bacchetti et al., 2008). Such methods can focus on cost efficiency, or selecting the sample size that minimises the average cost per subject. Most commonly used methods focus on minimising the sample size such that a certain result can be detected to some level of accuracy.

In clinical settings, sample size calculations are very important. From an ethical standpoint, trials should only not be conducted with patients when they are not scientifically justified (WMA, 2001). A trial that does not have scientific justification or worth cannot provide a benefit, and thus any potential risk of harm to patients through action or inaction cannot be ethically justified (Rutstein, 1969). This scientific worth comes not only from the validity of the hypothesis being tested, but also the manner in which it is being tested.

Clinical equipoise refers to when a researcher does not believe that any treatment in any arm of a clinical trial is superior to another. Should a researcher have reason to believe one treatment is superior to the others, then ethically they should only provide their patients with the superior treatment (Freedman, 1987). Some suggest an alternative view should be taken to allow for treatments with different perceived effectiveness. For example, Shamoo (2008) suggests clinical equipoise may not be adhered to in Phase 1 trials. Phase 1 trials occur early in the process of treatment development, meaning few patients are enrolled and the outcomes are less well known. De Meulemeester et al. (2018) propose that randomised clinical trials should have a clear hypothesis for which there is still uncertainty, and that the uncertainty can be established through a systematic review. In a review of literature, they found that over half, a total of 56%, of randomised clinical trials did not fulfil these criteria.

The sample size of a clinical trial determines how many patients will be provided with a new treatment. Exposing an unnecessarily high number of patients to treatment, especially a less well understood treatment such as in a Phase 1 trial, is also unethical. In order to minimise risks to patients, only the number which is required to should receive the treatment.

This problem is not a new one for clinical trial statisticians, and so many methods have been developed to design trials that balance the ethical considerations with collecting strong evidence. Adaptive trials, for example, allow for researchers to adjust the allocation of patients to treatments based on initial results, while still maintaining the scientific quality of the results (Laage et al., 2017). Other options can include designing trials to ensure all patients receive an effective treatment or ending trials early if it becomes clear the treatment is not effective.

In general, the requirements to conduct a sample size calculation are similar irrespective of the approach taken. The type of analysis to be used to analyse the data should be known prior to calculating the sample size. In addition, information about the effect size being investigated is also required. This information, which is used as an input to the sample size calculation, can be an estimate of the true effect size, or a minimum clinically relevant effect size.

Herein lies one of the main difficulties in calculating sample sizes. If the effect size is already known with a high level of certainty, then a study is redundant. If the effect size is not known, then the calculations may not accurately reflect data gathered in the study, and thus provide a poor estimate for the sample size needed.

The most common method of calculating the required sample size is to use statistical power based on the test to be conducted in the analysis of the trial. In the following sections, we will first consider statistical power, before reviewing a Bayesian alternative known as assurance.

2.3 Power

When conducting hypothesis tests, there are two common errors (Cohen, 2013). Type I errors, or false positives, occur when a null hypothesis is rejected when it was in fact true. Type II errors, or false negatives, occur when a null hypothesis should have been rejected, but was not.

Figure 2.1 provides a visual depiction of these errors. We take an example with a hypothesis test, with a simple null and alternative hypothesis. The distribution of the test statistic under the null hypothesis, here in a dashed line, has a critical value displayed by the vertical line. This critical value is the cutoff where the null hypothesis is rejected if the test statistic is greater than the critical value. Conversely, if the calculated test statistic is below this critical value, the null hypothesis is not rejected.

The probability of a Type I error is the area of the distribution that is greater than the critical value, shown in red. This represents the probability of rejecting the null hypothesis, even if it is true. This area is also the significance level of the test, or α, which should be chosen prior to the analysis.

The second distribution in the plot represents the distribution of the test statistic under the alternative hypothesis, here displayed as a solid line.

The probability of a Type II error is the area under this curve which is below the critical value, shown in dark blue. This represents the probability of not rejecting the null hypothesis, even though the alternative hypothesis is true. This area is denoted by β.

The power of the test is $1 - \beta$, and is displayed as the light blue area in the plot. The power of the test is the probability that the null hypothesis is rejected, if the null hypothesis is not true.

There is a trade-off between the selected values of α and β. As the critical value changes, α and β will both change, in opposite directions to each other. For example, in Figure 2.1, if the critical value was increased, the value of α would decrease while the value of β would increase.

Commonly used values for α include 0.05 and 0.01, while commonly used values for $1 - \beta$ include 0.8 and 0.9. Generally, studies are to be designed around these values as they provide a compromise between the minimisation of both errors. The selection of target values can sometimes be influenced by the importance of avoiding Type I or Type

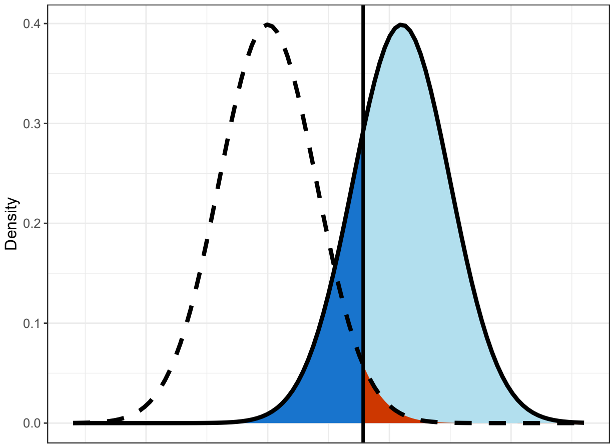


Figure 2.1: Type I and II errors. The dashed line represents the distribution of the test statistic under the null hypothesis, and the solid line the distribution of the test statistic under the alternative distribution. The vertical line represents the test statistic critical value for significance, the red area is the Type I error, the dark blue area is the Type II error, and the light blue area is the power.

II errors. For example, in cases where stronger evidence is required, a lower value of α may be used, while early-stage trials may use larger α values such as 0.1.

As the true effect is unknown, the value of $1-\beta$ relies on estimates of the size of the effect and sample standard deviation. The standard deviation of the estimate of the effect, σ_{effect} is related to the sample size, n, and sample standard deviation, σ, via

$$\sigma_{effect} = \frac{\sigma}{\sqrt{n}} \tag{2.1}$$

As the significance of an effect is related to its standard deviation, an increase in the sample size will result in smaller effect sizes being found statistically significant, and vice versa. This relationship allows for the sample size which provides a set level of power for a given effect size to be calculated (Meinert, 2009; Friedman et al., 2010). Increasing the sample size decreases the standard deviation of the estimate of the effect, which in turn will narrow the confidence intervals for the estimate.

Problems surrounding sample sizes and p-values are commonly labelled as major contributors to the replication crisis. The replication crisis refers to the lack of consistency between initial study results and replications, where findings are not repeated in subsequent experiments (Ioannidis, 2005). These concerns are largely focused in psychology articles, though they affect many other experimental-based disciplines (Maxwell et al., 2015). Part of the issue could be as a result of Type I errors, as an α of 0.05 suggests a 5% probability of rejecting the null hypothesis when it is true.

Properly powered replications could, likewise, fail to find a significant effect when one is truly there. For trials with low power the probability of failing to find an effect that

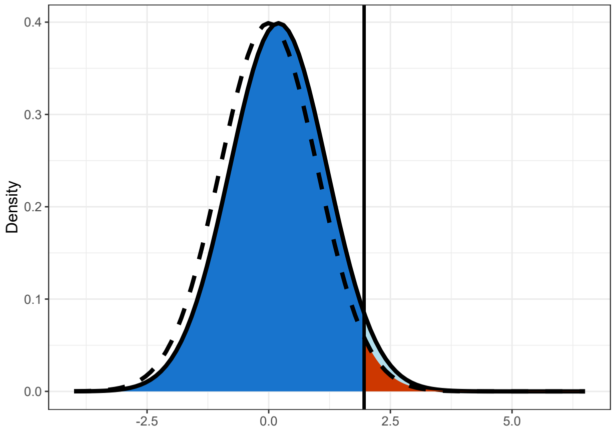


Figure 2.2: The dashed line represents the distribution of the test statistic under the null hypothesis, and the solid line the distribution of the test statistic under the alternative distribution as observed. The vertical line represents the test statistic critical values for significance, the red area is the probability of a significant result, the dark blue area is the Type II error, and the light blue area is the power.

does exist can be very high. Figure 2.2 provides an example where the distribution of the test statistic given the true effect size, shown as a solid curve, is very close to the distribution of the test statistic under the null hypothesis, the dashed curve, which has led to a power very close to the significance level. Equivalently, this means the probability of a Type II error is close to $1 - \alpha$. We further discuss consequences of low powered tests in Section 2.3.3.

Button et al. (2013) estimate the median statistical power in neuroscience articles to be between 0.08 and 0.31. They give an example of a trial that is attempting to replicate a previous result.

Consider an initial trial that finds an effect size with an associated p-value of 0.05. This corresponds to an effect size of $1.96\sigma_{effect}$.

If this effect is assumed to be the true effect, and a second trial attempts to replicate it, there is only a probability of 0.5 of a significant result being found. Figure 2.3 shows why. For simplicity, we will use $\sigma_{effect} = 1$, and thus an effect size of 1.96. As the trial will obtain a significant result if an effect is found with a corresponding p-value of less than 0.05, then only the light blue area in the plot will provide a significant result. The dark blue area then corresponds to a non-significant result.

In this scenario, the replication will not find a significant effect half of the time. This would be true even if the effect size of 1.96 was correct, as the power of the test is 0.5. By increasing the sample size in the second trial compared to the first, the estimate for σ_{effect} can be reduced, which in turn will increase the power, thus increasing the probability to replicate the effect.

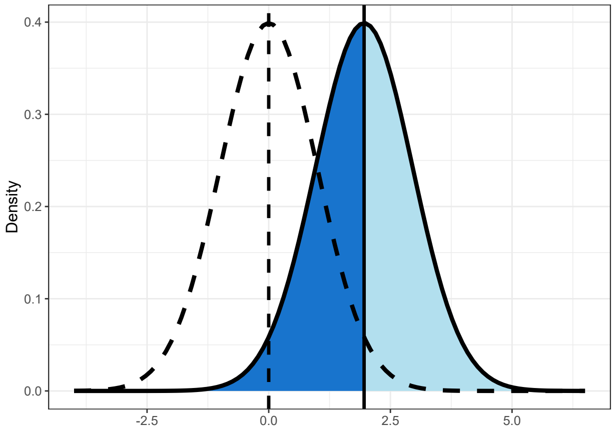


Figure 2.3: The dashed lines represent a null hypothesis, and the solid line the true effect of 1.96 (with a standard deviation of 1). The light blue area is the probability of a significant result, and the dark blue is the probability of making a Type II error.

If a trial is to be successful, and provide results that are conclusive and valid, then it is important it has a high level of power. The following sections will outline how this can be done through an example power calculation, and details about how the parameters required for a power calculation can be chosen.

2.3.1 Power Calculations

Power can be defined as

$$\text{Power } = P(\text{reject } H_0 \mid H_1 \text{ true}) \tag{2.2}$$

As such, a power calculation is dependent on the type of statistical test which will be used in the analysis stage. Machin et al. (2009) provide details for various tests.

We first consider a simple case, where a one sample Z-test is being used to test whether the mean of a population is equal to a particular value. We can state the null hypothesis $H_0 : \mu = \mu_0$, and suppose that the sample mean is $\bar{x}$, the known population standard deviation is σ, and the sample size is n. Then, the critical value for significance will correspond to z, given by

$$z = \mu_0 + 1.96\frac{\sigma}{\sqrt{n}} \tag{2.3}$$

The power is then the probability of observing a value greater than or equal to z, under the alternative hypothesis $H_1 : \mu = \mu_1$.

$$\text{Power} = P(Z > z) = \frac{z - \mu_1}{\frac{\sigma}{\sqrt{n}}} \tag{2.4}$$

It is more often the case that the alternative hypothesis is specified as an inequality, such as $H_1 : \mu > \mu_0$. In these cases, a value needs to be selected for μ_1 by the researcher. The typical approach is outlined in the following section.

2.3.2 Specification of Parameters

The choice of parameter values in power calculations is an important one. The required sample size is entirely dependent on the inputs to the power calculations, and so the researcher calculating a sample size needs to consider them carefully.

Cook et al. (2014) reviews clinical trial articles, and identifies a number of methods used to estimate effect sizes. They identify two aims of these methods, to specify important differences or to specify realistic differences.

The Anchor method evaluates a Minimal Clinically Important Difference (MCID). Not all effect sizes are meaningful in a clinical setting, regardless of statistical significance. The MCID is the smallest meaningful benefit to patients that a treatment would need to provide in order for it to be deemed effective (Jaeschke et al., 1989; McGlothlin and Lewis, 2014). A detected effect size of less than the MCID may still be statistically significantly different from zero, but does not represent a noticeable clinical benefit to patients.

The MCID can be based on the difference between the new treatment and no effect, a control, or the current standard treatment. It can also be calculated as the mean difference in effect between patients who had a clinically important change, and those who did not. This approach is a little different as the 'no clinically important change' group could still have some change. Likewise, the 'clinically important change' group could have a range of effect sizes exceeding the minimum level of difference which is classified as clinically important.

Variations can consider within-patient change, instead of between-patient change, when such information is available. The scale of the MCID can vary. Some trials will use a percentage change or improvement, while others will use an absolute change to measure the difference.

By using the MCID as the effect size in a power calculation, the sample size is chosen so that an effect at least as big as the MCID can be detected. As a smaller effect size is not of interest to the researchers, it is determined that the trial does not need to be powered to detect it.

This is not always easy to implement. As shown in Halme et al. (2015), the choice between different health-related quality of life measurements, and their corresponding MCIDs, could have a large influence on the sample size required. For a two sided test, the example study could require 52, 172 or 500 patients. Van Walraven et al. (1999) surveyed physicians for MCIDs and found the calculated sample sizes ranging between 116 and 3015. It seems clear that the value of the MCID is therefore somewhat subjective. The choice of scale and particular experts consulted will both impact the MCID.

This can sometimes be avoided. MCIDs are often published, to set a standardised target for a particular treatment or test. The MCID or MCID selection advice can also be set in protocols, guidance or funding requirements.

Another method to identify important differences is the distribution method, which considers the imprecision in the measurement of a treatment's effect. The chosen effect size is then a value which is greater than the imprecision that would be expected in the trial, and thus a result which has clinical importance.

A simple case would be to consider a 95% confidence interval for the mean, taken to be Normal, with a standard error of $\hat{\sigma}$. The distance from the estimate of the mean, $\hat{\mu}$, to the upper bound, CI_{upper}, is

$$CI_{upper} - \hat{\mu} = 1.96\sqrt{\hat{\sigma}^2} \tag{2.5}$$

If the observed difference is greater than this value, then the result can be considered clinically important. The value of $1.96\sqrt{2\hat{\sigma}^2}$, or approximately $2.77\hat{\sigma}$, varies in its use. For example, a value of $1\hat{\sigma}$ may be used for some disease-specific quality of life measures.

Important differences can also be identified using health economics methods. These consider the costs for a new treatment, and what the required outcomes would need to be in order for the new treatment to be cost-efficient. This can consider the costs involved with running the trial, or the costs of implementing and supplying the new treatment.

An example of how this could be applied would be to consider the Net Monetary Benefit, or NMB, which is calculated using

$$\text{NMB} = R_C \mathit{\Delta} E - \mathit{\Delta} C \tag{2.6}$$

where $\mathit{\Delta} E$ is the expected difference in effectiveness between treatments, $\mathit{\Delta} C$ is the difference in costs between treatments, and R_C is the willingness for the decision maker to pay for a unit of effectiveness. The decision is in favour of the treatment if NMB > 0, or is greater than a required level of improvement.

Cook et al. (2014) found that this method was rarely used in their review, with only 13 of the 777 trials implementing it.

In order to specify an estimated effect size, methods rely on using past studies or expert opinions as a source of information. We review elicitation techniques and the aggregation of multiple experts' opinions in Chapter 3.

For clinical trials in later phases especially, pilot studies may be available to guide the choice of effect size for the power calculation. The pilot study in this case can be run specifically for that purpose, or could be an earlier study into the same treatment which can be used as such.

A wider reaching alternative to a single pilot study is to use a review of the evidence base. This involves performing a literature review or meta-analysis of results for the outcome of interest. For treatments with many previous studies, this can allow for a more accurate and robust estimate of the effect size.

Browne (1995) demonstrates that using values directly from a pilot study often leads to the future study having lower power than the sample size calculation would suggest. As such, it is important to include inflation due to epistemic uncertainty on any estimates from a pilot study, or within the wider literature, when using them to inform trial design.

Clark et al. (2013) reviews research protocols for sample size calculation methods used in randomised clinical trials. Nearly a quarter of the protocols reviewed used previous studies to form an estimate for the effect size, while under 10% cited literature reviews and less than 1% used a meta-analysis. The use of previous studies or literature was much more common than the MCID, which was reported 12% of the time. Over half of the protocols did not mention how effect sizes were estimated for sample size calculations.

Effects can be compared across studies by standardising them to ensure they are on the same scale, for example using Cohen's d (Cohen, 2013). For example, in the case of an independent samples two-sample design and Normal observations, this would take the form

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \tag{2.7}$$

where $\bar{x_i}$ is the sample mean, s_i is the standard deviation, and n_i is the sample size of group i. The value of d then represents a standardised mean difference which can be directly compared to a value of d for a different trial comparing the same treatments (Lakens, 2013). Standard guidelines suggest d values of 0.2, 0.5 and 0.8 represents differences of small, medium and large magnitudes respectively (Cohen, 2013). This can be particularly useful in cases with multiple pilot studies or reviews of evidence as it allows results with different sample sizes and standard deviations to be compared. While they can be used as comparisons to interpret other results, these standard effect sizes should not be used as the basis of power calculations (Lenth, 2001).

2.3.3 Misspecification of Parameters for Power calculations

In order to choose a sample size to appropriately power a trial, it is important to use appropriate parameter values in the calculation. Poorly chosen inputs can lead to insufficient sample sizes, which may result in an inability to find effects that are present.

It is not easy to determine whether inputs have been misspecified. While power calculations could be repeated once the data is collected, using values gained from the data, this post hoc method of power calculation is generally advised against. Hoenig and Heisey (2001) demonstrate how observed power, using inputs from the observed data, should not be used to make judgements on whether there is support for the null hypothesis. The

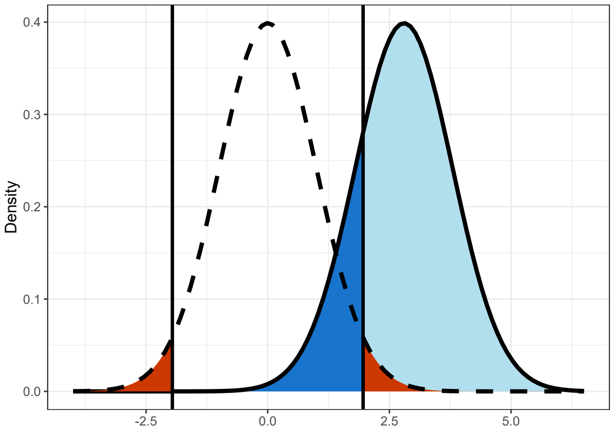


Figure 2.4: The dashed distribution is the null hypothesis, the solid distribution is the test statistic under the alternative hypothesis, the vertical lines are the critical values. The red areas represent the probability of committing a Type I error, the dark blue area the probability of committing a Type II error, and the light blue area is the power.

estimate of the effect has uncertainty around it, which means the point estimate might not give an accurate depiction of the power (Gelman, 2019). A confidence interval gives a range of values for the effect size supported by the data, and the post hoc power calculated will vary across this interval.

Overestimating the effect size in a power calculation can lead to missed opportunities in the analysis stage. For example, consider a trial which will conduct a two-sided hypothesis test on a mean, with a null hypothesis that the population mean is equal to zero, and using a significance level of $\alpha = 0.05$ and a power of $1 - \beta = 0.8$. Assuming that the standard deviation is $\sigma = 1$, a simple alternative hypothesis would be that the mean is equal to approximately 2.8, ie $H_0 = 2.8$. Figure 2.4 provides an example of this case.

If the population mean is much closer to zero than anticipated, a number of problems can occur (Gelman and Carlin, 2014). For example, if the true mean was actually 0.28, one tenth of the mean previously used to determine the sample size, then the final results can be misleading. As the critical values for determining a significant result are calculated using the null hypothesis, these values will not change. Figure 2.5 shows the distribution of the test statistic under the true mean, and critical values for the null hypothesis that $\mu = 0$. The area in red is the probability that a significant result will occur with the correct sign, in this case a positive mean, and the yellow area represents the probability that a significant result will occur with the incorrect sign, in this case a negative mean when the true mean was positive.

Gelman and Carlin (2014) labels these errors as Type M, for magnitude (coloured in red), and Type S, for sign (coloured in yellow). They argue that as a significant result in a trial is much more likely to be published, then those results which are published could

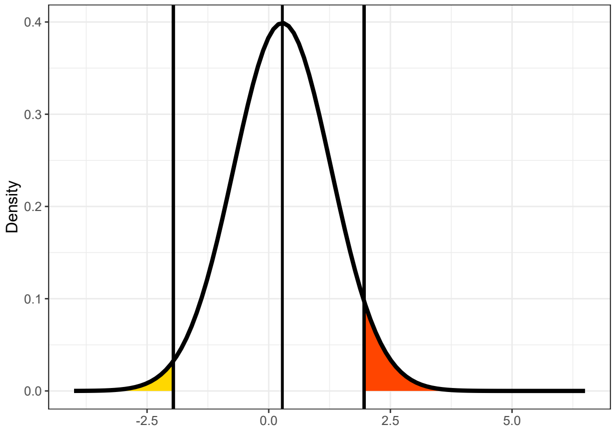


Figure 2.5: The solid distribution represents the test statistic under the alternative hypothesis. The thin vertical line represents the true mean, the thicker vertical lines represent the test statistic critical values for significance from Figure 2.4, the red area is the probability of a significant result with the correct sign, and the yellow area is the probability of a significant result with an incorrect sign.

contain these two other errors.

In our example, the Type M error indicates that for a positive mean to be detected, it must be seven times larger than the true mean. This is due to the critical value of 1.96 being seven times larger than the true mean, 0.28. For the given sample size, the true effect will not be accurately determined if a significant result is found.

The Type S error has a probability of 0.013, compared to the Type M error probability of 0.046. This means that conditional on a significant result being found, there is approximately a 20% chance of the mean having an incorrect sign. In practical terms, this could likely correspond to a negative estimate being found when there was actually a positive effect, or vice versa.

If the estimate used in the power calculations corresponded with the true mean, then the test would be underpowered. As such, the sample size would need to be increased in order to reduce the standard deviation until an appropriate level of power was reached. This would help reduce the Type M error, and almost certainly remove any chances of Type S error occurring.

By specifying inaccurate inputs to the power calculation, we have been left with a case where the test would be severely underpowered to detect the true effect size. Furthermore, if we did find a significant result, it would either overestimate the mean, or provide an estimate with an incorrect sign. Vasishth and Gelman (2017) showed that the focus on publishing statistically significant results filters the body of published work in a way that leads to over-optimistic results. While there is a large probability that such an experiment would lead to a non-significant result, the fact that significant results are over-represented in published literature (Leggett et al., 2013; Cristea and Ioannidis, 2018) means these errors are more likely to appear in scientific literature.

These problems can occur when powering a test using inputs based on a best estimate, and when using an MCID or similar figure to ensure the test can detect a particular difference. It does seem, however, that when using MCIDs that deliberately do not reflect the likely effect size, but rather a larger desired effect, that this issue may become more prevalent. If the MCID used as an input is greater than the true effect, then calculating a sample size with the MCID will lead to a trial that is likely to overestimate the effect.

We will discuss in Section 2.5 some ways this can be accounted for.

2.4 Assurance

Though Bayesian approaches have a long history in statistics, their uptake in clinical trials has been much slower.

Bayesian methods allow for the inclusion of information from other sources. Areas such as early phase clinical trials, especially Phases 1 and 2, are more popular for the inclusion of Bayesian techniques (Rosner, 2020). van Rosmalen et al. (2018) demonstrate that including historical data can improve the power and precision of analysis. They also suggest this may be particularly useful in early stage trials, in order to reduce the number of control patients required.

While there is a vast array of literature on Bayesian methods for clinical trials, it appears that their use is not necessarily put into practice at a similar rate. Campbell (2020) suggests that there are many cases where Bayesian clinical trials are not reported as such. They also suggest that there is a perceived resistance to Bayesian methods in reviews by organisations such as the Center for Drug Evaluation and Research (CDER) and the Food and Drug Administration (FDA). While this resistance does have some historical basis, many organisations now have guidance for the use of Bayesian methods.

The level of guidance for clinical trial development with respect to Bayesian methods has increased over recent years. Earlier guidance provided in Europe in 1998 mentioned Bayesian statistics as an option, but provided no further details (ICH, 1998; Lewis, 1999). The FDA has more recently provided guidance specifically for the use of Bayesian methods (FDA, 2010). While the statistical basis for Bayesian methods and their use in clinical trials has seen much academic progress, its application has relied on approval from regulatory bodies.

In a review of survival analyses in clinical trials, Brard et al. (2017) found that Bayesian approaches were used in the final analysis of four out of twenty-eight trials. While Bayesian approaches were also used in a number of secondary analyses or re-analyses, information about the use of the prior was scarce. There was only one case where expert knowledge was used to form a prior, and another three where an informative prior was used.

One common argument against the use of Bayesian methods in clinical trials is their lack of objectivity. The inclusion of an informative prior, and the name assigned to subjective Bayesian analyses, lends itself to a view that these approaches do not have the same level of rigour or impartiality as frequentist methods. Berger and Berry (1988) demonstrates that the choice of how a frequentist test is run can influence the outcome. They provide the following example.

Consider an experiment into whether vitamin C provides better relief from cold symptoms than a placebo. Collected data show that in 17 pairs of patients vitamin C provided more relief, and in four pairs of patients, the placebo did. If this sample is treated as a collection of patients with a sample size of 21 pairs, then the p-value testing the null hypothesis of vitamin C having no effect is 0.049. If the trial was conducted by the researcher collecting data until they had received four pairs where the placebo was better, then the corresponding p-value of the test would be 0.021. Furthermore, with just the data available, a third party would not be able to differentiate between these cases.

They argue that this demonstrates a subjectivity inherent in frequentist statistics. The way the researcher chooses to collect their data can influence the results. In addition, the specific null hypothesis and test statistic to determine significance are choices for the researcher to make, even though they have standard approaches. By defining priors, Bayesian methods encode the subjective decisions more directly and, hopefully, with more explanation and justification.

Bayesian methods have been used, especially in the design stage, both in sample size calculations and other areas of experimental design. Assurance is a Bayesian alternative to power in sample size calculations (Spiegelhalter and Freedman, 1986). The assurance is the unconditional probability that the results of a trial will provide a statistically significant result. Spiegelhalter et al. (2004) define assurance mathematically as follows, where D_1 is the outcome where the null hypothesis is rejected.

$$p(D_1) = \int p(D_1 \mid \theta) p(\theta) d\theta \tag{2.8}$$

where $p(D_1 \mid \theta)$ is the power function, $p(\theta)$ a prior distribution on the unknown parameter (for example the treatment effect) and $p(D_1)$ is the probability of rejecting the null hypothesis, or in other terms, returning a statistically significant result. This idea of assurance can be extended to the situation where the analysis to be conducted following the trial is Bayesian, in which case $p(D_1)$ is the prior probability of declaring that the new treatment is superior to the old treatment. The value of $p(D_1 \mid \theta)$ is then the probability that the Bayesian analysis will result in a successful outcome.

Assurance calculations vary from the power calculation approach. In a Bayesian assurance calculation, the treatment effect is represented as a probability distribution, rather than a single point estimate (M'lan et al., 2008). This allows for the uncertainty in the treatment effect to be accounted for in the calculations. In addition, the probability that a trial will lead to a statistically significant outcome is potentially more easily interpreted than the power of the test.

In order to calculate assurance, a prior distribution must be specified for the model parameter(s). This prior should be elicited from experts, based upon their knowledge (Chaloner et al., 1993). Ren and Oakley (2014) consider assurance in survival analysis using an Exponential or Weibull distribution.

O'Hagan and Stevens (2001) define both a design and an analysis prior for assurance calculations when the final analysis to be conducted will be Bayesian. This specifies separate priors for use in the design of the experiment and the analysis following the experiment. By doing this, the analysis can be conducted with a weak prior while the design can be obtained with a much stronger prior. Additionally, these two priors will represent the beliefs of different individuals or groups. The results of the analysis are more strongly influenced by the data, while the design, when the data are not yet available, can still be obtained to provide an indication of the likely success of a trial given current knowledge. This is discussed further in Section 2.4.2.

Assurance has been used in applications in the clinical trial literature. It is sometimes referred to as the Probability of Success (Bertsche et al., 2019; Takazawa and Morita, 2020) or the average power (Kowalski, 2019). Kunzmann et al. (2021) discuss that while these terms are often used interchangeably, the probability of success refers to the case where the probability of rejecting the null hypothesis is integrated over the entire range of the prior density, while the average power is a weighted average specifically in the region where $\theta > MCID$.

Carroll (2013) suggests assurance is useful in moving from Phase 2 to Phase 3 trials, though notes the difficulties with using and interpreting assurance may require a statistician to assist researchers with its use.

2.4.1 Assurance Calculations

Like power, assurance requires the analysis to be undertaken once the data have been observed to be determined before calculation. This choice is dependent on the research question, and the data that will be collected.

As opposed to more standard power calculations, assurance often involves some level of simulation in its calculation. For cases where the assurance cannot be found analytically, a straightforward procedure can be followed. The procedure to calculate assurance is as follows, for a sample size n and over I iterations.

1. Simulate a sample of the model parameters from the design prior distribution.
2. Find the power of the hypothesis test for sample size n and set of parameter values. In the case of a Bayesian analysis of the trial, find the probability of meeting the criterion for the trial to be a success, which will depend on the posterior distribution based on the analysis prior.
3. Repeat the process I times.
4. Calculate the assurance by averaging over I.

It may not always be possible, or easy, to calculate the power analytically in Step 2. If a closed-form power function cannot be used, the minimum required test statistic for a significant result may be determined. Then, the probability of observing at least that test statistic under the design prior can be used as an estimate of the power. Alternatively, power can be approximated by simulation.

The value of I should be chosen to be large. As is common for Monte Carlo methods, larger values of I will provide more accurate results at the cost of an increased computation time. Steps 1-3 of the algorithm here are parallelisable, and so can be run in parallel if a faster computation time is required.

If a Bayesian analysis is to be conducted in Step 2, it is beneficial to do so if a conjugate prior can be used. If not, the posterior probability of success will need to be inferred for each parameter sample, involving MCMC runs on each of a set of simulated data realisations. This could result in large calculation times and a build-up of Monte Carlo error. Section 2.4.2 discusses the use of Bayesian priors in Step 1 and Step 2 in further detail.

In the following sections, we will consider continuous observations and discrete observations in the form of counts.

Continuous data

O'Hagan et al. (2005) considers a one-sided hypothesis test. Considering two samples of size n, $\underline{x}_1 = (x_{1,1}, \ldots, x_{n,1})$ and $\underline{x}_2 = (x_{1,2}, \ldots, x_{n,2})$, where $x_{ij} \sim N(\mu_j, \sigma_j^2)$, the difference $\bar{x_1} - \bar{x_2}$ is of interest. This difference has a sampling distribution of $N(\delta, \tau^2)$, where

$$\delta = \mu_2 - \mu_1 \text{ and } \tau = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.9}$$

For a one sided hypothesis test, this corresponds to a power equal to

$$\text{Power} = \Phi\left(\frac{\delta}{\tau} - z_{1-\alpha}\right) \tag{2.10}$$

A prior distribution can be placed on δ, either directly or from priors for μ_1 and μ_2. Setting $\delta \sim N(m, s^2)$, the unconditional distribution of the differences in sample means is $\bar{x_2} - \bar{x_1} \sim N(m, \tau^2 + s^2)$.

In a one sided hypothesis test, this value will be compared to a Z-score, with a critical value of Z_α, above which a significant difference would be found. The assurance is then given by

$$P(\bar{x_2} - \bar{x_1} > \tau Z_\alpha) = \Phi\left(\frac{-\tau Z_\alpha + m}{\sqrt{\tau^2 + s^2}}\right) \tag{2.11}$$

Chapters 5 and 6 contain further assurance calculations.

When it is necessary to use simulation to evaluate the power, the required critical value of the test statistic for a frequentist test to provide a significant result can often be easily calculated. For example, $\mu + 1.96\sigma/\sqrt{n}$ gives the upper bound of a 95% confidence interval for a given variance. Values for μ and σ can be determined from the null hypothesis for the test. The probability the test statistic gives a value of greater than $\mu + 1.96\sigma/\sqrt{n}$ is then the assurance, and so the algorithm will estimate this using a large number of draws from the design prior.

The case of a Bayesian analysis may be more difficult, as the analysis may involve a non-conjugate prior. In this situation, the data is simulated and a model fit using an MCMC method of choice. The posterior can then be used to form a conclusion. Repeating this process over a large number of data simulations will give a probability of meeting the criterion for a successful trial.

Binomial data

We consider assurance for trials with binomial observations. In general, the assurance in this case can be stated as the probability of observing at least s successes out of a number n of Bernoulli trials, where s is the minimum number of successes in order for a successful outcome for the trial.

O'Hagan et al. (2005) provides an example where the outcome variable is binary, such as when a clinical test provides a positive or negative result. We assume that inconclusive results would not occur, or may be treated as a negative result. The example considers a two-sided test of proportions in which the null hypothesis is that the probabilities of success in the two treatments, θ_1 and θ_2, are equal. Using a Z-test approximation, this hypothesis would be rejected if $\mid Z \mid > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the critical value of a standard normal distribution at a significance level α. The test statistic Z is given by

$$Z = \frac{\hat{\theta_2} - \hat{\theta_1}}{\sqrt{\hat{\theta_1}(1 - \hat{\theta_1})/n_1 + \hat{\theta_2}(1 - \hat{\theta_2})/n_2}} \tag{2.12}$$

where n_1 and n_2 are the numbers of trials in the two groups, and $\hat{\theta_1}$ and $\hat{\theta_2}$ the estimates of θ_1 and θ_2 respectively.

The power function, $P(R_1 \mid \theta_1, \theta_2)$ is then given by

$$P(R_1 \mid \theta_1, \theta_2) \approx \Phi\left(-Z_{\alpha/2} + \frac{\theta_2 - \theta_1}{\sqrt{\theta_1(1-\theta_1)/n_1 + \theta_2(1-\theta_2)/n_2}}\right) \tag{2.13}$$

where R_1 is the event: Reject the null hypothesis.

To obtain the assurance, this is multiplied by a prior distribution for θ_1 and θ_2 and integrated over θ_1 and θ_2, as per Equation 2.8. Thus, the assessment of a suitable prior distribution for θ_1, θ_2 is a crucial step in the assurance calculation.

As in the continuous case, it may not always be possible to assess the assurance based on the analytic form of a power function, as above. The assurance calculation can also be performed using simulation.

To do so we need to find the critical number of successes, s. We generate a dataset with k successes and $n - k$ failures, firstly setting $k = 0$ and applying the relevant hypothesis test. If the test result is not significant, we increase k by one and rerun the test, repeating this procedure until the minimum number of required successes are found, $k = s$.

Once the value of s is determined, we can calculate the probability that the number of successes, x, is greater than or equal to s. This probability, $P(x \geq s \mid \theta)$, can be calculated using the binomial distribution.

If a Bayesian analysis is used, then this step will include an analysis prior.

The assurance can then be calculated as

$$\int_{\theta=0}^{1} \sum_{i=s}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} f(\theta) d\theta \tag{2.14}$$

where $f(\theta)$ is the design prior distribution. If $f(\theta)$ is a beta distribution, then the assurance can be found analytically.

2.4.2 Design and Analysis Priors

When conducting a Bayesian analysis and design of a trial, prior distributions are used at both stages.

The design prior, or sampling prior, is used in the design of the trial, where an MCID or anticipated effect size would be used in a corresponding frequentist calculation. This prior is used alongside the model which will be used for the analysis (Psioda and Ibrahim, 2019). Once collected, the data will be analysed based on a separate prior, referred to as the analysis prior.

While these two priors are on the same parameters, they do not need to be the same

(Wang and Gelfand, 2002). O'Hagan and Stevens (2001) suggests that while the design stage should reflect the researchers' beliefs, the analysis needs to be presented to and convince other parties, such as regulatory bodies or pharmaceutical companies. If the prior beliefs used in the analysis stage do not represent the beliefs of these parties, then they may not accept the results as convincing or sound. By using a different prior for the analysis, the researchers can present a convincing analysis while still accounting for their own beliefs when designing the trial.

Similarly, in cases where a frequentist trial is designed using Bayesian methods, the design prior distribution is unrelated to the null hypothesis which is to be tested. In this case, only a design prior is required to determine the sample size. Care should be taken, however, as a design prior inconsistent with the null hypothesis could result in a calculation where the prior provide no probability of the treatment being successful. In such a case, no sample size will fulfil the requirements of the sample size calculation.

As discussed in Section 2.2, there is an ethical argument that equipoise is an important consideration in trial design. As such, the researchers designing the trial should only be doing so if they believe that the new treatment will have a similar effect for patients. The design prior, therefore, will typically reflect the researchers' view that the treatment will have a positive effect for patients.

Chapter 3 reviews elicitation techniques and the aggregation of multiple expert opinions. These methods are suitable for the construction of design priors, as they incorporate the beliefs of the researchers organising the trial.

For analysis priors, Spiegelhalter et al. (1994) identify multiple types of potentially suitable priors to be considered in clinical trial settings.

The first type is a reference prior, designed to provide a minimal level of prior information. The simplest option here would be a uniform prior over the range of the parameter, which provides an equal probability for each possible value.

A clinical prior is another option which may be considered. These priors represent the opinions of informed experts, and usually would be elicited using techniques such as those mentioned in Chapter 3.

Using such a prior, however, may give the appearance of bias or partiality to any analysis. In such cases, it may be sensible to use a sceptical prior, which assigns a low probability to larger effect sizes.

It is important to consider how to represent a sceptical prior. While a natural assumption might be to use a flat prior so as to remain uninformative, this is not always sceptical. For example, given a binomial likelihood, a prior in the form of a $Beta(1,1)$ distribution may be selected, which is uniform over the range $(0,1)$. If a parameter is given this as a prior distribution then the implied beliefs are that each value in the range is equally likely. If the current standard treatment had a known effect size of 0.2, then this prior states there is an 80% probability the new treatment will outperform the current standard. Such a prior is, in this case, very enthusiastic and more informative than the researcher may intend.

Another common reference prior is Jeffery's prior, which likewise may be selected with the intention of using an uninformative prior or a sceptical prior. Jeffery's prior, $f_J(\theta)$ is defined as

$$f_J(\theta) = \sqrt{-\left[\frac{d^2}{d\theta^2}\log(f(x \mid \theta))\right]} \tag{2.15}$$

where $f(x \mid \theta)$ is the likelihood function.

For a binomial likelihood, with observation y, sample size n, and parameter θ,

$$p(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \tag{2.16}$$

$$\log(p(\mid y, \theta) \propto y\log(\theta) + (n-y)\log(1-\theta) \tag{2.17}$$

$$\frac{d^2}{d\theta^2}\log(p(\mid y, \theta) \propto -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta^2)} \tag{2.18}$$

As the expected value of y is $n\theta$, for a sample size of n, the expected value is

$$E\left[\frac{d^2}{d\theta^2}\log(f(x \mid \theta))\right] \propto \frac{n}{\theta(1-\theta)} \tag{2.19}$$

The Jeffery's prior then takes the form

$$f_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} \tag{2.20}$$

which is equivalent to a $Beta(\frac{1}{2}, \frac{1}{2})$ distribution.

Similarly to the uniform case, this prior places a large amount of probability on the values of $\theta > 0.2$, approximately 70%.

While commonly used as non-informative priors, these priors both provide some informativeness in this setting. If they are used in an assurance calculation, they do not behave as sceptical priors. This is further discussed in Chapter 5.

Sceptical beliefs should be represented as an informative prior, but one with the majority of the probability density on little or no effect being present. There are a number of approaches that could be used to define such a prior. A simple option would be to use a positively skewed distribution, where the majority of the weight is concentrated on lower values.

For example, a $Beta(1, 10)$ prior has approximately 90% of its density below the value of 0.2. This represents much more sceptical beliefs than either reference prior above.

Another option would be to use combinations of uniform priors. A uniform distribution could be fit to either side of a chosen effect size. For example

$$f(\theta) = \begin{cases} \frac{1}{ES-a}, & \text{for } \theta < ES \\ \frac{1}{b-ES}, & \text{for } \theta \geq ES \end{cases} \tag{2.21}$$

for a distribution with a minimum value of a, a maximum value of b, and a chosen or assumed effect size ES. This prior would represent beliefs where it is equally likely for the parameter to have a value less than or greater than the effect size.

This prior has a level of artificiality to it, as the non-continuous jump in the value of $f(\theta)$ at $\theta = ES$ does not necessarily represent beliefs that may belong to an expert. The difference in a person's beliefs between $f(ES - 0.0001)$ and $f(ES + 0.0001)$, for example, might not sensibly contain such a jump. While perhaps not realistic, this prior does still allow for a level of informativeness over the remaining values. This may be desirable in cases where impartiality is important to display.

Furthermore, it will typically place more weight on extreme values than other distributions.

Another option is to use a spike and slab prior. Such a prior places a probability mass on a single value, and the remaining probability over the remaining range of values encapsulated by the prior.

A sceptical spike and slab prior may place a large proportion of the prior weight on the effect size being zero. For example, we could choose a prior distribution to represent the belief that there was a 90% probability that the treatment had no effect, and then place a uniform distribution on the remaining range of possible values.

$$f(\theta) = \begin{cases} 0, & \text{with probability } 0.9 \\ 0.1, & \text{for } 0 < \theta \leq 1 \end{cases} \tag{2.22}$$

As the prior has only 10% of its density in the range $0 < \theta \leq 1$, the parameter has a smaller probability of taking any value greater than zero compared to a standard uniform prior on [0,1]. In the previous example, where a current standard treatment had a known effect size of 0.2, this spike and slab prior would give a probability of 8% that the parameter is greater than 0.2, rather than a probability of 80% as was the case for a uniform distribution over [0.1].

This spike and slab prior has some similarities to a hypothesis test setting. In both, most of the prior weight is placed on no effect either via the null hypothesis or the spike, and only if the data gives a strong signal that this is not true will a significant result be found. This may be a more natural choice for those familiar with hypothesis testing, or settings in which hypothesis tests are the norm.

Ultimately, if a sceptic, whose beliefs form a sceptical prior, can be convinced that the new treatment is effective by the results in the trial, then this is strong evidence there is an effect present.

Similarly, an enthusiastic prior may be used in determining when to stop a trial. If the lack of evidence in the data for a positive effect from the treatment can outweigh this enthusiastic prior, then there is strong evidence there is no effect and the trial can be stopped.

An enthusiastic prior could be constructed as the reverse of a sceptical one, where only a small amount of probability is assigned to parameter values suggesting the treatment will not work.

The enthusiastic and sceptical priors can be used together to answer different questions prior to a final analysis. For example, Ye et al. (2020) utilises sceptical and enthusiastic priors for analysis, and compares them over a range of possible true effect sizes for a given sample size. For cases when the true effect is smaller, the enthusiastic prior is still not strong enough to overcome the lack of data supporting the new treatment, and the recommendation would be to stop the trial due to futility. When the true effect size is large, the sceptical prior is convinced by the data that there is an effect, and the recommendation would be to stop the trial as efficacy has been found.

This use of a sceptical prior appears in the clinical trial literature, for example in Tan et al. (2003), Goligher et al. (2018), Pedroza et al. (2018), and Charkos et al. (2020). It is used as an analysis prior, rather than a design prior, as it represents the view that the treatment is most likely not effective. If it were used as a design prior, it would result in very small assurance values, even for large sample sizes. As such, the final sample sizes would be very large, as the researchers would be attempting to detect an effect which they are stating is very unlikely to exist.

Separate analysis and design priors have seen some use in practice. Walley et al. (2015) use informative design and analysis priors in a demonstration of Bayesian-only trial design and analysis, and Psioda and Ibrahim (2019) investigate using historic data to specify analysis priors.

2.4.3 Standardised Assurance

As a sample size increases, the power or assurance also increases. However, while power tends towards one as n gets large, the value the assurance will converge towards can vary (O'Hagan et al., 2005). This has implications for selecting a target assurance, as will be discussed in Section 2.5. It also means that assurance calculations and curves cannot always be easily compared.

One Bayesian method which provides an assurance-like value is often referred to as expected power (Kunzmann et al., 2021). The expected power uses a prior distribution which is conditional on the parameter being above the target value. While this method returns a consistent scale, it no longer incorporates information in the prior about the probability of a parameter being non-significant, or below the target value. We use an alternative method for rescaling the assurance as follows.

If we consider a sample of independent binary random variables, which could be modelled using a binomial distribution, a reasonable choice of prior is a beta distribution due to conjugacy. For a random sample $\underline{X} = (X_1, \ldots, X_n)$ of binary variables, we have n trials and $s = \sum_{i=1}^{n} X_i$ successes. Suppose we have an analysis prior of $\theta \sim Beta(\alpha, \beta)$, then the posterior distribution will be

$$P(\theta \mid \underline{X}) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + x)\Gamma(\beta + n - s)} \theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1} \tag{2.23}$$

which follows a $Beta(\alpha + s, \beta + n - s)$ distribution. The analysis will then make inferences using $P(\theta \mid \underline{X})$, such as calculating $P(\theta > \theta_0 \mid \underline{X})$, where θ_0 is an MCID or some other quantity of interest.

As n gets large, the effects of α and β on the posterior distribution become increasingly small in comparison to the values of s and n. The distribution will eventually become predominantly defined by the ratio of s to n.

At the design stage, there is not yet any data to determine the value of s. Instead, s can be simulated from the design prior. For a given draw of θ, which we will label θ_i, from its design prior distribution $P_D(\theta)$, s can then be simulated as

$$s_i \sim Binomial(\theta_i, n) \tag{2.24}$$

The expected value of s under the design prior is given by

$$E[s] = \theta n \tag{2.25}$$

This means that, based on samples from the design prior, the $Beta(\alpha + s, \beta + n - s)$ posterior distribution tends towards a posterior distribution of $Beta(\theta n, (1-\theta)n)$ as n gets large.

The maximum assurance is then the probability of a successful result under this posterior distribution. This could represent the prior probability a drug will be successful in treating a patient, or a diagnostic test correctly identifies a patient. This can be estimated using a Monte Carlo approach.

Consider an analysis where a significant result is defined as $P(\theta > \theta_0 \mid X) > 0.95$, for example. For $j = 1 \ldots J$ draws from the distribution of θ, $P_D(\theta)$, which we will label θ_j, then

$$P_D(\theta \mid X, \theta) = \frac{1}{J} \sum_{j=1}^{J} I_j \tag{2.26}$$

where I_j is an indicator variable which is equal to one when $\int_{\theta_0}^{1} \frac{\Gamma(n)}{\Gamma(s_i)\Gamma(n-s_i)} x^{s_i-1}(1-x)^{n-s_i-1}dx > 0.95$, and is otherwise zero.

This value then represents the maximum possible assurance that can be achieved for the given design prior and analysis method. This value can also be used to demonstrate the prior probability assigned to the treatment actually being effective. As such, it could be used to check design or analysis priors to ensure they are reflecting appropriate information about the treatment of interest.

In order to place assurances on a consistent scale, we use a standardised version of assurance, referred to as scaled assurance in Alhussain and Oakley (2020). For a sample size n, we will define this standardised assurance, SA_n, as the calculated assurance for n, A_n, divided by the maximum possible assurance under the chosen design prior A_{max}.

$$SA_n = \frac{A_n}{A_{max}} \tag{2.27}$$

The standardised assurance represents the proportion of the possible assurance given n. For example, $SA_{45} = 0.5$ would mean that for a sample size of 45, 50% of the maximum assurance for the chosen design has been achieved.

If there are different design priors or analysis methods to be considered, the standardised assurance allows for comparisons to be made. This may allow for a more nuanced comparison between different choices available to researchers when designing a trial.

It also places the assurance on the same scale as statistical power. Care should be taken if a comparison is made between the two. It is also important to take care when using standardised assurance to select a sample size. The standardised assurance can inform on the potential improvement to the total assurance possible by increasing sample size. It does not, however, provide the probability the trial will be successful. For example, a trial could have a standardised assurance of 90%, but if the maximum assurance possible was only 0.1, then there would be a small probability of the trial concluding successfully. As such, standardised assurance should be considered alongside he usual un-standardised assurance.

Section 2.5 outlines how standardised assurance could be used in determining a sample size.

2.5 Determining Sample Size

Power and assurance can both be used to select a sample size for a trial. The general process is the same for both methods.

The power or assurance is calculated for a chosen sample size. If the value is smaller than the required power or assurance, the sample size is increased and the values recalculated. This process can be repeated until the power or assurance has reached its required

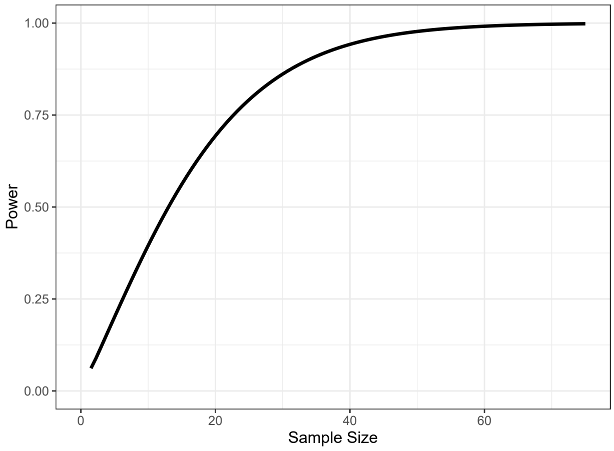


Figure 2.6: An example power curve, showing the power values over a range of sample sizes.

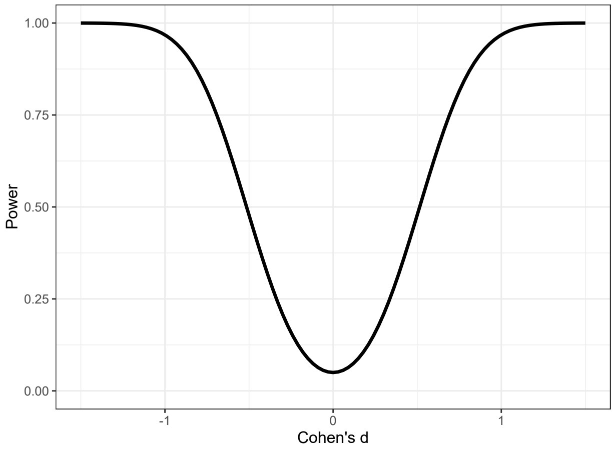


Figure 2.7: Power values of a two sided hypothesis test over a range of effect sizes.

level if possible. Alternatively, when an analytic solution is available, the sample size may be calculated directly for a chosen power or assurance.

Other than the parameter inputs for the power or assurance calculation, this method also needs the required level of power or assurance to be chosen. For power calculations, this is usually taken to be 0.8 or 0.9. A plot of the power for a given effect size, standard deviation and significance level can be plotted, such as that in Figure 2.6. This plot demonstrates the required sample size to achieve any level of power. Power curves tend towards one as the sample size increases.

Often, a power analysis will include different effect sizes, in case the true effect is larger or smaller than expected. Figure 2.7 provides a plot showing the relationship between the power and a change in effect size for a two sided hypothesis test. Often, only the region greater (or less) than zero would be considered. This example shows the power of a one sample t-test, with $\alpha = 0.05$.

Table 2.1: Clinical Trial Success Rates (Wong et al., 2019)

Trial Phase	Probability to reach next phase	Probability to be approved
Phase 1	38.8%	6.9%
Phase 2	38.2%	28.8%
Phase 3	59.0%	59.0%

Plots such as Figure 2.7 allow for differences required to achieve specific powers to be compared. For example, the effect size for a power of 0.8 and 0.9 can be seen. Plots such as these have a lower bound of α, and an upper bound of one.

Ultimately, a number of power calculations can be conducted, varying the effect size and power levels. In practice, the researchers should at least consider an optimistic, most likely and pessimistic view of possible effect size outcomes, as well as varying levels of power. This can provide a range of sample sizes.

Unlike power, assurance does not have a common value for comparison against. When interpreting the assurance, it is also important to consider what a reasonable probability for a successful clinical trial is. Wong et al. (2019) reviewed success rates of clinical trials, finding around 6.9% of Phase 1 trials pass future phases and are successfully approved. Further success rates are given in Table 2.1.

The success rates of clinical trials vary across different areas of medicine. For example, Hay et al. (2014) found that 64.5% of Phase 1 trials progressed to Phase 2, and 10.4% of them were approved after Phase 3, with values ranging from 58.3% to 72.2%, and 6.7% to 18.2% respectively, for trials in different disease areas. Mullard (2016) gives a 9.6% probability of approval for a Phase 1 treatment, with a range of 5.1% to 26.1% for different types of disease. Travessa et al. (2017) found a total success rate of 3.5% for trials looking at treatments for Huntington's disease, with a success rate of Phase 1 trials at 25%.

Consider an assurance calculation for a Phase 1 trial which gives the maximum probability of success as being 30%. This might suggest the trial is unlikely to succeed compared to average Phase 1 trials. However, if the trial was looking into a treatment for a disease such as Huntington's disease, which historically has a low trial success rate, then this result may be relatively promising. This is noticeably different from power, which always takes a value between zero and one and for which the bounds, accordingly, do not influence interpretation.

It is important to recognise that these average success rates should be used to put assurance values in perspective, rather than used as strict guidelines. These rates of success do not consider every trial planned, as many potential treatments do not even reach Phase 1. Furthermore, the rate of success will vary depending on the disease being studied. Rare diseases and diseases with short survival times are naturally harder to study, and are more likely to have a lower probability of success.

This difference in success probabilities means that a single assurance cut off value that can be used in the majority of trials is not likely appropriate. Instead, consideration should be taken on a case by case basis.

One way of quantifying this would be to use statistical decision theory. An example of this approach can be found in health economic modelling. The Quality Adjusted Life Year (QALY) is a measure of the relative quality of a person's life with respect to health, and can be used to measure the benefit a new treatment may provide (Ogden, 2017). It can be calculated using the following equation

$$QALY = \text{Expected number of years of life} \times \text{Quality of Life utility value} \tag{2.28}$$

The Quality of Life (QoL) utility value is set to one when the person would have perfect health, and at lower values for less than perfect health. The difference between the QALY without treatment and the QALY with a treatment gives a value for the benefit provided by the treatment.

The National Institute for Health and Care Excellence states that an improvement of 1 QALY is cost-effective if it can be achieved for less than £20,000, while those costing between £20,000 and £30,000 can be cost-effective in some cases (NICE, 2012, 2013). It is acknowledged that this method does not capture all information, and is not the only part of the decision making process.

The estimated QALY can be used along with the assurance, and other relevant values such as the costs involved with designing, testing, and providing the treatment to patients, as a basis for a decision-theoretic approach to sample size calculation. Such an approach focuses on the cost-benefits of the trial, and would select a sample size which optimises the outcomes based on these quantified costs and benefits. However, this approach may require further information about the wider pathway to fully capture the effects implementing a new treatment or diagnostic test may have. This may include considering post-treatment or test outcomes, follow-ups, or further treatments or tests which may result from the proposed change to the current standard methods.

Further attributes could be considered within a full decision analysis, choosing the decision which maximises the expected utility of a multi-attribute utility function, for example. Figure 2.8 shows a typical assurance curve. As the sample size increases, the assurance increases. The assurance will asymptote to its maximum value as n increases. The assurance curve's second derivative is negative, which is to say the rate of change is decreasing across the curve. As such, each additional unit increase in the sample size provides a decreasing amount of additional assurance.

Instead of selecting an assurance as a cutoff, a researcher could instead determine how much additional assurance an extra unit of sample size is worth. For example, if increasing

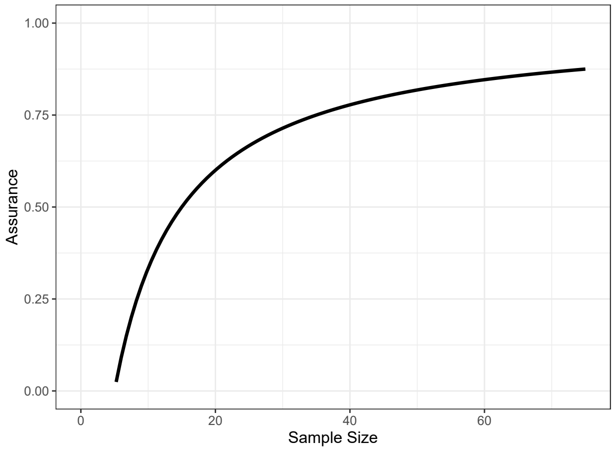


Figure 2.8: An example assurance curve, showing the assurance values over a range of sample sizes.

the sample size by one is not deemed worthwhile if the associated assurance increase was 0.01, then this value could be used to determine the corresponding sample size.

Alternatively, a specific value of the slope could be chosen. A gradient of $\varDelta A$, or the change in assurance per increase in sample size, can easily be determined for all points on the curve, and so if a specific gradient is selected, the corresponding sample size can be found. For example, a gradient of one in a curve similar to Figure 2.8 represents the point where the sample size begins increasing faster than the assurance. This might be an appropriate cutoff for some assurance curves, but due to the different scales of the two axes, for many it will provide a very low assurance.

2.5.1 Patient Dropouts

When conducting a clinical trial, it is likely that not all patients will remain in the trial for the entire time period. Patients may leave the trial due to worsening condition, personal circumstances or other reasons.

This may mean that the initial sample size provided the correct power or assurance, but the final sample does not. To ensure this is not the case, initial sample sizes are often inflated to account for potential dropouts.

Dropout rates vary from trial to trial. For example, Cooper and Conklin (2015) found a dropout rate of 20%, and Dixon and Linardon (2020) a 28% dropout rate, for trials concerning mental health issues. For anti-psychotic drug trials, Wahlbeck et al. (2001) found a dropout rate of a third, which was similar to the rate found in Santarlasci et al. (2003), which reviewed schizophrenia treatment trials. If such a dropout was unaccounted for, then the sample size required for a certain power or assurance will not be achieved.

The potential for patients to drop out of a trial can be accounted for by inflating the sample size. In order to account for a 10% dropout rate, the sample size would need to be divided by 0.9. Adjustments could also be made to the power or assurance requirements in a post hoc manner. For example, if a trial was designed to achieve a power of 0.9, but received higher dropout rates than expected, the sample size may be re-calculated using a lower power of 0.8.

Dropouts can also be managed over the course of the trial. Little et al. (2012) provides a summary of how dropouts can be limited through trial design and planning, and analysis methods of adjusting for missing data. These methods involve targeting patients less likely to leave the study, and ensuring it is as easy as possible for patients to remain in the study for as long as required.

In terms of analysis, methods such as imputation can be used when there is missing data. Simple imputation methods often assume that patients drop out at random, which is not always a valid assumption. For example, patients who are clearly responding positively to a treatment may be more likely to continue than patients who do not see an improvement, or who have negative side effects. As such, it may be required to consider more comprehensive methods.

By managing patient dropouts well, the sample size can be achieved without having to recruit too many additional patients.

2.5.2 Other issues

It may not always be the case that a sample size is achievable or reasonable.

It could be determined that there is not a large enough cohort available and that it is not feasible to conduct a study, such as in Breckenkamp et al. (2009). In these cases, the trial may not be conducted until a feasible way of doing so is found. However, it is not always the case that trials can be put off until they become feasible. Rare diseases, or diseases with short survival times, may not provide access to large enough cohorts to satisfy standard sample size calculations.

Miller et al. (2018) suggests that assurance or a decision-theoretic approach may be useful in sample size calculations for rare diseases. They point out that it may be useful to calculate values based on a number of these methods, and advise taking multiple calculations into account when determining sample size.

Other options allow for trials to be designed to account for small possible sample sizes (van der Lee et al., 2008). For example, sequential designs can allow for changes to the design to be made during the trial. If early results show one treatment is performing better than others, it might be sensible to adjust the proportion of patients assigned to each treatment. This could potentially reduce the total number of patients needed compared to initial calculations before any data was collected. Likewise, a boundaries design can allow a conclusion to be reached early if results cross a particular preset boundary. For example, if an early test statistic is particularly large, or a test statistic half way through the trial is particularly small, then the trial could conclude a significant or non-significant result early.

Power and assurance calculations can also be used to back fit an effect size based on a pre-chosen sample size. This may be useful if a trial has a limit to the number of patients that can be recruited, and the researcher wishes to know what size effect could feasibly be detected.

Another issue which may need to be considered in the case of diagnostic test studies is disease prevalence. Depending on how a trial is designed, the total sample size may need to account for the prevalence of the disease. For example, if a disease is relatively rare, a larger sample may be required to allow the diagnostic test to be tested on a sufficient number of positive cases. In such cases, the sample size may determine the number of positive or negative patients required using power or assurance, and then the total sample scaled according to the prevalence of the disease. Alternatively, a prior distribution can be placed on the prevalence, such as in Wilson et al. (2021).

2.6 Conclusion

In this chapter we have considered approaches to sample size determination. We have reviewed statistical power and Bayesian assurance as two methods of selecting a sample size.

The methods outlined in this chapter will be compared, through simulations in Chapter 5 and in an application involving a clinical trial in Chapter 6.

Part III

Elicitation for Assurance

Chapter 3

Elicitation for Assurance

3.1 Introduction

In this chapter, we review elicitation protocols and detail their application to a case study.

We begin by outlining the elicitation process in general. We then review common cognitive biases which can influence how experts provide their judgments, and the elicitation techniques that can be used to address them. We then review elicitation aggregation techniques for combining the judgments from groups of experts. These are categorised as mathematical aggregation methods, which use a predefined mathematical rule to aggregate distributions, and behavioural aggregation methods, which provide a framework to allow the group of experts to form a consensus view among themselves.

In the context of a case study into a novel diagnostic test for Motor Neurone Disease, we discuss the implementation of the above techniques to two elicitations, each of which aims to compare popular mathematical aggregation methods to a popular behavioural aggregation method. The first included in-person meetings and discussions, while the second was conducted entirely online.

3.2 Elicitation

Informative prior distributions can be created using information from a variety of sources. For example, information from previous studies can be codified into a probability distribution for use as a prior. However, it is often the case that the information required is not easily accessible in a quantified form. This is especially the case when investigating novel methods for the first time, or in early stage research. In these cases, it is possible to utilize expert knowledge to form prior distributions. The elicitation process also has the benefit of gathering additional information and added transparency through the formalised process (Dallow et al., 2018).

The process of converting expert knowledge into subjective probability distributions is commonly referred to as Expert Elicitation (Mikkola et al., 2021) . Throughout this thesis, the people providing their knowledge for use in specifying prior distributions will be referred to as experts, and the person running the elicitation as the elicitor.

As experts do not often think directly in terms of parametrised probability distributions, these need to be elicited from them. Transforming the valuable expert knowledge and experience into distributions allows the knowledge to be incorporated into a model. This process is not always simple, as the experts may have little mathematical training or experience, and can be influenced by a number of psychological biases.

French (2011) identifies three contexts under which elicitations might be carried out: the case where the decision maker (DM) and experts are separate groups of people, known the expert problem; the case where the DMs and experts are the same people, known as the group decision problem; and the case where there is no DM and the experts are giving judgments for yet undefined circumstances, known as the textbook problem. Under the context of assurance, the DM is the person for whom the sample sizes are calculated, and so we will focus on the expert problem.

Elicitation has previously been used as a basis for sample size calculations, especially in medical settings (Lenth, 2001; Cook et al., 2014, 2018). Alhussain and Oakley (2020) provides an example of elicitation for assurance calculations, as discussed in Chapter 2, which demonstrates that elicited priors can play a large role in the calculated sample size. Many guidelines exist on how to elicit the values, though in general, these guidelines are common with those eliciting values in other fields or for other reasons. The elicited priors can be used alongside, or be updated with, previously collected data (Mayo and Gajewski, 2004).

3.2.1 The Elicitation Process

While the process of elicitation varies given the specific scenario in which it is conducted, there are a common set of steps that tend to be employed.

Bojke et al. (2019) present an elicitation protocol for healthcare settings. First, the relevant variables are identified. These variables could be probabilities, counts, times-to-event, diagnostic accuracies, minimum clinically important differences or effectivenesses. The variables chosen should be fit for purpose, so they can be used to properly inform the DM (Choy et al., 2009). They should also reflect the experts' knowledge and how they are familiar with thinking about the problem.

Elicitation questions can be direct or indirect. Direct questions address things the experts can comment on specifically, such as the number of patients they may treat in a year. Some information is not as easily gathered, especially that which is more complicated, entwined with other information, or is of a sensitive nature. Indirect elicitation questions attempt to elicit this more difficult to reach information (Hudlicka, 1996). Instead of directly asking about the quantity of interest, it is inferred based on responses to other questions. For example, Browne et al. (2020) indirectly elicits information on the effects of gambling that participants may not recognise themselves or accurately divulge.

In terms of the quantities which should be asked about, Speirs-Bridge et al. (2010) suggest a four step elicitation process. The first and second steps are to ask for the lowest and highest realistic values. Thirdly, a best estimate or median is requested, and then finally a probability level that the true value will fall within the stated interval.

It was found in Teigen and JØrgensen (2005) that allowing experts to define their own probability to intervals led to much better calibrated intervals than asking for an interval with a set probability. They also suggest that the probability assigned for an interval has little impact on the width, highlighting the difficulty experts can have with the elicitation process and why it needs to be carefully managed. Soll and Klayman (2004) also found respondents who were asked to specify a minimum and maximum in two separate questions tended to give better calibrated intervals than those asked to specify an interval in a single question. Furthermore, a three point estimate (minimum, maximum and median) led to better calibration than a two point estimate (minimum and maximum).

Experts then need to be selected. The provided judgements will vary by expert, and so it is best to have a heterogeneous group of experts to properly gather the full range of views (Verdolini et al., 2015). Differences can, for example, be due to physical location or employment background (Nemet et al., 2017). Australian Government: Department of Health (2016) states that a random or comprehensive group of experts is preferred over an alternative such as an advisory board, which may not have as general or wide-reaching range of knowledge.

The number of experts is also a consideration. Bolger (2018) finds that many suggestions published on this issue are based more on opinion rather than theory or evidence, however some literature (such as Budescu and Chen (2015)) does show that a greater number of experts improves performance as long as the additional experts bring a positive contribution. Alternatively, Mannes et al. (2014) showed that when there is a range of expertise in a group, a subset of the best performing experts (or even the single best performing expert) can give a better outcome than the entire group together. While there is room for further research in this area, it seems clear that the ideal group of expertise is well calibrated and is heterogeneous. It should also be noted that constraints on which experts are available to participate can often be a deciding factor.

Experts should receive some training before the elicitation begins. Experts cannot be assumed to be familiar with concepts such as quantiles, probability distributions, or specifying their beliefs numerically, and so the training should be aimed to address these areas of unfamiliarity. Furthermore, as the statistical accuracy of expert judgements is related to their informativeness and calibration, helping the experts improve on how well they can quantify their beliefs will improve the quantities provided (Aspinall et al., 2016).

The elicitation session can then be held. It is important the elicitor is aware of the cognitive biases outlined in Section 3.2.2, in order to actively ensure they are minimised. The elicitor should assist the experts with the elicitation questions, by clarifying any issues they have and helping to draw out the best quantifications of the experts' knowledge. In order to not influence the experts themselves, this usually involves asking guiding questions to encourage the experts to further explore the issue. For use in Bayesian statistics, probability distributions are often fit to the expert judgements. While these will often be parametric distributions, the elicitation can also be designed to elicit for non-parametric distributions (Oakley and O'Hagan, 2007).

After the elicitation is complete, there are a number of activities that can be completed. Elicited responses may be fit to distributions or aggregated together. The experts may also be provided feedback, to help them improve for future elicitations.

3.2.2 Cognitive Biases

In elicitations, experts are often already operating outside of their usual framework. Experts, and people in general, rarely think directly in the terms of parametrised probability judgments. They may also be unfamiliar with probability intervals, especially when it comes to assigning numbers based on their own knowledge and experiences. It is important that an elicitation makes this process straightforward for the experts, and allows them the best opportunity to translate their knowledge to numbers.

The content of the questions asked of the experts is determined by the statistical models which will be applied. There are, however, a vast number of options as to how these questions can be constructed. This is a very important consideration, as the way in which an expert is asked a question can have a large impact on their response.

The specific wording and design of a question have been shown to influence the responses provided. A common example is that from Tversky and Kahneman (1981), which offered participants a scenario in which 600 people had been infected with a dangerous disease. Two options were presented, one in which 200 people could be saved for sure, and another in which there was a $\frac{1}{3}$ chance everyone would be saved, otherwise they would die. The majority of participants chose the first option. For a second group, the options were presented in a different way. The two options became one in which 400 people would die, and another in which there was a $\frac{2}{3}$ chance everyone would die, otherwise they would all be saved. In this case, the majority of people selected the second option.

Despite the two sets of options being the same, the change of framing from lives saved to deaths occurring changed how the participants responded. It was suggested that people tend to be more risk-averse when talking about gains, and more risk-taking when talking about losses, and that changing how the question was asked changed how the participants considered and answered it.

If the framing of a question can completely alter the responses to it, then the way elicitation questions are asked should be carefully considered.

One of the main causes attributed to these influences are cognitive biases. Cognitive biases, or cognitive heuristics, are often attributed to the brain avoiding long information processing times by making shortcuts or using rules of thumb (Haselton et al., 2015). Unfortunately, these heuristics often break down when trying to quantify judgements, and can lead to inconsistencies or errors. Well-designed elicitation questions take cognitive biases into account in their construction, and minimise the effects they may have (Montibeller and von Winterfeldt, 2018).

Some of the more common cognitive biases are explained below, along with their implications for elicitation question construction.

Anchoring

Anchoring occurs when an expert estimates a value based upon an initial value, or anchor (Tversky and Kahneman, 1974). If a question asks an expert to start with a specific value, or consider the probability of being below or above a specific value, then their future estimates will be affected by the previous value.

O'Hagan (2019) provides an example, in which participants in a study were asked to estimate probabilities concerning the number of Muslims in the 2011 United Kingdom Census. Participants were either asked to estimate the probability the total number was over 8 million, or asked if it was over 2 million and then asked if it were over 8 million. Those in the first group gave a higher probability to the value being greater than 8 million. This was because the second group had been anchored on the 2 million value, thus making 8 million seem larger and less likely by comparison. Likewise, those anchored on the 8 million value first gave a higher probability to the total being over 2 million, as it seemed more likely compared to 8 million.

The more significant figures an anchoring value has, the stronger the effect tends to be (Loschelder et al., 2014). This has been shown to affect experts as well as amateurs (Loschelder et al., 2016), suggesting that it is an important consideration when constructing elicitation questions.

Anchoring can also occur in groups (de Wilde et al., 2018). Groups that were cooperating were shown to be more susceptible to anchoring than competitive groups, which further outlines the potential for this bias in an elicitation setting.

To avoid anchoring, elicitation questions should be written to avoid providing any values the experts could use as an anchor. Unless there is a strong reason to do so, asking experts to give estimates of probabilities that a quantity will be above or below a specific value can be avoided. By asking the expert to specify the values themselves, the elicitor avoids 'leading the witness' and influencing the expert.

Availability

The availability bias occurs when people judge probabilities based on the ease of remembering specific examples. For example, Tversky and Kahneman (1973) asked people to judge how often a letter would occur as the first letter of a word compared to the third. For all letters included in the study, participants said they were far more likely to occur at the start of the word. As it is much easier to recall words that start with a particular letter, rather than recall words where a particular letter comes third, the participants were overestimating the ratio of occurrences between the two groups.

Availability bias can be more pronounced when considering rare events, or rates of occurrence (Meyer and Booker, 2001). Rare events are generally more memorable, while all of the occasions in which they didn't occur are less so. By focusing solely on what is more memorable, the experts can miss less memorable events and occasions where the event did not happen at all.

To assist the experts in avoiding availability bias, they can be asked to consider their full range of experience (Morgan, 2014). By reflecting on not just recent or memorable events, experts can take more information into consideration when making their judgements. When considering rare events, experts can also be reminded to consider the total possible time frames in which the event could have occurred but didn't.

Representativeness

Representativeness can be an issue when judging the probability that something belongs to a particular class or category. Tversky and Kahneman (1974) give the following example, where participants were provided with the following description of a person.

> "Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail."

If asked about his occupation (for example, from a list of farmer, salesman, airline pilot, librarian, or physician), people tend to focus on which he is most representative of. This ignores the base rate at which the occupations occur. In the given example, they suggest that because this character fulfils more stereotypes of a librarian than the other occupations, people tend to assign a higher probability to that choice.

They continue with an experiment in which people are asked to estimate whether a personality description belonged to an engineer or lawyer. The participants were either told these belonged to a group of 70 engineers and 30 lawyers, or 30 engineers and 70 lawyers. Despite the change in base rates, the two groups assigned extremely similar judgements, basing their decisions on which stereotype the profiles fell into.

This bias can also apply when an option is representative of neither class. From the same study, the following profile was provided.

> "Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues."

This description was created to provide a case where the profile was not representative of either group. Regardless of which base rate the participants were provided, they tended to provide a probability of 0.5 of Dick being an engineer or lawyer. It is suggested that as he was representative of both groups equally, then the representativeness probability of 0.5 overrode the base rates between the groups of 0.3 and 0.7.

To avoid the representativeness bias from impacting expert judgements, it is important to have the experts consider the underlying base rates closely. The elicitor can achieve this through the elicitation questions, or could discuss them before the questions are asked.

Range-Frequency

When presented with a number of categories to assign probabilities to, it is often the case that experts will try to spread the probability equally between them. This can result in the final judgments being a compromise between the experts' actual beliefs about the probabilities of each event, and an even split between them (O'Hagan, 2019).

In an experiment in Fischhoff et al. (1978), participants were asked to consider the reasons why a car wouldn't start, and assign probabilities between possible causes. In the first round, participants could assign probabilities to six named causes or 'Other'. In the second round, the number of named causes was reduced. While the total probability across the named causes common in both rounds was expected to be constant, and the probabilities from the removed causes to be included in 'Other', this was not the case. Instead, it was found that participants would increase the probabilities assigned to the named causes, resulting in the probabilities assigned being more equally split between the offered groups.

Fox and Clemen (2005) conducted a similar experiment, asking participants to assign probabilities to which MBA program they thought would be ranked highest. They assigned probabilities to six categories, and then a collapsed version with just two. The median probability assigned to the university with its own category both times changed from 0.3 to 0.6. They also showed that listing alternative options within the 'Other' category does not change the results. This suggests the difference is not due to the availability bias, where the participants may not have considered the same range of possibilities contained within 'Other'.

To avoid this bias, elicitation questions can be constructed to avoid asking experts for a number of categories at once. By asking for each value one at a time, the experts can consider each individually.

Probability Laws

Kynn (2008) summarises a further issue, which is internal coherence. This refers to whether the answers given by the expert follow an internal consistency, such that the probabilities provided are valid under the laws of probability. If experts are providing a number of probabilities that are related, perhaps as a joint or conditional probability, it is important that they follow the laws of probability.

The most straightforward case is when probabilities should sum to one. This could be the case when asking about a probability and its complement, or when asking about mutually exclusive events. By definition, $P(A) = 1 - P(A^C)$ for an event A. While the case of the complement of a probability is well known and able to be avoided by the experts, in cases where there are a large number of possible events, it may become harder to ensure the probabilities remain consistent.

In these cases, the elicitors could choose to infer one set of probabilities from the others. For example, if there are three mutually exclusive events that represent the only possible cases, eliciting the probability of two of them allows the elicitor to infer the probability of the third occurring. It may also be advisable to feed any inferred probabilities back to the experts as a check.

Another case which may occur is when considering joint probabilities. By basic probability laws, $P(A \cap B) \leq P(A)$ for events A and B. However, there is evidence that this is not always intuitive. A common example from Tversky and Kahneman (1974) provided participants with the following passage about a woman called Linda.

> "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

They were then asked if it was more probable that Linda worked as a bank teller, or that she worked as a bank teller and was involved in the feminist movement. The majority of participants stated the second was more likely, despite the laws of probability stating it can at most be equally likely as the first option.

This effect disappears, however, if the question is rewritten in a different way (Fiedler, 1988; Gigerenzer, 1991). If instead of asking participants which is more likely, they are asked to consider 100 people similar to Linda and consider how many they believed belonged to the two groups, the proportion of inconsistencies drops from over 80% to around 20%.

As such, this issue can be accounted for by the elicitor carefully considering the questions they are asking. Asking about the number of occurrences out of 100, for example, may help experts to stay consistent as the consider different proportions. Alternatively, in cases where events are conditionally independent, joint probabilities can be calculated using just the individual probabilities. In cases where the events are dependent, a joint probability could be calculated by asking the expert about the conditional probability instead.

One final case to consider is Bayes rule, which given events A and B states

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3.1}$$

Also considering that $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, an expert who is asked about a number of related probabilities can easily end up contradicting themselves. For example, if an expert provides values for $P(A)$, $P(B)$ and one joint or conditional probability relating A and B, then the remaining joint and conditional probabilities can be calculated. This means that if the expert provides any of the remaining probabilities, there is a high chance they may contradict themselves.

Many probabilities can be calculated using the full range of probability laws once a few have been elicited. It is important to note that any probabilities calculated should be checked with the expert before they are used, to ensure they agree. While an extrapolated value may be mathematically correct, experts cannot be assumed to consistently align with them. It may be useful to show extrapolated values and allow the experts to modify their initial values to ensure they agree with the full set of probabilities.

Overconfidence

Overconfidence occurs when experts specify probability intervals that do not correspond with the observed occurrences of true values. For example, if an expert provides a series of 50% probability intervals, we would expect the true value to lie within them 50% of the time. An overconfident expert would have less than 50% of the true values within their 50% probability intervals. Overconfidence has been shown to be advantageous in many settings, and so it would not be unexpected for it to present in elicitations of judgment (Johnson and Fowler, 2011). It has also been shown to be present cross-culturally, demonstrating it is a widespread bias (Acker and Duck, 2008).

Lichtenstein and Fischhoff (1977) found participants' 100% intervals had an actual probability of between 55% and 95%. They also found overconfidence somewhat negatively correlated with expertise, with those who answered the most questions wrong tending to be the most overconfident. Additionally, participants who correctly answered over 80% of the questions started showing signs of underconfidence. Soll and Klayman (2004) suggest that overconfidence is higher in this style of interval estimates than with assigning probabilities to point estimates.

Plous (1995) asked participants to specify 75% or 90% intervals, and found both groups had similar overall accuracy. This suggested that the participants were not aiming to be 75% or 90% accurate, but perhaps aiming for 100% accuracy. When questioned, many participants felt that giving a wider interval would be uninformative, useless, cheating, make them look ignorant, or just 'a cop out'. This suggests that some elements of overconfidence can be due to deliberate decisions, and not just inaccuracies in judgments.

Overconfidence can affect groups as well as individuals. Schuldt et al. (2017) showed a case where pairs of experts gave more confident judgements together than they did individually, despite no increase in accuracy.

Overconfidence can also be found in estimates of prediction of an individual's own future performance, and not just abstract judgments (West and Stanovich, 1997). This suggests overconfidence can be a problem regardless of the content of the questions, and may be better to be addressed through the terms in which the question is asked. For example, it may be possible to reduce overconfidence by framing questions in frequencies rather than probabilities (Cesarini et al., 2006).

Group Biases

During group meetings and discussions, all of the previous cognitive biases can affect experts, as well as some specific to group dynamics.

Groupthink is when a group's thinking is dominated by concurrence seeking rather than a more realistic perspective of alternative options (Janis, 1972). While in the context of elicitation it may sound beneficial to have a group of experts with strong consensus-finding behaviour, groupthink tends to narrow the group's range of views. For example, if an expert is influenced heavily by groupthink they may state their agreement with the group, regardless of any other opinions or disagreements, in order to conform.

Symptoms of groupthink include collective rationalisations, pressure to conform, self-censorship, and an illusion of unanimity (Janis and Mann, 1977). Experts can feel pressure to conform to the group's opinion, either by adjusting their views to match or avoiding contradicting views. Experts can also self-censor to avoid going against the group. Another case is that of collective rationalisation, when the group accepts reasoning without proper consideration, leading to a common misconception. Finally, the illusion of unanimity can lead the members of the group to think that the group's opinion is held by everyone, overlooking any disagreements or the potential self-censorship.

There are a wide variety of case studies showing the effects of groupthink, many of which are compiled in Esser (1998) and Janis (1991). These examples often include groups of experts with extensive expertise or in positions of power, demonstrating this effect is widespread.

Another issue to consider is group tenure. The longer a group works together, the less pronounced their differences tend to become. The effects of demographic diversity have been shown to weaken the more a group works together (Chatman and Flynn, 2001; Harrison et al., 2002). Group tenure also leads to an increase in confidence, potentially leading to overconfidence in their judgements (Meissner et al., 2018).

Methods involving elicitation from groups of experts are designed with these group biases in mind. Often, they have experts work individually on the elicitation questions before bringing them together in a group. The individual stage will ensure the experts avoid the group biases with their first judgements, and help to narrow the influence of this bias in the group stage. By showing the group of experts what they thought individually, then it encourages them to consider all group members' uncensored views and the full range of opinions.

3.3 Elicitation Aggregation

When we elicit a prior distribution from an expert, it is representative of that expert's knowledge, experiences and beliefs. By eliciting a probability interval, we can account for the uncertainty present within the expert's views and understanding of previous data. This uncertainty surrounding a future study can be epistemic or aleatory in nature (O'Hagan and Oakley, 2004).

Epistemic uncertainty refers to the expert's lack of complete knowledge. Each individual will have a unique set of knowledge, experiences and beliefs, and the difference in these between individuals is the cause of their differences in epistemic uncertainty. The epistemic uncertainty is aimed to be captured by eliciting a probability distribution rather than a single point estimate.

Aleatory uncertainty refers to the uncertainty which is due to statistical variation. If we are eliciting the likely outcome of a future trial, for example, there is uncertainty about what will occur simply from the fact that the sample of patients will be random. An expert may believe a treatment has a certain level of effectiveness in the population of patients, but as the trial only takes a sample from the population then there is uncertainty as to the outcome of the trial.

While aleatory uncertainty will always be present, epistemic uncertainty can be reduced by using multiple experts. By combining the views of multiple experts, the uncertainty due to a specific expert's knowledge, experiences and beliefs can potentially be reduced.

The resulting distribution can provide a better overview of the current, wider state of knowledge within the field, which in turn provides a more nuanced prior distribution.

The problem of how to represent the views of multiple experts in a single analysis has many possible solutions. Each expert's beliefs could be represented by their own individual prior, and the analysis completed as many times as there are experts. Alternatively, the experts' beliefs could be aggregated to form a group prior, allowing for a single analysis (West, 1984). Group priors have been shown to be more informative and have better probabilistic calibration than priors of individual experts (Clemen, 2008; Lin and Cheng, 2009).

There are many methods of aggregating a number of expert prior distributions into a single group prior. Broadly speaking, the two widely used approaches for aggregation are either mathematical rule based, where individual priors are combined according to a predetermined mathematical rule, or behavioural, where experts work together to form a group prior. Some methods do combine these two approaches, using both behavioural and mathematical rules during the aggregation process. Common methods are discussed in detail in EFSA (2014).

While each category has methodological differences that provide benefits in different circumstances, their statistical performance relative to each other has not previously been directly compared. In this thesis, an elicitation for a clinical trial into diagnostic tests for Motor Neurone Disease (MND) has been used to form a range of priors and used as a basis for comparing the following aggregation methods in Chapter 4.

3.4 Mathematical Aggregation Methods

Mathematical aggregation methods combine experts' distributions using a predefined, mathematical rule.

There are a number of benefits that come from using mathematical aggregation methods. Firstly, each distribution can be elicited separately. The experts do not need to meet, or even know each other, for these methods to be used. Mathematical aggregation methods can easily include elicitations from different times and locations, allowing experts more flexibility to complete the elicitation.

Additionally, the predefined rule ensures that there is an impartial method of assigning weights to the experts. This ensures that the final distribution will not be swayed by individual experts' personal interests, and that the final group prior will not have been deliberately chosen to favour a particular individual or viewpoint. Groups of experts can also have imbalances in occupational position or power, or experts who push their opinions more forcibly than others. The predefined rule can assist in these cases to ensure no expert has an unreasonable or unjustified sway over the final decision.

Most mathematical aggregation rules combine probability distributions by assigning each expert a weight, and then calculating a weighted combination of the distributions. For n experts, each expert $i = 1, \dots, n$ has prior probability distribution $f_i(\theta)$ for a parameter of interest, θ.

Each expert is assigned a weight w_i by the selected aggregation rule, such that all of the weights sum to one and are non-negative ie. $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$.

Once the weights have been obtained, they can be used to calculate aggregate priors. There are two common ways of doing this.

This first method is simple linear pooling. In this case, the individual priors are multiplied by the experts' weights, and summed (O'Hagan, 2006). This results in a prior where $f(\theta)$ is the weighted arithmetic mean of each expert's prior distribution, $f_i(\theta)$ at all values of θ.

$$f(\theta) = \sum_{i=1}^{n} w_i f_i(\theta) \tag{3.2}$$

Alternatively, a log linear pooling method can be used instead. This results in a prior where the value of $f(\theta)$ is the weighted geometric mean of each expert's prior distribution, $f_i(\theta)$ at all values of θ.

$$f(\theta) = k \prod_{i=1}^{n} f_i(\theta)^{w_i} \tag{3.3}$$

where k is a rescaling factor that ensures $f(\theta)$ integrates to 1.

In comparison to a linear pool, a log linear pool emphasises the sections where the experts have the most agreement. Figure 3.1 shows an example of two experts' distributions being aggregated with linear and log linear pooling. The linear pooling method, in purple, forms a bi-modal distribution, where the modes match the locations of the experts' modes. The log-linear pooling method, in orange, places the mode at the point with the highest agreement between experts.

In cases with more than two experts, such as the four experts in Figure 3.2, both pooling methods result in similar styles of distributions. The linear pooling method, in purple, creates a multi-modal distribution, with high density around the modes of each expert's distribution. In this case, the modes of the three distributions with higher means are close, and so the aggregated prior has not formed distinct modes for each. Likewise, the log linear pooling method, in orange, has created a unimodal distribution with a mode representing expert agreement.

While the log linear pooling method creates a simpler distribution, it does have a number of issues.

For any value of θ for the pooled distribution $f(\theta)$ where one expert has provided a probability of zero, i.e. $f_i(\theta^*) = 0$, then the aggregated distribution will have a value of

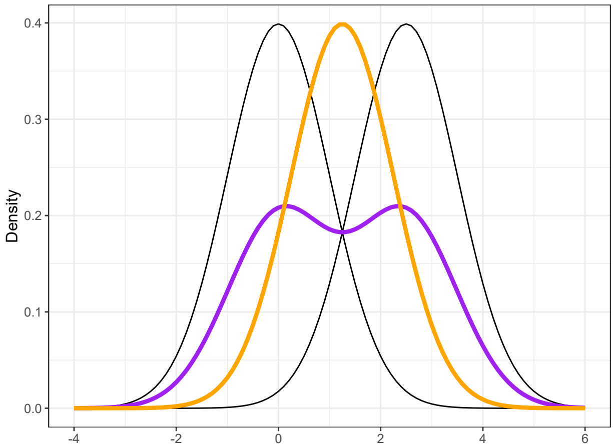


Figure 3.1: Linear (purple) and Log linear (orange) pooling for two experts (black) with equal weights.

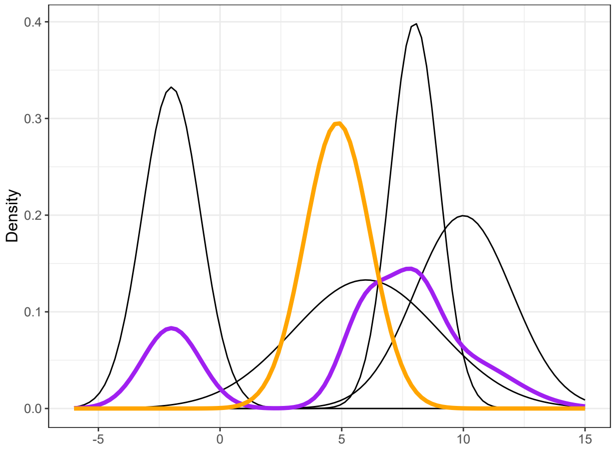


Figure 3.2: Linear (purple) and Log linear (orange) pooling for four experts (black) with equal weights.

zero at that location, $f(\theta^*) = 0$. This can cause issues when an expert provides a minimum and maximum value of θ, where they assign a probability of zero to everything outside the interval. In this case, the bounds provided by any expert become hard cutoffs, so that the final aggregated distribution can only have a range between the largest minimum and smallest maximum. Furthermore, if two experts' distributions do not overlap at all, the aggregated prior will have a density of zero for all values of θ.

A second issue that can arise with log linear pooling is when there is little overlap between the densities of experts' distributions. In these cases, the aggregated distribution will place high density on the small overlap, and comparatively little on the remaining values. This results in a distribution with high probabilities in areas no individual expert has deemed likely, and low probabilities in areas the experts do deem likely.

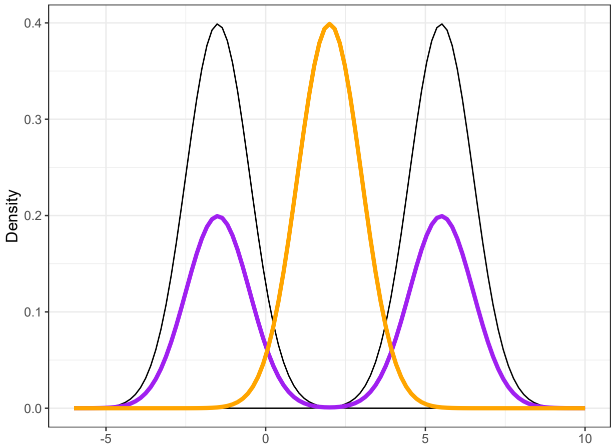


Figure 3.3: Linear (purple) and Log linear (orange) pooling for two experts (black) with equal weights.

Figure 3.3 shows an example of this. While the linear pooling method, in purple, splits the aggregated prior between the two experts' distributions, the log linear pooling method, in orange, places the aggregated distribution between them. This area is that with the most expert overlap, however neither expert has deemed it likely for the parameter to lie in this region. The aggregated prior has formed a compromise which then represents neither of the experts' views, and its practical interpretation may not be realistic.

While the above issues can be identified before the aggregation is run, and addressed by using a linear pooling method instead, the log linear pooling method is much less commonly used. We therefore focus on approaches that use the linear pooling method to determine how to assign weights to each expert.

3.4.1 Equal Weights

The Equal Weights (EW) aggregation method assigns the same weight to each expert. For each of the n experts, they are assigned a weight of

$$w_i = \frac{1}{n}. \tag{3.4}$$

This method treats all experts equally. By doing so EW ensures impartiality between experts, which may be advantageous in some circumstances. EW is also the easiest weighting system to use. As it does not require any additional information about or from the experts, it does not add any additional questions or work during the elicitation process. This ensures the elicitation is as short as possible. Additionally, EW also allows for easier aggregations across multiple sources. By not requiring specific information, any elicitation with multiple experts can be aggregated using it. As such, it is a relatively standard comparison to compare against more complex aggregation techniques.

3.4.2 Classical Method

The Classical Method (CM), or Cooke's Method, assigns experts weights based on calibration and informativeness scores (Cooke et al., 1988). These scores are calculated using a number of seed questions, for which the elicitor knows the answers but the experts do not.

The weight of expert i, $i = 1 \ldots n$, is given by

$$w_i = \frac{w_i^*}{\sum_{i=1}^{n} w_i^*} \tag{3.5}$$

where the w_i^* terms are given by

$$w_i^* = C_i \times I_i \times \alpha_{C_i} \tag{3.6}$$

and C_i is the calibration score, I_i is the informativeness score, and $\alpha_{C_i} = 1$ if the calibration score is above a cutoff, or otherwise $\alpha_{C_i} = 0$.

The calibration and informativeness scores reflect the performance of the experts in the two different ways, comparing their seed question intervals against the ideal case (e.g. perfect calibration and highly informative).

The calibration score is a measure of how well an expert specifies a probability interval. For example, an expert's 25% probability intervals would be well calibrated if 25% of the repeated intervals contained the true value. Likewise, a well calibrated expert's estimates of the median would be above the true value 50% of the time, and below the true value 50% of the time. A poorly calibrated expert may still give responses that are close to, or include, the true value, but their probability intervals would not represent where the true value falls upon repeated measurements. The calibration score is given by

$$C_i = P(2qI_i \leq x) \tag{3.7}$$

where q is the number of seed questions (Cooke, 1991). The value of $P(2qI_i \leq x)$ is approximated using the χ^2 distribution, with $q - 1$ degrees of freedom. The value of $\chi^2_{q-1}(2qI_i)$ is bounded between zero and one.

The informativeness score is a measure of how much information the expert has provided. An informative expert provides narrow probability intervals, which provide more information than wider intervals. For example, for a parameter with a range between zero and one, an expert who provides a 90% probability interval of (0.4,0.5) has provided a more informative interval than one who provides an interval of (0.1,0.8). A less informative expert may still provide responses that are well calibrated, or centred around the true value, but their priors would have more uncertainty than an informative expert. The information score is calculated by

$$I_i = \sum_{j=1}^{r} s_j \ln \frac{s_j}{p_j} \tag{3.8}$$

where s_j is the expected proportion of seed questions values that fall within a probability interval, p_j is the observed proportion of seed questions values that fall within the probability interval and r is the number of probability intervals elicited for each variable.

The α_{C_i} term provides a mechanism for ensuring only well calibrated experts are included in the final aggregated prior. Experts with a calibration score less than the cutoff are given a weight of zero, excluding them from the aggregated prior.

Seed Questions

The Classical Method requires the experts to complete a number of seed questions alongside the elicitation. The seed questions should be questions the experts have uncertainty about the answers to, while the elicitor has, or will have, access to the answer. For clarity, the questions in an elicitation that are used to create the individual experts' prior distributions will be referred to as the elicitation questions to distinguish them from the seed questions. It is important to carefully construct the seed questions to ensure they reflect the performance of the experts in the elicitation questions.

Quigley et al. (2018) classifies seed questions as either predictions or retrodictions, and either domain or adjacent to the field of expertise.

A prediction type question asks the experts about an event which has not yet occurred, been recorded, or been released. Prediction questions are preferable over retrodiction questions as they remove the possibility that the expert knows the answer. They are also more similar in nature to the elicitation questions, which tend to also be predictions about a value not yet measured. In order to use prediction questions, the elicitor will have to wait until the value is available in order to use it in CM calculations. There is also a risk that the values are not released on time, or at all. While this narrows the available time frame for which the questions can reasonably come from, it does allow the experts to make prediction style judgments as they do for the elicitation questions.

A retrodiction type question asks the experts about an event or value which has happened in the past. These questions need to be carefully selected so that the experts will not know the exact answer already, which may be a risk as the question should come from their field of expertise. There are a number of advantages to retrodiction questions. Firstly, they allow the CM calculations to be carried out immediately following the elicitation. They also tend to be easier for the elicitor to write as there tends to be a wider range of previous datasets than there are upcoming values.

Domain questions directly reflect the topic of the elicitation questions and may ask the same, or a very similar, type of question as to the elicitation questions. They are preferable over adjacent questions as they are closer in style and topic to the elicitation questions, and the the experts would be expected to perform similarly. A domain question should ideally match the elicitation in both the field of question and in the dimension of the question. For example, if the experts are to be asked to give percentage values for the elicitation, a domain question would ideally ask about percentages as well.

Adjacent questions are less related to the field of expertise than domain questions and can be less similar to the elicitation questions. Adjacent questions tend to be easier to develop, as there is a wider range of possible sources of data to base the questions on. In cases where the elicitation is being completed due to a lack of relevant information available to form an informative prior, there are clearly already limits to the construction of domain seed questions. While they provide less of a reflection of the experts' knowledge of the field, they are still valuable in examining their ability to specify probabilities.

While the difference between prediction and retrodiction is more distinct, the difference between adjacent and domain questions is less binary. A question which is not directly in the field of expertise can still be valuable if it is of a similar dimension to the elicitation questions, and vice versa. Ultimately, the more similar the seed questions are to the elicitation questions, the more similar we would expect the experts to perform in both sets of questions. As we only measure the performance of the experts in the seed questions, we want them to be as close to elicitation questions as they can be.

Other than the content of the seed questions, the number of seed questions provided to the experts also needs to be considered. A greater number of seed questions will allow for experts' informativeness and calibration scores to be better estimates (Eggstaff et al., 2013). The addition of each extra seed question provides diminishing returns to the final aggregated prior, and it is suggested that there may be an upper bound to the number of useful seed questions.

Clemen (2008) finds that the required number of seed questions to ensure the CM performs well is likely greater than 10. Bolger and Rowe (2015) show that to statistically identify a difference of 0.1 between the calibrations of two experts requires 1546 judgments, and thus suggest that 10 seed questions are insufficient. A greater number of seed questions is clearly ideal, but is not always a reasonable choice to make.

The cost of adding seed questions must be considered when developing the elicitation. Each seed question increases the amount of time an expert spends on making judgments. A longer time required to complete the elicitation then limits the number of experts who may be able or willing to commit such time, and may lead to a deterioration in judgments due to tiredness. The trade off between time and the accuracy of the informativeness and calibration scores needs to be considered, and will vary by the circumstances of a specific elicitation.

3.4.3 Other Mathematical Aggregation Methods

While we focus primarily on the mathematical aggregation methods, there are many other methods above. We briefly review some of the published material on alternative methods below.

Self and Peer Ratings

Self rating methods ask experts to rate their own abilities on a scale, and form weights based on the responses. While originally developed for use with the Delphi Method (Dalkey et al., 1970), they have since fallen out of favour due to issues such as inconsistency between experts, and overconfidence (Brockhoff et al., 1970; Aspinall and Cooke, 2011).

When the experts involved know each other, an alternative could be to request anonymised ratings of each other. This peer rating system hopes to remove personal bias, reduce the inconsistency between how experts rate and their overconfidence (Degroot, 1974; Bordley et al., 1986). However, there is still strong evidence that these ratings do not accurately represent the performance of experts. Burgman et al. (2011) found that when experts were asked how well they and their peers will perform, their responses correlated with years of experience and publication record, but not performance. Brockhoff et al. (1970) found the number of years of experience in a field also does not correspond with performance.

Rating the experts could also be achieved by directly measuring their years of experience or publication record, given that those factors have been shown to heavily impact the experts' own ratings. Cooke (2008) investigated a weighting scheme using the number of papers an expert has published which had been cited by other experts involved in the elicitation. This was shown to be outperformed by the Classical Method.

While perhaps easier to implement than the Classical Method, the poor performance of expert determined ratings means they are not worth implementing. A mathematical weighting method that includes a measure of the experts' performance through measurement rather than personal factors is preferable.

Moment Methods

Instead of combining probability distributions, a Bayes linear approach can be taken to combine moments. These moments can be difficult to elicit directly, so may need to be calculated based on expert judgments. Instead of using the conditional probability rule defined by Bayes theorem, Bayes linear analysis instead conditions on the observations using linear fitting (Bedford et al., 2010). For a vector of quantities of interest $\vec{X}$ and observed data $\vec{D}$,

$$E_D(\vec{X}) = E(\vec{X}) + \text{cov}(\vec{X}, \vec{D})(\text{var}^{-1}(\vec{D}))(\vec{D} - E(\vec{D})) \tag{3.9}$$

$$\text{var}_D(\vec{X}) = \text{var}(\vec{X}) - \text{cov}(\vec{X}, \vec{D})(\text{var}^{-1}(\vec{D}))\text{cov}(\vec{D}, \vec{X}) \tag{3.10}$$

The elicited values can then be combined, with Wisse et al. (2008) suggesting an analogous approach to the linear pooling method. For each expert i and variable v, derived estimates of the mean, $\mu_{i,v}$, and variance, $\sigma^2_{i,v}$, can be used to calculate a penalty function using the following equation.

$$\psi_i = \sum_{v=1}^{V} c_1(\theta_v - \mu_{i,v})^2 + \sum_{v=1}^{V} c_2(\theta_v^2 - \sigma^2_{i,v})^2 \tag{3.11}$$

where θ_v is the realisation of the variable of interest and V is the total number of variables. It is suggested to set $c_1 = 1$ and c_2 such that

$$0.5 = \frac{\sum_{v=1}^{V} c_2(\theta_v^2 - \sigma^2_{i,v})^2}{\psi_i} \tag{3.12}$$

They also show the method outperforms an EW linear pool, and can perform similarly to the Classical Method when using the same seed questions to form weights, while assigning weights to a greater range of experts.

Supra-Bayesian Method

Another approach is to consider the prior from the decision maker's points of view, with it being informed by the experts' elicited judgements (French, 1981). This supra-Bayesian approach can start with an uninformative prior, learn about the parameter of interest through Bayesian updating using expert priors as data, and then use the resulting posterior as the prior in future analyses. In order to achieve this aggregation of prior distributions under a Bayesian framework, the experts' judgments can be treated as data and an appropriate likelihood formed to represent them (Hartley and French, 2018).

For a quantity of interest θ, and a series of expert statements Q,

$$P(\theta \mid Q) \propto P(Q \mid \theta)P(\theta) \tag{3.13}$$

where $P(\theta)$ is the decision maker's prior distribution and $P(\theta \mid Q)$ is a posterior distribution representing the updated beliefs of the decision maker.

The difficulty of this approach comes from specifying a likelihood $P(Q \mid \theta)$. We will look at two examples, a Multivariate Normal approach and a Copula method.

Multivariate Normal Bayesian Method

Roback and Givens (2001) provide an example of how a multivariate normal distribution can be used to aggregate priors for two experts. Consider a quantity of interest, θ, with judgments made about it by two experts, who provide a value μ_i for the mean and σ_i for the standard deviation. The posterior distribution of the decision maker's beliefs, incorporating the beliefs stated by the experts, is

$$P(\theta \mid \vec{\mu}, \vec{\sigma}) \propto P(\vec{\mu} \mid \vec{\sigma}, \theta) P(\vec{\sigma} \mid \theta) P(\theta) \tag{3.14}$$

It is assumed that $\vec{\sigma}$ does not depend on θ, so $P(\vec{\sigma} \mid \theta) = P(\vec{\sigma})$.

We then let $P(\vec{\mu} \mid \vec{\sigma}, \theta) \sim N_2(\vec{\mu}, \mathit{\Sigma})$ where

$$N_2(\vec{\mu}, \mathit{\Sigma}) = N\left(\begin{bmatrix} \alpha_1 + \beta_1\theta \\ \alpha_2 + \beta_2\theta \end{bmatrix}, \begin{bmatrix} \gamma_1^2 s_1^2 & \rho_{12}\gamma_1 s_1 \gamma_2 s_2 \\ \rho_{12}\gamma_1 s_1 \gamma_2 s_2 & \gamma_2^2 s_2^2 \end{bmatrix}\right) \tag{3.15}$$

The α, β, and γ terms reflect biases in the means and standard deviations, which would be inferred based on the expert's judgments rather than elicited directly.

If we place an improper flat prior on θ, then the posterior distribution can be calculated as

$$P(\theta \mid \vec{\mu}, \vec{\sigma}) \propto N\left(\frac{\beta^T \mathit{\Sigma}^{-1}}{\beta^T \mathit{\Sigma}^{-1} \beta}(\mu - \alpha), (\beta^T \mathit{\Sigma}^{-1} \beta)^{-1}\right) \tag{3.16}$$

where $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

Winkler (1981) outlines a more general method that uses a multivariate normal distribution to aggregate multiple normally distributed expert priors.

As each expert's prior is a normal distribution, it will have a parameter for the mean μ and variance σ^2. The mean from each individual normal distribution forms the elements, μ_i, of a vector of means for the multivariate normal distribution. The covariance matrix $\mathit{\Sigma}$ is constructed using the experts' variances on the diagonal, and covariances $\sigma_{i,j}$, $i, j = 1, \ldots, n$ and $i \neq j$, for the other elements.

The multivariate normal distribution can then be simplified down to a univariate normal distribution for a known covariance matrix. An improper diffuse prior density is placed on the parameter of interest θ, and μ is normally distributed with a mean of zero for all elements. Note, the means modelled are the difference between the experts' best estimates and the parameter of interest.

The posterior density of θ is then:

$$p(\theta \mid \mu) \propto \phi((\theta - \mu^*)/\sigma^*) \tag{3.17}$$

where ϕ is the standard normal density function. The mean and standard deviation of the combined distribution are then,

$$\mu^* = \frac{1^n \Sigma^{-1} \mu}{1^n \Sigma^{-1} e} \tag{3.18}$$

$$\sigma^* = \frac{1}{1^n \Sigma^{-1} e} \tag{3.19}$$

where $1^n = (1, \ldots, 1)^n$.

The mean of the combined distribution, μ^*, can be expressed as a linear weighting of each experts' mean. The weights in this case are:

$$w_i = \frac{\Sigma_{j=1}^n \alpha_{ij}}{\Sigma_{m=1}^n \Sigma_{j=1}^n \alpha_{mj}} \tag{3.20}$$

where $\alpha_{ij} = \Sigma_{ij}^{-1}$.

This method requires values for the covariances $\sigma_{i,j}$. A simplifying assumption could be made that the experts are uncorrelated, and so set the non-diagonal elements of Σ to zero. Alternatively, the covariances could be obtained using seed questions, such as those in the Classical Method. For a vector of point estimates for the seed questions from expert i, X, and the values from expert j, Y, along with the vector of true values, T,

$$\sigma_{i,j} = \text{cov}(X - T, Y - T) \tag{3.21}$$

It may not always be possible to calculate the covariance between experts, and in these cases a prior can be placed on σ. An inverse Wishart distribution provides conjugacy in the above Normal distribution.

Copula

While the Multivariate Normal method above is simplified for non-correlated experts, copula methods offer an alternative when dealing with correlation between expert judgements (Jouini and Clemen, 1996). Wilson (2017) found there was evidence of between-expert dependence, and that fitting a copula for aggregation provided a more appropriate distribution than an equal weight pooling.

A copula is a joint cumulative distribution, which we can define for multiple experts as

$$C(u_1, \ldots, u_n) = P(F_1(\theta) \leq u_1, \ldots, F_n(\theta) \leq u_n) \tag{3.22}$$

where $F_i(\theta)$ is the cumulative probability distribution corresponding to the elicited prior $f_i(\theta)$, and $u_1, \ldots, u_n \sim Uniform(0, 1)$.

Once the copula is defined, it can be used to aggregate the experts' distributions using the following equation (Wilson and Farrow, 2018).

$$f(\theta \mid D) \propto c(F_1(\theta), \ldots, F_n(\theta)) \prod_{i=1}^{n} f_i(\theta) \tag{3.23}$$

where D is the set of elicited distributions, and c is the probability density function corresponding to C, i.e.

$$c(u_1, \ldots, u_n) = \frac{d}{du_I \ldots du_n} C(u_1, \ldots, u_n) \tag{3.24}$$

The copula induces dependence between the expert priors, while maintaining the correct marginal distributions for each expert.

IDEA

The IDEA protocol aims to bridge the gap between mathematical and behavioural aggregation methods (Hanea et al., 2017). The acronym IDEA stands for Investigate, Discuss, Estimate and Aggregate, the four steps of the method.

Experts begin by providing estimates for each question anonymously. By having experts estimate the quantities prior to their first meeting, it helps to avoid groupthink and invites a wider range of assumptions, reasoning and possible interpretations to the question.

The estimates are shared between the experts, before they meet to discuss the questions. Any assumptions or questions can be discussed and clarified at this meeting, which is led by a facilitator who can direct the conversation if need be. While the estimates can be discussed at this stage, it is aimed to keep anonymity as to who provided which values. This is to avoid any expert having an unintended influence over others due to the experts' perceptions of each other.

The experts then provide another round of estimates. Hemming et al. (2018) found that experts updated their probability intervals for a median of 7 questions out of 9, and their best estimates for a median of three questions out of nine. They also found the majority of those who updated their values improved their accuracy (in the form of an average log-ratio error, or ALRE, score), and updating the best estimate improved the score by around twice as much as updating the probability interval.

The second round of estimates are then aggregated together using a mathematical aggregation method. This avoids the experts having to formulate a consensus themselves. After applying the IDEA protocol, Hanea et al. (2018) found only a marginal improvement by using Classical Method weights over equal weights. The group of experts used, however, had been selected as they had previously been shown to perform well, and so these results may not replicate on less well calibrated experts.

For eliciting probability distributions, the IDEA protocol suggests a set of four questions. First, the experts are asked to provide the lowest and highest plausible values. Then, they make their best estimate in between these bounds. Finally, they are asked how confident they are that the interval has captured the true value. The best estimate is taken as the median, while the lower and upper interval bounds are taken as quantiles corresponding to the experts' confidence in their interval.

3.5 Behavioural Aggregation Methods

Behavioural aggregation methods provide a framework for the group of experts to form an aggregated prior through agreement, rather than a formalised rule. The aim is to create a consensus among the experts about a sensible prior distribution that represents the view of the group.

One benefit of a behavioural aggregation method is it makes clear whose beliefs the prior represents. For the situations elicitation is used in, provided in French (2011) and mentioned in Section 3.2, the experts and decision maker may not be the same individuals. In the case where the decision maker is a separate person, a mathematically aggregated prior could be argued to be the decision maker's prior, based upon the evidence or data which are the expert priors. In the other cases, however, it becomes less clear.

A mathematically weighted average of a group of experts' priors carries no guarantee that any of the experts will agree with the final outcome. The aggregated prior represents a mathematical weighting of the group's views, rather than the group's views themselves. Furthermore, if some or all of the experts disagree with the final aggregated prior, it is hard to argue that it represents all members of the group's beliefs.

Behavioural aggregation methods, however, rely on experts forming their own consensus. The resulting prior is one they all agree to, as best representing the views of the group. There is, then, a clear group who this prior represents, which brings with it a level of transparency and accountability that may otherwise be missing.

It is important to be aware that behavioural aggregation methods do not guarantee that the experts will agree upon a single consensus prior distribution. In some cases, the elicitor may be able to assist in the discussion and move them towards a distribution that includes all perspectives. This could involve suggestions which the experts may not have thought of themselves, such as creating a multi-modal distribution rather than a uni-modal one. Alternatively, the elicitor could encourage the experts to consider the other points of view further, and try to create a balanced prior which includes all viewpoints.

It may be the case, however, that the experts are unable to come to a single consensus among themselves. In this case, multiple aggregated distributions can be formed to represent the views of the different subgroups of experts. For example, in a case where there are two prevalent hypotheses among the experts, they may wish to split up and form two separate aggregated priors.

When multiple priors are formed, there are a number of options as to how they can be used in a model. One option is to use them separately, and run the model multiple times. If the outcome of the model is very similar each time, then the difference between the priors clearly does not have a major practical difference to the final outcome. If there is a difference, then either multiple outcomes will be reached, or the prior may need to be aggregated beforehand to ensure there is a single outcome.

The way such a decision is made would be very circumstance specific. The specific details of the study and the stakeholders involved may influence what type of outcome is appropriate, and what methodologies can be used.

3.5.1 Delphi

The Delphi method was one of the earliest behavioural aggregation methods developed (Dalkey and Helmer, 1963). There are many variations to the Delphi method and its application, but implementations fit a similar pattern. Experts are first asked for their judgments on a number of questions individually. These judgements are then shared anonymously between the group of experts, along with other comments they may have. The experts then make a new set of judgments taking the other experts' views into account. This process of sharing results and updating judgements is repeated a number of times.

The Delphi method is often anonymised between experts, and is usually conducted without experts discussing their views in person (EFSA, 2014).

The Delphi method provides a framework for assisting in the flow of information between experts. The original method does not prescribe a particular set of numbers to be elicited, which has led to variation in what is asked of the experts. Most elicitations will ask for a best estimate, interval and an associated probability, such as in Filyushkina et al. (2018).

It would seem advisable, then, to follow general elicitation advice, such as that outlined in Section 3.2.1.

Unlike other elicitation methods, which focus on extracting quantitative information, the Delphi Method has seen use in qualitative decision making as well (Brady (2015); Meijering and Tobi (2016), for example). These methods can ask the experts different questions in different rounds, such as focusing more on brainstorming possible ideas in early rounds, and analysing and reviewing the ideas in later rounds. Okoli and Pawlowski (2004) provides an example where experts first brainstorm ideas over two rounds by listing all relevant factors. They then narrow down the list by each selecting the most important factors, of which the commonly chosen ones were retained. Finally, the experts rank the factors. These rankings are then shared and the experts are asked to adjust their rankings, with this final step being repeated until a consensus is reached or no further changes are made.

One issue with the Delphi method is the formation of a consensus. Humphrey-Murto and De Wit (2019) identifies both defining a consensus, and building a consensus as under-researched areas surrounding this methodology. In three studies, Landeta (2006) found that 55-65% of experts updated their judgments based on feedback, updating an average of 25% of their values. They also found convergence of opinion between the first and second rounds of elicitations in each case. While this shows it is likely the majority of experts will update their views, it does show that many will not. Unless their judgments overlap already, if the experts do not change them they cannot reach a consensus. In such cases, a mathematical aggregation such as an equal weight linear pool could be used to aggregate between the different responses. More recent behavioural aggregation methods allow the elicitor more opportunities to help the experts' judgements converge and to reach a consensus.

3.5.2 SHELF

The Sheffield Elicitation Framework (SHELF) is a structured method of eliciting probability distributions from a group of experts (Oakley and O'Hagan, 2016). Similarly to the Delphi method, SHELF aims to help experts form a consensus between themselves. Unlike the Delphi method, this is done in an in-person meeting, where the experts try to provide a distribution that incorporates all of their views. The SHELF group meeting contains a number of steps, outlined in Gosling (2018).

To begin, the experts are trained in the elicitation process. This training includes ensuring they have an understanding of probability and statistics, in which they may have little previous training. It is also recommended that elicitors run the experts through a toy problem, in which they can practice the elicitation and specification of probabilities. This training is aimed at combatting the various cognitive biases, and ensuring the experts have a common understanding of the terminology and area of discussion.

The next stage of the elicitation is for the experts to share information about themselves. The experts are asked about potential vested interests, their expertise, and the sources of evidence they are using to base their judgments on. The information here helps alert the facilitator to potential biases, and strengths and weaknesses of the group, while also reminding the experts of their full range of knowledge.

The experts are then asked to provide their individual judgments on each of the quantities of interest. It is recommended that the facilitator ask the experts to provide minimum and maximum values, median, and lower and upper quartiles. The order of these values aims to address potential anchoring bias and overconfidence. Other methods, such as the roulette method, can be used instead (Gore, 1987; Johnson et al., 2010). Though the experts will be in the same meeting, these values are elicited individually and, at this stage, privately.

After each expert has specified their individual judgments, distributions are fit using their values. This is usually done using a least squares algorithm, though the exact distribution types to be considered will depend on the parameter of interest. If experts have provided their opinions in the form of a roulette method histogram, this may just be carried on to the next step.

The individual distributions are then shared with the group, and the experts are encouraged to use them as a basis for creating a group aggregation. The elicitor can use the similarities and differences of the individual distributions to start a conversation, or probe at the experts' judgments. It is noted that there may be difficulties in having experts come to an agreement on a single prior distribution given the potential for a large range in different views and experiences. SHELF recommends experts are instead asked to provide a prior for a rational impartial observer (RIO), an imaginary person who has observed their discussion and all of their evidence. This neutral viewpoint aims to help experts to avoid bias, personal investment and interpersonal difficulties.

Once the experts have formed a consensus, they are then given feedback on the distribution. This can include probabilistic statements from the distribution, or practical interpretations of the parameter. They are given an opportunity to change their distribution if they do not agree with each statement, and the final aggregated prior is the one resulting from the feedback loop.

Oakley and O'Hagan (2016) provides guidance, templates and an R package to assist with the application of SHELF. These documents and templates assist with the management of evidence and definitions for the experts, and the collection of additional information. The R package contains an interactive Shiny application that provides real time feedback of fitted distributions given elicited values. Further applications have been developed for specific studies, for example Truong et al. (2013).

3.6 Implementation of First Elicitation

An elicitation was developed to assist in the design of a trial, the Case Study, which is described in Chapter 1. The case study made a comparison between a reference test (RT), namely the Awaji criteria, currently used in the diagnosis of Motor Neurone Disease (MND) with a novel experimental test (ET), the BIMC test as an addition to the Awaji Criteria. The elicitation aimed to gather information about the effectiveness of both tests in order to perform appropriate sample size calculations. Secondly, it was also designed to allow a comparison between mathematical and behavioural aggregation methods.

The three experts who agreed to take part had been involved in the development and design of the new medical diagnostic test and the clinical trial. As this new test had limited previous data, an elicitation was the best way to gather the necessary information to design the trial. As the experts had experience in developing the new diagnostic test, they were suitable choices to elicit from.

Due to the availability of the three experts to meet at the same time for a meeting of the required length, parts of the elicitation were conducted individually through a pre-elicitation survey. The second part of the elicitation was held in person, in a meeting with all three experts and an elicitation team.

The following sections will outline the development of the elicitation protocols used in the two rounds of elicitations.

3.6.1 Expert Profiles

The three experts had between eight and twenty-five years of experience of research in MND. Table 3.1 outlines the experts' self assessed knowledge and expertise. All of the experts believed they were at least as knowledgeable about BIMC as they were about the Awaji Criteria.

The experts were also asked to consider their strengths and weaknesses in terms of MND and providing information during the elicitation. The experts listed research expertise, background knowledge of the disease and Awaji Criteria, a good understanding of electrophysiology, and clinical experience as their strengths. Their self assessed weaknesses included distance from front-line neurology care, a relative lack of on-going lab based research and a shorter amount of practice as a neurologist and clinical neurophysiologist.

In general, the weaknesses of each expert tended to be covered by the strengths of the other experts. This suggested that although the experts had a similar level of experience with the new diagnostic test, they had more varied backgrounds in the broader research area.

Table 3.1: Experts' Self Assessed Knowledge and Experience

	Expert 1	Expert 2	Expert 3
Please rate your knowledge on the disease from 1 (least) to 5 (most)	4	3.5	4.5
Please rate your knowledge on the Awaji Criteria from 1 (least) to 5 (most).	5	3.5	3
Please rate your knowledge on BIMC from 1 (least) to 5 (most).	5	4	3
Years of experience studying MND	8-10	10	25

3.6.2 Classical Method Seed Questions

The first half of the pre-elicitation tasks sent to the experts contained a number of questions designed to facilitate the use of the Classical Method. Alongside this and the elicitation questions, a document of supplementary material was also provided to the experts.

The questions were provided to the experts in a text document, which they filled in and returned via email.

Table 3.2 outlines the seed questions included in the survey.

Seed Question Number	Seed Question	Answer	Citation
1	For the years 2006 to 2009, what was the incidence rate per 100,000 people of ALS in the Netherlands?	2.77	Huisman et al. (2011)
2	For the years 2006 to 2009, what was the prevalence rate per 100,000 people of ALS in the Netherlands?	10.32	Huisman et al. (2011)
3	For the years 1995 to 1997, what percentage of people with ALS in Ireland were male?	67.5	Traynor et al. (1999)
4	For the years 2002 to 2003, what percentage of people with ALS in Uruguay were male?	60.5	Vázquez et al. (2008)
5	For the years 1995 to 1997, what percentage of new ALS diagnoses in Ireland were male?	57.6	Traynor et al. (1999)
6	For the years 2002 to 2003, what percentage of new ALS diagnoses in Uruguay were male?	66.6	Vázquez et al. (2008)
7	For the years 1985 to 2006, what percentage of ALS diagnoses in New Zealand were familial?	4.1	Byrne et al. (2011)
8	For the years 1989 to 1992, what percentage of ALS diagnoses in Hong Kong were familial?	1.2	Byrne et al. (2011)
9	For the years 1987 to 2009, what was the incidence rate per 100,000 people of ALS in the Faroe Islands?	2.6	Joensen (2012)
10	For the years 1987 to 2009, what was the prevalence rate per 100,000 people of ALS in the Faroe Islands?	8.2	Joensen (2012)

Table 3.2: The full list of seed questions.

The seed questions contain a mix of questions. As the case study involves the diagnosis of MND, of which Amyotrophic Lateral Sclerosis (ALS) is the most common type, the seed questions are all of a similar topic. While domain related questions are preferable, the new diagnostic method did not have enough previous data to use as a basis for the seed questions. The data available was also predominantly gathered by the experts involved in the elicitation, meaning they would likely know the correct answers to many of the questions. Additionally, there were no topics that could be used for prediction by the experts on the new diagnostic method, as they were the group developing and researching it.

Another factor that was considered when developing the seed questions was the availability bias. In order to assist the experts in considering the breadth of their knowledge, and the variability within the trial's patients, the questions were chosen to span a number of different populations. This was achieved by asking about populations from a range of different time periods and locations. Additionally, the questions also ask the experts to consider how differences in sex and family history can affect diagnoses, with the intention that it encourages them to consider demographic factors within the patient group.

Questions 3-8 ask the experts to make judgments in percentages, while questions 1-2 and 9-10 ask about rates per 100,000. The elicitation questions asked about proportions, and so bounded questions involving percentages and rates give the experts a similar style of questions.

The analysis of the responses to these questions is located in Chapter 4.

A number of additional background questions were included, as shown in Table 3.3.

Question Number	Questions
1	What is your background in researching MND?
2	How long have you been involved in MND research?
3	What sources of information is your knowledge of MND based on?
4	What are your strengths and weaknesses regarding this topic?
5	Please list any sources of quantitative information about BIMC tests you are aware of. This information will be shared with other participants
6	Please rate your knowledge on Motor Neurone Disease from 1 (least) to 5 (most).
7	Please rate your knowledge on the Awaji Criteria from 1 (least) to 5 (most).
8	Please rate your knowledge on the Beta-band intermuscular coherence tests from 1 (least) to 5 (most).
9	What is the smallest percentage increase in correct positive diagnoses from using BIMC you would need to see to implement it in diagnoses?

Table 3.3: The full list of background questions.

The additional questions 1-5 were included to encourage the experts to further consider their total range of knowledge, in order to help combat the availability bias. By directly asking about the experts' backgrounds, it encourages them to consider their past history in MND research and practice.

Questions 6-8 ask the experts to rate their own knowledge of the topics surrounding the trial. The responses to these questions can be used as a basis for creating weights using a self-rating method, as outlined in Section 3.4.3. It was also intended that this question might help the experts reflect further on their own expertise, and help them adjust any overconfidence.

The results for these questions are presented in Section 3.6.1.

Question 9 was used to determine values suitable for the Minimal Clinically Important Difference, which is discussed further in Chapter 2.

Additional and Supplementary Materials

The supplementary material document provided definitions of all statistical and medical terminology, to ensure the experts would have the same interpretation of the questions. The full document is provided in Appendix A.1, and a summary is presented here. Table 3.4 provided initial definitions and acronyms which were used later in the document.

	Definitions
MND	Motor Neurone Disease, which involves the progressive degeneration of motor neurones in the cerebral cortex, brainstem and spinal cord.
AC	Awaji Criteria
Positive Awaji Diagnosis	A diagnosis using the Awaji Criteria leading to a patient being assigned treatment for MND.
BIMC	Beta-band intermuscular coherence test
Positive BIMC Diagnosis	A diagnosis using the BIMC test and Awaji Criteria leading to a patient being assigned treatment for MND.
Median	The value where an outcome is equally likely to occur above or below.
Best Estimate	The median.
Lower 25% Quartile	Assuming the outcome will occur below the median, this quartile is the value where an outcome is equally likely to occur above or below.
Upper 25% Quartile	Assuming the outcome will occur above the median, this quartile is the value where an outcome is equally likely to occur above or below.

Table 3.4: Supplementary Material Definitions

Following this, the documents provided a brief reminder of the design of the trial. Then, the Awaji criteria were provided, alongside further information about BIMC from the trial designers' previous works. This information would have been seen before by the experts in this elicitation, but it was included as a reference and to remind them of the previous results they had gathered.

The experts had also provided the data from their previous work, which included measurements from two pairs of muscles in each of the legs and arms of patients known to either have, or not have, MND. From this, a number of plots were created to provide further information. The data was first plotted as histograms and boxplots, an example

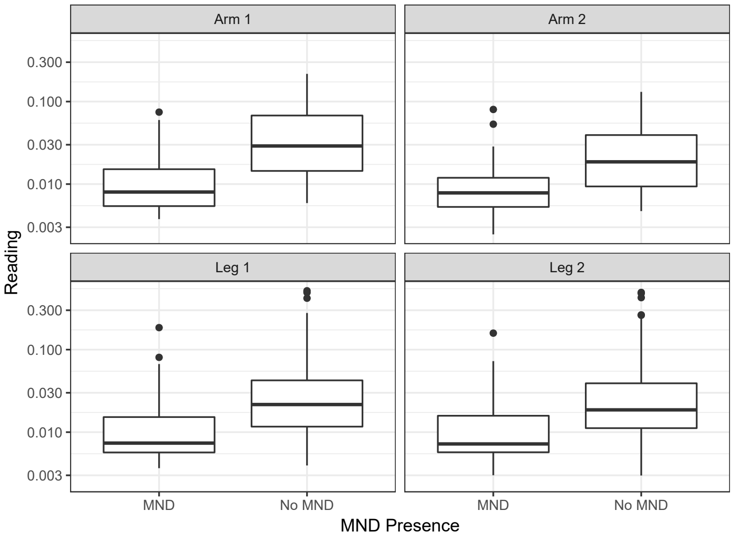


Figure 3.4: Example Supplementary Material Boxplot

of which is provided in Figure 3.4.

While the experts had likely seen the data plotted this way before, it may have been some time since they had seen it. The boxplots and histograms were included to allow the experts to refamiliarise themselves with the data. As the measurements collected had a strong positive skew, each plot was provided on both an untransformed and logarithmic scaled axis.

The readings in each limb were also plotted against each other, such as in Figure 3.5. This was done to help prompt the experts to further consider the diagnostic abilities of their new method. While the histograms and boxplots showed that the mean recorded measurements from the two groups of patients were different, the scatter plots gave a more nuanced view of how patients could be categorised using the measurements. These plots were only presented with logarithmic transformed axes to ensure all points were readable.

The final set of plots created for the experts were receiver operating characteristic curve, or ROC curve, plots. Figure 3.6 provides one example. ROC curves plot the relationship between the sensitivity and specificity of a classifying test as the cutoff for classification changes. A diagnostic test can have a sensitivity of one if it gives a positive result for all measurements, but will have a specificity of zero as it will fail to identify anyone who does not have the disease. As the cutoff is changed, the test decreases its sensitivity as it begins to misclassify positive patients, but improves its specificity by correctly identifying negative patients.

A final summary table was also presented alongside the ROC curves, Table 3.5. The area under the ROC curve is commonly used as a measure of how well a test performs. A perfect test will have an area of 1, and a completely uninformative test (which appears

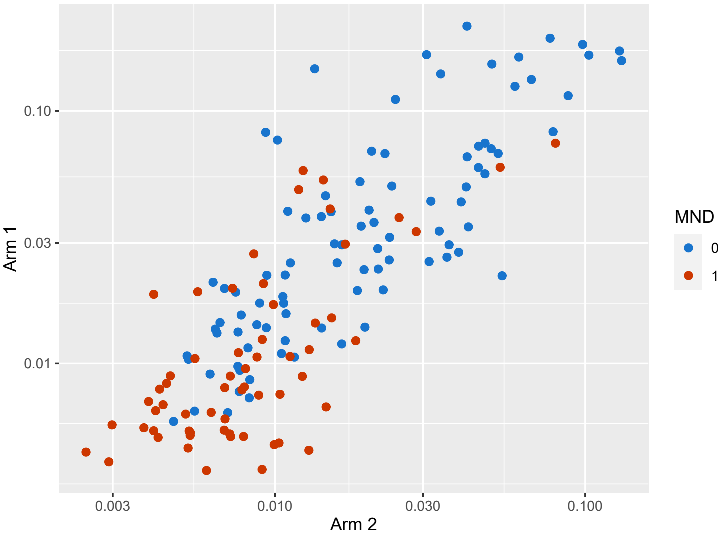


Figure 3.5: Example Supplementary Material Scatter Plot

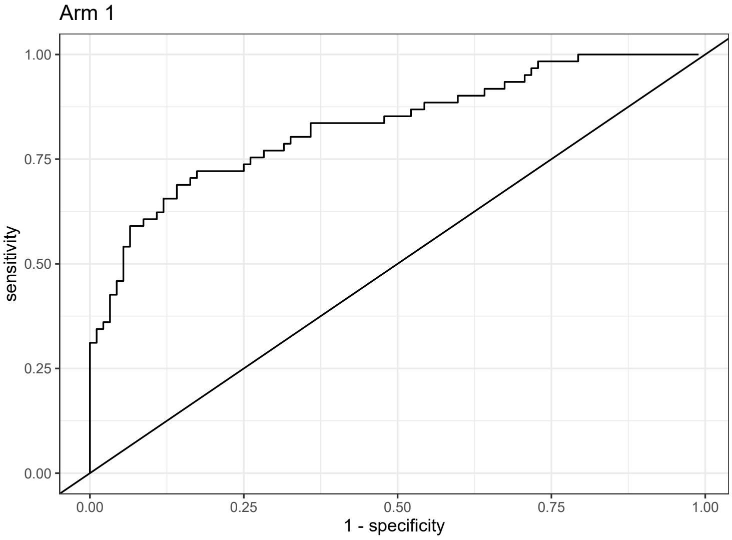


Figure 3.6: Example Supplementary Material ROC curve

as a diagonal line from (0,0) to (1,1)), will have an area of 0.5. The table also provides estimates for the PPV, the positive predictive value, and NPV, the negative predictive value. The PPV and NPV values are calculated as follows.

$$PPV = \frac{TP}{TP + FP} \tag{3.25}$$

$$NPV = \frac{TN}{TN + FP} \tag{3.26}$$

where TP refers to the number of true positive results, FP refers to the number of false positive results, TN refers to the number of true negative results, and FN refers to the number of false negative results.

BIMC Comparison	Area under curve	PPV	NPV
Leg 1	0.782	0.931	0.315
Leg 2	0.759	0.931	0.239
Arm 1	0.820	0.902	0.391
Arm 2	0.775	0.967	0.250

Table 3.5: Supplementary Material Summary Table

The experts were all familiar with ROC curves and the quantities in the summary tables, and had experience using them. As the ROC curves were their natural way of investigating and analysing a diagnostic method, this provided them with a summary of their previous work in a way they would be comfortable using. It also meant that they were thinking of the new diagnostic test in terms they may be familiar with when considering other tests in.

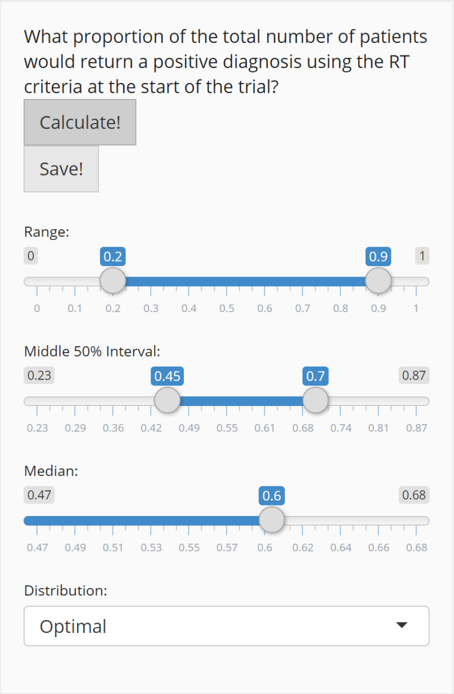
3.6.3 SHELF Shiny Application

SHELF typically conducts the individual and group components of the elicitation in person. However, due to time constraints and expert availability, the SHELF protocols needed to be modified to fit a shorter meeting. The individual component of the elicitation was instead held prior to the group elicitation, in an online setting. After the experts completed the initial seed questions, they then proceeded to an interactive online application for the elicitation questions.

The online application was developed using the R Shiny package Chang et al. (2015).

The initial page required the expert to enter a name, so that if they did not complete the full elicitation they could return at a later time and retrieve their previous answers. The following pages asked about the parameters of the model. For each, the experts were asked to enter upper and lower limits, quartiles and a median. Figure 3.7 shows how the experts inputted these values.

The experts could input values using the sliders. The range available for the quartiles was constrained by the inputted minimum and maximum, and the range for the median was constrained by the quartile values. In order to ensure a distribution could be fit, the available ranges were slightly narrower than the answers inputted in the preceding row. This was done to ensure that experts could only enter mathematically consistent values, for example such that the quartiles and median were within the interval bounded by the minimum and maximum. These values were used to fit a distribution that approximated the expert's views.



What proportion of the total number of patients would return a positive diagnosis using the RT criteria at the start of the trial?

Range: 0.2 0.9

Middle 50% Interval: 0.45 0.7

Median: 0.6

Distribution: Optimal

Figure 3.7: R Shiny application input box.

Once the values were inputted, clicking on the ‘Calculate’ button would fit a distribution to the values. A number of different distributions were considered, as outlined later in this section, and each were fit using a least squares method. The parameters of the distributions, ϕ_1 and ϕ_2, were estimated by calculating S, using

$$S = \sum_{i=0}^{4} (q_i - \theta_i)^2 \tag{3.27}$$

where q_i refers to the ith quartile provided by the expert, and θ_i is the ith quartile from a distribution with parameters ϕ_1 and ϕ_2. An optimisation algorithm minimised the value of S by searching over the possible values of ϕ_1 and ϕ_2. The use of minimum and maximum values are discussed later in this section.

The fitted distribution with the lowest S was then displayed to the expert as the best fit for their answers. Additionally, the values of the provided quantiles and those from the fitted model were overlaid on a plot of the distribution to give the expert a guide as to how well the distribution was fitting to their beliefs, as demonstrated in Figure 3.8. A table was provided beneath the plot, outlining the summary statistics of the fitted distribution as further feedback.

If the expert felt this distribution corresponded with their beliefs, they could save it and continue to the next page. Otherwise, they could either select a specific distribution from the dropdown list available, shown below, or change their answers to the questions and refit new a distribution.

The distributions available for the experts to fit included the following.

Optimal The distribution with the minimal S value, so the experts could easily return to it. This option gave the same results as manually selecting the best fitting distribution below.

For the following Beta distributions, the values of ϕ_1 and ϕ_2 are denoted as α and β respectively.

Beta The Beta distribution, with the inputted minimum and maximum taken to be the 1% and 99% quantiles. The beta distribution has a probability density function

$$f(x) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1} \tag{3.28}$$

Beta Rescaled A Beta distribution, rescaled to lie within the minimum and maximum provided. The original Beta variable x is rescaled using

$$y = (u-l)x + l \tag{3.29}$$

where y is the rescaled beta variable, l is the lower bound of the new range and u is the upper bound of the new range.

The probability density function then takes the form

$$f(y) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \frac{(y-l)^{\alpha-1}(u-y)^{\beta-1}}{(u-l)^{\alpha+\beta-1}} \tag{3.30}$$

Beta Truncated A Beta distribution, truncated at the minimum and maximum provided. The probability density function takes the form

$$f(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{F(b)-F(a)}, & \text{for } a \leq x \leq b, \\ 0, & \text{for } b < x \end{cases} \tag{3.31}$$

where a is the minimum, b is the maximum, and $F(b) - F(a)$ is the cumulative area under the distribution between a and b. If $a = 0$ or $b = 1$, then the truncation would only occur on one side of the distribution, otherwise if $a > 0$ and $b < 1$ then the truncation would be on both sides of the expert's distribution.

The truncated Beta distribution was fit in R using the truncdist package (Novomestky and Nadarajah, 2016)

For the following distributions, the values of ϕ_1 and ϕ_2 are as denoted as μ and σ respectively.

Normal Truncated A Normal distribution, truncated at the minimum and maximum provided. The probability density function takes the form

$$f(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{1}{F(b)-F(a)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, & \text{for } a \leq x \leq b, \\ 0, & \text{for } b < x \end{cases} \tag{3.32}$$

where a is the minimum, b is the maximum, and $F(b) - F(a)$ is the cumulative area under the distribution between a and b. As the prior being elicited is bound between 0 and 1, $a \geq 0$ and $b \leq 1$. Note that due to the truncation, the μ and σ parameters will no longer represent the mean and standard deviation of the distribution.

The truncated Normal distribution was fit in R using the truncdist package (Novomestky and Nadarajah, 2016).

Log Normal A Log Normal distribution, with the inputted minimum and maximum taken to be the 1% and 99% quantiles. This distribution provided a good fit for when the experts gave positively skewed answers.

The probability density function takes the form

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp(-\frac{(ln(x) - \mu)^2}{2\sigma^2}) \tag{3.33}$$

where μ and σ are the mean and standard deviation of the original Normal distribution.

Logit Normal A logit Normal distribution, with the inputted minimum and maximum taken to be the 1% and 99% quantiles. This distribution provided a good fit for when the experts gave negatively skewed answers.

The Logit Normal distribution was fit in R using the logitnorm package (Wutzler, 2018).

The probability density function takes the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x(1-x)} exp(-\frac{(logit(x) - \mu)^2}{2\sigma^2}) \tag{3.34}$$

where μ and σ are the mean and standard deviation of the inverse of the logit of x.

Uniform The uniform distribution would only take the values of the minimum and maximum into account, and fit a uniform distribution between them.

The probability density function takes the form

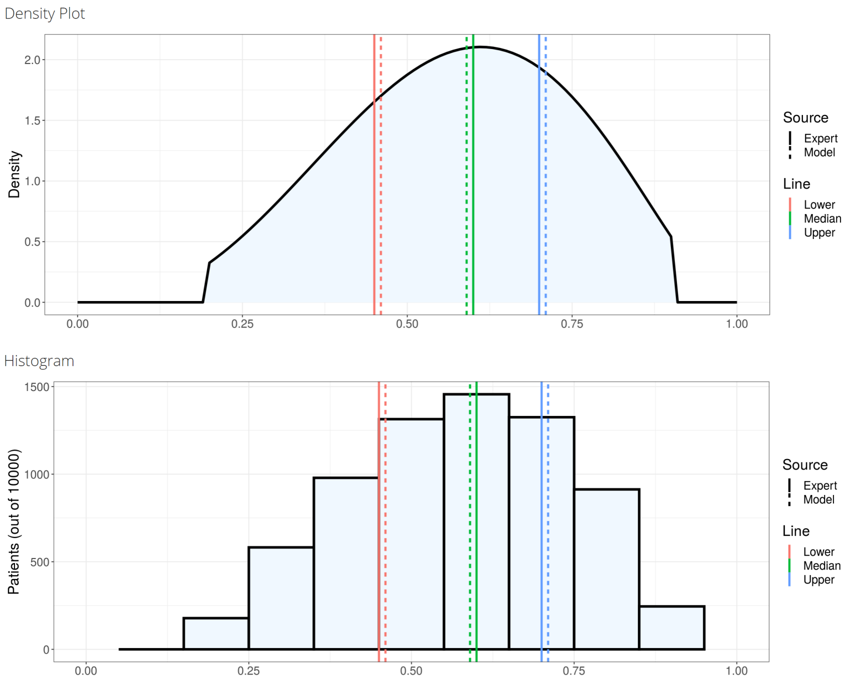


Figure 3.8: R Shiny distribution plots.

$$f(x) = \begin{cases} 0, \text{for } x < a, \\ \frac{1}{b-a}, \text{for } a \leq x \leq b, \\ 0, \text{for } b < x \end{cases} \tag{3.35}$$

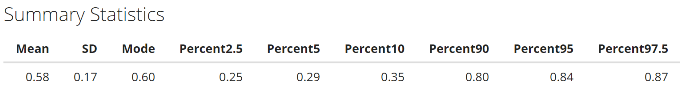
with a minimum bound of a and a maximum bound of b.

While this option tended to perform very poorly when calculating a value of S for the fit, it was included to allow the experts to fit a uniform distribution if they believed it best represented their knowledge.

Once a fitted distribution was selected, either by the algorithm or the user, it was displayed to the experts in a density plot and a histogram. Figure 3.8 shows an example distribution as provided to the experts.

A number of coloured lines were overlaid to help guide the experts as to how well the distribution fitted the values they entered. The solid lines represent the expert's values, and the dashed lines represent the corresponding values from the fitted distribution. The lower quartile is given in red, the median in green and the upper quartile in blue.

In order to provide further feedback, a summary statistics table was included, as shown in Figure 3.9. This table included the mean, standard deviation, mode and the 80%, 90% and 95% probability intervals of the fitted distribution. These values were included to provide a wider range of feedback, and to ensure the experts had access to whichever set



Summary Statistics

Mean	SD	Mode	Percent2.5	Percent5	Percent10	Percent90	Percent95	Percent97.5
0.58	0.17	0.60	0.25	0.29	0.35	0.80	0.84	0.87

Figure 3.9: R Shiny distribution summary.

of summary statistics they were most used to working with.

If the experts felt that the fitted distributions were consistent with their opinions, then they could continue to the next page. If they were not happy with the fitted distribution, they could continue to change the inputted values or manually select a distribution.

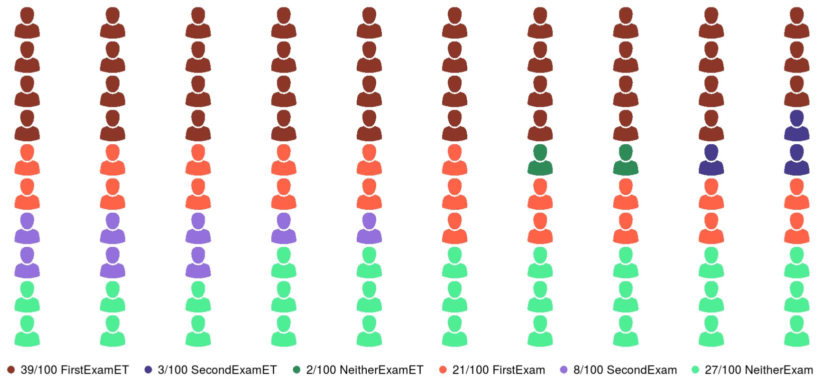
After the experts had completed the five elicitation parameter pages, they moved on to a pair of checks. These checks were included as a way for the experts to confirm they approved of the extrapolated interpretations.

The first check calculated an estimate and probability interval for the total proportion of patients with a positive BIMC result at the start of the trial and a positive Awaji Criteria result after six months, or $(1-\eta)\psi$. As this value was not directly elicited, but rather extrapolated from the experts' other distributions, it provided a check to ensure that the elicited values make sense in a new context. If the experts felt this value was incorrect, a number of options were provided as to which elicited values would need to be modified to raise or lower it. An example of the information provided to the experts is as follows.

> Given the values for previous parameters, it suggests that the expected proportion of positive ET diagnoses is 0.45 with a 90% probability interval of (0.24 , 0.68). If these values do not represent your beliefs on the ET, you may adjust them in the following ways.
>
> To decrease the proportion of positive ET results, you could increase RT, decrease RT Round 2, decrease ET/RT, decrease ET/RT Round 2, or decrease ET/Negative RT.
>
> To increase the proportion of positive ET results, you could decrease RT, increase RT Round 2, increase ET/RT, increase ET/RT Round 2, or increase ET/Negative RT.

The second check used the median estimates from each parameter to build an example sample of 100 patients. This represented the 'most plausible' trial outcome for the expert's values. An example is provided in Figure 3.10.

The shade of each patient icon represented their BIMC results, and the colour represented their Awaji Criteria results. A table summarised these and further proportions. If these summaries did not represent an expert's beliefs about the results of the trial, then it would identify which parameters needed to be modified further.



RT	0.60
RT Round 2	0.25
ET/RT	0.65
ET/RT Round 2	0.25
ET/Negative RT	0.09
ET out of total	0.44
RT Round 2 out of total	0.10

Figure 3.10: Elicitation check, providing an example sample of 100 patients using an expert's medians. Positive BIMC results are coloured in a darker shade, negative BIMC results are in a lighter shade. Patients with a positive Awaji Criteria in the first round are represented by red. Patients with a positive Awaji Criteria in the second round are represented by purple. Patients with a negative Awaji Criteria from both rounds are represented by green. Further scenario specific details were provided to the experts.

By displaying the information in a new format, both as a new type of plot and in terms of a trial with 100 patients, the experts are forced to consider their answers in a different way. If this plot does not seem reasonable to them, it suggests the values they provided are not consistent with their actual judgments, and that they should change their inputted values.

In both the first and second checks, the abbreviations RT and ET were used in place of the Awaji criteria and BIMC test respectively. This served two purposes. Firstly, it was used to match specific tabs in the application to allow the experts to more easily make changes if required. Secondly, it acted as a reinforcement to ensure the experts were considering which test was currently the standard method of diagnosis.

SHELF Meeting

The elicitation for the trial was held on the 21st of March, 2019. The meeting contained three experts, and three facilitators. In order to run the elicitation meeting, the facilitators took on different roles.

The first facilitator, myself, acting as the elicitor, guided the discussion between the experts. Their responsibility was to lead the meeting, assist the experts in coming to a consensus view, and ensure all questions were completed within the available time. A second facilitator took notes during the meeting. This served two purposes. The first was as a backup record of the elicited values, in case there were any unforeseen issues with the software. Secondly, they recorded the reasoning behind the experts' judgements, and how they came to their final decisions. The third facilitator recorded the experts' responses in the R shiny application, navigated between the pages as the experts wished, and provided live feedback based on the elicited values.

While fewer facilitators could have run the meeting, by spreading the responsibilities among the three it allowed for the lead facilitator to focus on guiding the discussion with the experts, while still ensuring as much information was recorded as possible.

The elicitation meeting began with introductions, as not all of the experts had met the facilitators in person. After this, a review of BIMC was covered, including prior data provided by the experts. This allowed time for the experts to review their knowledge and discuss any new ideas they had. Additionally, information about the SHELF procedure, including the idea of the rational impartial observer, was also covered.

The first half of the elicitation discussed the seed questions. This was presented as a warm up for the main elicitation, and as practice for all parties involved in the elicitation process. For each, a slide with the anonymous responses from each expert was displayed, along with printed hardcopies. The experts discussed each question, and came to consensuses between themselves. They also ensured to encourage all experts to give their perspectives and to incorporate each persons' views. The results they provided were entered into tables presented on a projector.

The second half of the elicitation covered the elicitation questions. These were presented to the experts in a similar Shiny application to the one they had previously used. The updated shiny application included the three experts' anonymised distributions before each question was asked. At this step, the group was asked to look at the individual answers and think about how this might affect their views. The group also completed the same checks as they had individually, which led to a modification of a previous response.

During both parts of the elicitation, the elicitor helped to guide and focus the discussion using specific questions. For each question, the elicitor would ask about specific values, such as how sure the experts were that no values would lie above their maximum or below their minimum. Then, once a distribution was presented, they would ask whether the experts were happy with the fit, the width, and the shape of the distribution. Once the questions were completed, they explained the checks to the experts and investigated whether the experts were happy with the final conclusions.

The elicitor also compared the group responses to the individual responses, to encourage the experts to do the same. They asked the group whether any differences between the group values and the individual values were important, or should be considered. In some cases, specifically in the seed questions, the experts gave a group answer which was different to any of their individual responses. The elicitor made sure to question why they had changed their minds, and to ensure that their decision had a justification the group felt was satisfactory.

3.7 Implementation of Second Elicitation

After the first elicitation was complete, a second round was planned with a new group of experts. The aim was to conduct the second elicitation with an identical format to the first to allow a direct comparison between them. It also aimed to recruit experts who had not been directly involved in the development of the novel diagnostic test, in order to see if there were systematic differences between the beliefs of the two groups. However, the spread of coronavirus meant that in-person meetings were impossible to conduct. On top of this, as the experts required for the elicitation were in medical fields, and often employees of the NHS, it became increasingly difficult to organise meetings for groups of experts.

As such, the second round of elicitations was modified. In order to elicit from experts who were not available at the same time as others, a modified Delphi approach was used, instead of SHELF, as the behavioural aggregation method. Furthermore, all parts of the elicitation were moved to an online setting, where the experts could complete them in their own time.

While this change meant the possible comparisons would be different, it allowed for further focus on conducting online elicitation. Many elicitation methods focus on in-person meetings, which even outside pandemic settings can still add difficulties. Online elicitations allow for a wider range of experts to be reached, and potentially less time constraints due to the need for a common meeting time.

The experts for this elicitation were sourced through mailing lists provided by the experts involved in the first elicitation. A total of seven experts responded, with three providing responses to the seed questions alone, and the remaining four responding to both seed and parameter questions.

This section covers the changes and implementation of the second elicitation.

3.7.1 Expert Profiles

Background information about the second group of experts is provided in Table 3.6. This group was made up predominantly of consulting clinicians, all of which had expertise in the diagnosis of MND in a clinical setting. The additional seven experts had a wide range of experience in researching MND, with between zero and fifteen years of experience.

Table 3.6: Second Group Experts' Self Assessed Knowledge and Experience

Expert	1	2	3	4	5	6	7
Please rate your knowledge on the disease from 1 (least) to 5 (most)	3	5	4	4	5	3	4
Please rate your knowledge on the Awaji Criteria from 1 (least) to 5 (most).	2	4	4	1	5	3	4
Please rate your knowledge on BIMC from 1 (least) to 5 (most).	2	1	2	1	1	1	2
Years of experience studying MND	5	-	7	15	5	10	0

There was also more variation within this group of experts' self assessments. While their knowledge about MND was assessed to be similar to the first group of experts, those in this group had a wider range of expertise for the Awaji Criteria. In general, they also believed they were more knowledgeable about the Awaji Criteria than BIMC, with only three experts rating their knowledge of BIMC higher than 1.

Most experts identified their strengths in this area as being related to MND diagnosis in a clinically setting, and experience using the Awaji Criteria in practice. The commonly self identified weakness of the experts was their purely theoretical understanding of BIMC, and subsequent lack of practical experience of its use. Two experts also identified themselves as lacking a current involvement in research.

Overall, it appears that the second elicitation group had a stronger focus on clinical and MND diagnosis experience than the first group. While the first group had a stronger knowledge of BIMC, the second group of experts are still able to provide valuable information about MND diagnosis and the Awaji Criteria's effectiveness.

3.7.2 Classical Method Seed Question

For the second set of elicitations, the classical method survey was updated.

In order to make direct comparisons between the two, the questions in the second round were chosen from those in the first. However, in order to decrease the length of the survey, a reduced number of seed questions were included. From the results of the first elicitation, it was noted that for certain pairs of questions each expert answered the same for both. In these cases, one of each pair of questions was removed. This resulted in a total of seven questions included in the survey.

The seven questions were also presented in a different order, as shown in Table 3.7. Additionally, the experts were provided further guidance when they completed Question 2. This was to ensure they had a good understanding of what the statistical terms meant, and how the values being elicited should best be provided.

Each question was presented on its own page of the survey, with an explanation of each of the five required inputs: minimum, maximum, median, lower quartile and upper quartile. The experts were guided through these in the listed order, and given instructions

Original Seed Question Number	Seed Question	Answer	Citation
2	For the years 2006 to 2009, what was the prevalence rate per 100,000 people of ALS in the Netherlands?	10.32	Huisman et al. (2011)
3	For the years 1995 to 1997, what percentage of people with ALS in Ireland were male?	67.5	Traynor et al. (1999)
9	For the years 1987 to 2009, what was the incidence rate per 100,000 people of ALS in the Faroe Islands?	2.6	Joensen (2012)
1	For the years 2006 to 2009, what was the incidence rate per 100,000 people of ALS in the Netherlands?	2.77	Huisman et al. (2011)
6	For the years 2002 to 2003, what percentage of new ALS diagnoses in Uruguay were male?	66.6	Vázquez et al. (2008)
8	For the years 1989 to 1992, what percentage of ALS diagnoses in Hong Kong were familial?	1.2	Byrne et al. (2011)
10	For the years 1987 to 2009, what was the prevalence rate per 100,000 people of ALS in the Faroe Islands?	8.2	Joensen (2012)

Table 3.7: The list of seed questions used in the second elicitation.

as to how to make their judgements.

This increased level of detail was designed to assist experts who were less familiar with the terminology to complete the survey. As there was limited contact between the elicitors and experts, it was important that the experts could complete the survey on their own. As such, this second elicitation was designed to provide more support and guidance throughout.

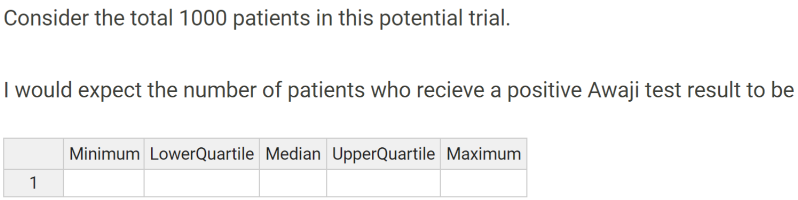
While the wording of the seed questions remained the same, an update in the wording of the question asking about the minimal clinically important difference was made. The updated wording is as follows.

> Consider 1000 patients with motor neurone disease (MND), who are to be diagnosed using the Awaji Criteria. How many additional positive diagnoses would BIMC have to offer in order for you to use it within the Awaji Criteria in the diagnostic procedure?

During the first elicitation survey, two of the experts did not respond to the MCID question. The wording was changed in an attempt to clarify the question, and to ask it on the same scale as the later Delphi Method questions.

3.7.3 Delphi Shiny Application

As the in-person SHELF meetings could not be conducted and it became increasingly difficult to arrange meetings between groups of experts, a Delphi approach was used as

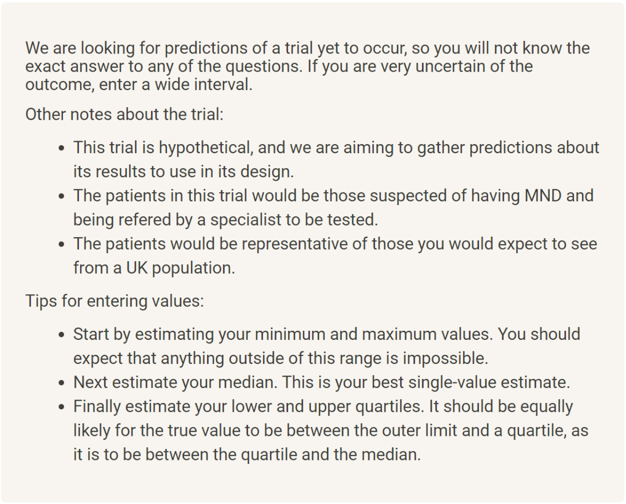


Consider the total 1000 patients in this potential trial.

I would expect the number of patients who recieve a positive Awaji test result to be

	Minimum	LowerQuartile	Median	UpperQuartile	Maximum
1					

Figure 3.11: R Shiny input table



We are looking for predictions of a trial yet to occur, so you will not know the exact answer to any of the questions. If you are very uncertain of the outcome, enter a wide interval.

Other notes about the trial:

- This trial is hypothetical, and we are aiming to gather predictions about its results to use in its design.
- The patients in this trial would be those suspected of having MND and being refered by a specialist to be tested.
- The patients would be representative of those you would expect to see from a UK population.

Tips for entering values:

- Start by estimating your minimum and maximum values. You should expect that anything outside of this range is impossible.
- Next estimate your median. This is your best single-value estimate.
- Finally estimate your lower and upper quartiles. It should be equally likely for the true value to be between the outer limit and a quartile, as it is to be between the quartile and the median.

Figure 3.12: R Shiny help

a behavioural aggregation technique for the further rounds of elicitation. The Delphi method allowed for the experts to all complete their elicitations in their own time, while also allowing for a behavioural aggregation to be made.

An online R shiny application was developed to elicit values from the experts. As this elicitation would include no direct contact it was simplified compared to the first round of elicitations.

As shown in Figure 3.11, the experts inputted their values in tables. Distributions were not fitted in this application, and would instead be fitted at a later date.

If the value entered in the median box was outside the possible range, either below zero or above the value in the question, then the application would not allow the expert to continue. Otherwise, once they were satisfied with their responses they could move to the next model parameter.

The first question started with a total of 1000 patients, and each of the following questions used the medians from previous responses to change the value given to build up a potential trial. A value of 1000 was chosen as the starting point to allow for more precision than a 100 patient trial, while still being low enough to be reasonably considered by the experts.

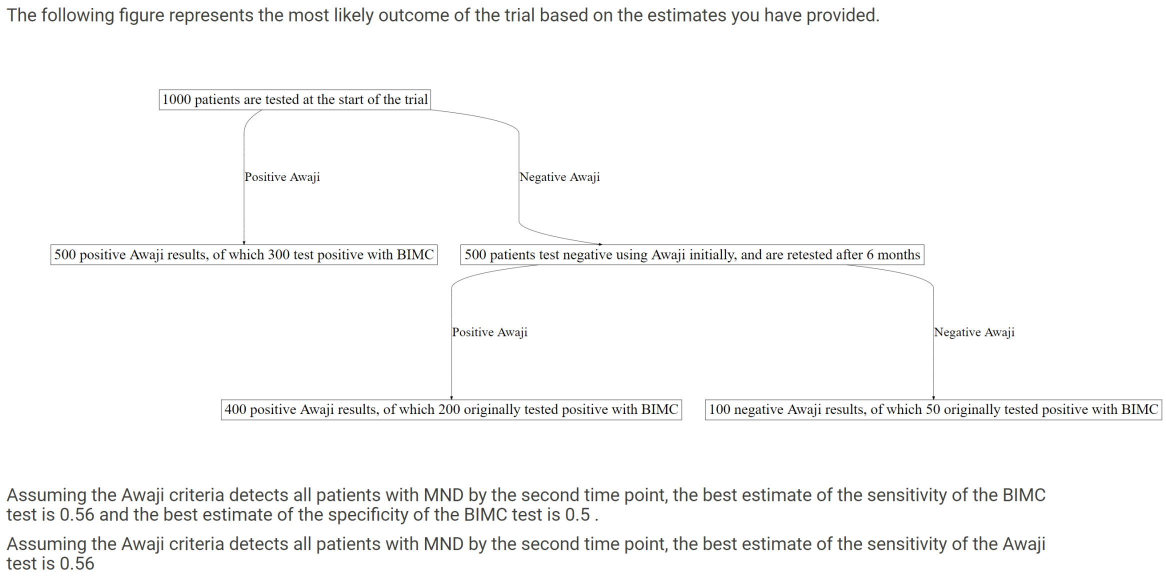
Each page also provided a summary box, reminding experts how to fill in the table and some important information about the trial. This is provided in Figure 3.12. This additional help was included in the Delphi application to assist the experts to complete the elicitation on their own. As they would receive no in-person contact, and more limited online support, this additional information was provided to ensure they had more assistance.

Additionally, an information page was included in the application between the questions about the Awaji Criteria and BIMC. This page repeated information found in the supplementary materials, to remind the experts about BIMC. As the experts using this application had not previously worked with BIMC, this acted as a reminder of the details and data currently available about it. The experts could return to this page or the supplementary material at any time.

A different check was provided in this application for the experts once they had completed the parameter questions. Figure 3.13 shows the flow chart summary provided to the experts. This flowchart represents how the trial would proceed given their estimates.

Additionally, the experts were provided estimates of the sensitivity and specificity of the two tests. They may have knowledge about the sensitivity and specificity of the Awaji Criteria, and so can use this to check their previous responses. While they may not be aware of these values for BIMC, they can still check the estimates provided to see whether they are sensible and in line with their general experiences.

After the elicitation was complete, the results were used to fit appropriate distributions. Details about the elicited distributions are presented in Chapter 4.



The following figure represents the most likely outcome of the trial based on the estimates you have provided.

1000 patients are tested at the start of the trial

500 positive Awaji results, of which 300 test positive with BIMC

500 patients test negative using Awaji initially, and are retested after 6 months

400 positive Awaji results, of which 200 originally tested positive with BIMC

100 negative Awaji results, of which 50 originally tested positive with BIMC

Assuming the Awaji criteria detects all patients with MND by the second time point, the best estimate of the sensitivity of the BIMC test is 0.56 and the best estimate of the specificity of the BIMC test is 0.5 .

Assuming the Awaji criteria detects all patients with MND by the second time point, the best estimate of the sensitivity of the Awaji test is 0.56

Figure 3.13: R Shiny summary flowchart of trial outcome, based on an expert's best estimates.

3.8 Conclusions

In this chapter we have reviewed the elicitation aggregation literature, and outlined the more common methods. We have also seen how cognitive biases shape the elicitation process and how we should ask for experts' judgments. We have detailed the development of two elicitations, the results of which are presented in a later chapter.

The information collected from the elicitations outlined in this chapter forms the basis for the comparison between different aggregation methods, and the application of assurance techniques, in Chapters 4 and 6.

Part IV

Investigation and Comparisons of Elicitation Techniques

Chapter 4

Investigation and Comparisons of Elicitation Techniques

4.1 Introduction

Throughout this chapter, we will consider three groups of experts. The first, containing experts labelled 1 to 7, are experts with no direct involvement with the study. Experts 8 to 10 make up the second group, who are experts involved with designing the study. In addition, we also consider a third group, in which all experts are combined.

As detailed in Chapter 3, the two original groups of experts were asked a different number of seed questions. The first group were asked seven seed questions, and the second group ten seed questions. When considering the two groups separately, all seed questions asked are used. However, when the two groups are combined, we use the seven questions in common among all experts.

4.2 Individual Elicitation Results

In order to obtain prior distributions for the BIMC study, two rounds of expert elicitations were held. The details of these elicitations are provided in Chapters 1 and 3.

The final set of results included responses from ten experts for the seed questions. Three of these experts were asked the full list of ten seed questions, while the remaining three responded to a subset of seven questions due to time constraints. Seven experts also completed the parameter questions. The three experts who answered all ten seed questions provided responses to all parameter questions, while of the other seven, two provided responses to the first two parameter questions and a further two responded to all parameter questions.

For many of the following aggregation and scoring methods, a density function is required for each elicited distribution. In order to model the experts' prior beliefs, split-normal distributions were fitted based on their elicited values. A split-normal distribution is made up of two half-normal distributions to be fit, with different variances on either side of the mode. This allows any asymmetry in the experts' elicited values to be captured by the distribution, while still ensuring the experts' best estimates were directly included.

The probability density function for the split-normal distribution is given by

$$p(x \mid \mu, \sigma_1^2, \sigma_2^2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right) & \text{if } x \leq \mu \\ \frac{1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x-\mu)^2}{2\sigma_2^2}\right) & \text{if } x > \mu \end{cases} \tag{4.1}$$

where μ is the mode of the distribution, σ_1 is the lower tail's standard deviation, and σ_2 is the upper tail's standard deviation. It is important to note that while μ represents the median and mode of the distribution, the mean of this distribution is given by $\mu + \sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1)$. Furthermore, if the expert's distribution has no skew, such that $\sigma_1 = \sigma_2$, this distribution will be a normal distribution with a single variance.

To fit a split-normal distribution, the value of μ is taken to be the expert's best estimate, or median. The standard deviations for each side of the distribution are calculated as $\sigma_1 = \frac{q_{25}-q_{50}}{\Phi(0.25)}$ and $\sigma_2 = \frac{q_{75}-q_{50}}{\Phi(0.75)}$. Note, some minor truncation may also occur to ensure the resulting distributions stay within the bounds of the relevant question, in particular ensuring $x \geq 0$ for the seed questions, and $0 \leq x \leq 1$ for the parameter questions. Furthermore, in the case of the Bayesian aggregation, the aggregation occurs on the logit scale, ensuring this condition is satisfied.

The values presented within this section are those elicited before a distribution is fitted, and as such represent the exact values provided by the experts.

4.2.1 Seed Questions

For each seed question, a minimum, lower quartile, median, upper quartile, and maximum value were elicited from each expert. These have been plotted as boxplots and provided in Figure 4.1. The answer for each seed question has been identified with a horizontal line, and those boxplots that include this value within their range are coloured blue. Figure 4.2 provides the same data, with the boxplots containing the answer within their middle 50% interval coloured blue.

The majority of elicited distributions are overlapping, showing the experts had similar beliefs about the values of the quantities of interest. There were some exceptions, such as Experts 5 and 8 for Question 1. The responses which vary largely from the other experts tend to be from Experts 3, 5, and 8. While in many cases these outliers move further away from the true value, the response from Expert 3 in Question 8 is the closest to the true value.

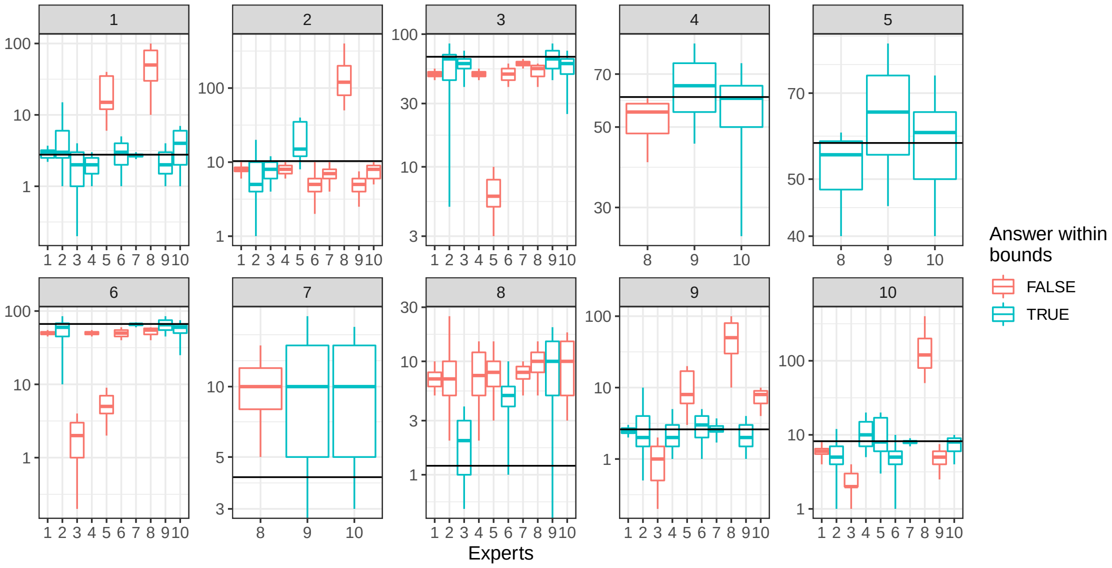


Figure 4.1: Boxplots of the individual responses to each seed question, with a log scaled y axis. Blue boxplots signify cases where the true value falls within the experts' minimum and maximum values, and red when the answer falls outside this range.

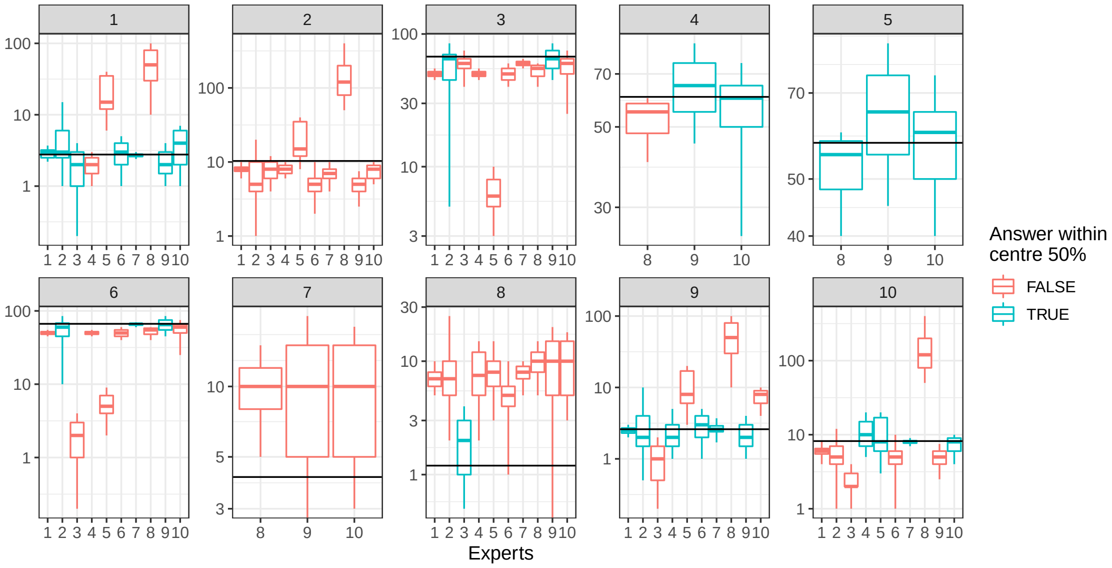


Figure 4.2: Boxplots of the individual responses to each seed question, with a log scaled y axis. Blue boxplots signify cases where the true value falls within the experts' 25th and 75th quartile values, and red when the answer falls outside this range.

Table 4.1: A table showing how often seed question answers fell within an expert's minimum and maximum

Expert	Out of Expert's bounds	Within Expert's bounds	Percent Within
1	5	2	29%
2	1	6	86%
3	3	4	57%
4	4	3	43%
5	5	2	29%
6	3	4	57%
7	3	4	57%
8	9	1	10%
9	2	8	80%
10	3	7	70%

Table 4.2: A table showing how often seed question answers fell within an expert's 25th and 75th quartiles

Expert	Out of Expert's bounds	Within Expert's bounds	Percent Within
1	5	2	29%
2	3	4	57%
3	5	2	29%
4	5	2	29%
5	6	1	14%
6	5	2	29%
7	3	4	57%
8	9	1	10%
9	4	6	60%
10	6	4	40%

While the range of the elicited values should contain all values the experts' believed were possible for the answer to lie within, it is clear that the coverage does not match this. It is often the case that the experts' ranges do not contain the true value within them. While the ranges provided could represent the views of the experts, it is likely that many experts did not provide wide enough bounds to properly quantify their views. Table 4.1 outlines the number of times the true answer fell within the ranges for each expert, and Table 4.2 outlines the number of times the true answer fell within the central 50% interval.

For example, Expert 9 had the true value within their bounds 80% of the time. If the expert had been providing 80% probability intervals, then they would appear to be well calibrated. Given the number of seed questions, a well calibrated expert may not have the probability exactly match the proportion within the interval due to random variation. Furthermore, the calibration of an expert will not in general match the expert's probability judgements. An expert with 80% coverage in their intervals, for example, may still believe their intervals to be 100% probability intervals.

Considering this table, it is reasonable to conclude that the experts with a higher proportion of intervals that include the true value are better calibrated in their judgements. As such, we would expect aggregation methods that take the experts' calibration into account to give these experts a higher weight. Notably, Experts 2 and 9 were particularly well calibrated, while Experts 1, 5, and 8 were less so. The reasons for these differences in calibration are not discernable, so it cannot be commented on as to whether specific experts were overconfident, had difficulties expressing their views numerically, or were just unfamiliar with the topic for a particular seed question.

Figures 4.1 and 4.2 also identify the questions experts performed better and worse on. For example, only three experts had the true value within their bounds for question eight, and all estimated the median to be at least twice as large as the true value. For question one, however, the majority of experts centred their distribution close to the observed value, and only two experts did not include it in their range. This could be due to a particular question being more difficult than others, requiring more obscure knowledge to answer, having an unexpected or surprising answer, or the experts sharing a common misunderstanding or bias.

We can also observe correlations between the experts within a number of other questions. For questions two, three, and six, the vast majority of experts who do not include the true value within their range are underestimating the value. Furthermore, the majority of medians are also underestimating the value. This pattern is also present in question eight, where the responses are overestimating the value. In such cases, it appears there is an agreement between the experts away from the true value. The reasons for this could be similar to those mentioned previously.

We may also wish to consider the relative mean squared error (RMSE) of the experts' medians compared to the true values, while accounting for the different scales of the seed questions. Table 4.3 shows this for each expert, calculated as

$$RMSE_i = \sum_{q=1}^{Q} \frac{(m_{i,q} - o_q)^2}{o_q} \tag{4.2}$$

where $RMSE_i$ is the RMSE for Expert i, $m_{i,q}$ is Expert i's median value for question q, and o_q is the true value for question q.

This table provides a guide as to which experts provided medians closer to the true values than others. Experts 5 and 8, in particular, had high RMSEs. This suggests their elicited medians were often further away from the true value than the other experts, meaning they were providing less accurate central estimates.. Experts 2 and 6 performed better than the other experts by this measurement, suggesting they provided more accurate central estimates..

While Experts 3, 6, and 7 seemed similarly calibrated in Table 4.1, each including the true value within their range four out of seven times, Table 4.3 shows that there were differences when considering the RMSE of their best estimates. Ideally an expert will be well calibrated, such that they provided probability intervals that provide the appropriate coverage, and informative, by providing a narrow interval, which in turn ensures their median is close to the true value. As such, Expert 6 appears to be performing better than Experts 3 or 7 over this set of seed questions.

It should be noted that in both tables, the Group 2 experts' values consider the full ten seed questions available, while the Group 1 experts only consider seven questions. As the appropriate context is provided in Table 4.1, and the values presented in Table 4.3 are averaged across questions, results should be comparable across groups. Of course, if the Group 1 experts had provided answers to the additional questions, these performance metrics would change.

It is important to note that while the experts' performance can be compared over the seed questions, the results will not necessarily follow in the model's parameter questions. As there is some difference between the topics of the two sets of questions, an expert may be more or less knowledgeable in the second part. However, if an expert's strong performance in the seed questions is due to being able to provide well calibrated distributions that accurately reflect their own uncertainty, then their performance may likely carry over to the second set of questions regardless of subject matter.

Table 4.3: The relative mean squared error between the experts' medians and true values.

Expert	RMSE
1	5.41
2	4.70
3	10.06
4	6.12
5	31.27
6	3.54
7	5.78
8	443.74
9	7.88
10	8.69

4.2.2 Model Parameters

The second part of the elicitation was to elicit prior distributions for the 5 model parameters. The η and μ parameters refer to the proportion of patients who will receive a positive test result from the reference test in the first and second rounds of testing respectively. The θ_1, θ_2, and θ_3 terms refer to the proportion of patients who test positive with the experimental test out of those who tested positive from the reference test in the first, second, or in neither round respectively.

The following plot, Figure 4.3, outlines the elicited values from each expert. We show all responses provided, noting that some experts did not provide elicited values for all questions.

The parameter questions refer to two diagnostic methods. The η and μ terms refer to the Awaji Criteria, while the θ terms refer to the BIMC test. It would be expected that the experts were more familiar with the η and μ terms, as the Awaji Criteria is routinely used to diagnose MND, whereas the BIMC test is a novel diagnostic method.

There appears to be expert agreement for the η term, with most experts placing their median between 0.5 and 0.7, and their upper quartile around 0.75. There is less concordance within the distributions for μ, though most experts have provided a distribution with lower values than they did for η.

There are also strong similarities between θ_1 and θ_2. Practically, this means the experts believe that the BIMC test should have similar performance whether the Awaji Criteria provides a positive test result at the first or second time point. As such, this provides an indication that the experts have a positive view of the BIMC test, believing that it will likely be able to detect MND cases at an earlier time point than the Awaji Criteria alone. This interpretation was verified with the experts in the SHELF elicitation meeting. The

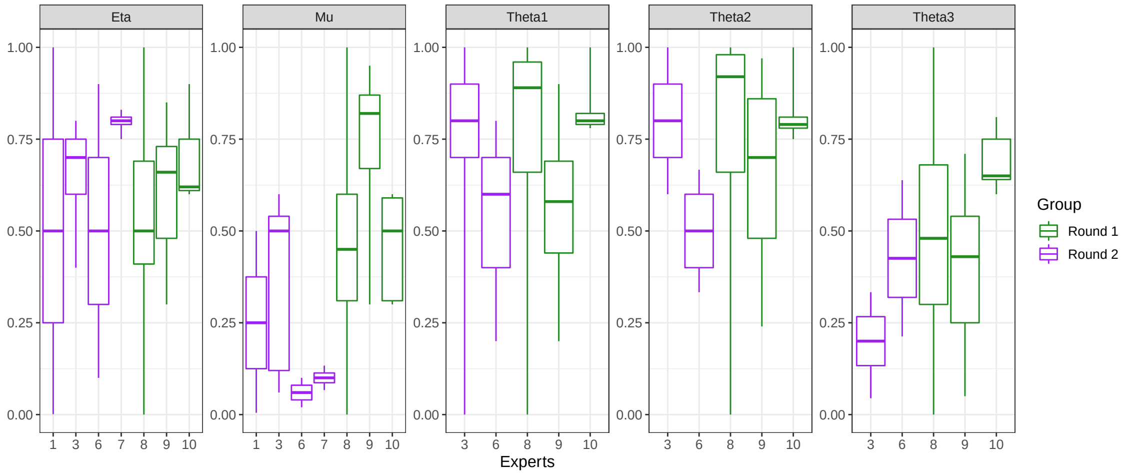


Figure 4.3: Boxplots of the elicited distributions for each parameter in the model.

θ_3 distributions are all shifted lower than the distributions of θ_1 and θ_2 for each expert.

It would be expected then that we would observe similar patterns in the aggregated priors if they are to provide distributions which have a similar practical interpretation.

As Experts 8, 9, and 10 were involved with the development of the diagnostic test, and the design of the study, they would appear to have a vested interest in the study's success, and may be more likely to overestimate how well the BIMC test will perform. Whether or not this is true, it is important to consider the views of additional experts in order to include viewpoints which are more impartial.

The elicited values from both groups of experts appear to cover similar ranges, suggesting the second group of experts may not be noticeably more optimistic. For each parameter, the best estimates of all Group 1 experts lie within the bounds of at least one Group 2 expert. In addition, the elicited values for the θ terms, which are those directly related to the effectiveness of the BIMC test, have a large amount of overlap between the two groups.

It can also be noted that for all experts, all possible values of each parameter have been included within the range of Expert 8. This suggests that many of the later aggregation methods will result in aggregations with a range between zero and one. Even excluding Expert 8, the remaining experts still place density on the majority of the unit interval. As such, the experts appear to believe that there is still reasonable uncertainty as to how the BIMC test may eventuate.

The combination of overlapping expert priors and relatively large variances, resulting in most values being plausible, has a number of positive implications. Firstly, it suggests that the experts are not overly confident in terms of their probability intervals, as they are mostly considering large ranges of potential outcomes. It also means that the aggregated prior distributions will likely have a strong overlap with the individual experts' priors. This ensures that the aggregated distributions are sensible from a practical standpoint, as at least one expert, if not more, will have stated that any values given weight are plausible.

4.3 Seed Questions

In this section we form aggregated priors for each seed question, using the other seed questions to provide the weights for individual experts where required. The seed questions are listed in full in Chapter 3.

As mentioned previously, Experts 1 to 7 were asked seven seed questions, while Experts 8 to 10 were asked the full ten seed questions. Our aggregations will consider three groups: Group 1 containing Experts 1 to 7, Group 2 containing Experts 8 to 10, and Group 3 containing all Experts, but only considering the common seven seed questions.

These three groups represent different types of expert who may be asked to provide elicited values. Group 1 represents experts who are knowledgeable about the field of interest, but are not directly involved in the design of the study or treatment. As such, these experts may be less familiar with the new treatment, but are likely to be considered more impartial. Group 2 represents the experts involved with the design of the study and treatment, and thus are likely to be more familiar with it. However, they may also be considered potentially biased, as they may have a personal interest in the study's funding and success. Finally, Group 3 represents a mix of experts from both groups. This group represents the wider range of knowledge within the field.

For the mathematical aggregation methods, we fit a split-normal distribution to each individual experts' quantiles before aggregating.

In this section, we first provide the aggregated priors for each aggregation method. We then discuss the performance of each aggregation method.

4.3.1 SHELF

SHELF is a behavioural aggregation method, and was reviewed in Chapter 3. Only the Group 2 experts were aggregated using this method, due to the change of format of elicitation, as discussed in Chapter 3.

Before eliciting the parameter values during the SHELF meeting, the experts re-examined the seed questions. This served both as a warm-up and practice for the experts to get used to the SHELF procedure, but also to provide further values for comparisons.

Figure 4.4 shows the elicited values for the seed questions in the SHELF session, as well as the individual seed question elicitations from the three experts who participated as a comparison.

The SHELF priors tended to be consistent with the individual experts' priors. There is only one case, Question 5, where the aggregated prior did not include the true value in its range. Conversely, there are many other questions where the aggregated prior did include the true value where individual experts did not.

It is also noted that the aggregated SHELF priors are often narrower than the individual experts' priors. This may be due to reduced uncertainty between the experts in the group setting, or overconfidence as discussed in Chapter 3.

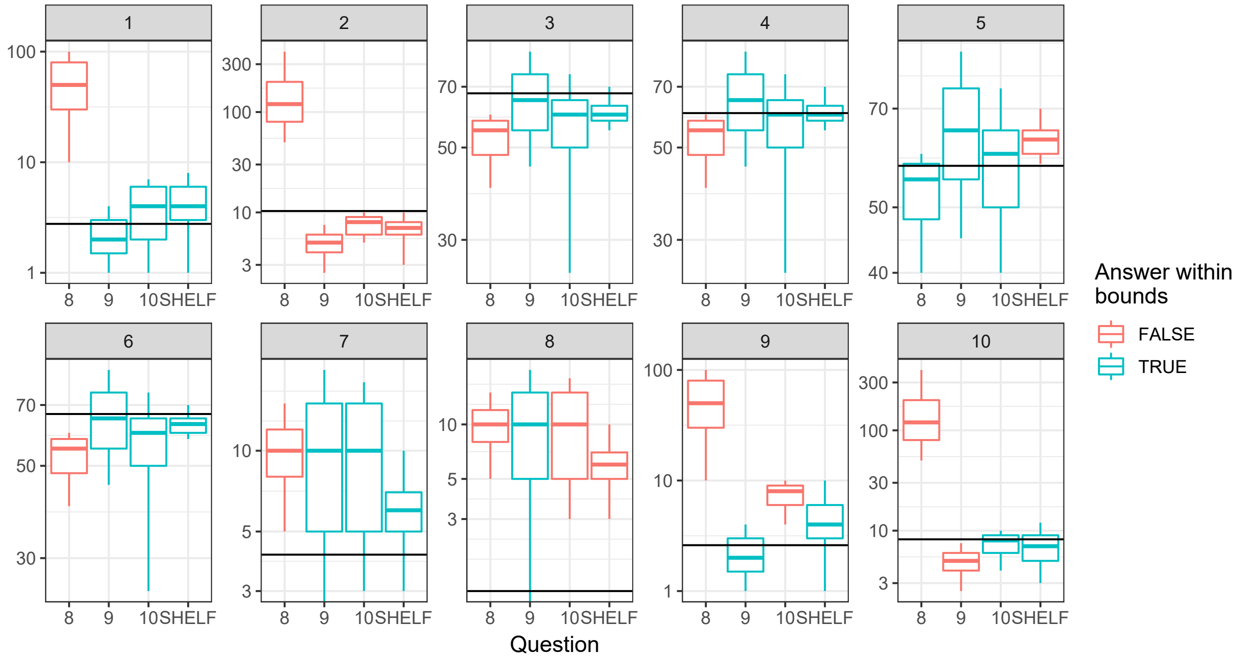


Figure 4.4: Boxplots for SHELF aggregations of the seed questions, from experts in Group 2.

4.3.2 Equal Weights

Figure 4.5 provides violin plots of the equal weights aggregation across each question, aggregating within each of the three groups. The horizontal line on each plot is the true value.

As Group 2 contains the fewest experts, and thus an aggregation of the smallest number of distributions, its aggregations tend to be more smooth and with fewer modes than the other aggregations.

Each of the aggregated priors contains the true value within its range. With the exception of Question 8, the true values tend to fall in areas where the aggregated priors have a larger density. As such, it appears that this aggregation method has performed better than most, if not all, of the individual experts.

For Questions 1, 2, 9, and 10, it can be seen that there is a very wide range on the aggregated priors for Groups 2 and 3. As can be seen in the individual experts' distributions, shown in Figure 4.1, Expert 8 provided very wide quantiles for these questions. The equal weights aggregation has given this expert an equal weight in the final distribution, and so there are large tails in these priors.

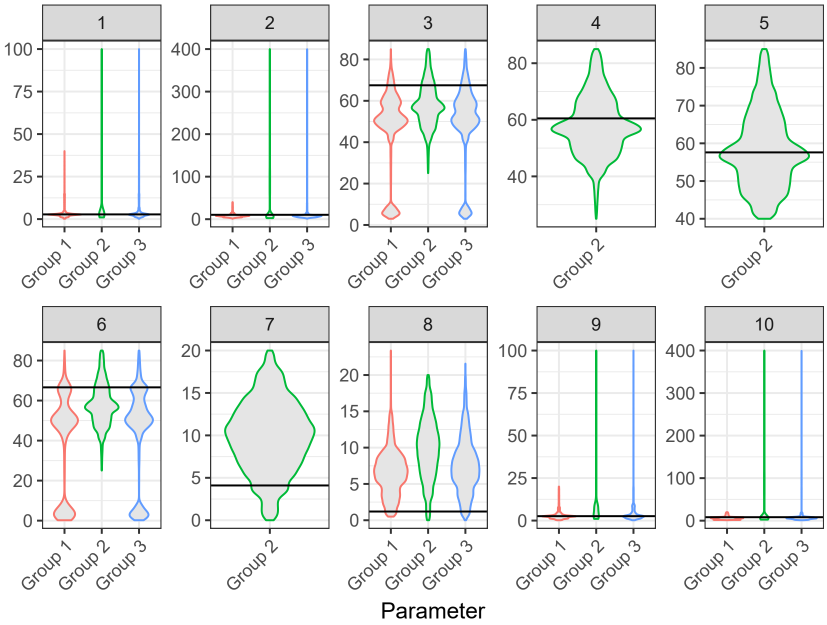


Figure 4.5: Violin plots for equal weight aggregations of the seed questions.

4.3.3 Classical Method

For Classical Method aggregation, the weights for each expert are calculated using the seed questions. Chapter 3 outlines the calculations used within this method.

Table 4.4 provides the weights for each of the ten experts. We consider weighting based on all seed questions asked to each group of experts when considering the two groups, and the seven common questions when combining the groups. As such, the Group 3 weights take into account a subset of the questions asked to Group 2, and accordingly result in different weights.

Table 4.4: Classical Method Weights

Group	1	2	3	4	5	6	7	8	9	10
Group 1	0.00	0.98	0.01	0.00	0.00	0.01	0.00	-	-	-
Group 2	-	-	-	-	-	-	-	0.00	0.73	0.27
Group 3	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00

The Classical Method has concentrated its weight primarily on Expert 2. This is perhaps unsurprising, as Expert 2 had the true value within their intervals more often than the other experts, as Table 4.1 shows, and had one of the lower relative mean squared errors in Table 4.3. While Experts 9 and 10 share the weight assigned in Group 2, Expert 2 again is given the majority of the weighting in Group 3.

Figure 4.6 provides the Classical Method's aggregated distributions for each seed question.

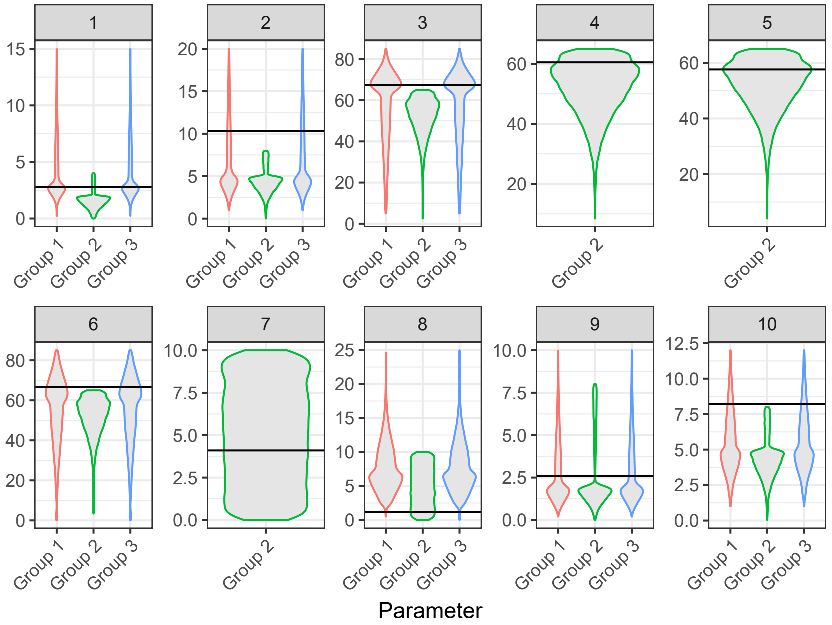


Figure 4.6: Violin plots for Classical Method aggregations of the seed questions.

In comparison to the equal weight aggregations in Figure 4.5, the Classical Method aggregations tend to have fewer modes. This is due to the fact that fewer experts have weight assigned to them. The Group 1 and 3 distributions are all very similar to Expert 2's, given the high weighting, but with some small differences due to the other experts included in the aggregations.

We can also note a difference in the coverage of the aggregated distributions. In contrast to the Equal Weights aggregation, Group 2's Classical Method aggregation does not contain the true value for all seed questions. While the Classical Method has led to distributions with less uncertainty, this has come at a cost of a decreased proportion of intervals containing the true value. This has not been an issue for Group 1 or 3, suggesting that the inclusion of Group 1 experts has led to better calibrated Classical Method aggregations.

4.3.4 Bayesian Aggregation

The Bayesian aggregation method used, based on work in Hartley and French (2021), consists of two stages. The first stage involves inferring parameters which stretch or narrow the elicited quantiles to improved calibration, based on responses to the seed questions. The second stage aggregates the recalibrated quantiles of the experts to create a single prior for each seed question or parameter.

For the Bayesian aggregation of the seed questions, we first fit a split-normal distribution to the experts' judgements for each question, as defined earlier. This allows for any

Table 4.5: Recalibration Parameter Posterior Medians

Group	Expert	Lower Interval Multiplier	Upper Interval Multiplier
Group 1	1	2.81	4.09
	2	1.54	0.73
	3	0.87	17.50
	4	1.37	3.74
	5	2.14	16.29
	6	1.57	2.68
	7	4.38	1.46
Group 2	8	2.07	2.09
	9	0.87	1.78
	10	1.12	1.00
Group 3	1	2.78	4.11
	2	1.54	0.73
	3	0.89	17.48
	4	1.37	3.72
	5	2.13	16.16
	6	1.57	2.67
	7	4.39	1.47
	8	2.10	2.48
	9	1.39	1.77
	10	1.38	1.12

asymmetry in experts' elicited quantiles to be captured by the model.

For seed question s and expert i, seed question answer y_s, lower quartiles $q_{25,s,i}$, medians $q_{50,s,i}$, and upper quartiles $q_{75,s,i}$, we consider a recalibration model as follows.

$$\mu_{25,s,i} = q_{50,s,i} - \alpha_{lower,i}(q_{50,s,i} - q_{25,s,i}) \tag{4.3}$$

$$\mu_{75,s,i} = q_{50,s,i} + \alpha_{upper,i}(q_{75,s,i} - q_{50,s,i}) \tag{4.4}$$

$$\tag{4.5}$$

where $\mu_{25,s,i}$ is the recalibrated lower quartile, $\mu_{75,s,i}$ is the recalibrated upper quartile, and $\alpha_{lower,i}$ and $\alpha_{upper,i}$ are the recalibration parameters adjusting the lower and upper variances respectively. The $\alpha_{lower,i}$, and $\alpha_{upper,i}$ parameters each have vague priors centred on one placed on them. These terms then define how the experts' intervals should be adjusted in order to better calibrate them. This component of the model is fit alongside Equations 4.6 to 4.12.

Table 4.5 provides the posterior medians of the $\alpha_{lower,i}$ and $\alpha_{upper,i}$ parameters calculated from the seed question data, defining how the elicited quantiles are recalibrated for each expert. For example, a value of three corresponds to a new interval half-width three times as wide as the original, or a value of 0.5 would correspond to a new interval with a half-width 50% of the original half-width. A value of one represents an interval that is unchanged.

As before, Group 1 consists of Experts 1 through 7 who were each asked seven ques-

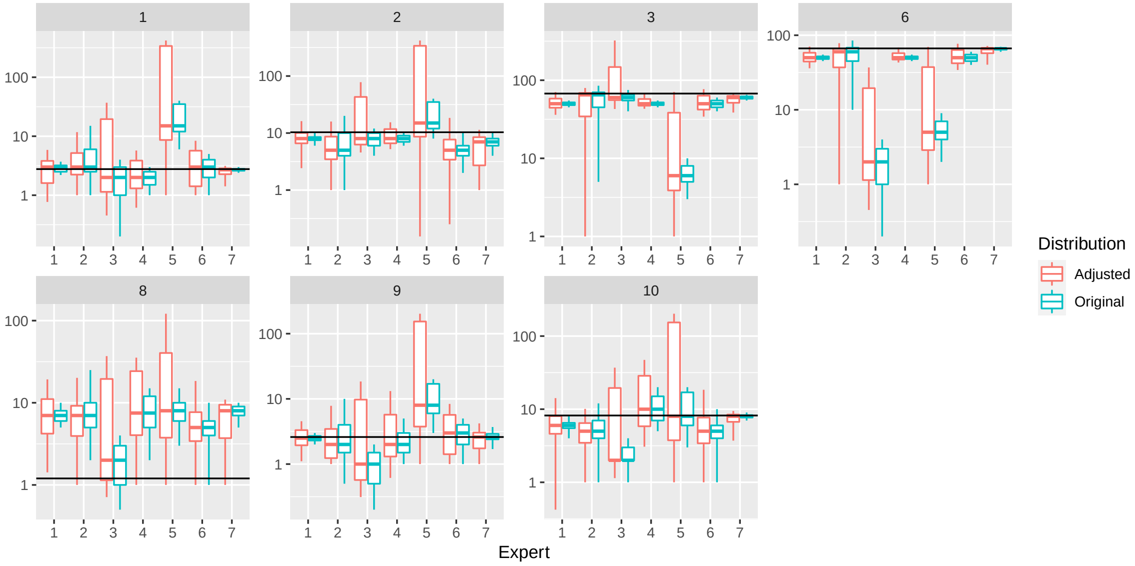


Figure 4.7: Rescaled experts' seed question distributions, from experts in Group 1.

tions, Group 2 consists of Experts 8 to 10 who each answered ten questions, and Group 3 contains all experts but only considering the common seven questions.

The majority of experts had quartile recalibration parameters greater than one, suggesting they were overconfident. Overconfidence, here, refers to quantiles that are too narrow as judged by the seed questions. This is consistent with findings in the literature, as discussed in Chapter 3. Notably, Expert 10 had recalibration parameters closest to one for both intervals, suggesting they were the best calibrated of all the experts.

Figures 4.7, 4.8, and 4.9 presents the original and recalibrated distributions for each seed question for experts in Groups 1, 2, and 3 respectively. As they demonstrate, many of the recalibrated distributions, in red, have been stretched wider to provide better calibrated quantiles. This is particularly noticeable in seed questions 3 and 6, where some elicited intervals that previously did not contain the true answer have been recalibrated so that they do.

Experts 3 and 5 had very large multipliers for their upper intervals. As demonstrated in these plots, this was driven by a small number of seed questions, specifically 3 and 6, where the experts greatly underestimated the true answer. The large difference has led the Bayesian recalibration to stretch the upper interval to better fit the true value for these questions, while balancing against the smaller change required for the other seed questions. As such, the recalibrated distributions only just include the true answer, if at all.

Using the inferred $\alpha_{lower,i}$, and $\alpha_{upper,i}$ parameters, the following model can then be used to determine an aggregated distribution for each seed question. We consider an alter-

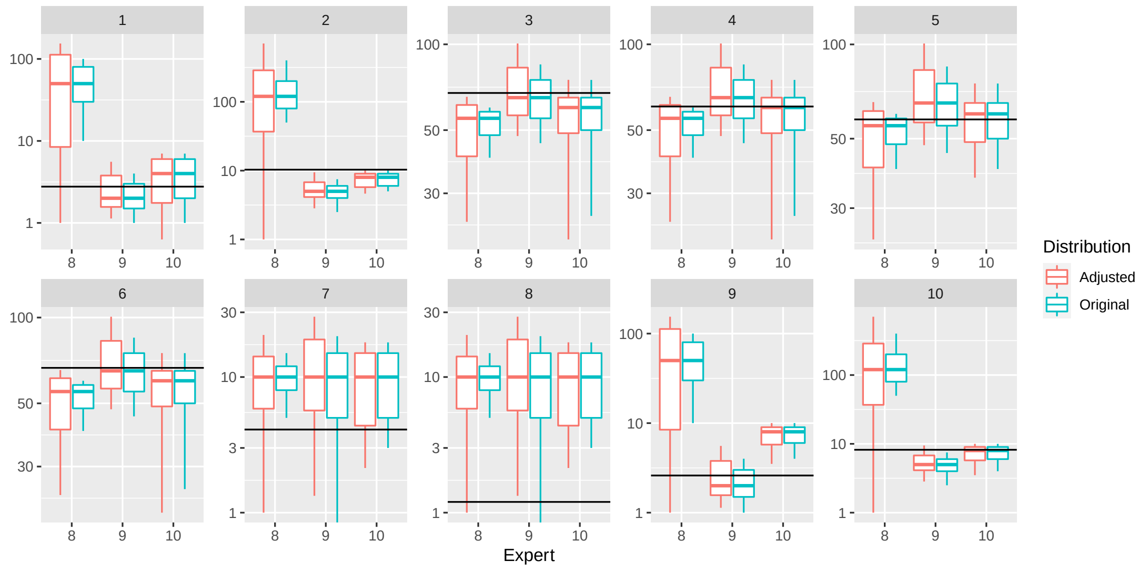


Figure 4.8: Rescaled experts' seed question distributions, from experts in Group 2.

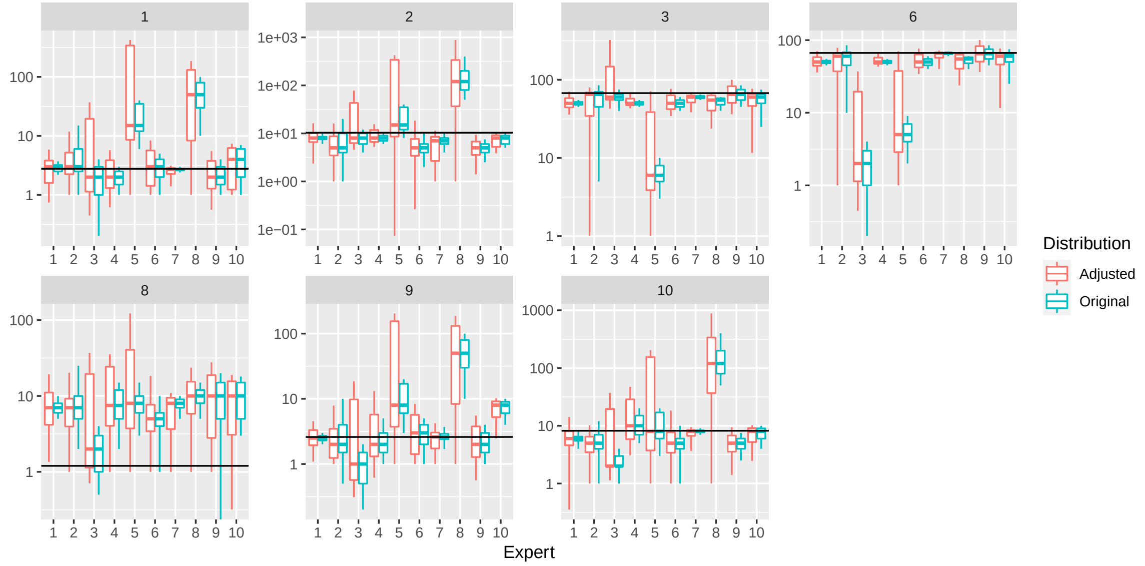


Figure 4.9: Rescaled experts' seed question distributions, from experts in Group 3.

native model structure to Hartley and French (2021), instead modelling and aggregating the lower and upper quartiles directly, and set $\mu_{50,s,i} = q_{50,s,i}$.

$$y_s \sim Split - Normal(\mu_{50,s,i}, \sigma^2_{lower,s,i}, \sigma^2_{upper,s,i}) \tag{4.6}$$

$$\sigma_{lower,s,i} = \frac{\mu_{50,s,i} - \mu_{25,s,i}}{\Phi(0.75)} \tag{4.7}$$

$$\sigma_{upper,s,i} = \frac{\mu_{75,s,i} - \mu_{50,s,i}}{\Phi(0.75)} \tag{4.8}$$

$$\mu_{50,s,i} \sim N(M_s, S_s^2) \tag{4.9}$$

$$\mu_{25,s,i} \sim N(L_s, Sl_s^2) \tag{4.10}$$

$$\mu_{75,s,i} \sim N(U_s, Su_s^2) \tag{4.11}$$

This model calculated values for the Split-Normal distribution's variances based on the adjusted quartiles values, as recalibrated using the $\alpha_{lower,i}$ and $\alpha_{upper,i}$ parameters.. These adjusted values are then included in a hierarchical model structure in order to aggregate them. Vague priors are placed on M_s, L_s, U_s, S_s, Sl_s, and Su_s. During computation, we also constrain the values of $\mu_{25,s,i}$ and $\mu_{75,s,i}$ to be respectively less than and greater than $\mu_{50,s,i}$. This ensures the lower and upper quartile values proposed by the MCMC are always on the correct side of the median.

M_s, L_s, and U_s then represent the aggregated mean, lower and upper quartile respectively for each seed question. The aggregated priors for the standard deviations for the lower and upper halves of the split-normal distribution can then be calculated using these as follows.

$$\sigma_{l,s} = \frac{M_s - L_s}{\Phi(0.75)} \tag{4.12}$$

$$\sigma_{u,s} = \frac{U_s - M_s}{\Phi(0.75)} \tag{4.13}$$

The final model for each aggregated seed question then takes the form of a $Split - Normal(M_s, \sigma_{lower,s}, \sigma_{upper,s})$ distribution, which we can sample from based on posterior samples of the parameters.

Figure 4.10 provides the aggregated distributions for each group, where the red bar represents the true value. In each case, the true answer is within the bounds of the probability distributions. This suggests the aggregate distributions are covering sensible ranges of values.

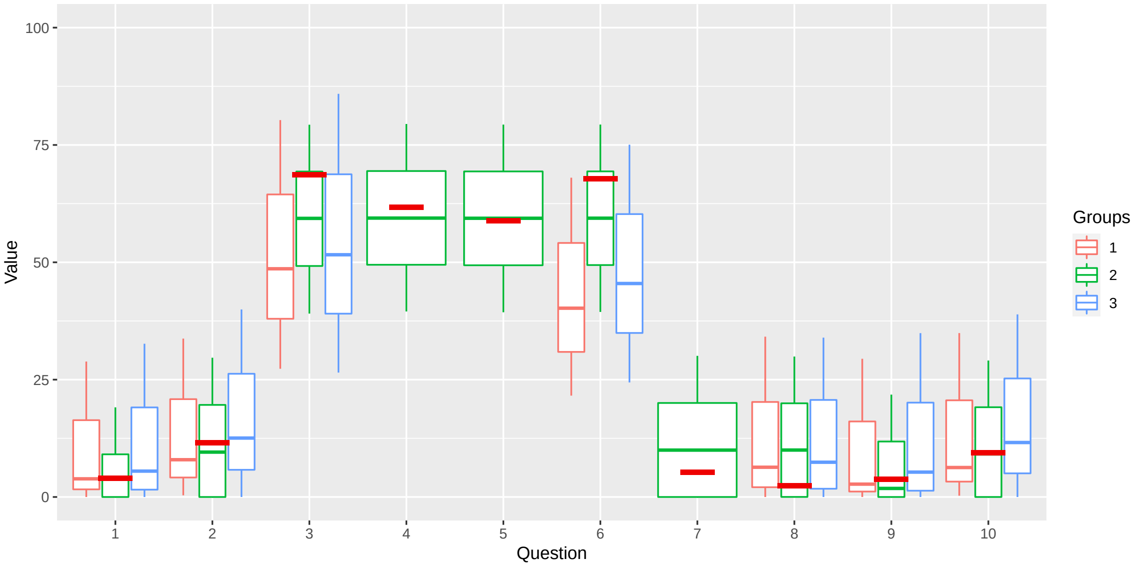


Figure 4.10: Boxplots for Bayesian aggregations of the seed questions, from all experts.

Model Extension of Recalibrated Central Estimate

This Bayesian model can also be extended to recalibrate the experts' median values. This allows for judgements by experts whose medians consistently over or under-estimate the true values of the seed questions to be adjusted.

For seed question s and expert i, seed question answer y_s, expert medians $q_{50,s,i}$, lower quartiles $q_{25,s,i}$, and upper quartiles $q_{75,s,i}$, we consider an extended recalibration model as follows.

$$\mu_{50,s,i} = \beta_i q_{50,s,i} \tag{4.14}$$

$$\mu_{25,s,i} = \mu_{50,s,i} - \alpha_{lower,i}(q_{50,s,i} - q_{25,s,i}) \tag{4.15}$$

$$\mu_{75,s,i} = \mu_{50,s,i} + \alpha_{upper,i}(q_{75,s,i} - q_{50,s,i}) \tag{4.16}$$

$$\tag{4.17}$$

where $\mu_{25,s,i}$ is the recalibrated lower quartile, $\mu_{50,s,i}$ is the recalibrated median, $\mu_{75,s,i}$ is the recalibrated upper quartile, β_i is a multiplier adjusting the median, and $\alpha_{lower,i}$ and $\alpha_{upper,i}$ are recalibration parameters adjusting the lower and upper variances respectively. The β_i, $\alpha_{lower,i}$, and $\alpha_{upper,i}$ parameters each have vague priors centred on one placed on them. These three parameters then define how the experts' intervals should be adjusted in order to better calibrate them.

Table 4.6 provides the posterior medians of these three parameters calculated from the seed question data. The interpretations of the α parameters the Lower Interval Multi-

Table 4.6: Recalibration Parameter Posterior Medians

Group	Expert	Lower Interval Multiplier	Median Multiplier	Upper Interval Multiplier
Group 1	1	3.07	1.32	0.63
	2	1.54	1.01	0.74
	3	1.50	1.42	22.77
	4	1.35	1.33	0.49
	5	2.12	1.08	18.23
	6	2.06	1.30	1.83
	7	3.09	1.00	1.67
Group 2	8	1.30	0.66	5.66
	9	0.94	1.07	1.88
	10	1.09	1.00	1.10
Group 3	1	3.05	1.32	0.63
	2	1.55	1.01	0.74
	3	1.49	1.42	22.84
	4	1.35	1.33	0.48
	5	2.11	1.08	18.11
	6	2.06	1.30	1.83
	7	3.08	1.00	1.66
	8	0.23	0.15	11.86
	9	1.43	1.37	1.70
	10	1.28	1.05	1.09

plier and Upper Interval Multiplier, have the same interpretations as previously. For the elicited median, the Median Multiplier, β, is multiplied by the elicited median to create the recalibrated median. As such, a number greater than one corresponds to an increase in the recalibrated median, and a value lower than one corresponds to a decrease in the recalibrated median.

The size of the quartile recalibration parameters demonstrates both the overconfident experts. The majority of experts had quartile recalibration parameters greater than one, suggesting they were overconfident. Notably, Expert 10 still had recalibration parameters close to one for both intervals and the median, suggesting they were the best calibrated of all the experts.

In comparison to the previous model, the recalibration parameters for the intervals change most when the median's multiplier is further away from one. For those experts whose median were recalibrated, their intervals would then be recalibrated in a different way to account for the difference in the central estimate.

Expert 8 had very low median recalibration parameters, suggesting they tended to overestimate the true value. This can be seen in Figure 4.1, where they vastly overestimated the answer to a number of questions. To a lesser extent, Experts 1, 3, 4, and 9 appeared to be underestimating the answer more often, as demonstrated by their median recalibration parameters being greater than one.

Expert 3 was also notable, in their recalibration parameters all tending to be away from one. The high median recalibration parameter suggests a tendency to underestimate, while the very large upper recalibration parameter suggests this was combined with insufficiently wide intervals.

While Expert 5 had a median multiple close to one, suggesting they were not consistently over or underestimating the seed question responses, their interval widths were both greater than one. This suggests the expert was equally likely to provide estimates above or below the true answer, together with intervals that were too narrow on average.

Figures 4.11, 4.12, and 4.13 presents the original and recalibrated distributions for each seed question for experts in Groups 1, 2, and 3 respectively. As these plots demonstrate, many of the recalibrated distributions, in red, have been stretched wider to provide better calibrated quantiles. This is particularly noticeable in seed questions 3 and 6 of Group 1, where some elicited intervals that previously did not contain the true answer have been recalibrated so that they do.

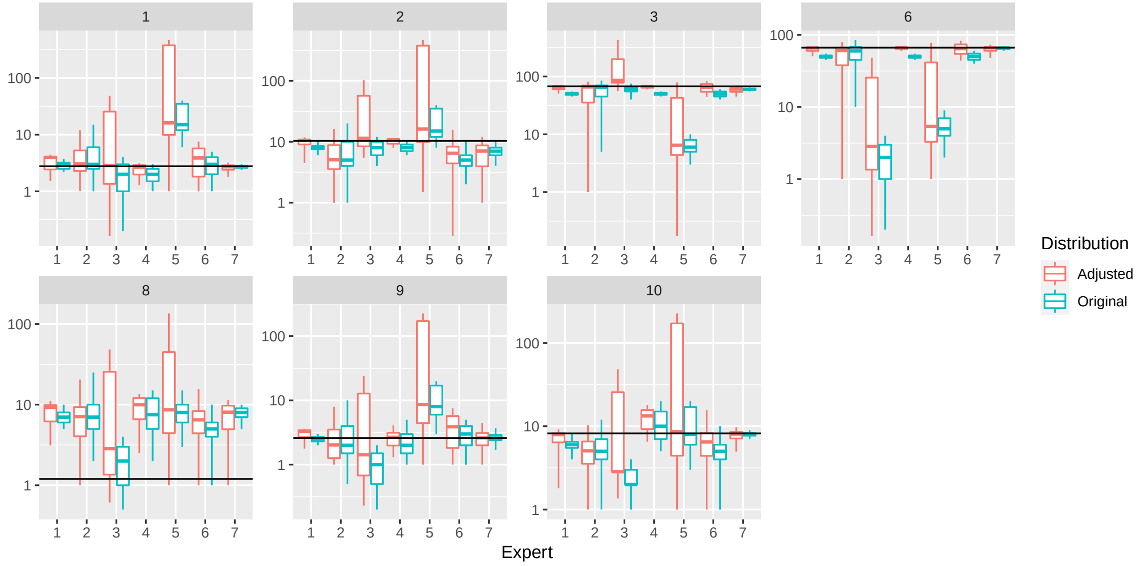


Figure 4.11: Rescaled experts' seed question distributions, from experts in Group 1.

In comparison to the previous method, it can also be seen where the median values were adjusted as well. For example, Expert 8 in Group 2 has rescaled medians lower in value than those provided, as the Bayesian model has determined. While in some cases this has led to an adjusted median closer to the true value, in other questions it has moved the median further away.

We use the same aggregation model as above after recalibration, using the newly inferred β_i, $\alpha_{lower,i}$, and $\alpha_{upper,i}$ parameters.

Figure 4.14 provides the aggregated distributions for each group, where the red bar represents the true value. As before, for each case the true answer is within the bounds of the probability distributions. We also note that the aggregation distribution medians

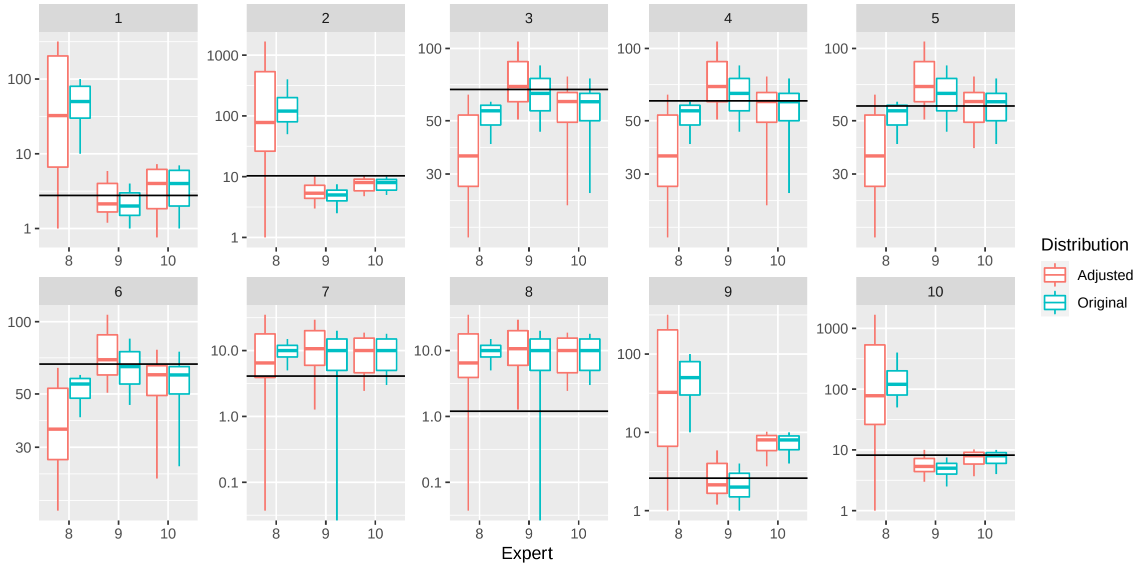


Figure 4.12: Rescaled experts' seed question distributions, from experts in Group 2.

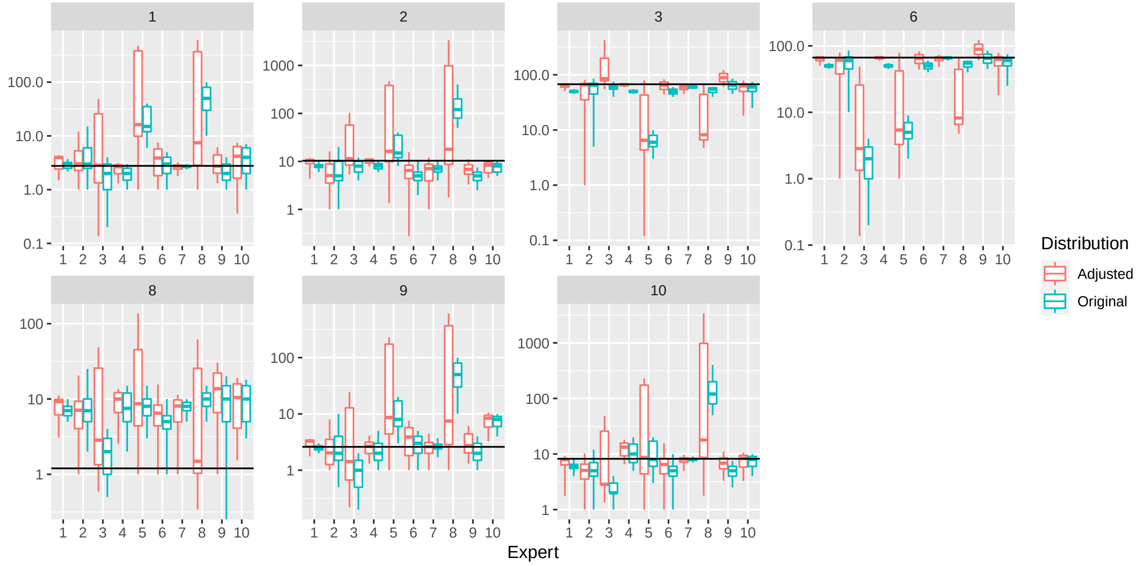


Figure 4.13: Rescaled experts' seed question distributions, from experts in Group 3.

appear to be more accurate in the previous model than this extension, suggesting that recalibrating the medians may not have improved performance. We discuss this further in Section 4.3.9.

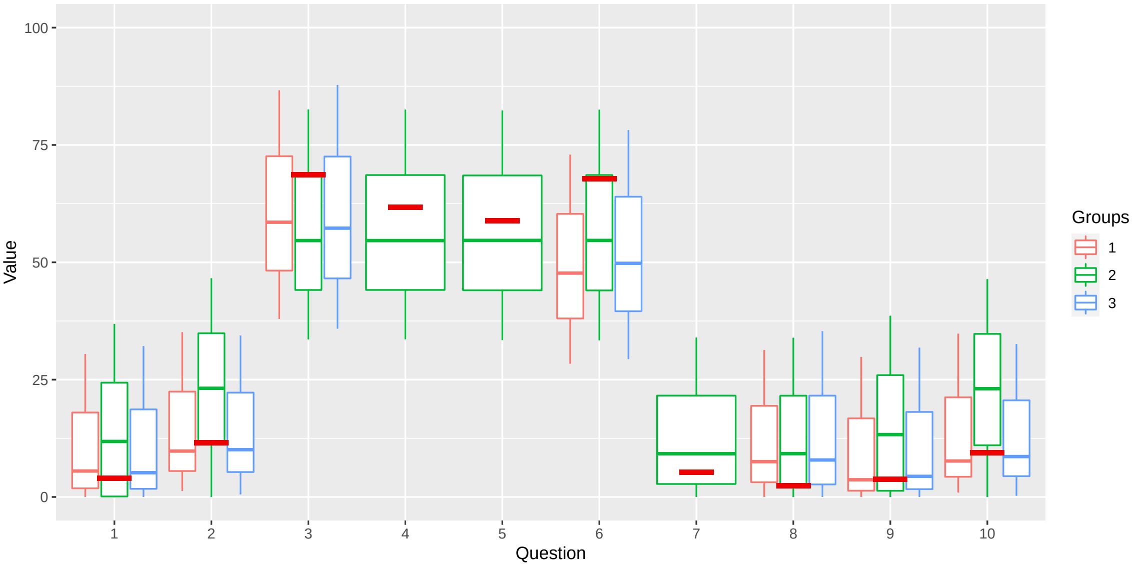


Figure 4.14: Boxplots for Bayesian aggregations of the seed questions, from all experts.

4.3.5 Scoring Rules

Three proper scoring rules were selected to evaluate the individual responses and aggregated distributions. The Brier, Logarithmic and Quadratic scores each reflect the informativeness or calibration of the experts (Winkler, 1996; James E. Matheson, 1976; Gneiting et al., 2007). For the scoring rules that require a density function, a split-normal distribution has been fitted, to allow for different variances on either side of the mode.

The logarithmic scoring rule calculates the negative of the natural logarithm of the density at the location of the true value. At the value r,

$$L(r) = -\ln f(r) \tag{4.18}$$

As only quartiles were elicited, in order to determine the value of the density at a given point a distribution was fitted to each question. The split-normal distribution that best fit the provided quartiles was selected. This score is undefined for values where $f(r) = 0$, and this results would represent very poor performance.

The second method used was a Brier Score. This calculates the squared error between the known value, and the distribution's median. For the known value r, and an expert's median m,

$$B(r) = (m - r)^2 \tag{4.19}$$

The quadratic score calculates the integral of the distribution squared, and subtracts it from the value of the density at the location of the true value multiplied by two.

Table 4.7: Individual and Aggregated Distributions Scores

Method	Expert/Group	Brier	Logarithmic	Quadratic
Individual	1	78.22	6.52	0.24
	2	15.30	2.56	0.14
	3	534.63	Inf	0.20
	4	78.88	5.38	0.10
	5	977.97	Inf	0.28
	6	79.38	3.12	0.19
	7	14.24	2.13	- 0.74
	8	2676.95	3.88	0.07
	9	21.42	2.55	0.14
	10	23.11	2.25	- 0.01
SHELF	Group 2	12.84	2.20	0.06
Equal Weight	Group 1	74.45	2.19	- 0.22
	Group 2	28.97	2.96	- 0.05
	Group 3	60.07	2.10	- 0.08
Classical Method	Group 1	16.44	2.59	-0.10
	Group 2	19.32	3.14	-0.05
	Group 3	15.66	2.58	-0.10
Bayesian Aggregation	Group 1	155.37	2.45	-0.05
	Group 2	25.12	Inf	-0.04
	Group 3	109.14	2.76	-0.02
Bayesian Aggregation Extended	Group 1	69.59	2.42	-0.04
	Group 2	102.43	3.14	-0.01
	Group 3	63.01	2.48	-0.04

$$Q(r) = -\left(2f(r) - \int_{-\infty}^{\infty} [f(\theta)]^2 d\theta\right) \tag{4.20}$$

For ease of comparison, we take the negative of the quadratic scoring rule to ensure that the score consistently provides lower scores for better performing experts.

An expert who is well calibrated and informative will receive a low value for each score. While both calibration and informativeness are reflected in each score, it is expected the Brier score will more strongly reflect the informativeness of the experts, the Logarithmic score will more strongly reflect the calibration of the experts, and the Quadratic score will provide a more balanced position between them.

Table 4.7 presents scores from each individual expert's seed questions, and from each of the aggregation methods.

4.3.6 Expert Performance

The individual experts' performances can be considered using the scores in Table 4.7. Experts 2, 7, 9, and 10 each had notably lower Brier scores compared to the other experts. This was also the case for the Logarithmic score, noting that Experts 6 and 8 also performed reasonably well. Under the Quadratic score, Expert 7 was the best by a large margin, with Experts 2, 4, 8, 9, and 10 all performing similarly. Noting that the Classical

Method gave high weight to Expert 2, and preferred Experts 9 and 10 over 8, it appears these three experts performed particularly well.

Experts 3, 5, and 8 performed the worst in terms of the Brier score, had the largest RMSE, from Table 4.3, and also received the largest recalibration parameters in Table 4.8, as part of the Bayesian aggregation method. It appears then that these experts were poorly calibrated. This does not mean they did not have the necessary understanding or knowledge, but rather did not provide probability intervals with the expected level of coverage.

Another consideration is the performance of each group. Across the seed questions, we would expect both groups of experts to perform similarly, as the difference between the groups is related to their work on the novel diagnostic method, and thus the model parameters, rather than their wider knowledge of MND. Group 1 had a mean Brier score of 254.09, and median score of 78.88, while Group 2 had a mean Brier score of 907.16, and median score of 23.11. In terms of the Logarithmic score, Group 1 had an infinitely large mean score, and median score of 5.38, while Group 2 had a mean score of 2.89 and median score of 2.55. In terms of the Quadratic score, Group 1 had a mean score of 0.06, and median score of 0.19, while Group 2 had a mean score of 0.07 and median score of 0.07.

The mean scores are easily influenced by large individual scores, such as Expert 8's Brier score. As such, the median scores provide a more sensible measurement of performance. Group 2's median Brier and Logarithmic scores were much lower than Group 1's, while both median Quadratic scores were similar. Comparing the range of values across all three scores, there does not appear to be strong evidence that either group strongly outperformed the other.

4.3.7 Aggregation and Individual Expert Distributions

Table 4.7 also outlines the benefit of using an aggregated prior distribution over a single expert's prior.

The mean Brier score of the individual experts was 450.01, and the median was 78.55. As such, randomly selecting an expert would give an expected Brier score greater than any aggregated prior distribution. While some experts had better scores than some aggregation methods, as those experts would not be identifiable prior to the elicitation there is no guarantee they would be selected.

The median Logarithmic score for the individual experts' prior distributions was 3.5. The calculation for the mean for this score includes two infinite values, and so is not suitable for comparison. However, the median score is still higher than any of the aggregated distributions for any group.

The individual experts' Quadratic scores had a mean of 0.06 and median of 0.14. While the SHELF method's Quadratic score was the worst of the aggregated methods, it was still competitive in comparison to the individual experts', and an improvement over the median score. Given the improvement SHELF provided under the other scoring rules, this provides further evidence that aggregation of individual experts' prior distributions results in better calibrated and more informative distributions.

This result is also seen in the aggregated distribution plots. In each case, the true value tends to lie within the aggregated distributions, whereas the individual experts provide a much lower success rate.

As reviewed in Chapter 3, the benefit of aggregated prior distributions over individual experts' prior distributions has been previously demonstrated in the literature. These results are consistent with previous findings. Should a single expert be elicited from, the suitability of their prior distribution relies on both their knowledge of the field and their ability to specify probabilities. An expert who does not perform well in the elicitation may provide distributions that do not reflect the final outcome particularly well, yet that same expert can be included within a group that performs better than any individual expert.

It is demonstrated in Marti et al. (2021) that experts' performance in seed questions is representative of future performance, and not due to random chance. As such, the inclusion of multiple experts with different levels of weights can improve the performance of the aggregated distribution.

Furthermore, during a single elicitation with a single expert it is hard to judge their ability, as there are no other experts to compare their scores against. This makes it difficult to determine whether a well calibrated expert has been selected or not. Should a single expert be elicited from, the decision maker would have little knowledge of whether the resulting priors are accurate. While a score could be calculated for an individual expert, as was done in Table 4.7, the resulting score is only contextualised by the scores of the other experts.

4.3.8 Comparison of Aggregation Methods

We can also compare the performance of aggregation methods using the scores in Table 4.7. Note that scores should be compared within expert groups when making comparisons of aggregation methods. As such, the SHELF method, in which Group 2 were the only experts who undertook the required meeting, should be compared to scores from other aggregations of Group 2's priors. Comparing across both expert group and aggregation method simultaneously fails to account for the differences in performance between experts.

The Classical Method and SHELF are both widely used aggregation techniques, and the comparison of seed questions shows they both perform well. Both methods consistently outperformed any individual expert, suggesting the use of either would result in a more informative and better calibrated prior distribution. Equal Weights aggregation offers a quicker and simpler method of aggregation, though there is less of an improvement over individual experts. Still, without seed questions with which to judge experts to select those that performed best, using an Equal Weights aggregated prior often performs better than randomly selecting an expert.

No aggregation method uniformly outperforms the others across all scores. The SHELF method has the best Brier score and Logarithmic score, performing better than nearly any individual expert or Group 2 aggregation method. This suggests the SHELF method results in a relatively well calibrated and informative distribution. However, its Quadratic score was worse than many of the other aggregation methods'.

Previous studies comparing mathematical aggregations and individual experts tended to find similar results as found here. Cooke (2008) found the Classical Method outperformed Equal Weights and the best expert. Ganguly et al. (2014) showed that over a group of 48 datasets, Equal Weights' error was 3% higher than the Classical Method, while the majority of the time the Classical Method was better calibrated than Equal Weights. Flandoli et al. (2011) suggested that any of the aggregation methods would perform better than selecting a single expert at random, and that none of the aggregation methods tested ever performed worse than the single best expert. Likewise, Lin and Cheng (2009) found aggregation performs better than a single expert, and that the Classical Method and Equal Weights perform similarly. The effectiveness of the best expert appears to depend on how well the best expert has performed, as Hammitt and Zhang (2013) found the best expert performed better than either aggregation method. In this case, the Classical Method still outperformed Equal Weights. While these comparisons looked at various aggregation methods, the comparison of SHELF to the Classical Method has not been previously assessed.

The resulting distributions also show some differences. The Equal Weights aggregations tend to be multi-modal, with long tails when individual experts have provided larger ranges of values. The Classical Method, however, tends to provide distributions with fewer modes due to the inclusion of fewer experts, and those experts providing particularly large ranges being assigned no weight. Across both methods the true value always falls within the ranges, with the exception of Group 2's answer to seed questions 2 and 3. The plots for SHELF and the Bayesian aggregation method are all uni-modal, though this is a direct result of the methods themselves.

In terms of mathematical aggregation methods, the Classical method has performed better than Equal Weights and the Bayesian Aggregation Methods for each group in terms of Brier score, however the Equal Weights aggregation performed better under the Logarithmic score. Both the Bayesian and Classical Method aggregations performed similarly using the Logarithmic score. There does not seem to be a consistently better aggregation method using the Quadratic score. Overall, it appears that the Classical Method results in the best calibrated distributions of the mathematical aggregation methods. This is expected, as it assigns higher weights to better calibrated experts.

The distributions also show how the differences between Group 1 and 2 affect the final aggregations for these methods. For example, the Bayesian aggregation results in similar distributions for both Group 1 and Group 2. The Equal Weights method on the other hand, results in distributions with large differences in their ranges, and with peaks in the density around different locations. The Classical Method falls between the two, tending to have fewer distribution modes than the Equal Weights distributions. As such, it appears that the Bayesian method is less sensitive to differences between experts, while the Equal Weights method is most sensitive.

In behavioural aggregations, the discussions involving a range of experts have the potential to reduce the variability in their individual abilities to specify probabilities, and knowledge of the domain of interest. Ideally, the discussion can take advantage of the most knowledgeable experts' for each question, and best calibrated experts' ability to specify probabilities, to form informative and well calibrated distributions. Ideally the range of experts would also account for implicit biases, such as those relating to personal interests in the outcome of a study, but these cannot always be avoided. The group setting used in SHELF facilitates this type of discussion, though does have some potential downsides. As seen in the seed question example, SHELF resulted in narrower distributions than the other aggregation methods. While this could simply reflect a decrease in uncertainty, it could also represent overconfidence in the quantities provided by the group.

In terms of the practicalities of implementing each of the aggregation methods, each has a different level of effort and complexity involved for both the experts and elicitor.

From the elicitor's perspective, the Equal Weights method is the simplest, as it only involves eliciting distributions individually from each expert. The SHELF method, although requiring more time with the experts and some facilitation skill, does not require additional questions. The Classical Method presents a further level of work, in developing appropriate seed questions. The Bayesian Aggregation method also involves a more complex algorithm than the Classical Method, and an understanding of how to implement an MCMC method.

From the expert's perspective, the Equal Weights method is the simplest, requiring just elicitation of prior distributions. Both the Classical Method and Bayesian aggregations will mostly appear the same to the experts, as they require both elicitation questions on parameters of interest and seed questions. Finally, the SHELF method can present a larger challenge, as it requires all experts to meet at the same time, and multiple rounds of elicitation.

Given the performance of each method, it is then advisable to at least consider an Equal Weights aggregation when multiple experts can be elicited from. Though results are limited to a single expert group, it appears that the SHELF elicitation is the preferable aggregation method. The Classical Method is clearly preferable to Equal Weights, and presents a good alternative to SHELF when external factors may prevent, or make difficult, the required meetings.

4.3.9 Bayesian Recalibration

The Bayesian aggregation method, while similar to the Classical Method in terms of implementation difficulty, has performed worse across the three scores. There is a slight difference between the initial Bayesian aggregation and the extended model. For the Brier score, the extended model offered an improvement for Groups 1 and 3, but a worse score for Group 2. The Logarithmic and Quadratic scores were very similar between both methods. As such, for the two Bayesian aggregation methods presented, there does not appear to be a clear improvement across all scores when extending the recalibration to include an adjusted median.

One advantage the Bayesian aggregation method presents, however, is through the recalibration of the individual experts' priors. Table 4.8 presents the scores from the recalibrated experts as individuals. In comparison to Table 4.7, the individual experts' recalibrated distributions tend to have improved scores. For example, Expert 1 originally had a Brier score of 78.22, Logarithmic score of 0.652, and Quadratic score of 0.24. Each of these scores reduced dramatically once the recalibration was completed. By accounting for an expert's tendency to under or over-predict, and be under or over-confident in their interval estimates, the Bayesian aggregation method can improve their performance.

Furthermore, there does seem to be an advantage to the extended recalibration when considering individual experts. With the exception of Expert 9, all experts have seen a noticeable improvement in their Brier score in the extended model over the original one. There are also improvements in the other two scores, though these do not occur in all cases.

For Expert 2, who the Classical Method identified as the best performing in the seed questions, there was still an improvement in all scores. While the improvement was not as large as it was for other experts, this suggests that this method can still improve individual experts who do well in the seed questions. It is, however, important to note that this method is altering the prior distributions in ways the experts who provided them may not agree with.

We can also compare the recalibrated individual expert scores in Table 4.8 with the aggregated prior distributions in Table 4.7. While the original expert scores were predominately worse than the aggregated priors, the recalibrated distributions are more likely to perform as well as the aggregated distributions.

This suggests that in cases where aggregation is not possible, an individual expert's

Table 4.8: Recalibrated Distribution Scores

Group	Expert	Recalibration			Extended Recalibration		
		Brier	Logarithmic	Quadratic	Brier	Logarithmic	Quadratic
1	1	78.22	1.95	-0.02	8.66	1.49	-0.11
1	2	15.30	2.13	0.02	13.56	2.12	0.01
1	3	534.63	Inf	0.29	552.70	Inf	0.05
1	4	78.88	3.51	0.05	13.03	1.30	-0.15
1	5	977.97	49.22	0.06	971.90	49.33	0.06
1	6	79.38	2.22	0.08	7.42	1.90	0.00
1	7	14.24	1.48	-0.18	13.52	1.27	-0.22
2	8	2676.95	3.57	0.00	1334.10	3.81	0.03
2	9	21.42	2.75	0.22	36.24	2.52	0.12
2	10	23.11	2.29	-0.01	22.91	2.28	- 0.01
3	1	78.22	1.95	-0.02	8.66	1.49	-0.11
3	2	15.30	2.13	0.02	13.56	2.12	0.01
3	3	534.63	Inf	0.29	552.70	Inf	0.05
3	4	78.88	3.51	0.05	13.03	1.30	-0.15
3	5	977.97	49.22	0.06	971.90	49.33	0.06
3	6	79.38	2.22	0.08	7.42	1.90	0.00
3	7	14.24	1.48	-0.18	13.52	1.27	-0.22
3	8	3671.83	3.77	0.01	890.32	77.01	-0.18
3	9	15.72	2.10	0.06	141.57	1.87	- 0.07
3	10	26.67	2.14	-0.02	20.13	2.12	- 0.02

prior distributions may be improved through a Bayesian recalibration. We also suggest that the poorer performance of the Bayesian aggregation in comparison to other methods is due more to the aggregation step, rather than the recalibration.

4.4 Parameter Aggregations

In this section, we will form aggregated priors for each parameter in the model, for use in assurance calculations. As previously discussed, the model contains five parameters.

The θ_2 term is of particular importance, as it represents the improvement provided by the experimental test. In practical terms, $100\theta_2\%$ of patients who would not otherwise be diagnosed until the second round would instead be diagnosed in the first round by the experimental test.

The η and μ terms are also important in terms of assurance calculations, as they will help determine the proportion of total patients who receive a second round test. Of all patients, $100(1-\eta)\mu\%$ will receive a positive test in round two. It follows then that $100(1-\eta)\mu\theta_2\%$ of the total patients will see an improvement in diagnosis from the use of the experimental test.

Note that while all experts completed the seed questions, only a subset provided values for the parameter elicitations as well. This was a clear disadvantage of methods that require higher levels of input from experts. Those who are busy, and unable to commit larger time periods, may not be able to complete the full elicitation. In total, Experts 3, 6, 8, 9, and 10 completed all questions from both the seed and parameter elicitations. Experts 1 and 7 completed all seed questions, and the η and μ parameter questions, while the remaining experts were only able to complete the seed questions.

4.4.1 SHELF Elicitation

During the SHELF elicitation meeting, Experts 8, 9, and 10 worked together to determine appropriate quartile values for the five parameters. Figure 4.15 provides density plots of the elicited values from the SHELF meeting. Only the experts from Group 2 took part in the SHELF aggregation, as detailed in Chapter 3.

The density plots show that the experts felt that η and μ had a similar range of values, with μ likely being slightly lower. Both θ_1 and θ_2 also strongly overlap, suggesting the experts thought the two parameters would have very similar values. The θ_3 distribution gives weight to lower values than either θ_1 or θ_2.

In terms of the experts' confidence, the η, μ, and θ_3 distributions have similar ranges, while the experts were more confident about the range of values for θ_1 and θ_2.

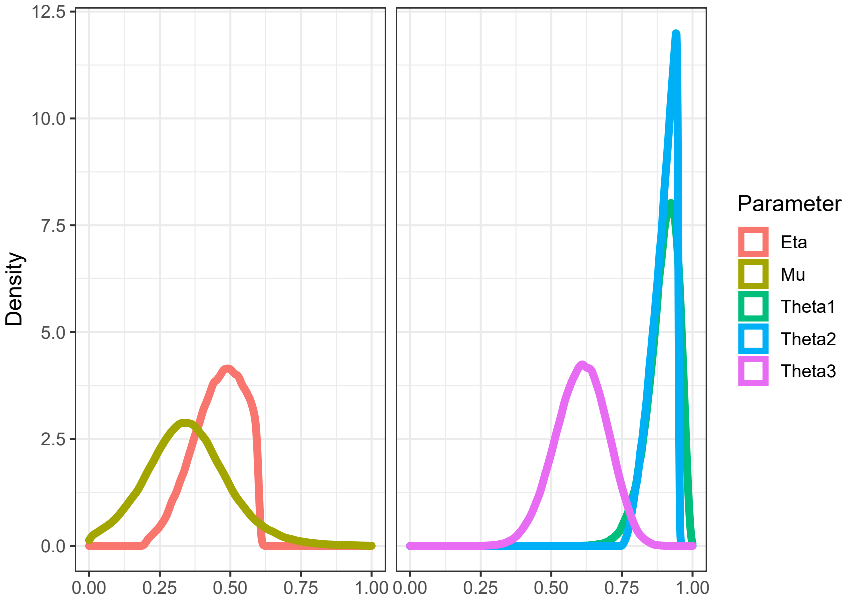


Figure 4.15: Density plots for SHELF aggregation of the parameters.

For comparison with further aggregation methods, we also provide a violin plot for the SHELF elicitation in Figure 4.16. The violin plot presents the density for each parameter in a format similar to a boxplot. It is clear that the SHELF method has provided uni-modal distributions for each parameter, which is to be expected as a uni-modal distribution was fitted in each case to the elicited values. This is in contrast to many of the aggregation methods.

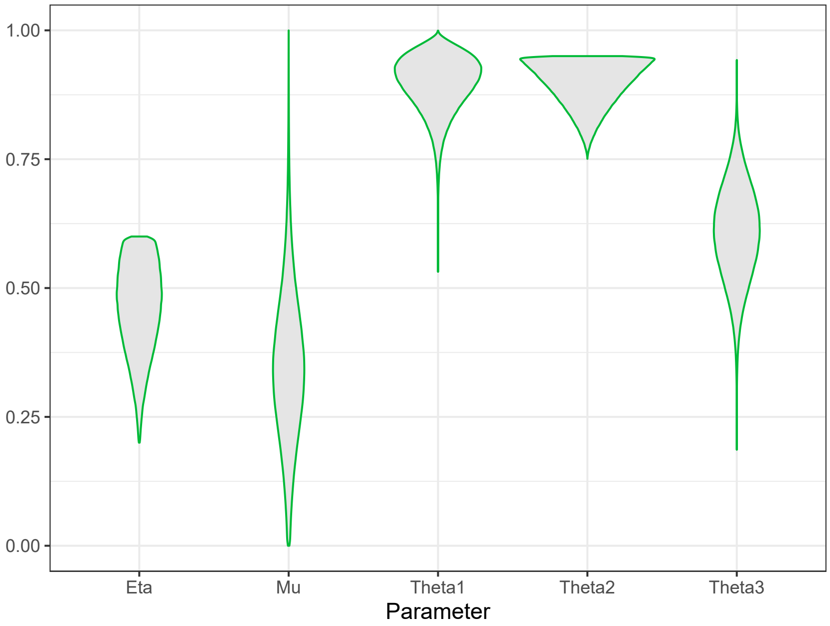


Figure 4.16: Violin plots for SHELF aggregation of the parameters.

4.4.2 Equal Weights

We next consider an Equal Weights aggregation. In this case, equal weight is provided to each expert who provided responses for a parameter.

Figure 4.17 provides the Equal Weights aggregated distributions for the parameters.

Each of the aggregated distributions is multi-modal, though each tends to have one main peak. For θ_1 and θ_2, there seems to be agreement across all experts that the values should be in a higher range. The Group 2 experts have presented more confidence in this, with higher peaks, reflected in both the Group 2 and Group 3 distributions.

There appears to be less cohesive agreement across the other parameters. As such, the combined Group 3 distributions tend to have a peak matching one of the other groups, and then a considerable amount of density spread across a wider range.

Figure 4.18 provides a further comparison of all Equal Weights aggregations in a violin plot.

The η parameter presents a clear example of how the three groups are related. The Group 1 and Group 2 priors have distinct areas of high probability density, around 0.8 and 0.6 respectively. The Group 3 prior, which incorporates all experts across the groups, then gives reasonable probability to both of these possibilities.

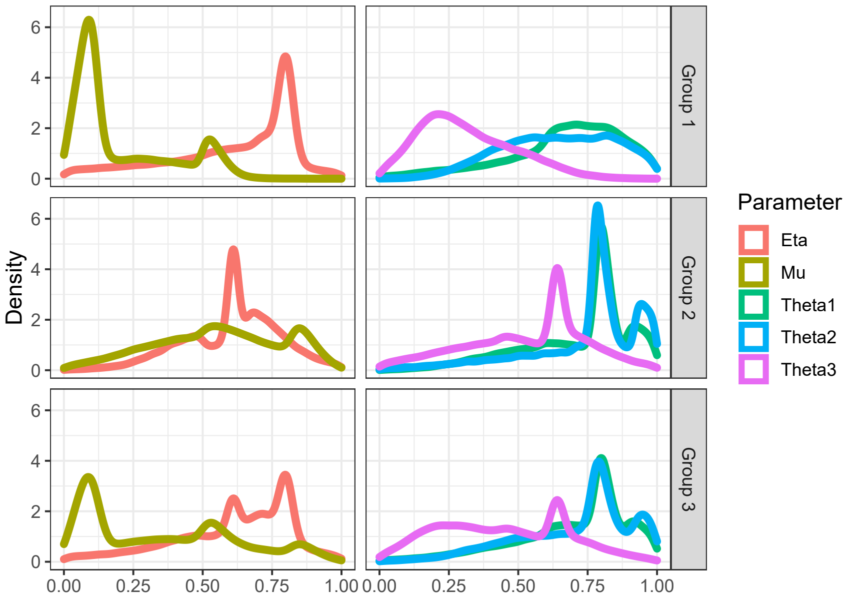


Figure 4.17: Density plots for equal weight aggregations of the parameters.

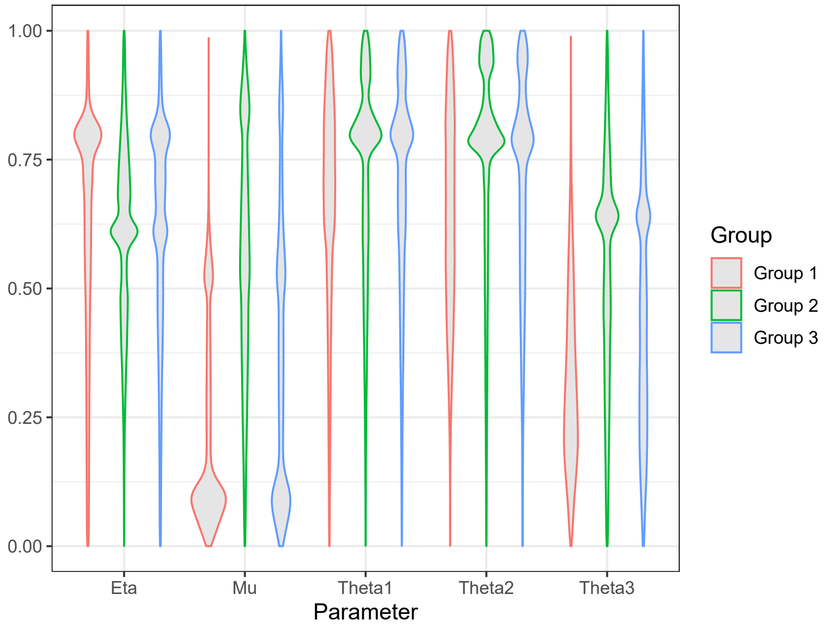


Figure 4.18: Violin plots for equal weight aggregations of the parameters, from all groups.

4.4.3 Classical Method

The Classical Method calculates weights for each expert based on their seed question performance. While the previous analysis of the seed questions took into account all experts, we have recalculated Classical Method weights for the experts who responded to

Table 4.9: Classical Method Weights for Parameter Aggregation

Group	1	3	6	7	8	9	10
1	0.00	0.61	0.24	0.15	-	-	-
2	-	-	-	-	0.00	0.73	0.27
3	0.00	0.22	0.18	0.05	0.00	0.40	0.16

the parameter questions.

Table 4.9 provides the new weights for this subset of experts. As previously organised, Group 1 contains Experts 1, 3, 6, and 7, Group 2 contains Experts 8, 9, and 10, and Group 3 contains all experts.

In the previous Classical Method aggregation, Experts 2, 3, and 6 were assigned the non-zero weights for Group 1. It can be seen here that without Expert 2's presence, Expert 3 has been assigned a higher weight compared to Expert 6.

The weights for Group 2 are the same as in Table 4.4, as the same experts and seed questions are used in each.

For the Group 3 aggregation, the weight is split quite evenly between Group 1 and Group 2 experts. This suggests that both groups had members who performed well in the seed questions. Furthermore, and unlike the previous Classical Method aggregation, there is no one expert who dominates the aggregation with the majority of the weight. As such, this aggregation will take into account the views of a wider range of experts.

The Classical Method's aggregated distributions are presented in Figure 4.19. As shown, many of the distributions are multi-modal, as they are incorporating multiple experts' prior distributions. Those distributions with fewer modes form when the experts' individual distributions were largely overlapping.

There is a strong level of agreement between the θ_1 and θ_2 parameters. This was common in many of the experts' individual distributions, and so it would seem desirable for an aggregation method to reflect this too.

The distributions are also presented in a violin plot, shown in Figure 4.20. This plot tends to show a stronger similarity between Group 2 and Group 3 prior distributions. While there is a slight bias towards Group 2 in the weights, this is also likely in part due to Expert 9 having a larger weight than the other experts. As such, there are similar peaks in Group 3 as in Group 2, while the Group 1 experts have tended to increase the density in the tails of the Group 3 prior.

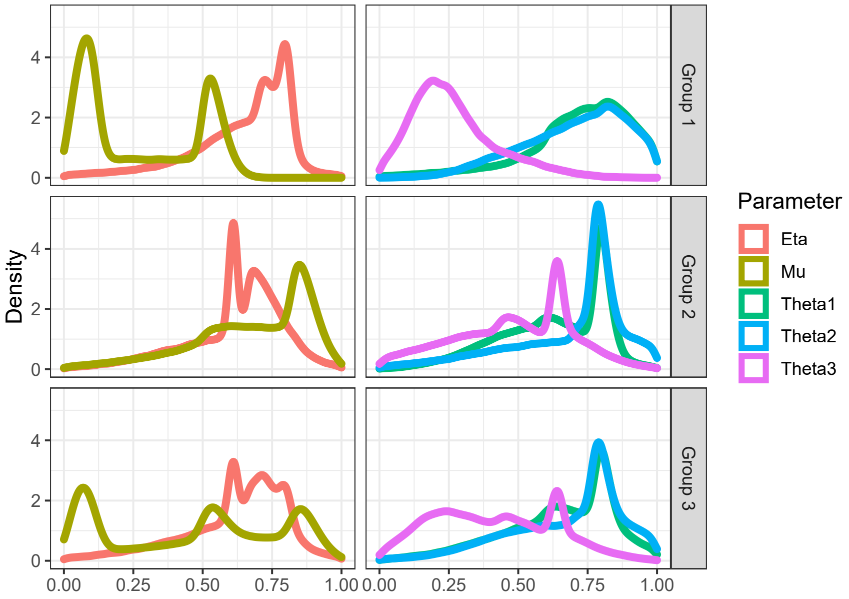


Figure 4.19: Density plots for Classical Method aggregations of the parameters, from experts in Group 1.

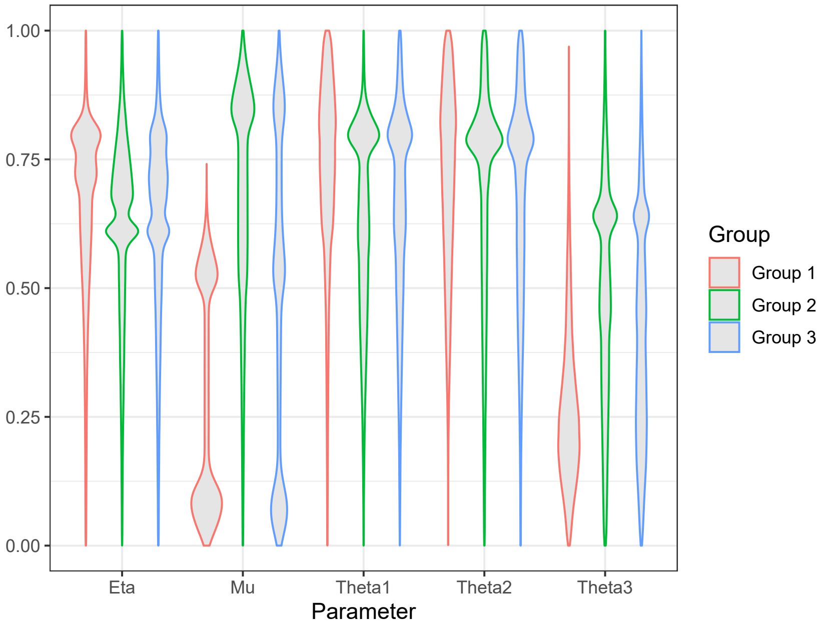


Figure 4.20: Violin plots for classical method aggregations of the parameters, from all groups.

4.4.4 Bayesian Aggregation

Aggregated distributions, representing the opinions of a decision maker, have also been found using the Bayesian aggregation method. As the parameters are all bounded between zero and one, it was important to take this into account in the hierarchical model. A logit transform was used to convert the experts' elicited values, p, to a scale between negative and positive infinity, X, as follows

$$X = \log\left(\frac{p}{1-p}\right) \tag{4.21}$$

The Bayesian aggregation was then performed on the rescaled values X, and then converted back to the original scale using the reverse transformation

$$p = \frac{e^X}{1+e^X} \tag{4.22}$$

This ensured that the final aggregation would be bounded between zero and one.

Figure 4.21 provides the density plots of the aggregated distributions for each group. There was agreement between each group for η, θ_1, and θ_2. All groups provided reasonably wide distributions for θ_3, with Group 1 providing a more positively skewed distribution ;than Group 2. Group 1's aggregated distribution for μ was also much more positively skewed than Group 2's. Overall the resulting distributions were consistent with patterns in the individual expert distributions, and other aggregations.

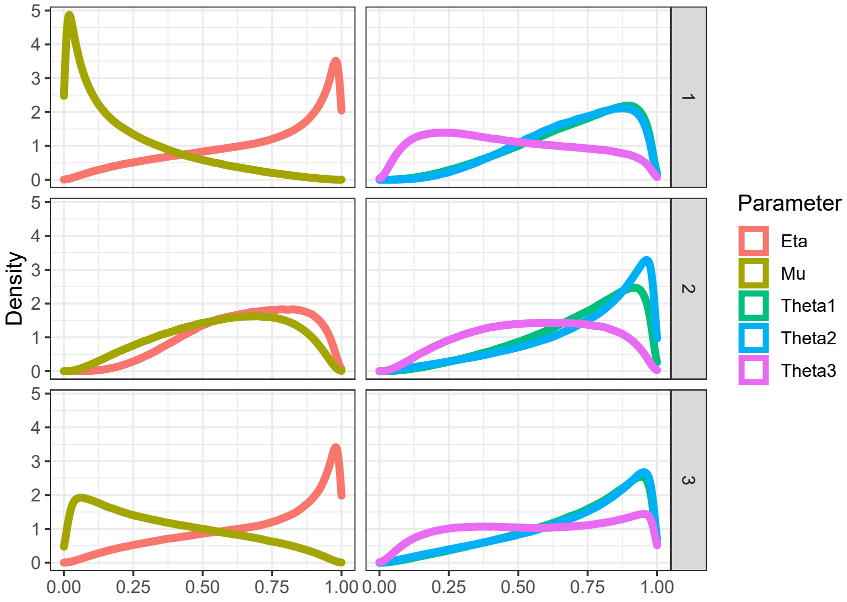


Figure 4.21: Density plots for Bayesian aggregations of the parameters, from all experts.

The distributions are also presented as violin plots in Figure 4.22. This plot highlights the similarities in the θ_1, θ_2, and θ_3 parameters across groups. It also shows that, in comparison to the other aggregation methods, the Bayesian aggregation has resulted in more dispersed distributions, without a sharp peak.

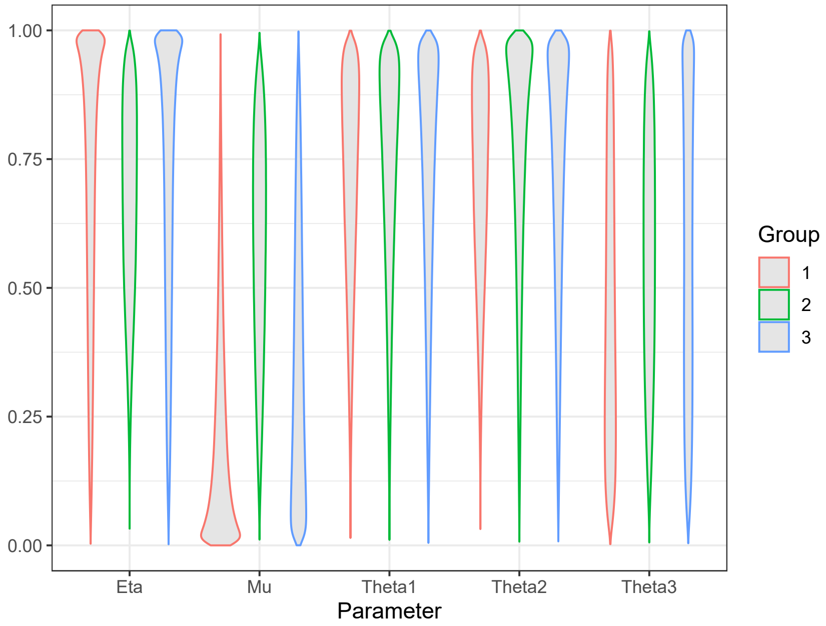


Figure 4.22: Violin plots for Bayesian aggregations of the parameters, from all experts.

Figure 4.23 provides the density plots of the extended Bayesian aggregated distributions for each group. There appears to be strong agreement between groups for the η, θ_1, and θ_2 parameters in this case. The μ and θ_3 parameters have opposite skews for Group 1 and Group 2. This has resulted in a more uniform distribution for θ_3. It appears that the mode in the aggregated distribution of μ from Group 1 was high enough that it carried over to Group 3, resulting in a similar skew, but with more uncertainty.

While the μ distribution for Group 1 has a higher density for lower values, overall the aggregated distributions appear similar to those from the original Bayesian aggregation model. Each distribution takes a similar shape, suggesting the extended model has not resulted in a large change in the final aggregations.

In comparison to previous parameter aggregations, these tend to present vaguer priors, with lower peaks. Each distribution is also uni-modal, which is to be expected as each is being modelled by a split-normal distribution which has a single mode.

The distributions are also presented as violin plots in Figure 4.24. This plot, especially in comparison to previous aggregations, again shows smooth uni-modal distributions for each parameter.

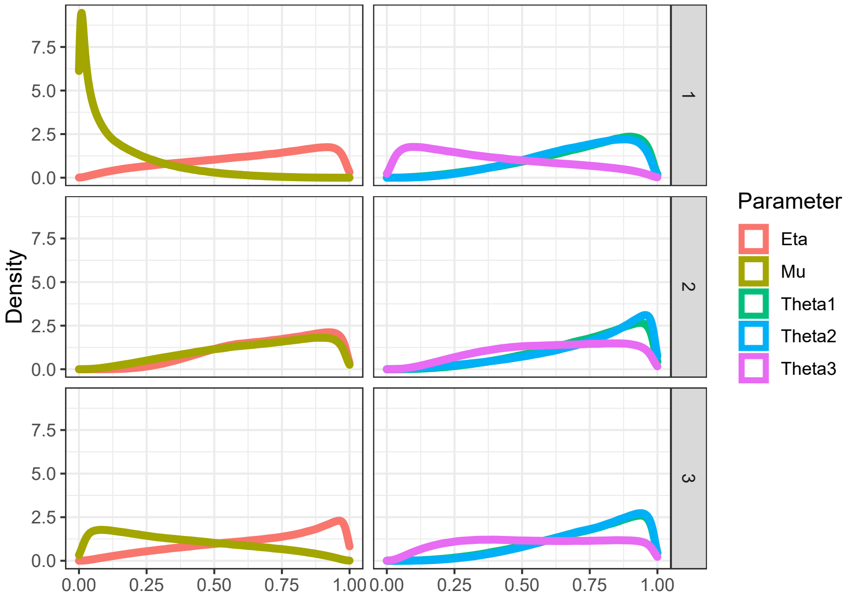


Figure 4.23: Density plots for Extended Bayesian aggregations of the parameters, from all experts.

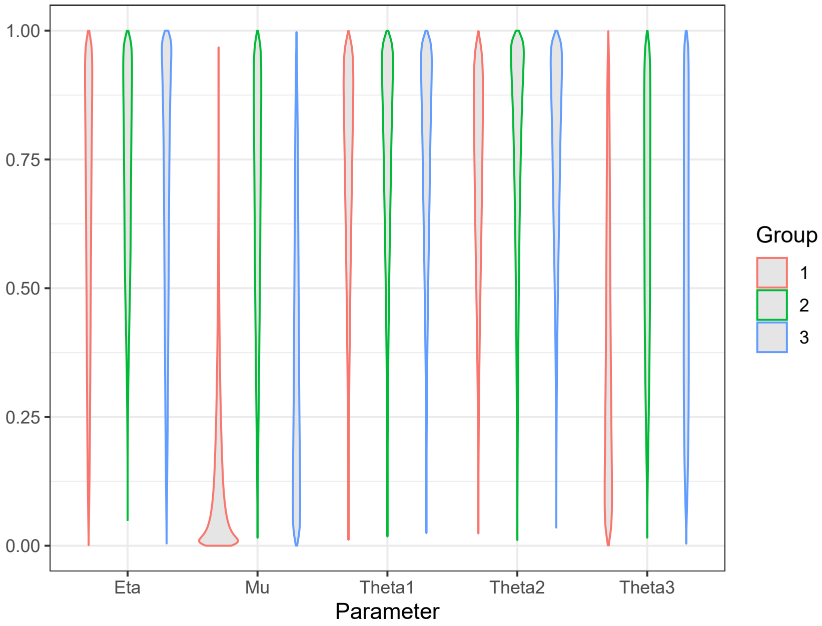


Figure 4.24: Violin plots for Extended Bayesian aggregations of the parameters, from all experts.

4.4.5 Aggregated Prior Distribution Comparisons

For ease of comparison, Figure 4.25 provides density plots of each of the above aggregation methods, separated by parameter and expert group. As discussed in Section 4.2.2, certain patterns were present between the individual experts' distributions for each parameter.

The η distributions tended to agree that the median lay between 0.5 and 0.7, and had an upper quartile around 0.75. The μ parameter tended to have medians and quartiles lower than the η parameter for most experts. For the θ parameters, θ_1 and θ_2 had very similar distributions for each expert, and θ_3 consistently took lower values than θ_1 and θ_2.

This was echoed in the SHELF aggregation, where the experts providing the elicited values were able to ensure the practical interpretations for each parameter made sense. As shown in Figure 4.16, the η distribution placed density on higher values than the μ distribution, and the θ_1 and θ_2 distributions were very similar, and both placed density on higher values than the θ_3 distribution.

These patterns between parameters at the individual level suggest the experts consistently felt there were certain relationships between the parameters. As such, if a mathematically aggregated parameter distribution did not match this pattern, it suggests there may be an issue with its practical interpretation.

Violin plots for the Equal Weights aggregation, Figure 4.18, Classical Method, Figure 4.20, and Bayesian aggregation, Figure 4.22, all strongly show this pattern for each parameter in Group 3. There are some minor differences, however, between Group 1 and 2.

Especially notable in Group 2 when aggregating via the Classical Method, μ has more density placed higher than η. This is reflective of Expert 9, who provided a prior for μ with density at higher values than their prior for ηas, having a greater influence on this aggregation.

As there are not currently results from the study, the distributions cannot be assessed against any results. However, it appears that from the practical interpretation of the aggregated distributions compared to the individual expert distributions, the aggregated distributions provide sensible and consistent prior distributions.

These aggregated prior distributions will be used in Chapter 6 to calculate assurance in order to determine appropriate sample sizes for the study.

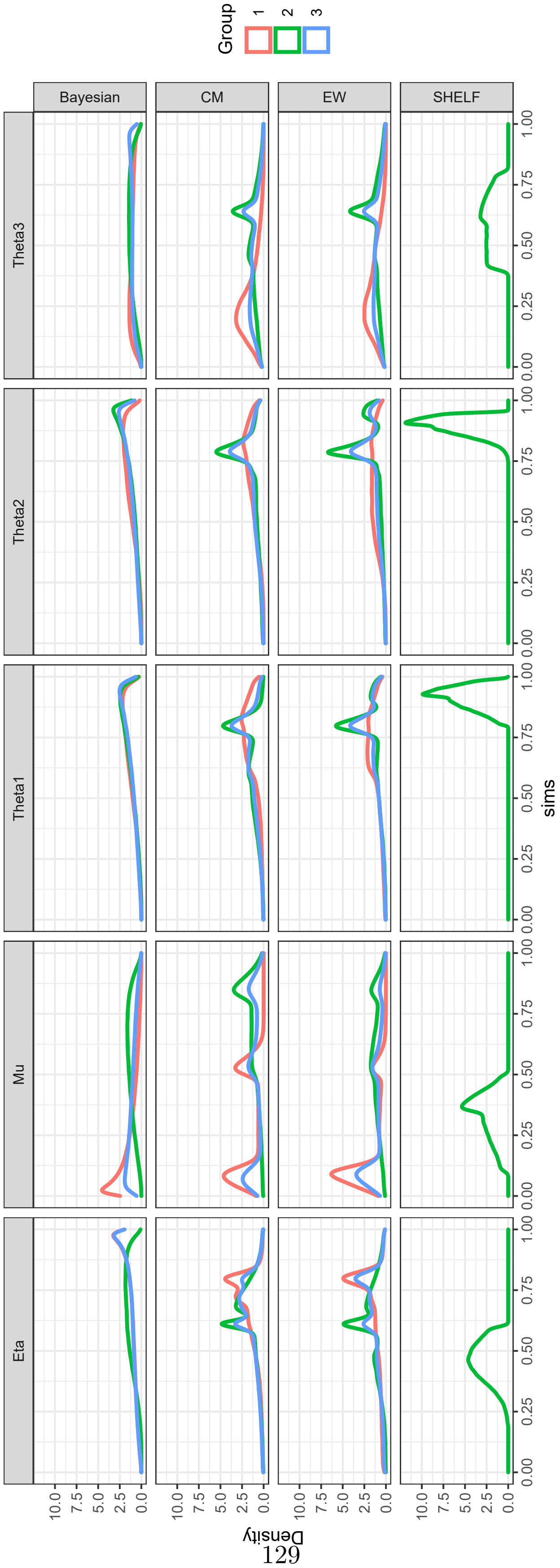


Figure 4.25: Density plots of all aggregations for each parameter, by each expert group.

4.5 Conclusion

In this chapter, we have presented the results of elicitations for the design of a clinical study into a novel diagnostic test for Motor Neurone Disease. As part of the elicitation, we elicited prior distributions on a number of seed questions and model parameters. We presented these results, and found that the experts involved in designing the novel diagnostic test and study appeared no more confident than the experts without involvement.

Using the seed question results, commonly used behavioural and mathematical aggregation methods were compared. It was demonstrated that any method of aggregation was preferable to eliciting from a single expert, and that the SHELF and Classical Method both performed better than other aggregation methods.

We also presented aggregated distributions for the model parameters, which will be used in Chapter 6 as the basis for assurance sample size calculations.

Part V

Comparing Assurance and Power through Simulation

Chapter 5

Comparison of Assurance and Power through Simulation

5.1 Introduction

In this chapter, we explore assurance through simulation and make comparisons to power.

We begin by defining the normal and binomial models to be used throughout this chapter. We then present some simulations of power and assurance, showing how they change as the sample size or other values change.

The following sections then further investigate the behaviour of assurance and power under different circumstances. We consider the following:

- How assurance and power are related for similar calculation inputs
- How the limits of assurance change for different inputs
- Assurance calculations with different analysis prior distributions
- The sensitivity of power and assurance to overestimated effect sizes
- Extending the normal model to consider unknown population variance

We conclude by considering the implications of these simulations for the use of assurance in practice.

5.2 Assurance and Power Simulations

Both assurance and power are used in a similar way when determining a sample size. There are, however, a number of differences in how the two methods perform in different situations.

Previous studies have compared assurance and power. Chen and Fraser (2018), for example, note that assurance is typically lower than power for similar cases, as it considers a wider range of effect sizes including those which represent a prior probability of the treatment being ineffective. Additionally, Ring et al. (2019) compare sample sizes determined by assurance and power, suggesting that sample sizes in the usual range for Phase 3 trials tend to have an assurance of around 70 to 80%.

In this section, we provide some initial examples of power and assurance curves, and how they relate to each other.

5.2.1 Terminology and notation

Throughout the simulations in this chapter, we will use certain values as standard. For frequentist tests and power, we focus on a significance level of $\alpha = 0.05$, and a power of 0.9 unless otherwise stated.

We also focus on cases where data comes from either a normal or binomial distribution, to correspond with standard simple scenarios in medical studies and the case study.

In the case of a random sample $X_1, \dots, X_n$ from a Normal distribution, we consider a one sample Z-test.

For a two sided Z test, the null and alternative hypotheses are

$$H_0 : \mu = \mu_0 \tag{5.1}$$

$$H_1 : \mu \neq \mu_0 \tag{5.2}$$

where mu is the population mean and μ_0 is the value of μ to be tested against in the null hypothesis, most commonly zero. The Z score is calculated as

$$Z = \frac{\bar{X} - \mu_0}{s} \tag{5.3}$$

where $\bar{X}$ is the sample mean, and s is the standard error of the mean. The value of s can be calculated given the population standard deviation σ, and the sample size n, using

$$s = \frac{\sigma}{\sqrt{n}} \tag{5.4}$$

The p-value of the test is then found by comparing the Z-score to a $N(0, 1)$ distribution.

If we consider the same random sample from a normal distribution, a suitable Bayesian model is as follows. We assume that the observations, X_i, $i = 1, \dots, n$, are normally distributed and give the mean a hyper-prior which is also normal.

$$X_i \sim N(\theta, \sigma^2) \tag{5.5}$$

$$\theta \sim N(\mu, \gamma^2) \tag{5.6}$$

We can calculate the posterior probability that the mean is greater than a value μ_0, $P(\theta > \mu_0 \mid X)$, and consider it a significant result when $P(\theta > \mu_0 \mid X) < 0.025$ or $1 - P(\theta > \mu_0 \mid X) > 0.975$.

We first assume the population standard deviation is known and equal to one, $\sigma^2 = 1$. We relax this assumption in Section 5.7 to explore the effect of an unknown population variance. It is important to note that there are two different standard deviations used in these calculations. The population standard deviation, which has been set to equal one, represents the aleatory variability in the observations, which we simulate as $X_i \sim N(\theta, 1)$. The standard deviation in the prior for θ, labelled γ, is our epistemic uncertainty, representing our lack of knowledge on the mean.

For power calculations, the effect size is chosen to be equal to μ when comparing against assurance.

For binomial observations, we consider an exact binomial test. Suppose we will make an observation $X \sim \text{Bin}(n, \theta)$, and have a null hypothesis that the binomial probability is equal to or less than a particular value, $H_0 : \theta \leq \theta_0$, and an alternative hypothesis $H_1 : \theta \geq \theta_0$, we can directly calculate the probability of observing k successes, or a more extreme result, assuming this null hypothesis is true. That is,

$$P(X \geq k) = \sum_{i=k}^{i=n} \binom{n}{i} \theta_0^i (1 - \theta_0)^{n-1} \tag{5.7}$$

While this test can be approximated using a normal approximation for larger sample sizes, as we wish to also consider smaller sample sizes, we will use this test throughout.

For a Bayesian analysis, we consider the following model setup. An observation, X, comes from a binomial distribution with parameter θ, on which we place a Beta analysis prior with parameters α and β.

$$X \sim Binomial(n, \theta) \tag{5.8}$$

$$\theta \sim Beta(\alpha, \beta) \tag{5.9}$$

We then consider the posterior probability that the parameter θ is greater than a chosen value θ_0, ie $P(\theta > \theta_0 \mid X)$, and consider the result significant when $\Pr(\theta > \theta_0 \mid X) \geq 0.95$.

Additionally, unless otherwise stated, we are considering a two sided Z-test and a one sided binomial test.

We refer to the input to power calculations as an effect size. This refers to the value used for $\bar{X}$ for the Z-test, and the estimate of θ for the binomial test. As noted in Chapter 2, there are a number of ways this value can be determined, such as the MCID.

We also refer to the design prior's mean as a best estimate. In cases where power and assurance are compared for a frequentist analysis, the same value is used for the effect size in the power calculation and the best estimate in the assurance calculation.

When simulations are used for assurance calculations, unless specified elsewhere, 100,000 replications were used to estimate the value.

5.2.2 Normal observations

For an initial comparison between power and assurance, we take the case of a Z-test. Figure 5.1 provides example power and assurance curves for five different values of the effect size and best estimate as the sample size n increases.

For the following simulations, we set a prior distribution for θ as $\theta \sim N(\mu, 0.5^2)$.

These plots show an example of the common shapes of assurance and power curves. For set population and prior standard deviations, a higher effect size or best estimate corresponds to higher power and assurance values regardless of n. Additionally, in both cases, as n increases, so too do the power and assurance.

While the power curves vary largely between the different effect sizes in terms of slope, the assurance curves follow a much more similar shape, albeit shifted up or down. This is likely due to the additional variability which is accounted for within the assurance calculation.

A case of interest is that where the effect size is equal to the value assumed under the

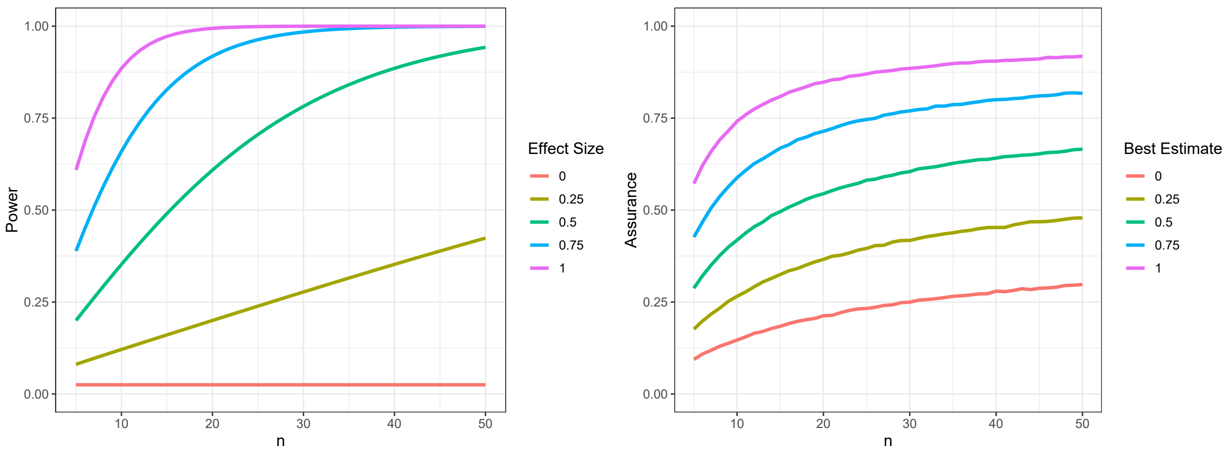


Figure 5.1: Example power (left) and assurance (right) curves for a Z-test.

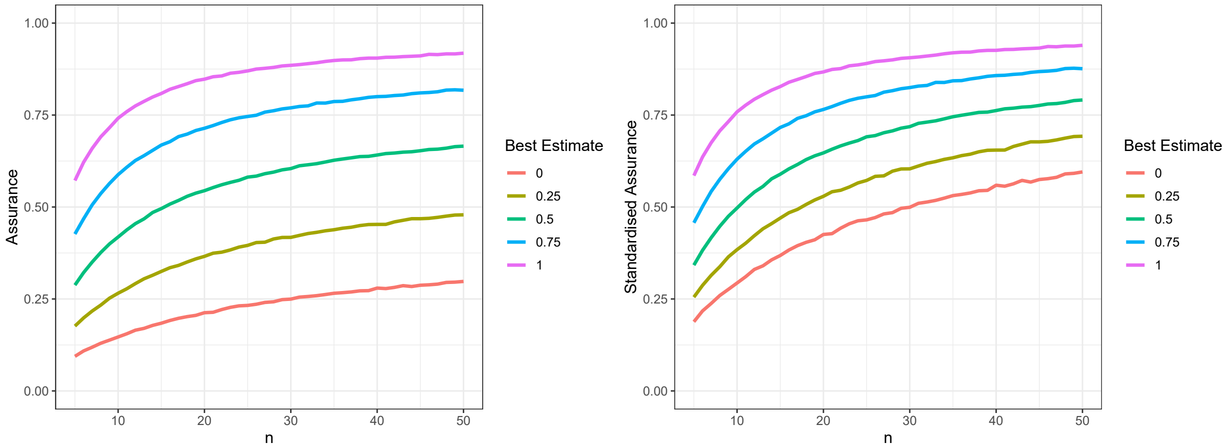


Figure 5.2: Example assurance and standardised assurance for a Z-test, with varying design prior means.

null hypothesis, in this case zero. Regardless of the sample size, the power of such a trial will always be equal to α, in this case 0.05. The assurance for such a case does change as n changes, as it also considers values other than zero as a result of the uncertainty from the prior distribution placed on the best estimate.

As mentioned in Chapter 2, we can also consider the assurance in terms of its maximum possible value. This standardised assurance is calculated by dividing the assurance by the maximum possible assurance, as defined by the design prior. Figure 5.2 shows the effects of standardising the assurance compared to the original unstandardised assurance.

These new curves represent how quickly the assurance is approaching its maximum. Each curve on the plot has its own maximum assurance, which is defined by the design prior it represents. As such, each curve is scaled by a different amount.

When a design prior leads to a maximum assurance of closer to one, such as the effect size of one in this plot, this transformation will have less of an effect. Those curves which tend towards values that are lower than one, such as the effect size of zero, will have a greater transformation effect.

The benefit of this type of plot is that it allows comparisons of assurances that were previously on different scales. While the initial plot showed the assurances were all converging to different values, it is difficult to tell which values they are converging to, and how quickly. The standardised assurance, however, shows that the higher the best estimate, the more quickly the convergence at lower values of n. However, for the larger values of n, the slopes for the higher best estimates become more flat more quickly than those of the lower best estimates.

We can also consider the effect of the prior standard deviation on assurance calculations. Figure 5.3 shows the assurance and standardised assurance curves for a range of

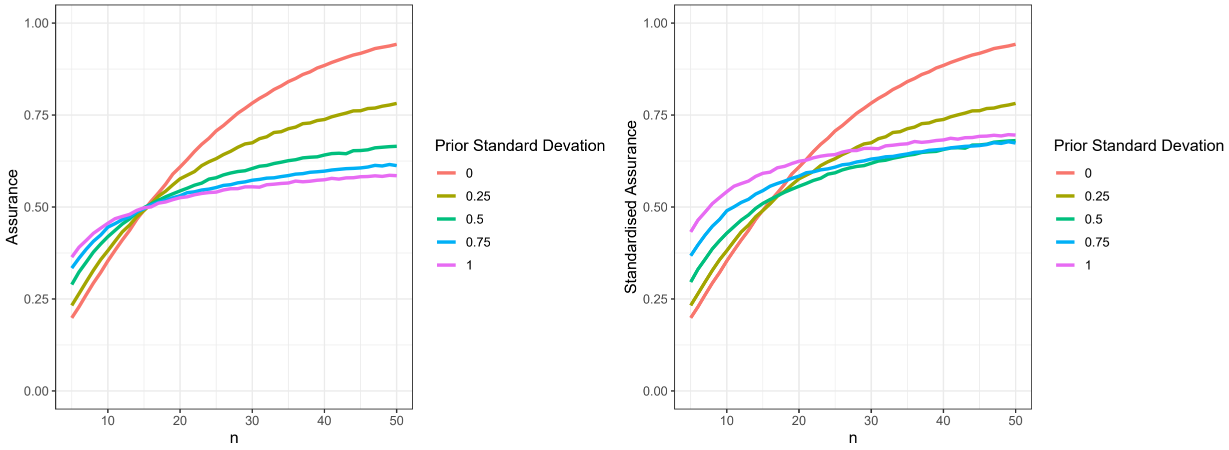


Figure 5.3: Example assurance and standardised assurance for a Z-test, with varying design prior standard deviations.

different prior standard deviations, with a prior mean of 0.5.

In this case, it can be seen that varying the standard deviation while holding the effect size constant changes the slope of the assurance curve.

The curve where the prior standard deviation is equal to zero will match the power curve, when the best estimate is equal to the effect size.

Of note is the point where the lines converge, at an assurance of 50%. This behaviour is consistent for different best estimates, though the associated sample size n varies. The reason this point exists is because it is the point where the critical value for the test is equal to the effect size.

For a two sided Z-test, the critical value is defined as $1.96\frac{\sigma}{\sqrt{n}}$ for $\alpha = 0.05$. When the value of n is approximately 16, this critical value is equal to 0.5. As the effect size is equal to this value, and the design prior is symmetrical, then half of the prior's weight will be above and below the critical value regardless of the prior's standard deviation.

These plots could be used to determine an appropriate sample size for a trial. For example, if a power of 0.8 was desired, then a sample size of 25 would be required for a corresponding effect size of 0.5. Likewise, the assurance gives a 60% chance of a trial of this size providing a significant result.

5.2.3 Binomial observations

We next consider binomial observations. The mean and variance of a binomial random variable are determined by the sample size, n, and probability of success, θ.

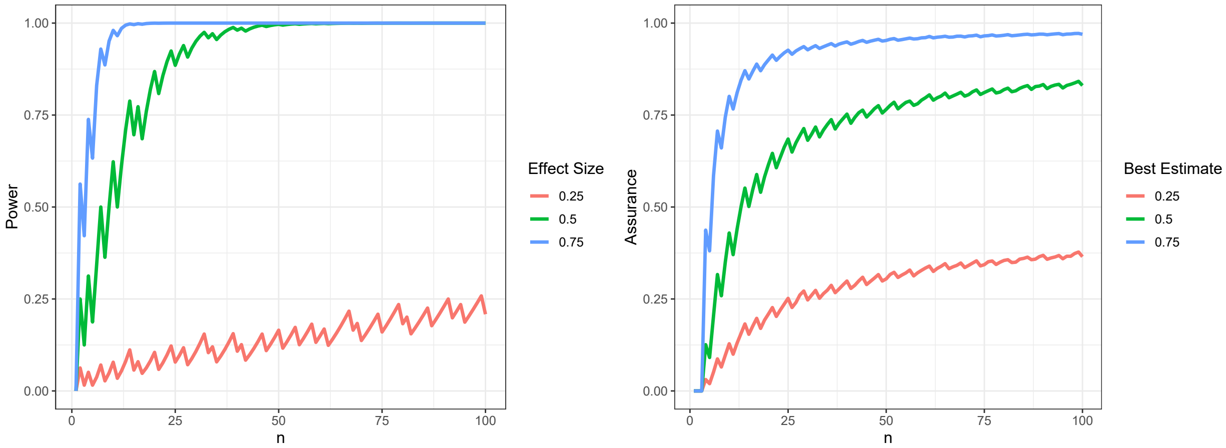


Figure 5.4: Example power and assurance curves for a binomial test.

$$\mu = n\theta \tag{5.10}$$

$$\sigma^2 = n\theta(1-\theta) \tag{5.11}$$

Figure 5.4 provides some example power and assurance curves for an exact binomial test. The exact binomial test has a null hypothesis of $H_0 : \theta = 0.2$, and a one-sided alternative hypothesis. The design prior takes the form of a $Beta(\alpha, \beta)$ distribution, where α and β are chosen to ensure the distribution has a mean given by the best estimate, and a standard deviation of 0.2. For a Beta distribution, this can be achieved using the following equations.

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)\mu^2 \tag{5.12}$$

$$\beta = \alpha\left(\frac{1}{\mu} - 1\right) \tag{5.13}$$

In both cases, a pattern of steps appears. In order for a significant result to be found for a binomial test, a certain number of positive observations must be observed. In some cases, increasing the sample size by one requires an extra ‘success’ to be observed, leading to a decrease in the probability of observing the required number of positive observations in comparison to the lower sample size. However, in other cases the required number of positive observations remains constant for an increase in total sample size by one, which then has a higher probability of being achieved with the larger sample size. Each step in the plot represents the point where an additional positive observation is required.

As the sample size increases, so too does the power. However, as demonstrated by the

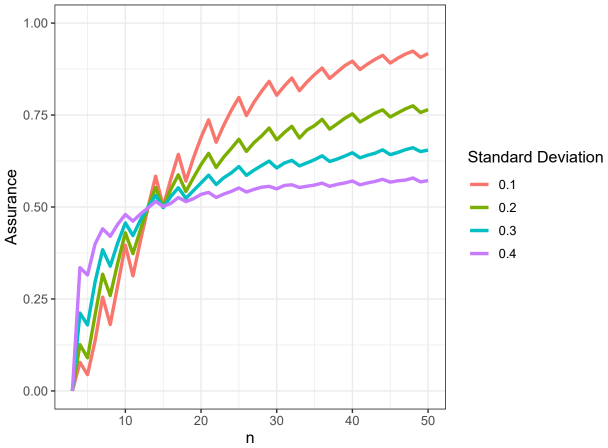


Figure 5.5: Example assurance curves for a binomial test.

red power curve, the convergence of power towards one varies with the effect size, and is not always quick.

Both power and assurance start at a value of zero for a small sample size. For the lowest effect size, assurance increases more rapidly than power. This is likely due to the incorporation of higher effect sizes in the prior distribution, which are not included in the power calculations.

The assurance plot shows the assurance curves converging to different locations, depending on the best estimate. Best estimates further away from the null hypothesis value lead to a trial having a greater chance of success. This is logical, as larger effect sizes should tend to be more easily detected than smaller ones.

We can also consider the effects of a change in the variance of the design prior. Figure 5.5 plots assurance curves where the design prior has a mean of 0.5, and standard deviations ranging from 0.1 to 0.4.

As seen in the Z-test case, the higher standard deviations correspond to lower assurance levels for larger values of n. Also similar is the overlap in assurance curves when assurance is equal to 0.5.

The size of the changes in the step pattern become smaller for the larger standard deviation. These priors are taking into account a wider range of values, and appear less sensitive to the requirement of an additional success.

A common pattern in all of these plots is that the jumps in assurance or power occur at the same sample sizes. This is because the required number of successes for a significant result is not dependent on the design prior. Instead, the design prior affects how likely it is to observe the required number of successes for each possible sample size.

5.3 How Power and Assurance are related

The assurance based on a hypothesis test is the average power over a prior distribution for the effect. As such, the selected prior distribution will influence how similar the assurance and power are to each other. In this subsection, we explore the relationship between power and assurance for different combinations of parameter values.

5.3.1 Normal observations

In order to examine this, we first consider a Z-test with varying sample sizes values, $n = 1, \ldots, 100$, and varying effect sizes and best estimates, between 0 and 1. The standard deviation of the design prior is set to 0.5, half of the population standard deviation. The results from these simulations are plotted in Figure 5.6.

As the plot shows, larger prior means relative to the standard deviation correspond with a more linear relationship between power and assurance. The beginnings of the lines in the lower-left corner of the plot correspond to the lowest sample size values. The sample size values then increase as the curves move up and towards the right. The diagonal black line represents the points where assurance and power are equal. In this case, it is clear that for low sample sizes, the assurance is greater than the power. The power is equal to assurance at 0.5, after which the power is greater than assurance.

While these curves demonstrate how a different effect size or best estimate can change the relationship between assurance and power for a set standard deviation, it is also important to consider the effect of varying the standard deviation. For an effect size set at 0.5 for the power calculation, and a design prior centred on 0.5 for the assurance, we again consider sample sizes between 1 and 100. We calculate the power and assurance for each, varying the design prior's standard deviation. This is provided in Figure 5.7.

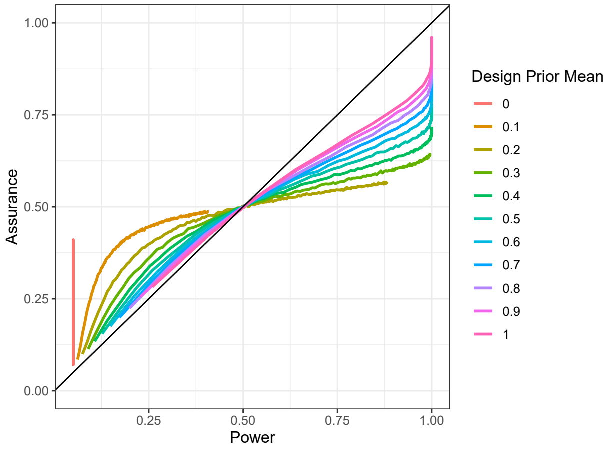


Figure 5.6: Power and assurance for a range of different effect sizes, over sample sizes 1 to 100.

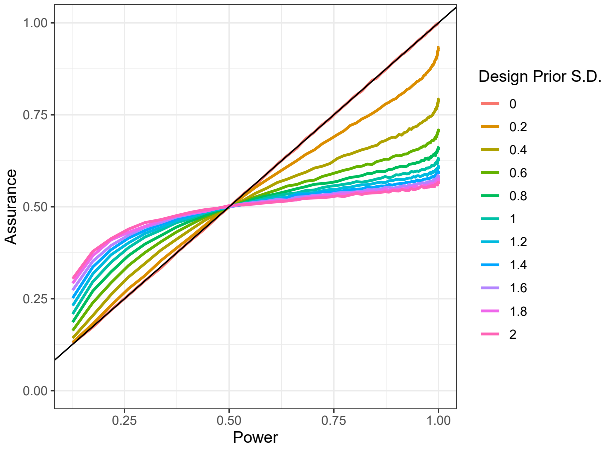


Figure 5.7: Power and assurance for a range of different design prior standard deviations, over sample sizes 1 to 100.

As this shows, when the design prior has a very low standard deviation, or the prior is very informative, the relationship between power and assurance is almost linear with a slope close to one. In the case where the design prior's standard deviation is equal to zero, then the assurance and power are equivalent.

The more uncertainty on the parameter of interest, the further the assurance changes from the power. This is because the prior is taking into account a wider range of values as possible effect sizes.

Common to both plots is that the power and assurance coincide at the value 0.5. As mentioned in Section 5.2.2, the assurance will have a sample size for which it has a value of 0.5, regardless of design prior standard deviation. As the power is equivalent to an assurance with a design prior standard deviation of zero, then it will be the case that the assurance and power will be equal at this point.

It is also noted that for values below this point, the value of the assurance will be greater than or equal to the power, and for values greater than 0.5, the power will be greater than or equal to the assurance.

By comparing the previous two plots, it can be seen that the difference between the power and assurance for a particular sample size varies based on the design prior used. While there is a similar shape in the curves when fixing either the design prior mean or standard deviation, the difference comes from the probability assigned to values above or below the null hypothesis.

Figure 5.8 shows a plot where the design prior mean is equal to the design prior standard deviation. In each case, the null hypothesis value of zero is one standard deviation away from the design prior mean and so, there is a consistent probability of the mean being greater than the null hypothesis value in the design prior. In this case, the relationship

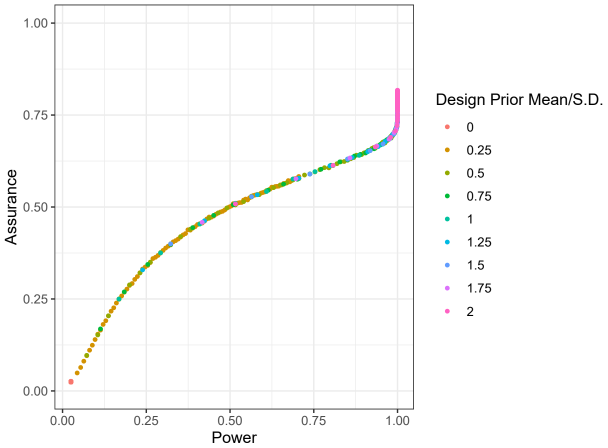


Figure 5.8: Power and assurance for a range of different design prior means and standard deviations, over sample sizes 1 to 100.

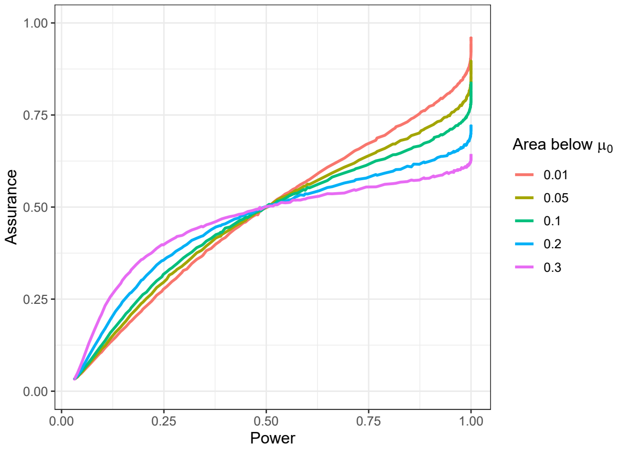


Figure 5.9: Power and assurance for a range of different design prior means and probabilities below μ_0, over sample sizes 1 to 100.

between power and assurance is the same for all design prior parameter combinations.

We can also consider the cases where the mean and standard deviation combinations are chosen to give a certain probability below the value in the null hypothesis, μ_0. Figure 5.9 demonstrates the relationship between power and assurance for design priors with different probabilities assigned to values less than μ_0.

This plot demonstrates that the more probability the design prior provides to values above μ_0, the more linear the relationship between assurance and power will be. It also further demonstrates the relationship between the design prior and the assurance. Design priors placing higher probabilities on a positive trial result in turn lead to higher assurance.

Overall, the more uncertainty within the design prior, the lower the assurance values will tend to be. If the design prior is very certain that there will be a successful result,

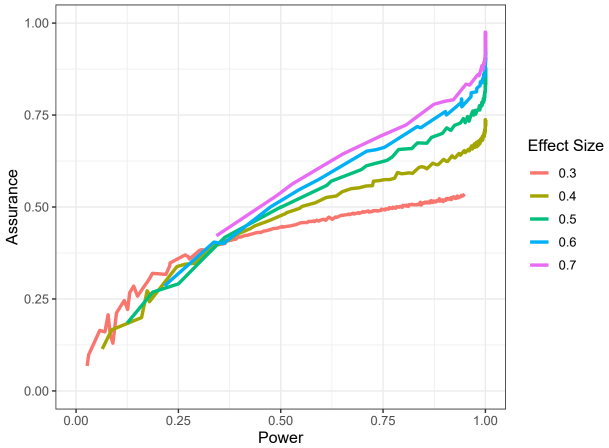


Figure 5.10: Power and assurance for a range of different design prior means, and a standard deviation of 0.2, over sample sizes 1 to 200.

then unsurprisingly, this will be reflected in higher assurance values.

5.3.2 Binomial observations

Similar results are obtained when using an exact binomial test. Figure 5.10 provides a plot of assurance and power for effect sizes with varying sample sizes.

Unlike the Z-test curves, in this case the relationship is less smooth. This is due to the binomial data being discrete, and the corresponding cutoff for a successful trial also being a discrete value.

As before, however, a pattern emerges of higher effect sizes or best estimates for a given sample size and power corresponding to a higher level of assurance.

When varying the standard deviation, Figure 5.11 presents relationships between power and assurance. For the right hand plot, the case where a mean of 0.5 is present, a similar relationship to the Z-test case is found, where assurance and power are equal at a value of 0.5. However, for a mean of 0.25, and other means not equal to 0.5, a different pattern is found. This is due to a symmetrical distribution present in the first design prior, similar to the normal distributions used for the Z-test. When the mean of a beta distribution prior is not equal to 0.5, the beta distribution is not symmetrical as it is bounded between zero and one.

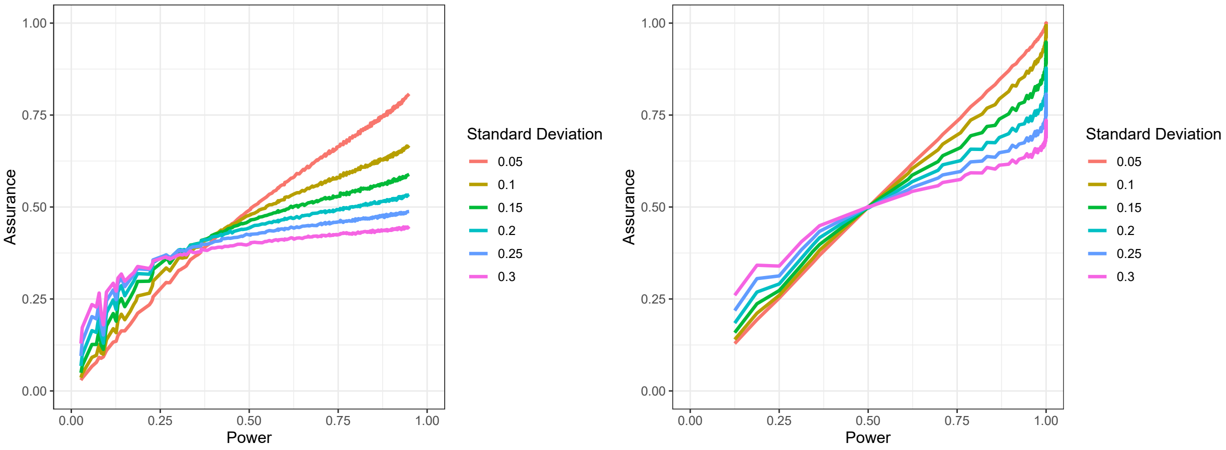


Figure 5.11: Power and assurance for a range of different design standard deviations, with means of 0.25 (left) and 0.5 (right), over sample sizes 1 to 200.

5.4 Assurance for Bayesian Analyses

When calculating assurance when a Bayesian analysis will be conducted, there are two steps in which a prior distribution is required. The design prior refers to the prior used in the assurance calculation, and the analysis prior refers to the prior used in the analysis stage.

As reviewed in Chapter 2, these two priors do not have to be identical. In many cases it will make sense to use different priors to represent the views of different individuals or groups.

5.4.1 Conjugate Analysis Priors

While the design prior can be chosen to represent the beliefs of the researchers about a future trial, the analysis prior is used to analyse the data. In order to ensure this analysis is acceptable to third parties such as funding or regulatory bodies, the analysis prior needs to be justifiable. If an informative analysis prior which is optimistic about the trial outcomes is used, it could be argued that any positive results from the trial are due to this prior rather than the data.

One option would be to use a sceptical prior. Such a prior represents the view of someone who is sceptical about the trial being a success, and thus places a low probability on such an event occurring.

Normal observations

To investigate the effects of the analysis prior on assurance calculations, we consider a Z-test.

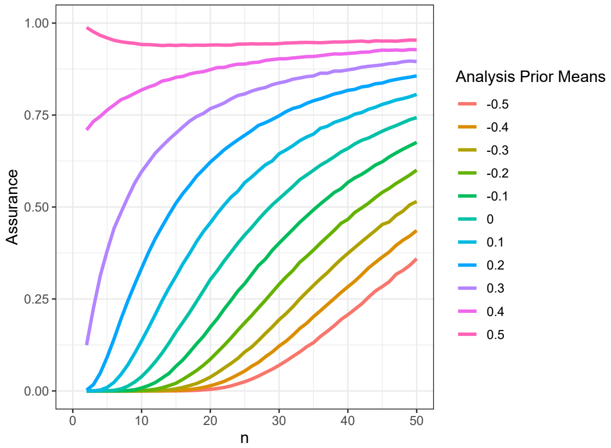


Figure 5.12: Assurance curves with analysis priors with various means.

First, we simulate the case where we vary the analysis prior mean. The design prior is set as a $N(0.5, 0.2)$ distribution, and the analysis prior as a $N(\mu, 0.1)$ distribution, where μ is varied. A successful result will be determined when 95% of the posterior distribution is above zero. While a larger analysis prior variance may be less likely to be used in practice than a larger design prior, it has been chosen here to better distinguish the effects of varying the analysis prior mean. Figure 5.12 shows the resulting assurance curves.

As the plot demonstrates, the analysis prior mean affects the rate at which the assurance approaches its maximum. In this case, as we are testing to see if the mean is greater than zero, analysis priors which place a larger amount of weight above this value provide higher assurance values. However, as the sample size increases and more data is available, the analysis priors which are more sceptical begin to catch up with those providing larger assurance values.

The analysis prior is also the determining factor as to whether the assurance starts at a value of zero or one when $n = 0$. In the case of no observations, the analysis is solely dependent upon the analysis prior, which will either give a significant result, and thus an assurance of one, or a non-significant result, and accordingly an assurance of zero.

Another consideration is the ratio between analysis prior mean and standard deviation. We calculate assurance for a number of simulated trials, with the same design prior and varying best estimates. For each best estimate, the corresponding standard deviation also varies, to ensure that 10% of the prior is below 0. Figure 5.13 shows the assurance curves from these simulations.

As the plot shows, the assurance curve varies for each best estimate and standard deviation combination. This means the effect of the analysis prior on the assurance is not simply determined by how much weight it places above and below the cutoff. Furthermore, the larger the mean and variance of the analysis prior, the faster the assurance curve tends

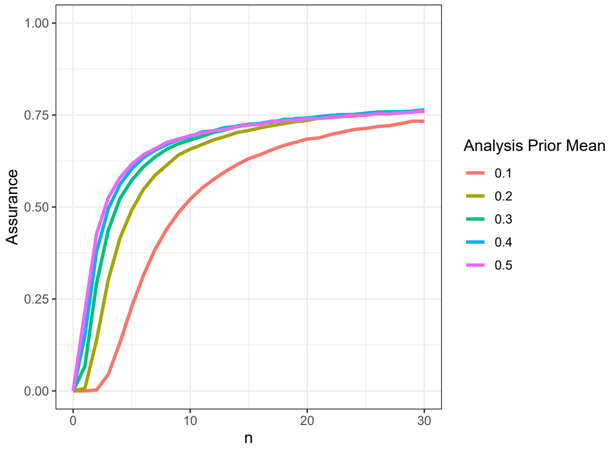


Figure 5.13: Assurance curves with analysis priors that have 10% weight below zero.

to grow.

Should a sceptical prior be chosen for the analysis prior, it is important to consider the probability the sceptical prior assigns to be greater than the cutoff, as well as the mean and variance of the chosen distribution. An obvious choice may be to place prior with 5% of the weight above the cutoff, in order to provide some level of consistency with Frequentist methods which most are more familiar with. However, it may still be useful to consider an appropriate mean for the distribution.

This can be compared to Figure 5.20, which selects standard deviations based on the mean for the design prior. In this second case, the assurance curves differ in slope for design priors with different standard deviations. This behaviour suggests that the assurance is more sensitive to changes in the design prior than in the analysis prior.

Binomial observations

We also look at the case of binomially distributed data. We use a $Beta(10, 20)$ distribution as the design prior, and consider the case where the critical value for a successful trial is 0.2. The analysis priors are Beta distributions, chosen to have a standard deviation of 0.2 and varying means. Figure 5.14 shows the resulting assurance curves.

As seen for the Z-test, the assurance curves all tend towards the same value. The differences in analysis prior means had led to different starting points when n is low, and different rates of convergence towards the maximum. The larger n is, the smaller the effect of the analysis prior becomes. This is to be expected, as the effects of a prior diminish as more data is observed.

Unlike changes in the design prior, changing an analysis prior for binomial observations does affect where the zig-zagging pattern appears. A common analysis prior will ensure these patterns occur at the same value of n, but when the priors vary so does the pattern.

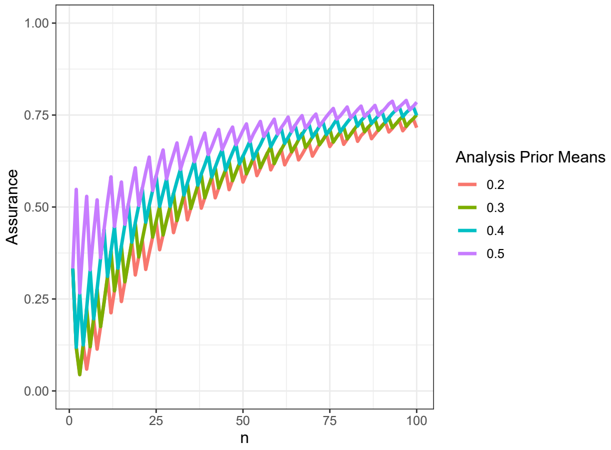


Figure 5.14: Assurance curves with analysis priors with various means.

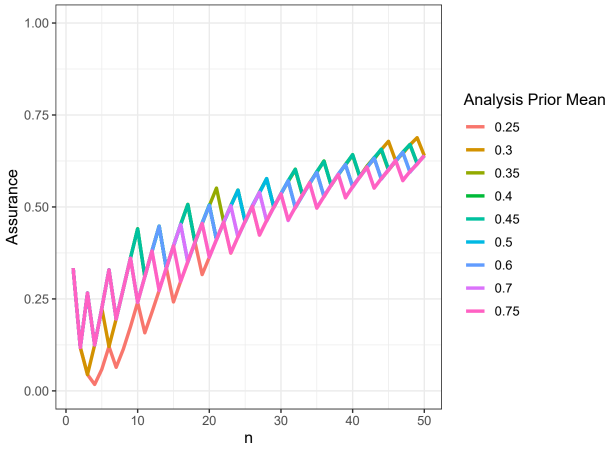


Figure 5.15: Assurance curves with analysis priors that have 20% weight below 0.2.

Figure 5.15 shows a second case with binomial observations and beta prior distributions, but where each curve places the same probability below the critical value for the posterior. The mean of the distributions varies, with the associated standard deviations chosen to ensure 20% of each analysis prior is below 0.2.

Similarly to the case with normally distributed observations, the assurance curves for analysis priors with a common probability of being below the value under the null hypothesis are similar. Due to the different analysis priors, the required number of observed successes change at different sample sizes, as evidenced by the different locations for the peaks. Once a curve drops from a peak, however, it then follows a common shape with the other assurance curves. This is because the design prior leads to the same probability of observing a certain number of successes out of n samples regardless of analysis prior.

5.4.2 Non-conjugate Analysis Priors

While mathematically and computationally more complex, it is also possible to use non-conjugate priors for the analysis prior.

This may be necessary when an aggregation of priors is required for the analysis prior, as these can often lead to mixtures of distributions which may not take the form of a conjugate prior. More generally, this gives additional flexibility in the specification of the analysis prior.

In order to include a non-conjugate prior in the analysis, the assurance calculations will involve an MCMC step, or equivalent method, in order to simulate from a posterior density. When searching for a sample size, this can involve rerunning the MCMC for each value of n to be considered, increasing the computational time considerably.

While the design prior needs to be selected carefully to represent the beliefs of the individuals deciding the trial design, there is more freedom of choice when selecting an analysis prior. As such, it would seem sensible that, unless we have a compelling reason otherwise, a conjugate prior be chosen. Such a conjugate prior can still be easily modified to allow for varying levels of scepticism and uncertainty, and so should allow for an appropriate analysis prior choice.

We consider a non-conjugate analysis prior in Section 5.4.3, in the form of a Spike and Slab prior.

5.4.3 Types of Analysis prior

While the choice of design prior for assurance calculations is an important choice, when conducting a Bayesian analysis it is also important to consider the prior which will be used in the analysis stage.

We consider four different categories of prior distribution. Firstly, the prior distributions can be designed to be informative or non-informative.

An non-informative, or reference, prior aims to provide minimal information about the parameter. In the case of a Beta distribution, this is often taken as a $Beta(1, 1)$ distribution which is uniform over the range of possible values. For a Normally distributed prior, a prior with a large variance is a common choice. While, as discussed previously, these priors are in fact informative, we will demonstrate the outcome of using them in this scenario.

An informative prior represents some level of prior information about the parameter. In the following simulations, we will use an analysis prior the same as the design prior to represent an informative prior centred on the same effect size.

We can also categorise priors based on how enthusiastic they are about a significant result being found.

A sceptical prior represents the views of a researcher who is sceptical of the effect size

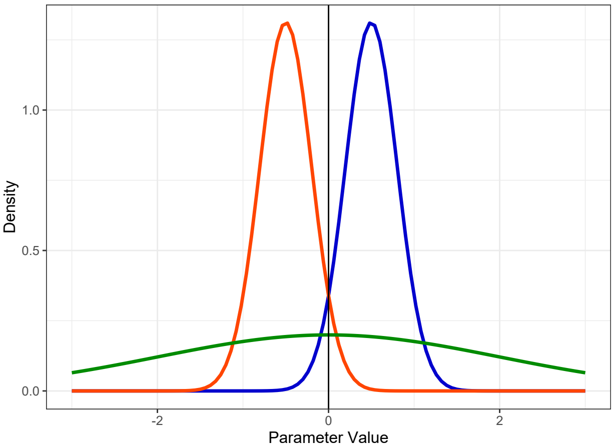


Figure 5.16: Example analysis prior distributions when considering a Bayesian analysis in which success is measured by the analysis posterior distribution having high probability above zero.. An optimistic prior is presented in blue, a sceptical prior in orange and an uninformative prior in green.

required for a successful trial. Such a prior is chosen by ensuring the probability assigned for significant parameter values is low.

An optimistic prior represents the views of a researcher who gives high probability to an effect size required for a successful trial. Such a prior will place the majority of its probability in the range of significant parameter values.

Both of these priors tend to be informative.

Figure 5.16 presents example distributions that could be used as analysis priors. The optimistic and sceptical priors, in green and orange respectively, are more informative than the vague, uninformative prior shown in blue. The example optimistic prior takes the form of a $N(0.75, 0.2)$ distribution, the sceptical prior a $N(-0.75, 0.2)$ distribution and the uninformative prior a $N(0, 2)$ distribution.

The corresponding assurance calculations will, assuming the same design prior is used in each, result in curves tending towards the same maximum value. The differences in the analysis prior will affect the rate at which they approach that maximum. Figure 5.17 presents the assurance curves for the three analysis priors shown in the previous plot.

As this demonstrates, the sceptical prior provides much lower assurance values for low sample sizes than the optimistic prior.

Additionally, the uninformative prior provides assurance values between the sceptical and optimistic analysis priors. In this case, this was due to the mean of the uninformative prior lying between the means of the two other distributions, and its probability of the parameter being above zero being in between the equivalent probabilities of the other distributions.

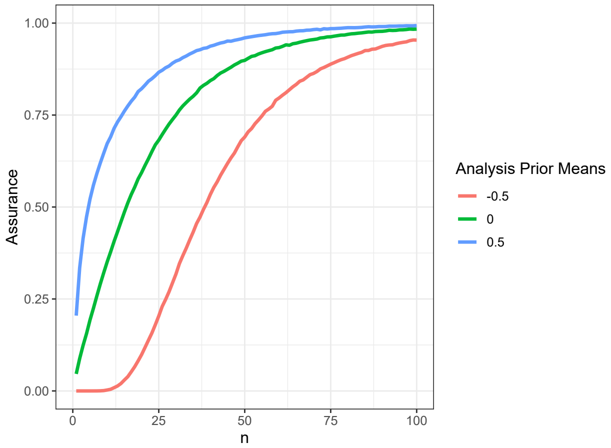


Figure 5.17: Assurance curves for the example analysis priors given in Figure 5.16.

Sceptical Priors

From a conceptual standpoint, it makes sense to select a sceptical analysis prior. This allows for a convincing argument to be made should a statistically significant result be found: that the results in the data are strong enough that they can convince a sceptic.

The question that follows this, is how sceptical should a sceptical prior be? Should the sceptical prior be chosen to be too sceptical then it may be near impossible to find a significant result. Likewise, should the prior not be sceptical enough, then the results may not be convincing enough to sceptical third parties.

As demonstrated in Section 5.4.1, one method of determining sceptical priors is based on the probability assigned to the parameter values which lead to a successful trial. As a sensible starting point, somewhat equivalent to Frequentist standards, we will first consider an analysis prior with a probability of 5% that the mean is above zero.

However, with such a constraint there is an infinite number of combinations of mean and standard deviation which will fulfil this for the normal distribution. Table 5.1 provides example values that could be used in such a case.

These combinations all provide different levels of assurance for each sample size, and each has a different impact on the posterior distributions. The following equations show the posterior parameters for a normal likelihood and prior distribution, for a known population standard deviation, here set to one.

$$\mu_{posterior} = \gamma^2_{posterior} \left(\frac{\mu_{prior}}{\gamma^2_{prior}} + \frac{\sum_{i=1}^{n} x_i}{1^2} \right) \tag{5.14}$$

$$\gamma^2_{posterior} = \left(\frac{1}{\gamma^2_{prior}} + \frac{n}{1^2} \right)^{-1} \tag{5.15}$$

Table 5.1: Normal distribution parameters that give $P(x > 0) = 0.05$.

Prior Mean	Prior Standard Deviation	Posterior Mean	Posterior Standard Deviation
-0.1	0.0608	0.0120	0.0597
-0.5	0.3040	0.1611	0.2191
-1	0.6080	0.2641	0.2805
-10	6.0796	0.3346	0.3158

For a given sample, $\sum_{i=1}^{n} x_i = \bar{x}n$, where n is the sample size, the posterior can easily be calculated based on the prior distribution parameters. If we consider a sample size of $n = 10$, and imagine a dataset with sample mean $\bar{x} = 0.5$, we can calculate the following posterior parameter values given in Table 5.1.

As the table demonstrates, a researcher needs to select more than just the prior probability of success when choosing a sceptical prior. Each combination of prior mean and standard deviation leads to a different posterior mean and standard deviation combination. While the ratio between the prior parameters remains constant due to the condition placed upon it, this is not the case for the posterior parameters. The greater the prior standard deviation, the greater the posterior probability above the cutoff value will be.

If the views of a sceptical third party are to be represented, then a prior mean closer to the cutoff may be more appropriate. A party such as a funding body would likely argue that for the trial to be approved, the treatments should be in equipoise, as mentioned in Chapter 2. A prior representing this would suggest there is no difference between the treatments. If a normal distribution was to represent the difference between two treatments, a value of zero would represent no difference.

As such, an analysis prior which includes a sceptical view, but also one which does not place strong probabilities on a difference in treatments may be appropriate. Such a prior would have a mean close to the cutoff, zero, and a relatively low standard deviation, such as that in the first row in Table 5.1.

Spike and Slab Prior

Another form of analysis prior which may be of use is a spike and slab prior. This type of prior places a large probability on a single value, the spike, and the remaining probability over a range of other values, the slab.

Such a prior may not always be an appropriate representation of an individual's beliefs. If a person believes there is a high probability that a parameter takes one particular value, it is likely their beliefs on the values surrounding that value will increase or decrease smoothly, rather than with a hard drop. For example, if an individual's beliefs form a density where $f(\theta = 0) = 0.4$, then the density at $f(\theta = 0.00001)$ would be expected to be close to 0.4, which will typically not be the case with this form of prior. However, for a more artificial prior which is not aiming to represent an individual's beliefs, but rather a generic sceptical viewpoint, this type of prior may be useful. In a sense this is closer than the sceptical prior to the situation in a hypothesis test, where the spike represents the value under the null hypothesis.

For a normal distribution representing the differences between two treatments, where a value greater than zero for the mean is considered a success, a sceptical prior could place a large probability on the mean being equal to zero, and a small probability spread across a range of positive values. This would again represent an equipoise between two treatments.

In this case, a spike and slab prior could be developed as follows. For a random sample on a random variable X, which is normally distributed with population parameters θ and σ, a spike and slab prior can be placed on θ. This prior has a normally distributed slab component, labelled below as θ_{slab}, centred on zero with a standard deviation of γ. The spike, θ_{spike}, is obtained by multiplying the slab component by a Bernoulli random variable, which is equal to one with probability $1 - p_{spike}$ and is equal to zero, the spike, with probability p_{spike}.

$$X \sim Normal(\theta, \sigma) \tag{5.16}$$

$$\theta_{slab} \sim Normal(0, \gamma) \tag{5.17}$$

$$\theta_{spike} \sim Bernoulli((1 - p_{spike})) \tag{5.18}$$

$$\theta = \theta_{slab} \times \theta_{spike} \tag{5.19}$$

This means that p_{spike} proportion of the distribution has a value of zero, forming a spike at that location. An example of this is provided in Figure 5.18, where $p_{spike} = 0.5$, and $\gamma = 0.5$.

The size of the spike at $\theta = 0$ is determined by how sceptical an analysis prior is required to be.

We run simulations to demonstrate the use of a spike and slab analysis prior. Figure 5.19 shows assurance curves from 5 different spike and slab priors, each with a $N(0.5, 0.2)$ design prior, varying levels of probability assigned to the spike, and the slab component formed by a $N(0, 0.5)$ distribution. In this case, we consider a one sided test, where the slab component is truncated to only include positive values. These curves are not completely smooth due to sampling variability.

As this demonstrates, the more probability assigned to the spike, or the more sceptical the analysis prior is, the lower the assurance will be for a given n. Note that the case where $p_{spike} = 0$ is where only the slab component's normal distribution is used, as there is no spike. Additionally, each of these assurance curves tend towards the same maximum

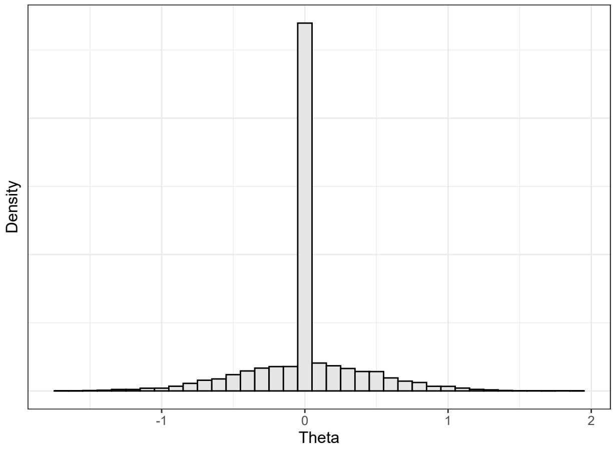


Figure 5.18: An example spike and slab prior distribution.

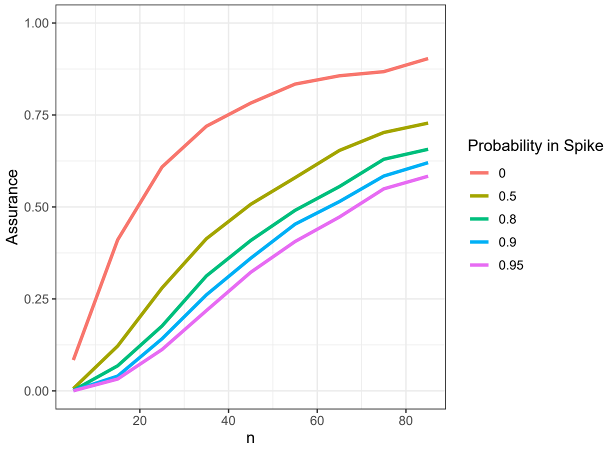


Figure 5.19: Assurance curves for spike and slab analysis priors.

value as n increases, the analysis prior is only affecting the slopes of the curves.

The difference between these analysis priors can be quite large, for example the sample size required for an assurance of 0.5 is 20 when $p_{spike} = 0$, but as high as 70 when $p_{spike} = 0.95$. As such, the level of scepticism needs to be carefully decided prior to the analysis.

One benefit of the spike and slab prior is that it can be interpreted in terminology which avoids statistical knowledge. By framing the value of p_{spike} as a question along the lines of "What is the probability the effect of the new treatment is equal to the current treatment?" or "What is the probability that both treatments are equally effective?", the prior can be easily explained or elicited in a practical manner. This may be beneficial when communicating to non-statistically trained audiences, and can present a convincing argument that the prior has not influenced the results in a way which favours the treatment being studied.

For example, if the analysis prior states there is only a 5% chance the new treatment will outperform the current treatment, and a significant result is found, it is clear that the data gathered has provided strong evidence in order to overcome this prior.

5.5 Assurance Limits

As discussed in Chapter 2, the maximum assurance is defined by the design prior. As assurance is a probability, it cannot go outside of the values of zero and one.

The minimum possible assurance is zero. This could be the case when the design prior assigns no probability to the probability of a successful trial, or when the analysis requires a reasonably large number of observations to produce a significant result. Additionally, a very small sample size may not be sufficient to provide a significant result even for optimistic design priors.

The maximum possible assurance is one. This can occur when a strongly optimistic analysis prior is used, or more usually when the design prior places no probability on a non-significant result being found. Such a design prior may lead to the requirement of a larger sample size to reach the assurance value of one if a more sceptical analysis prior is used.

While these values are hard limits for assurance, further constraints can exist depending on the design prior.

5.5.1 Normal observations

A change in the standard deviation of a design prior distribution can be related to the change in the mean.

If we consider the value under a null hypothesis for the effect, say μ_0, we have shown in Chapter 2 that the assurance will tend towards $P(|\theta| > \mu_0)$ as n increases.

The maximum assurance, A_{max}, can then be defined as

$$A_{max} = \Phi(\frac{\mu - \mu_0}{\gamma}) \tag{5.20}$$

where Φ is the cumulative distribution function for the standard normal distribution, and μ and γ the parameters of a normally distributed design prior for θ.

Rearranging this equation shows that the value of γ can be related to the value of $\mu - \mu_0$. To keep an equivalent maximum assurance, if the design prior mean is changed then the design prior standard deviation needs to change inversely proportional to the maximum assurance. Likewise, if the design standard deviation is changed, the value of $\mu - \mu_0$ will need to change proportionally to the maximum assurance.

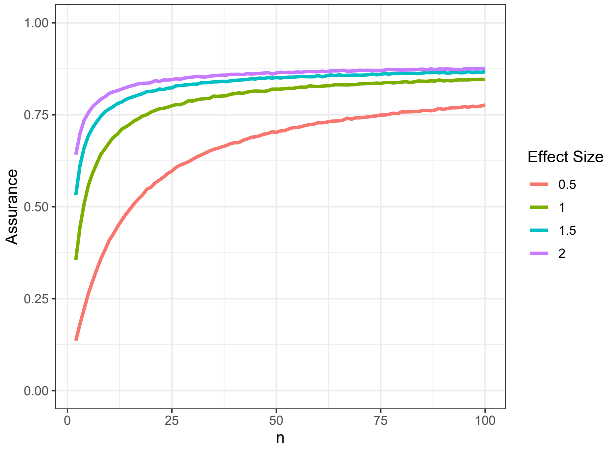


Figure 5.20: Assurance curves for design priors, each giving a 90% probability that the effect size is greater than zero.

$$\mu - \mu_0 = \Phi^{-1}(A_{max})\gamma \tag{5.21}$$

$$\gamma = \frac{\mu - \mu_0}{\Phi^{-1}(A_{max})} \tag{5.22}$$

We consider this in terms of $\mu - \mu_0$ to represent the difference between the value under the null hypothesis of the population mean and the mean of the design prior. For simplicity, we proceed using $\mu_0 = 0$, though the following results will generalise to other valid values of μ_0.

We consider a number of design priors, each with 90% of their probability above zero. Figure 5.20 provides the assurance curves. The design prior means are displayed as separate lines, and each has a standard deviation defined as $\frac{\mu}{\Phi^{-1}(A_{max})}$. In this case, $\Phi^{-1}(0.9) \approx 1.28$.

While each of the curves is converging to a probability of 0.9, they all converge at different rates. While each design prior has 10% of its area below zero, those with higher standard deviations spread that area over a wider range. While there is the same probability the design prior mean is below zero, those with larger mean and standard deviation combinations give a higher weight to larger effect sizes.

For example, a $N(0.5, 0.39)$ distribution provides less than 1% probability of an effect size being greater than 1.5, while a $N(2, 1.56)$ distribution gives a 62.6% probability, despite both distributions providing a 10% probability of an effect size less than zero.

This demonstrates that while two prior distributions can provide the same maximum assurance, they may still have different behaviour as n increases. It also has relevance when choosing design priors. As shown, a design prior should not simply be selected based on the probability a distribution assigns to be above or below a single value. Careful choice of a second parameter input is required to select the distribution which will provide the appropriate assurance curve.

5.5.2 Binomial observations

For a binomially distributed observation, the conjugate prior takes the form of a Beta distribution. While normally parametrised in terms of α and β, these parameters can be related to the mean, μ, and standard deviation, σ, using the following transformations.

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{5.23}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{5.24}$$

For a set mean and standard deviation, the corresponding α and β terms can be calculated using Equations 5.12 and 5.13.

We can find combinations of the mean and standard deviation, or α and β, which provide a given probability above a set value.

For example, we assign a 75% probability above the value 0.2, an analysis prior in the form of a $Beta(2, 8)$ distribution, and find α and β values of approximately (3.26, 7.62), (1.26, 1.89), and (0.70, 0.70) to correspond with means of 0.3, 0.4 and 0.5 respectively. Figure 5.21 shows the resulting assurance curves.

These assurance curves all tend towards the same value of 0.75. However, due to the different design priors, the rate at which they converge towards this value varies. As seen

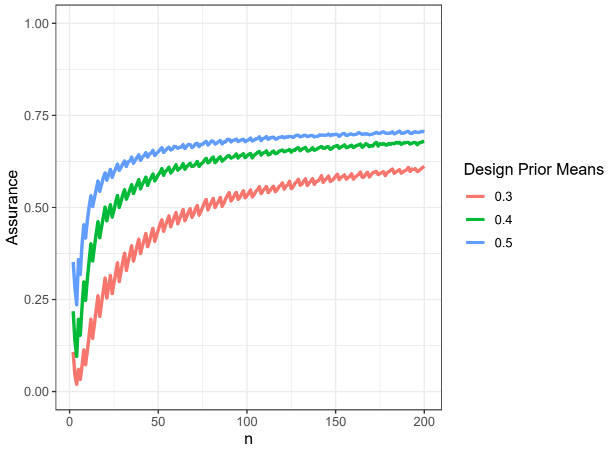


Figure 5.21: Assurance curves for design priors, each giving a 75% chance that the effect size is greater than 0.2.

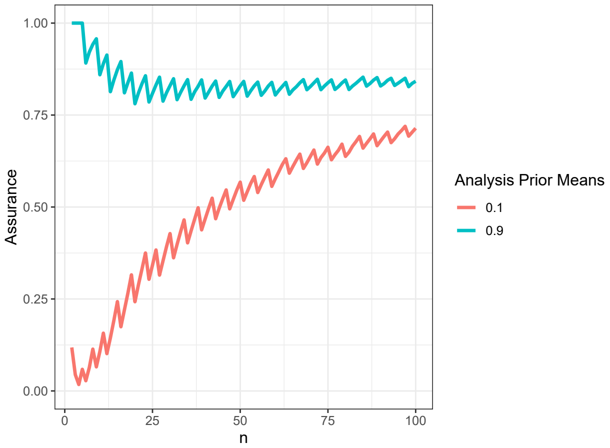


Figure 5.22: Assurance curves for two different analysis priors, with the same design prior.

in the normal example, those with higher means, and consequentially higher variances, are converging at a faster rate.

In cases where the analysis prior is more optimistic than the design prior, the assurance curve may decrease as n increases. While this behaviour is the opposite of what is typically seen with power and assurance curves, it would represent a case where an optimistic analyst is gathering data which does not support their conclusions. Figure 5.22 provides such an example, where the analysis prior with the larger mean is providing a higher probability of success compared to the design prior. This assurance curve then converges from above, rather than below. In the case of continuous data, this curve would be monotonically decreasing towards the maximum assurance.

This further supports the case for comparatively more sceptical analysis priors and more optimistic design priors, and is discussed further in Section 5.4.3.

Ultimately, the design prior informs the value to which assurance will tend as n increases, and the rate it approaches it. The analysis prior influences the rate of convergence, and also informs which direction the curve is converging from.

From a practical standpoint, an assurance curve that decreases as n increases due to an optimistic analysis prior is not appropriate for sample size calculations. Such a curve states that gathering more data makes it more likely a trial will not come to a successful conclusion. We interpret this as a sign that the analysis and design priors have not been chosen to provide sensible results.

5.6 Overestimated Effect Size

One benefit often associated with assurance is that it is more robust when the initial parameter inputs do not match the observed data. For example, Chen and Fraser (2018) perform a number of simulations, and demonstrate the sensitivity of statistical power to a difference in effect size used in power calculations compared to that estimated from the observations. They recommend assurance as an alternative.

In this section, we calculate sample sizes for given parameter specifications using both assurance and power. We then consider hypothetical outcomes in potential trials of the given sample sizes, both when the effect size and design priors accurately represent the population parameters, and also when they do not.

5.6.1 Normal observations

To demonstrate the potential issue, we first consider a simple case where a Frequentist two sided Z-test will be used. For an effect size of 0.5, a population standard deviation of 1, and a power of 0.9, the required sample size using a power calculation is 42.

However, while an effect size of 0.5 may have been used in the power calculation, it may not necessarily correspond with the true effect size. Figure 5.23 shows the power provided by a sample of 42, with differing effect sizes.

As this plot shows, if the true effect size is lower than the value used in the power calculations, the power will also be lower.

One way this can be accounted for is by using a power of 0.9, when a power of 0.8 may still be acceptable. In this case, an effect size of 0.44 will still provide a power greater than 0.8. This provides some flexibility as to the accuracy of the effect size. However, if the true effect size was only half that which was estimated, the power would then only be around 0.36.

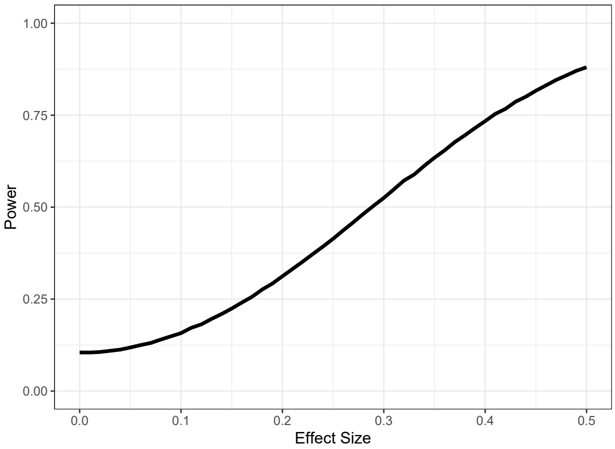


Figure 5.23: The actual power provided for different effect sizes, when the trial was powered on an effect size of 0.5.

This behaviour is consistent for different effect sizes. Figure 5.24 shows simulations using different effect sizes to determine the sample size, displayed in different colours.

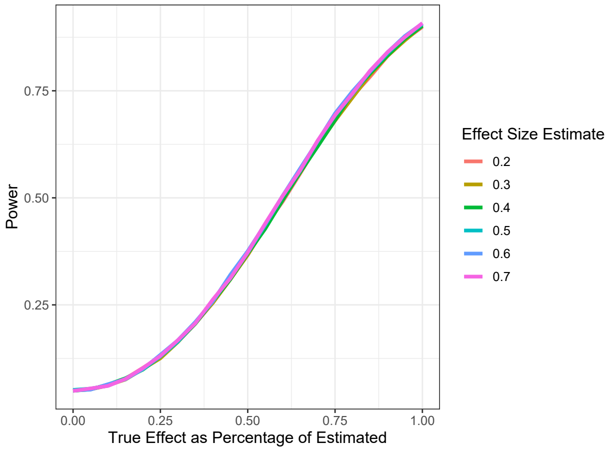


Figure 5.24: The actual power provided for different effect sizes, where the colour represents the effect size used in the calculation to determine sample size, and the x axis the true effect size as a percentage of the effect size used in the calculation.

Table 5.2: Assurance values by prior standard deviations

Prior Standard Deviation	Assurance
0	0.90
0.1	0.85
0.2	0.78
0.3	0.72
0.4	0.68
0.5	0.65
1	0.58

For true effect sizes, treated as percentages of the effect size used to power the trial, the relationship with power remains consistent. For example, when the true effect size is half of the estimated effect size, the power is approximately 0.36, and when it is a quarter of the estimated effect size, the power is approximately 0.13.

This issue may be predominant for lower specified effect sizes. A difference in effect size of 0.1, for example, is proportionally much larger for a smaller effect size than a larger one. For smaller values it may be more likely that effect sizes are rounded, leading to such errors.

We consider a similar case, and calculate the assurance. This will depend on the prior distribution placed on the effect size, θ. For consistency, we will use normal distributions with means equal to one, with varying design prior standard deviations. Table 5.2 shows the associated assurances for a sample size of 42.

As the table demonstrates, the equivalent assurance depends strongly on the chosen prior. The higher the effect's prior standard deviation is, the lower the calculated assurance is. While the power associated with this trial was set at 0.9, the assurance could have a wide range of values. Note that when the prior standard deviation is equal to zero, the

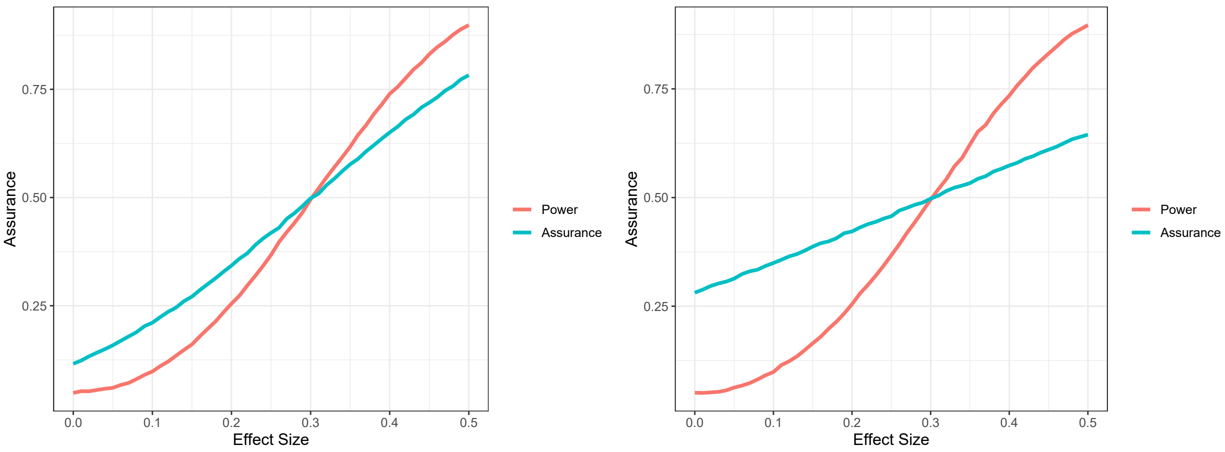


Figure 5.25: Assurance and power for different effect sizes, when the trial was powered on an effect size of 0.5. The assurance's prior on the effect size is a $N(\text{Effect size}, 0.2)$ distribution for the curve on the left, and a $N(\text{Effect size}, 0.5)$ distribution for the curve on the right.

assurance and power are the same. As this standard deviation increases, the assurance decreases.

As the population mean is equal to one, it would be reasonable for the prior distribution on the mean to have a prior standard deviation less than one. As an initial example, we take a prior standard deviation equal to 0.2. Figure 5.25 shows the change in assurance as the effect size changes, compared to the previous curve for power.

As the plot shows, the power curve is often steeper than the assurance curve, demonstrating a larger change to power for a change in effect size. This should be expected, as the prior distribution for assurance takes a range of values into account.

For a larger standard deviation in the prior, such as 0.5, the slope of the assurance curve becomes flatter. While larger standard deviations in the prior decrease the assurance's sensitivity to the mean, they also decrease the maximum assurance possible.

Type M and Type S Errors

We consider a more extreme case of prior misspecification. We start with a Z-test with the previous setup, with an effect size of 0.5 and population standard deviation of 1, leading to a sample size of 42.

We calculate the power, type M, and type S errors based on different true effect sizes (Timm et al., 2019). As covered in Chapter 2, the type M error represents the amount by which an estimate will overestimate the true effect if a significant result is found, and a type S error is the probability a significant result has the incorrect sign.

As Table 5.3 shows, if the true effect size is quite small in comparison to the effect size used in power calculations, there can be large errors associated with any significant

Table 5.3: The power, type M, and type S errors for Z-tests with different true effect sizes

True Effect	Actual Power	Type M Error	Type S Error
0.5	0.90	1.06	0.00
0.25	0.37	1.63	0.00
0.1	0.10	3.74	0.05
0.05	0.06	7.30	0.18
0	0.05	Inf	0.50

results. For example, for a true effect size of 0.05, one tenth of that used to determine the sample size, would correspond to a power of only 0.06. If a significant result was found, then the effect estimated would be larger than the true effect by a factor of 7.3 on average, and would have the incorrect sign 18% of the time.

While the errors which could be made in such a case are large, even if the true effect size was half that used in the sample size calculations, they can still occur. In such a case, for a true effect of 0.25, the power has dropped from 0.9 to 0.37, and the type M error is 1.63. This means a significant effect will, on average, overestimate the true effect size by approximately 63%.

Such an underpowered trial is more likely to provide a false negative than anticipated. Where an MCID is used as the effect size for a power calculation, it could be argued that this is not an issue. If the test has failed to detect an effect that is present, but lower than the MCID, then the effect was too small to be relevant anyway. The false negative here could be argued to be false in the sense that there is an effect, but true in the sense that there was not a large enough effect size to be determined.

The type M error, however, may still be an issue. Should a significant result be found, the effect size found would on average be around 60% larger than the true effect. This overestimation can lead to misleading results being published, and may have further implications for future studies. If a later trial was powered using the effect size found in this case, the overestimation would lead to an underpowered trial.

It is also important to note that underestimating the effect size will affect the power of a test as well. However, in these cases the power will be higher than expected, and thus the test's validity is not reduced. An underestimation of effect size, and the corresponding increase in power, means a lower sample size could have been used. Such an error may lead to increased costs for a trial, and it is unethical to recruit more patients to a trial than is required.

5.6.2 Binomial observations

We can also consider the case of binomial observations.

An exact binomial test, with a null hypothesis of $\theta_0 = 0.1$, and an effect size used for the power calculation of $\theta_a = 0.3$, requires a sample size of 32 to achieve a power of 0.9.

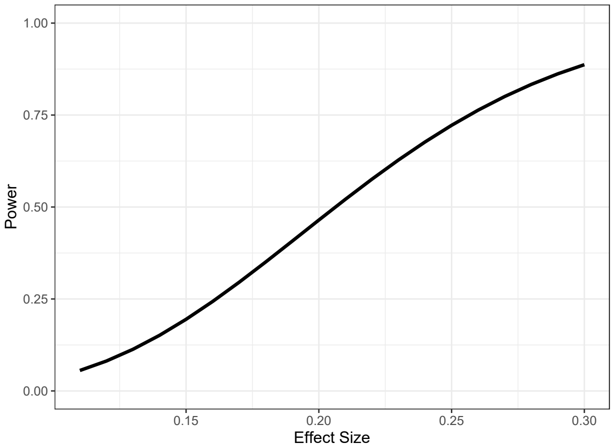


Figure 5.26: The actual power provided for different effect sizes, when the trial was powered on an effect size of 0.3.

Figure 5.26 shows the actual power for a trial with a sample size of 32, if the true effect size is less than the value used in this calculation.

If a power of 0.8 is deemed acceptable, then the corresponding effect size of 0.27 offers a lower limit on possible effect sizes for which this sample size is adequate. If the true effect size was below 0.27, then this trial would be underpowered.

If we consider this case, with a sample size of 32, we can also calculate the assurance. Figure 5.27 provides plots of the actual assurance a trial would have if the effect size was less than the one used to power it. We also present two different assurance curves with different standard deviations for the design prior.

As these plots show, the assurance decreases less than the power for a decrease in effect

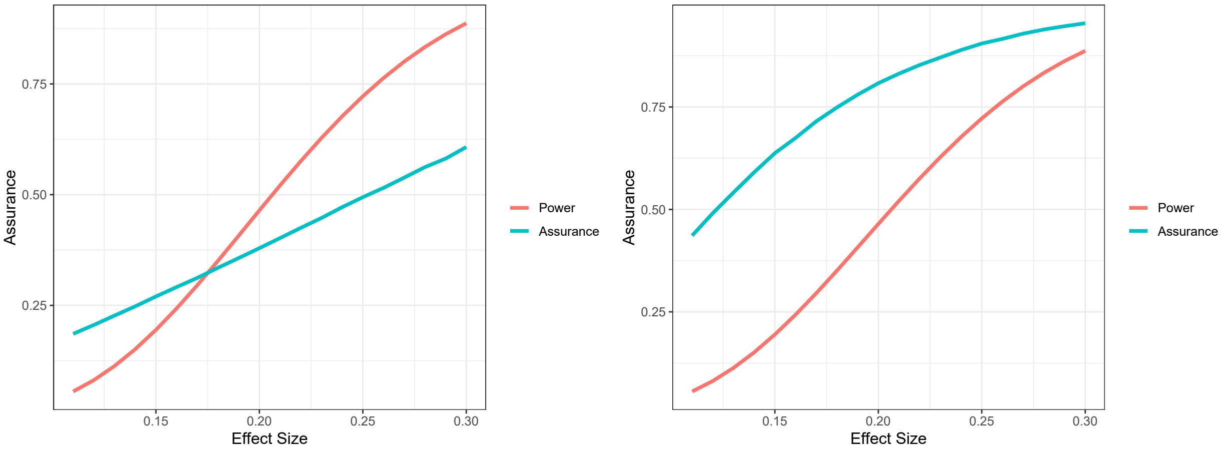


Figure 5.27: Assurance and power for different effect sizes, when the trial was powered on an effect size of 0.3. The assurance's prior on the effect size has a standard deviation of 0.1 for the plot on the left, and 0.2 for the plot on the right.

size. This demonstrates again that assurance is more robust to misspecified inputs than power.

This behaviour continues as the standard deviation increases. At higher standard deviations, the maximum assurance may decrease but the slope of the curve is still less steep then the slope of the power. As the standard deviation approaches zero, the assurance curve approaches the power curve.

5.7 Unknown Population Variance

In previous sections, we have considered normally distributed observations where the population variance is known. We chose the variance to be one for convenience, although the general results found hold more widely.

If our random sample comes from a Normal population with a population variance not equal to one, say v, we can transform the data by dividing by $\sqrt{v}$, which results in a variance equal to one, as per

$$Var(vX) = v^2 Var(X) \tag{5.25}$$

However, it will not typically be the case that the population variance is known, regardless of its value. Firstly, we can consider the case where the population variance is misspecified.

We consider a similar case to Section 5.6.1, where, for a power of 0.9, a sample size of 42 is required to find an effect of 0.5. An assurance calculation, under similar inputs, suggests a sample size of 46.

As Figure 5.28 demonstrates, if the population variance is different to the assumed value of 1, the actual assurance or power will also be different. Underestimating the population variance leads to smaller assurance or power than expected, which is detrimental to the validity of a trial. As such, it seems better to have overestimated the population variance than underestimated it.

One option for dealing with uncertainty around the true population variance could be to place a prior distribution on it. In this case, it would take the form of a design prior, similar to the prior on the mean.

Defining σ to be the population variance, the assurance is then given by

$$\text{Assurance} = \int\int P(\text{Success} \mid \mu, \sigma) P(\mu) P(\sigma) d\mu d\sigma \tag{5.26}$$

assuming that μ and σ are independent.

To explore the effect of a prior distribution placed on σ, we will hold μ constant. This prior can be elicited from an expert, by asking them about probabilities across a range of

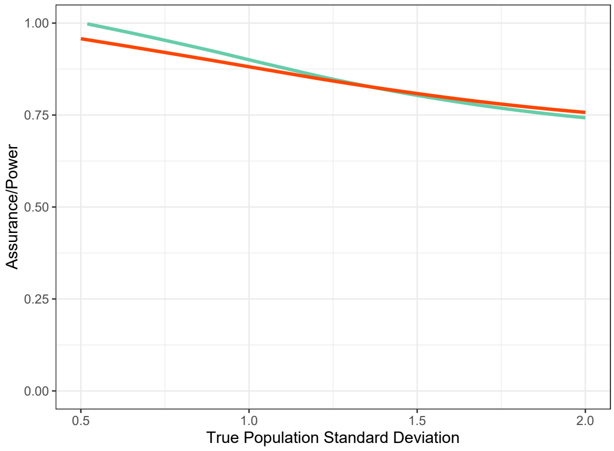


Figure 5.28: True assurance (red) and power (blue) levels for varying population standard deviations, given sample sizes calculated assuming a population standard deviation of one.

possible values and using these judgements to in turn fit parameter values (Alhussain and Oakley, 2017).

As σ must be greater than zero, we consider a gamma prior. The gamma distribution can be parametrised in terms of a shape, α, and rate, β, parameter, for which the mean is defined as $\frac{\alpha}{\beta}$ and the variance $\frac{\alpha}{\beta^2}$. Thus, $\alpha = \frac{E[X]^2}{Var[X]}$ and $\beta = \frac{E[X]}{Var[X]}$. Using this, we can choose parameters to have the same variance, and varying means, or vice versa.

Figure 5.29 demonstrates assurance curves for different priors on the population standard deviation, each with a standard deviation of 0.1. The higher the population standard deviation's mean, the lower the assurance for a given sample size. For a Z-test, a higher population standard deviation will increase the denominator of the test statistic, corresponding to a decrease in the test statistic itself. As such, it then becomes less likely that a statistically significant result will be found.

We can also consider the case where the population standard deviation prior's variance is changed. As Figure 5.30 demonstrates, increasing the uncertainty of the prior on the population standard deviation has an effect on the assurance for a set design prior. For lower values of n, a higher level of uncertainty corresponds to higher assurance values, but as n increases the assurance curves with lower levels of uncertainty provide higher values.

This behaviour is similar to that when the design prior mean's standard deviation is varied, and the reasoning is likely similar. At low n, a higher standard deviation allows for more cases which will lead to a significant result to be included. For larger n, however, the higher standard deviation then includes more cases which will lead to a non-significant result.

It is important to note that the variance of a binomial distribution is not defined as a separate parameter. Instead, it relies only on the probability of success and sample size

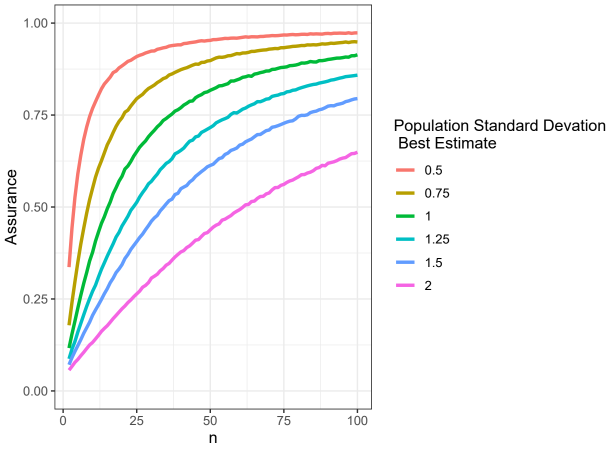


Figure 5.29: Assurance curves varying the mean for priors on the population standard deviation priors, with a design prior of $N(0.5, 0.2)$ on the mean.

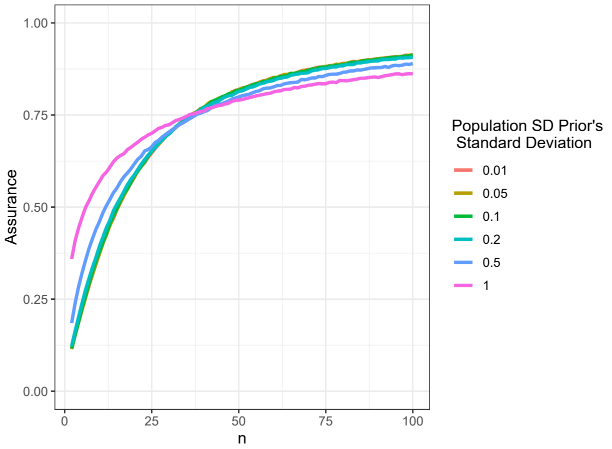


Figure 5.30: Assurance curves varying the standard deviation of population standard deviation priors, with a design prior of $N(0.5, 0.2)$ on the mean of the data.

$n\theta(1-\theta)$.

5.7.1 Bayesian Analysis Priors

Uncertainty about the population standard deviation can also be addressed in analysis priors for Bayesian analyses. In this case, there is a choice of prior on the mean and variance which leads to conjugacy. This can be achieved using a normal-inverse gamma distribution.

However, for ease of computation, this can be calculated using the precision, τ, instead of the standard deviation. The precision is defined as $\tau = \frac{1}{\sigma^2}$. In such a case, a normal-gamma distribution will be conjugate.

The normal-gamma distribution is a multivariate distribution with four input parameters. For a normally distributed random variable, X, with a mean of μ and precision of τ, then

$$X \sim N(\mu, \tau) \tag{5.27}$$

$$\tau \sim Gamma(\alpha, \beta) \tag{5.28}$$

$$(\mu, \tau) \sim NormalGamma(\mu_p, \lambda, \alpha, \beta) \tag{5.29}$$

As demonstrated in Murphy (2007), the posterior then takes the form

$$P(\mu, \tau \mid X) \sim NormalGamma\left(\frac{\mu_p\lambda + n\bar{x}}{\lambda + n}, \lambda + n, \alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\lambda(\bar{x} - \mu_p)^2}{2(\lambda + n)}\right) \tag{5.30}$$

As this form can be evaluated analytically, the calculations for assurance can be made quickly. Figure 5.31 provides an example assurance curve. The blue curve presents a case where there is no uncertainty on the population variance, and it is known to be equal to one. The red curve is when a prior is placed on the population variance. This prior has a mean of one.

As the plots demonstrate, including a prior distribution on the population variance decreases the assurance slightly. The plot on the right has a higher variance in the prior distribution on the population variance. Accordingly, the more uncertainty there is about the population variance, the smaller the assurance is for a given n.

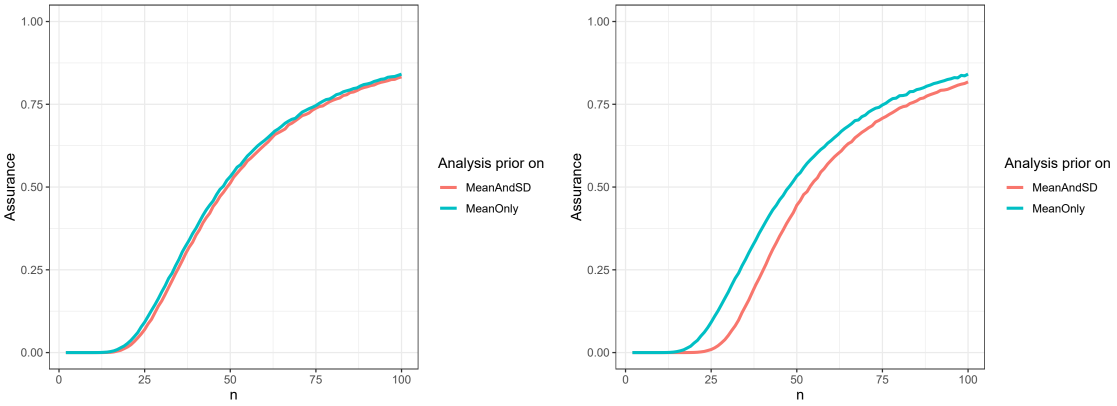


Figure 5.31: Assurance curves for known population variances (blue) and unknown population variances (red). The plot on the right has a higher level of uncertainty in the prior for the population variance.

5.8 Implications for Practice

While power and assurance can both be used in a similar way, there are a number of differences between them that have implications for their use in sample size calculations. It is important from an ethical perspective that sample sizes for clinical trials are chosen carefully. Too low a sample size can lead to insufficient data to observe a true effect, while too high a sample size can waste limited resources and expose additional patients to unnecessary risk. In order to best select a sample size, it is important to understand and correctly utilise the chosen method of sample size selection.

We have demonstrated how assurance can be more robust than power when the underlying true parameters in a trial vary from the values used in the sample size calculations. This is due to the incorporation of uncertainty through prior distributions, which better captures the uncertainty surrounding the future study. An overestimation of effect size, for example, can lead to a lower level of statistical power for a trial than that is intended. This means that the probability of making a Type 2 error, or failing to reject the null hypothesis when it should be rejected, is higher than anticipated. Assurance better accounts for this discrepancy between the anticipated effect size and observed effect, in turn reducing the probability of making an incorrect inference.

These simulations also outline additional potential difficulties in using assurance over power. The additional parameters required to specify prior distributions may be more difficult to obtain than a single point for the effect size in a power calculation. An assurance calculation with prior distributions centred on a certain effect size can result in different outcomes depending on the variance of the prior distributions, in comparison to a power calculation for which the same effect size will lead to a single output. As such, differences in both the prior distributions mean and variance can in turn lead to differences in assurance, and the subsequent sample size calculations. The specification of a suitable prior distribution is an important step.

Another issue demonstrated was the maximum possible assurance when conducting this type of Bayesian analysis. As there is not a consistent range of values for the assurance given varying design prior distributions, it is difficult to define a standardised rule to choose a sample size. As the simulations demonstrated, the maximum assurance value is set by the design prior distribution, while the assurance curve is determined by both analysis and design prior. As the design prior distribution will vary from study to study, so too will the maximum assurance.

One possible method for creating a target assurance across trials could be to utilise standardised assurance, which we previously defined as the assurance for a particular sample size divided by the maximum assurance for the chosen design prior distributions. However, this value no longer makes a direct statement about the overall probability the trial will successfully conclude with a significant result. As such, high standardised assurance values could still correspond with low assurance values given particularly pessimistic design prior distributions. In many cases, it may be more appropriate to select a target assurance based on the specific requirements of the trial, rather than a general rule.

While this is the case for a Bayesian analysis focusing on the posterior probability of exceeding a certain critical value, this issue may not be present for other types of analysis. For example, if the aim for the analysis is to achieve a credible interval width below a target width, then the assurance will have the same maximum value regardless of prior inputs. However, such methods may have further difficulties associated with them, as specifying a posterior credible interval target width alone does not assure the posterior of placing probability in any particular ranges of values.

These issues outline some of the difficulties with implementing assurance in practice. As the majority of sample size calculations currently conducted are based on power, with simple targets of 0.8 or 0.9, the more complex method of assurance, which requires further consideration of target values and specification of prior distributions, will likely be more difficult to implement in practice. To better allow non-specialist audiences to use assurance in their own sample size calculations, further development of assurance-based algorithms for commonly used statistical packages, including assistance for specification of prior distributions, would be beneficial.

However, another consideration is the requirements provided by organisations such as funding bodies and scientific journals. In many cases, trial designs are required to be accompanied by sample size calculations using statistical power. If these types of organisations do not have the necessary experience in assessing and judging Bayesian methods such as assurance calculations, it can be difficult for them to determine whether a Bayesian alternative is appropriate. Further development of guidelines which support and advise on the use of Bayesian sample size methods may also help facilitate a more widespread use of assurance in sample size calculations. Some examples, however, do demonstrate use of Bayesian assurance in industry usage (Crisp et al., 2018).

Further simulations have been provided in Appendix A.2.

5.9 Conclusions

In this chapter we have explored power and assurance, and how they behave in various circumstances. The behaviour of assurance we explored under different scenarios, and it was demonstrated that priors with the same maximum assurance could have differing assurance curves. We have demonstrated how the design prior incorporates additional uncertainty into an assurance calculation which is not taken into account in power calculations. We have also shown how this increased uncertainty can make the results more robust if the assumed effect size does not correspond to the true underlying value.

Part VI

Developing Assurance for a Motor Neurone Disease Diagnostic Study

Chapter 6

Developing Assurance for a Motor Neurone Disease Diagnostic Study

6.1 Introduction

In order to select a sample size when designing a clinical study, a consideration is made of the likelihood of success of the study given a particular sample size. In this chapter, we calculate power and assurance for use in sample size determination to compare their consequences for our case study into a diagnostic test for MND.

We first consider power calculations, providing details on the hypothesis test detailed in the study statistical analysis plan. We provide the sample sizes required to achieve a power of 0.8 and 0.9.

We then consider the calculation of assurance, for both the hypothesis test and the Bayesian analysis which we have developed to be used within the study. When a Bayesian prior distribution is required, we utilise the aggregated prior distributions from Chapter 4. We also consider a further sceptical prior distribution, which someone who is sceptical about the outcome of the study may hold. We note some additional considerations to be made when using assurance to determine sample sizes, including the use of distinct design and analysis priors, selecting a target assurance, and the maximum possible assurance.

We present assurance sample size requirements for both frequentist and Bayesian analyses. These results include a reasonable selection of inputs, and a secondary set of sample size calculations considering a wider range of possible inputs.

As the study designers believed the maximum number of patients who could be feasibly recruited was 120, we also calculate power and assurance for this sample size. This presents an idea of the achievable assurance or power for the study.

Finally, we discuss further considerations and results from the power and assurance calculations.

6.2 Assurance and Power Calculation Inputs

There are various sets of inputs that could be used for assurance and power calculations to determine sample sizes for the study following the elicitations, which are presented in Chapter 4. In this section, we briefly review those which will be used throughout this chapter.

Prior distributions elicited from experts were aggregated using a number of methods in Chapter 4. We now focus on these aggregated distributions, to consider how they affect assurance calculations. These prior distributions come from three groups of experts. Group 1 contains experts who were not involved in the design of the novel diagnostic test, and have expertise in MND. Group 2 contains the three experts involved with the design of the novel diagnostic test and the case study. Group 3 contains all experts from Groups 1 and 2. In power calculations where a single point-estimate is required, the median of the relevant prior distribution will be used as a best estimate.

During the elicitations, the experts were each asked to provide a Minimal Clinically Important Difference (MCID) for the proportion of patients with MND the new experimental test could identify six months earlier than the current reference test. The experts who provided a response to this question, provided values of 0.001, 0.05, 0.1, 0.15, 0.2, and 0.5.

The researchers designing the BIMC study also suggested that a sample size of 120 was deemed the maximum feasible number of patients who could be recruited to such a study. While this figure may not be a hard bound, we will use it in the following sections as a guide to which sample sizes may be appropriate.

6.3 Power Calculations

In this section, we calculate sample sizes for the BIMC study using statistical power. We consider three different ways of incorporating the MCID and elicited values into power calculations. For each case, we find the minimum sample size required to achieve a set power based on an analysis using McNemar's test.

The data can be structured as proportions in the form of Table 6.1, where the values of p are the proportion of patients who received each combination of positive or negative test results from both the BIMC test and the Awaji criteria. The sample size required for

Table 6.1: McNemar's Test

	BIMC Positive	BIMC Negative
Awaji Positive	$p_{1,1}$	$p_{1,0}$
Awaji Negative	$p_{0,1}$	$p_{0,0}$

McNemar's test with a significance level of α and power of $1 - \beta$ can be calculated using

$$n = \frac{(z_{1-\alpha/2}\sqrt{\xi} + z_{1-\beta}\sqrt{\xi - \varDelta^2})^2}{\varDelta^2} \tag{6.1}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are Z scores determined by the significance level and power respectively, $\xi = p_{1,0} + p_{0,1}$ and $\varDelta = p_{1,0} - p_{0,1}$ (Connor, 1987). As such, only the values of $p_{1,0}$ and $p_{0,1}$ are required from Table 6.1 in order to calculate sample sizes.

We consider three ways in which the values of $p_{1,0}$ and $p_{0,1}$ could be specified, each incorporating estimates of the effect size or MCID in different ways.

As discussed in Section A.2.2, it is possible for the MCID to be incorporated into the analysis or design stages, or both. In practice, the MCID is incorporated into power in the design step, as the size of the effect of interest. As such, for the following power calculations, we consider using MCID in this way.

6.3.1 Power Method One

We first consider the case where $p_{0,1}$ is set to be equal to the MCID, and $p_{1,0}$ is set to be equal to zero. In early, pre-elicitation meetings with the study designers, it was suggested that the reference test may provide no positive results when the experimental test was negative. As such, the assumption that the reference test is only positive when the experimental test is positive appears reasonable to consider.

$$p_{1,0} = 0 \tag{6.2}$$

$$p_{0,1} = MCID \tag{6.3}$$

Table 6.2 presents sample sizes for this case.

As expected, the lower the MCID, or higher the power, the higher the required sample size would be. Given that the maximum feasible number of patients who could be recruited was 120, this suggests that an MCID of 0.1 would be a sufficiently large value to power the study.

For a power of 0.9, a sample size of 120 will be able to detect a difference of 0.09 in this case. This difference would be small enough that two-thirds of the experts would consider it clinically important.

Table 6.2: Power Calculation Sample Sizes

	MCID					
Power	0.001	0.05	0.1	0.15	0.2	0.5
0.8	7847	155	76	50	37	13
0.9	10503	206	101	66	48	16

6.3.2 Power Method Two

We consider estimates for $p_{1,0}$ and $p_{0,1}$ based on the elicited values from the experts. For each aggregated distribution, the median values have been used as point estimates, to calculate the proportions as

$$p_{1,0} = \eta(1-\theta_1) \tag{6.4}$$

$$p_{0,1} = (1-\eta)\mu\theta_2 + (1-\eta)(1-\mu)\theta_3 \tag{6.5}$$

As such, $p_{1,0}$ represents a best estimate of the proportion of positive RT and negative ET results at the first time point, and $p_{0,1}$ a best estimate of the proportion of negative RT and positive ET results at the first time point.

Table 6.3 presents the sample sizes based on these inputs.

Given the estimates from the elicitations, it does not appear likely that the study can achieve a power of 0.8 with a sample size equal to the maximum feasible number of patients who could be recruited, with the exception of those elicited via SHELF from the Group 2 experts. It is notable that the estimates based on SHELF suggest the smallest sample sizes, followed by the Equal Weight's Aggregation. The Classical Method and Bayesian aggregations provide the highest sample sizes.

Furthermore, in comparison to Table 6.2, the sample sizes for both values of power tend to be higher. The sample sizes based on elicited values appear comparable to an MCID of below 0.05 using Method 1.

Table 6.4 provides the absolute values of $\varDelta$ calculated from the aggregations. As it demonstrates, many of the implied differences between $p_{1,0}$ and $p_{0,1}$ are small, with only Equal Weights and SHELF aggregations providing values of $\varDelta$ above 0.1.

As Method 1 suggested a difference of 0.09 as the minimum that could be detected with a sample size of 120, it is unsurprising that the sample sizes calculated here are much larger than 120. While the MCIDs provided by the experts tended to be higher, in this

Table 6.3: Power Calculation Sample Sizes with Aggregated Best Estimates

Aggregation	Group	n (Power = 0.8)	n (Power = 0.9)
EW	1	200	266
	2	159	212
	3	1534	2052
CM	1	340	454
	2	5567	7452
	3	215745	288820
BA	1	1103	419
	2	1949	1034
	3	8748	4299
SHELF	2	25	33

Table 6.4: Power Calculation Sample Sizes with Aggregated Best Estimates, $\mid \Delta \mid$ values

Aggregation	Group	$\mid \Delta \mid$
EW	1	0.11
	2	0.14
	3	0.04
CM	1	0.08
	2	0.03
	3	0.01
BA	1	0.05
	2	0.04
	3	0.02
SHELF	2	0.34

case we can see that there is a discrepancy between the individual MCIDs and values of Δ based on the aggregations.

6.3.3 Power Method Three

We also consider a combination of Methods 1 and 2. Given that the RT is a standard method of diagnosis, we suppose that the experts' knowledge of this is more likely to be accurate than their knowledge of the ET. As such, we use the median values from the aggregated prior distribution to inform us of the value of $p_{1,0}$. We then calculate the sample size required to observe an improvement of the MCID in addition to the value of $p_{1,0}$ to obtain $p_{0,1}$.

$$p_{1,0} = \eta(1-\theta_1) \tag{6.6}$$

$$p_{0,1} = \eta(1-\theta_1) + MCID \tag{6.7}$$

Table 6.3 presents the sample sizes calculated based on these inputs, for a power of 0.9.

These results provide a more informed calculation than Method One, while still utilising

Table 6.5: Power Calculation Sample Sizes

Aggregation	Group	0.001	0.05	0.1	0.15	0.2	0.5
EW	1	4409707	1966	541	261	158	34
	2	2749249	1301	375	188	117	28
	3	3217878	1489	422	208	129	29
CM	1	3715551	1688	471	231	141	31
	2	4664427	2068	566	273	165	35
	3	4305579	1924	530	257	156	34
BA	1	4294721	1920	529	256	155	34
	2	3455153	1584	445	219	134	30
	3	3863175	1747	486	237	145	32
SHELF	2	783837	515	178	100	68	20

the MCID. It can be seen that for most aggregation results, an MCID of 0.2 provides a sample size closest to the maximum feasible recruitment sample size, though the SHELF estimate suggests a difference between 0.1 and 0.15 could be detected with the appropriate power.

Unlike Method Two, there is less consistency between aggregation methods as to which provides a larger or smaller sample size. For example the Bayesian aggregation method has led to smaller sample sizes for Group 1, but larger sample sizes for Groups 2 and 3, in comparison to Equal Weight. The difference between the two methods comes from their differences in estimates for $p_{0,1}$. Those that previously were lower, but are now greater, likely had a more optimistic view on the benefits of the ET.

Overall, Method One suggests that a difference of 0.09 may be an appropriate target for a study with a sample size of around 120, while Method Three suggests a difference of 0.2, or higher, may be all that could be detected. However, Method Two suggests that, solely based on the experts' best estimates, a higher sample size would be required to detect the difference they expect. Given the aggregated distributions, the expected values of Δ were predominately less than 0.09, which suggests the experts may not expect the study to have a good chance of successfully observing the required difference.

This demonstrates how the sample size resulting from a power calculation can vary widely, depending on the inputs and method for determining the MCID. As different experts provided different MCIDs, as well as different implied estimates for Δ, there is large variability between the required sample sizes.

6.4 Assurance Calculations

In this section we outline how assurance is used to calculate sample sizes, and the methods and priors which will be used for comparison.

To calculate assurance, a measure of success must be defined for the study. As opposed to statistical power, assurance can be used regardless of whether a frequentist or Bayesian analysis will be conducted. We will consider both of these cases.

Firstly, we consider a study with a frequentist analysis, namely McNemar's test, at the end. The assurance calculation uses a statistically significant result from McNemar's test as the definition of a successful study.

Secondly, we consider the case where a Bayesian analysis will be performed. We present two possible methods of constructing a Bayesian model, and provide each of their definitions of a successful study below.

We present the methodology for the assurance calculations in this section, followed by the resulting sample size calculations and comparisons.

6.4.1 Assurance for Frequentist Analyses

We calculate the assurance by simulation using the following steps, where R is a large number of total replications, and $P_D(\theta)$ represents the design prior distribution on the parameter, or parameters, of interest. For a sample size n,

1. Draw R samples $\theta^{(1)}, \ldots, \theta^{(R)}$ from $P_D(\theta)$.

2. Calculate the power $\beta^{(j)}$ for $\theta^{(j)}$, $j = 1, \ldots, R$.

3. The assurance is then the average power $\frac{1}{R}\sum_{j=1}^{R} \beta^{(j)}$.

We focus on Methods 2 and 3 from the power calculations, as they include estimates of the effect size. These estimates will be replaced by draws from the design prior in the assurance calculations. As Method 1 incorporated only an MCID, it does not have an assurance equivalent.

6.4.2 Assurance for Bayesian Analyses: Method One

The first Bayesian method considers a model where we focus on the proportion of patients who receive positive experimental test results, and who would not receive a positive reference test for an additional six months. This proportion represents those patients who could then be diagnosed six months earlier. Of a total n patients in a study, $n \times (1-\eta)\mu\theta_2$ of them will belong to the group with early-diagnoses. We then define a successful study as one where the posterior probability, following the study, that this proportion is greater than a certain value is larger than 0.95. In this case, we will use the MCID as the value chosen. Letting $\lambda = (1-\eta)\mu\theta_2$, this can be mathematically stated as

$$P(\text{Successful Trial}) = P_D\left[P_A(\lambda > \text{MCID}) > 0.95\right] \tag{6.8}$$

where P_D denotes the design prior distribution and P_A denotes the analysis posterior distribution.

The value of 0.95 has been selected to correspond with a significance of $\alpha = 0.05$.

The aggregated prior distributions in Chapter 4 do not take the form of Beta distributions, and thus, are not conjugate to the binomial distribution. As such, the assurance will be calculated by simulation, rather than analytically.

For a given value of n, and an $MCID$, we can calculate the assurance as follows.

1. Sample R sets of parameters $\theta^{(1)}, \ldots, \theta^{(R)}$ from $P_D(\theta)$.

2. Simulate R sets of study results of size n, using the parameters $\theta^{(1)}, \ldots, \theta^{(R)}$.

3. Calculate posterior distributions using the likelihood from 2. and the analysis priors $P_A(\theta)$. If the analysis prior is conjugate to the likelihood this can be achieved analytically. In our case, we will use MCMC.

4. The assurance is then the proportion of posterior distributions from 3. where a successful result is found, i.e., the proportion where $\Pr_A(\lambda > MCID) > 0:95$.

The sample size which achieves a desired assurance is then found by searching for the lowest value of n which provides the required assurance.

This method incorporates the MCID into the analysis step. As mentioned in Section A.2.2, this means that as the MCID increases the assurance will tend to decrease. This pattern is the opposite of that in the power calculations due to the different method of incorporating the MCID.

We refer to this method in the following sections as Bayesian Method One.

6.4.3 Assurance for Bayesian Analyses: Method Two

We also consider an alternative framing of the model. The number of positive RT results at the first time point is given by $n\eta$, while the number of positive ET results at the first time point is given by $n(\eta\theta_1 + (1-\eta)\mu\theta_2 + (1-\eta)(1-\mu)\theta_3)$. We then compare the ratio of these two values, such that a ratio greater than one represents a higher diagnosis rate from the ET, a ratio lower than one represents a higher diagnosis rate from the RT, and a ratio of one represents an equal number of patients diagnosed under both methods.

Katz et al. (1978) demonstrates that the log-ratio between two binomial distributions, $X_1 \sim Binomial(n_1, p_1)$ and $X_2 \sim Binomial(n_2, p_2)$, can be approximated by a normal distribution of the form

$$log\left(\frac{X_1}{X_2}\right) \dot{\sim} Normal\left(\frac{p_1}{p_2}, \frac{1-p_1}{n_1} + \frac{1-p_2}{n_2}\right) \tag{6.9}$$

By setting X_1 to be the number of positive RT results at the first time point, X_2 to be the number of positive ET results at the first time point, $p_1 = \eta$, and $p_2 = \eta\theta_1 + (1-\eta)\mu\theta_2 + (1-\eta)(1-\mu)\theta_3$, we can then use this ratio to make inferences about whether the two proportions differ.

The ratio $\frac{X_1}{X_2}$ is constrained to be greater than zero, and thus, the log ratio is then unbounded. As such, a significant result is found when $P(\log\left(\frac{X_1}{X_2}\right) < 0) < 0.05$. The values of n_1 and n_2 will be equal in this case, as both are the total sample size.

To calculate the assurance under this case for a given sample size n, we simulate using the following algorithm.

1. Generate R samples from the design prior distributions for η, μ, θ_1, θ_2, and θ_3.

2. Calculate R sets of values for p_1 and p_2 using the draws.

3. Simulate R datasets of size n from the Normal distribution in Equation 6.9 using the samples of p_1 and p_2 in 2. to determine the parameters for the normal approximation.

4. Calculate posterior distributions using the analysis prior distributions and the simulated datasets. If the analysis prior is conjugate to the likelihood this can be achieved analytically. In our case, an MCMC method is required.

5. Calculate the assurance by determining the proportion of datasets that would result in a positive study result, i.e., the proportion for which $\Pr(log\left(\frac{X_1}{X_2}\right) < 0) < 0.05$.

We refer to this method in the following sections as Bayesian Method Two.

6.4.4 Design and Analysis Priors

When conducting a Bayesian analysis, the assurance calculation can take into account separate design and analysis priors.

For this case study, we consider different combinations of the two. Firstly, we consider the case where the design and analysis priors are the same. This scenario is one in which all relevant prior information is included into the design and analysis. This may be particularly useful in cases where there is limited opportunity to gather data, such as studies involving rare diseases.

Next, we consider the case where the Group 2 experts form the design prior, and the Group 1 experts form the analysis prior. This represents one scenario where the experts designing the study may not be felt to be suitable to provide the prior for the analysis, for example, or where the appearance of personal bias in the prior distributions needs to be avoided. A group of experts without any conflict of interest or personal stake in the study may provide a seemingly more impartial prior, allowing an informative prior distribution to be used in the analysis stage without the appearance of undue influence.

We also calculate assurance with an expert elicited design prior, and sceptical analysis prior. This scenario represents a case where the final analysis cannot incorporate an informative prior. If the information contained in the observations is strong enough to overcome the sceptical prior, thus being strong enough to convince a sceptic that an effect is present, then strong evidence will have been found in favour of the effect. Such a setup may be suitable in cases where Bayesian methods are less common, and the inclusion of an informative prior distribution may not be accepted.

The use of a sceptical analysis prior may better suit those more familiar with frequentist methods. In a hypothesis test, an assumption of no effect is made under the null hypothesis, and a significant result is found if the resulting test statistic is very unlikely to occur under this hypothesis. Likewise, a sceptical prior assumes that the effect is not present with high probability and only if there is strong evidence in the observations to support an effect will the posterior converge on their being an effect.

For the assurance calculations presented in this chapter, the sceptical priors are chosen such that there is a 90% probability that the parameter is below the relevant value. Specifically, for Bayesian Method One, this is a $Beta(1, \beta)$ distribution, where β is minimised under the condition that the probability of the parameter being less than the MCID is at least 90%. For Bayesian Method two, the sceptical prior takes the form of a $Gamma(\alpha, 1)$ distribution, where α is maximised under the condition that the probability of the parameter being less than the MCID is at least 90%.

6.4.5 Maximum Assurance

As discussed in Chapter 3, the maximum possible assurance value can be calculated.

Table 6.6 provides the maximum assurances for each design prior and MCID used with Bayesian Method One. In this case, the maximum value is the probability, under the design prior, of observing an effect larger than the MCID.

As the values demonstrate, the maximum possible assurance is dependent on both the design prior and MCID. However, it appears that the MCID has a greater impact than the design prior. This is because the MCID defines the critical value for the design prior.

For an MCID of 0.1, for example, the maximum assurance ranges from 0.1 to 0.89. This demonstrates the large range across design priors, and the effect selecting such a prior can have. If a sample size was required to reach a certain level of assurance, say 80%, this may be outright impossible for the majority of design priors provided. To achieve an assurance of 0.9, which could be selected to match a power of 0.9 as used in frequentist calculations, the MCID could only be 0.001, or 0.05 for some design priors.

Table 6.7 provides the maximum assurances for each design prior used with Bayesian Method Two.

This method provides a range of maximum assurances ranging from 0.36 to 1.00. The

Table 6.6: Maximum Assurances for Bayesian Method One

	MCID	0.001	0.05	0.1	0.15	0.2	0.5
Aggregation	Group						
EW	1	0.99	0.33	0.11	0.03	0.01	0.00
	2	1.00	0.90	0.70	0.42	0.21	0.00
	3	0.99	0.55	0.35	0.22	0.14	0.01
CM	1	0.99	0.45	0.24	0.12	0.06	0.00
	2	1.00	0.93	0.77	0.51	0.27	0.00
	3	0.99	0.65	0.48	0.32	0.21	0.01
BA	1	0.90	0.27	0.14	0.08	0.05	0.00
	2	1.00	0.78	0.56	0.38	0.26	0.01
	3	0.96	0.43	0.26	0.17	0.11	0.01
SHELF	2	1.00	1.00	0.89	0.63	0.10	0.00

Table 6.7: Maximum Assurances for Bayesian Method Two

Aggregation	Group	Maximum Assurance
EW	1	0.36
	2	0.71
	3	0.54
CM	1	0.39
	2	0.52
	3	0.47
BA	1	0.39
	2	0.58
	3	0.47
SHELF	2	1.00

range of maximum values represent the prior probability that each aggregation method is assigning to a successful outcome. Of note, the maximum assurance of 1.00 associated with the SHELF aggregation is because it is extremely optimistic about a positive outcome being achieved, and has assigned practically all of the probability to the associated values.

6.5 Assurance Sample Size Results

In this section, we present the resulting sample sizes using assurance.

Firstly, we present an example case, using the 'most plausible' set of input values. We then present an extended set of sample size calculation results, incorporating a wider range of possible inputs.

6.5.1 Example Case

We first consider the case where the Group 2 experts, those involved in the design of the novel diagnostic test and study, are used to construct the design prior. This group is representative of the types of experts who would likely be involved in an expert elicitation for a study, and specifically, those whose views are best represented through a design prior.

For assurance calculations in which a Bayesian analysis is used, we then present sample size calculations with the Group 1 experts' prior distribution, or a sceptical prior distribution. These priors represent two possible cases. The first, Group 1's prior, represents a case where an informative prior is used. While the experts involved in designing the study may provide a prior for use in the design, such a prior may appear to have bias in favour of the new test. As such, it could be perceived that an elicited analysis prior may unduly bias the statistical analysis in favour of the experts' personal interests. As such, an informative prior elicited from a secondary group may avoid this appearance.

The second analysis prior which will be used is a sceptical prior. This prior represents a sceptical viewpoint, that the new test or treatment will provide no additional benefit over the current one. Thus, if the evidence presented by the collected dataset is strong enough to overcome this sceptical prior, then the data has provided a strong level of proof in favour of the new test or treatment. For Bayesian Method One, the sceptical prior used is a $Beta(1, \beta)$ distribution, where β is minimised under the condition that the probability that the parameter is less than the MCID is at least 90%. For Bayesian Method two, the sceptical prior takes the form of a $Gamma(\alpha, 1)$ distribution, where α is maximised under the condition that the probability that the parameter is less than the one is at least 90%.

For the following calculations, the target assurance has been set to 50%. This means the sample sizes reflect a better than even chance of a successful study. Group 2's design priors lead to maximum assurances greater than 50%, allowing this target to be achieved under all aggregation methods.

Frequentist Analyses

We consider assurance calculations using two frequentist analyses. Table 6.8 presents these sample sizes calculated using assurance.

The sample sizes for Frequentist Method Two demonstrate the differences between the aggregation methods. Considering the maximum feasible sample size of 120, the Bayesian aggregation and SHELF priors both present sample sizes lower than this to achieve a 50% assurance. Both Equal Weights and the Classical Method priors, on the other hand, suggest a higher sample size is required.

Under the Frequentist Method Three of analysis, the sample sizes vary mainly under the change in MCID. The table suggests an MCID of 0.15 could be detected with sample sizes of less than 120, although differences closer to 0.1 may be detected based on some of the aggregations. The Group 2 experts indicate an MCID of 0.15 would be a desirable target for the study, which suggests that there is a better than even chance that the study will successfully find a clinically significant difference.

In comparison to Table 6.3 and Table 6.5, the sample sizes provided by the assurance calculations tend to be lower than those provided by power. While it is important to note that the selection of the target assurance has influenced this, and that selecting a different target would result in different sample sizes, this does provide an example of when assurance can provide a lower sample size.

Table 6.8: Sample Sizes using Assurance with a Frequentist Analysis

Aggregation Method	Method Two Sample Size	Method Three, MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	197	>5000	499	144	73	46	12
CM	229	>5000	546	156	78	49	13
BA	68	>5000	430	127	65	42	12
SHELF	48	>5000	575	163	81	51	13

Bayesian Analyses

We also compare our two Bayesian models within our assurance calculations.

For this section, analysis priors are used in addition to the design prior. In both cases, the design prior is taken from Group 2, and the analysis prior is either taken from Group 1 or a sceptical prior is used. The aggregation method used is common across both priors. It should also be noted that the SHELF aggregation was only conducted with Group 2, and as such is only used alongside a sceptical analysis prior as there is no equivalent Group 1 prior.

The resulting sample sizes for Bayesian Method One are provided in Table 6.9.

For a sample size to be equal to or less than the 120 maximum patients, an MCID of 0.05 or less would be required. Under this model, the sample sizes required for a posterior distribution to have a 95% probability of being above an MCID of 0.15 or higher are not feasible.

The difference between the analysis priors is demonstrated. The Group 1 analysis priors in this model are much more optimistic than the sceptical prior, and thus places more probability in the region required to conclude a successful study. As such, they require lower sample sizes to achieve a 50% assurance. If an informative prior of this type was used, less evidence would be required from the study itself in order to result in success.

The sceptical prior requires higher sample sizes in order to find a positive result. This would be expected, as additional evidence would be required to overcome the scepticism present in the analysis prior. Accordingly, a positive result under a sceptical prior may be viewed as more convincing than one under a more optimistic prior, such as that from Group 1.

The Classical Method and SHELF aggregations both resulted in assurance calculations that required lower sample sizes. This suggests that these methods were more optimistic about the novel diagnostic test's performance than the other methods.

Table 6.9: Sample Sizes using Assurance and Bayesian Method One

Aggregation Method	Analysis Prior	MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	Group 1	1	13	265	>5000	>5000	>5000
EW	Sceptical	50	86	298	>5000	>5000	>5000
CM	Group 1	1	2	63	>5000	>5000	>5000
CM	Sceptical	44	64	175	>5000	>5000	>5000
BA	Group 1	1	2	1033	>5000	>5000	>5000
BA	Sceptical	58	110	1375	>5000	>5000	>5000
SHELF	Sceptical	43	73	156	2854	>5000	>5000

Table 6.10 presents the sample sizes resulting from assurance calculations using Bayesian Method Two. In the majority of cases, the sample sizes required to achieve the 50% level of assurance are below 120. Only the Classical Method aggregation has led to a higher sample size requirement, and even so, it is still quite close to the 120 value.

The difference between the two analysis priors should also be noted. In this case, the

Table 6.10: Sample Sizes using Assurance and Bayesian Method Two

Aggregation Method	Analysis Prior	Sample Size
EW	Group 1	14
CM	Group 1	130
BA	Group 1	39
EW	Sceptical	14
CM	Sceptical	130
BA	Sceptical	38
SHELF	Sceptical	5

sample sizes resulting from both priors are very similar. This is due to the Group 1 prior distributions being implicitly sceptical about the difference between the two tests, when calculated using Bayesian Method Two.

6.5.2 Further Sample Size Calculations

While the previous section presents what we believe to be the most appropriate assurance calculations, we also consider further combinations of inputs into assurance calculations. We also consider a full range of design and analysis priors, with each expert group considered as both a design and an analysis prior.

6.5.3 Target Assurance

While sample size calculations using power typically aim to achieve a power of 0.8 or 0.9, there is no common target for assurance. As the results in the section have a wider range of inputs than the previous case, we consider two methods of setting a target assurance for the following calculations.

To begin, we consider a target assurance of 90% of the maximum assurance possible under the design prior. This method is similar to power, as it can be used without the need for additional information and can be implemented across different scenarios or prior distributions. It will also be consistent with power in cases where the maximum assurance is equal to one.

An alternative is to select a context specific target assurance, for example to set the target assurance based on the usual success rates for similar studies. If the study being designed has an equal or better chance to succeed than other similar studies, then the sample size may be appropriate. This option, however, does require additional effort, as success rates of similar studies need to be found.

Thomas et al. (2016) find Phase I clinical studies have a 9.6% chance of being approved. For Neurology in particular, this changes to an 8.4% chance to be approved. Success rates for studies involving neurology diagnostic tests do not appear to be as widely available, and so for the purposes of this study, we will use a target of 10%. This target assurance means the sample size selected will provide a better than average chance of the study being a success.

We also acknowledge that the target assurance can be selected to achieve a required probability of success. In such a case, a design prior which does not allow for this required target to be achieved would suggest the study should not proceed. A target assurance may be selected in this way to maximise a utility function. As demonstrated in Kunzmann et al. (2021), given an expected return for a successful study, and the associated costs for completing it with the relevant sample sizes, then the sample size can be chosen to maximise the expected utility.

Frequentist Analyses

We first consider assurance based on performing McNemar's test in the analysis. The calculated values are provided in Table 6.11.

For Method Two, sample sizes vary according to both the design prior and the target assurance. Group 2's prior distributions tend to lead to the largest sample sizes for each method, suggesting a less positive view on the new diagnostic test than that of Group 1.

There were also large differences based on the target assurance. In this case, the target assurance of 10% provides very low sample sizes, suggesting it is not necessarily a suitable target. The target of 90% of the maximum possible assurance provides more realistic sample sizes.

For Method Three, we note there is a large variation in sample sizes due to the MCID. In all cases, the sample size required to detect a difference of 0.001 is greater than 10,000. The exact values have not been calculated. Higher MCIDs in turn provide lower required samples sizes.

Table 6.11: Maximum Assurances for McNemar's test analysis

Aggregation Method	Design Prior Group	Target Assurance	Method Two Sample Size	Method Three, MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	1	90% max	125	>10000	413	153	89	62	19
	1	10%	7	>10000	18	8	5	4	3
	2	90% max	533	>10000	1359	390	194	121	29
	2	10%	23	>10000	57	18	10	7	3
	3	90% max	145	>10000	716	229	123	81	22
	3	10%	8	>10000	31	11	7	5	3
CM	1	90% max	138	>10000	540	185	103	70	20
	1	10%	7	>10000	24	9	6	5	3
	2	90% max	622	>10000	1489	422	209	129	30
	2	10%	27	>10000	62	19	10	7	3
	3	90% max	137	>10000	990	297	153	98	25
	3	10%	7	>10000	42	14	8	6	3
BA	1	90% max	105	>10000	335	133	81	57	18
	1	10%	6	>10000	15	7	5	4	3
	2	90% max	167	>10000	1155	338	172	108	27
	2	10%	8	>10000	48	15	9	6	3
	3	90% max	133	>10000	651	213	116	77	21
	3	10%	7	>10000	28	10	6	5	3
SHELF	2	90% max	127	>10000	1567	442	218	134	31
	2	10%	7	>10000	65	20	11	7	3

Bayesian Analyses

We also consider sample sizes required for a Bayesian analysis, as provided in Table 6.12. Sample sizes above 1,000 are not calculated exactly.

This table demonstrates the difference between the sceptical analysis prior, and those informed by the experts. In many cases, an expert analysis prior is already optimistic enough that further data does not need to be gathered in order for the assurance target to be reached. The sceptical prior, alternatively, always requires further data before the target assurance can be achieved.

The change in sample size as the MCID changes is much greater in this case, compared to the frequentist analysis. In many cases, the sample sizes change from 1 to >1,000 with a modest change in MCID. This is likely due to the design priors, as many assign the lowest 5% of their probability within the regions of the differences being compared. When the difference is lower than this 5% boundary, the required sample size is often very small. However, when it is closer to the boundary, the sample size requirement increases rapidly.

It is also noted that only the more optimistic combination of priors, alongside the more lenient 10% assurance target, allow for evidence to be found for the difference between diagnostic tests to be greater than 0.2. It is further noted that while many sample sizes in the table are reported as >1000, many may not actually be possible. For example, consider a design prior which places all of its weight below 0.5. No matter how large a sample is taken from this design prior, it is extremely unlikely that the posterior distribution will place a high enough probability above an MCID of 0.5 for the study to be considered successful.

While the 10% assurance target is unrealistic for many expert informed analysis prior distributions, it often presents more reasonable sample sizes when using a sceptical analysis prior. This is an important point to consider, as the sceptical prior with the 90% maximum assurance target rarely gives a sample size within a reasonable range. The choice of both target assurance and analysis prior clearly play an important role in determining the sample size. For example, if a target assurance is chosen to be 0.8 or 0.9, perhaps to align with power, then an overly sceptical prior may lead to an unreasonable sample size requirement.

Table 6.12: Maximum Assurances for Bayesian Method One

Aggregation Method	Design Prior Group	Analysis Prior Group	Target Assurance	MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	1	1	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	1	10%	1	9	>1000	>1000	>1000	>1000
	1	2	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	2	10%	1	1	>1000	>1000	>1000	>1000
	1	3	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	3	10%	1	1	>1000	>1000	>1000	>1000
	1	Sceptical	90% max	>1000	>1000	>1000	>1000	>1000	>1000
	1	Sceptical	10%	62	164	>1000	>1000	>1000	>1000
	2	1	90% max	1	109	>1000	>1000	>1000	>1000
	2	1	10%	1	3	36	136	715	>1000
	2	2	90% max	1	1	>1000	>1000	>1000	>1000
	2	2	10%	1	1	3	23	253	>1000
	2	3	90% max	1	1	1000	>1000	>1000	>1000
	2	3	10%	1	1	3	20	198	>1000
	2	Sceptical	90% max	161	434	>1000	>1000	>1000	>1000
	2	Sceptical	10%	27	29	40	74	343	>1000
	3	1	90% max	1	497	>1000	>1000	>1000	>1000
	3	1	10%	1	3	37	151	720	>1000
	3	2	90% max	1	294	>1000	>1000	>1000	>1000
	3	2	10%	1	1	2	25	272	>1000
	3	3	90% max	1	448	>1000	>1000	>1000	>1000
	3	3	10%	1	1	3	19	224	>1000
	3	Sceptical	90% max	829	>1000	>1000	>1000	>1000	>1000
	3	Sceptical	10%	27	34	46	88	363	>1000
CM	1	1	90% max	1	999	>1000	>1000	>1000	>1000
	1	1	10%	1	2	25	1517	>1000	>1000
	1	2	90% max	1	173	>1000	>1000	>1000	>1000
	1	2	10%	1	1	2	999	>1000	>1000
	1	3	90% max	1	750	>1000	>1000	>1000	>1000
	1	3	10%	1	1	3	>1000	>1000	>1000
	1	Sceptical	90% max	>1000	>1000	>1000	>1000	>1000	>1000
	1	Sceptical	10%	37	55	117	>1000	>1000	>1000
	2	1	90% max	1	61	747	>1000	>1000	>1000
	2	1	10%	1	2	7	42	293	>1000
	2	2	90% max	1	1	482	>1000	>1000	>1000
	2	2	10%	1	1	2	10	139	>1000
	2	3	90% max	1	1	644	>1000	>1000	>1000
	2	3	10%	1	1	3	5	141	>1000
	2	Sceptical	90% max	119	296	>1000	>1000	>1000	>1000
	2	Sceptical	10%	24	26	34	58	188	>1000
	3	1	90% max	1	174	889	>1000	>1000	>1000
	3	1	10%	1	2	8	39	137	>1000
	3	2	90% max	1	1	493	>1000	>1000	>1000
	3	2	10%	1	1	2	11	83	>1000
	3	3	90% max	1	1	>1000	>1000	>1000	>1000
	3	3	10%	1	1	2	8	43	>1000
	3	Sceptical	90% max	847	539	>1000	>1000	>1000	>1000
	3	Sceptical	10%	21	25	33	43	125	>1000
BA	1	1	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	1	10%	1	2	190	>1000	>1000	>1000
	1	2	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	2	10%	1	1	>1000	>1000	>1000	>1000
	1	3	90% max	1	>1000	>1000	>1000	>1000	>1000
	1	3	10%	1	1	79	>1000	>1000	>1000
	1	Sceptical	90% max	>1000	>1000	>1000	>1000	>1000	>1000
	1	Sceptical	10%	48	80	388	>1000	>1000	>1000
	2	1	90% max	1	420	774	>1000	>1000	>1000
	2	1	10%	1	2	7	26	85	>1000
	2	2	90% max	1	1	763	>1000	>1000	>1000
	2	2	10%	1	1	2	4	20	>1000
	2	3	90% max	1	240	861	>1000	>1000	>1000
	2	3	10%	1	1	3	1	46	>1000
	2	Sceptical	90% max	228	664	>1000	>1000	>1000	>1000
	2	Sceptical	10%	20	23	26	46	80	>1000
	3	1	90% max	1	952	1552	>1000	>1000	>1000
	3	1	10%	1	2	24	120	507	>1000
	3	2	90% max	1	844	>1000	>1000	>1000	>1000
	3	2	10%	1	1	2	28	149	>1000
	3	3	90% max	1	747	>1000	>1000	>1000	>1000
	3	3	10%	1	1	3	58	203	>1000
	3	Sceptical	90% max	>1000	>1000	>1000	>1000	>1000	>1000
	3	Sceptical	10%	29	39	58	79	311	>1000
SHELF	2	2	90% max	1	1	93	>1000	>1000	>1000
	2	2	10%	1	1	1	58	>1000	>1000
	2	Sceptical	90% max	77	171	619	>1000	>1000	>1000
	2	Sceptical	10%	21	31	61	164	>1000	>1000

6.5.4 Bayesian Method Two

We provide sample sizes for the assurance calculations using Bayesian Method Two in Table 6.13.

Unlike the previous methods, there is not a large difference between the use of a sceptical analysis prior and an expert aggregated analysis prior. Under this formulation, many of the expert priors are quite sceptical. In many cases, the Group 2 design or analysis prior tends to lead to smaller sample size requirements, again suggesting a more

Table 6.13: Maximum Assurances for Bayesian Method Two

Aggregation Method	Design Prior Group	Analysis Prior Group	Sample Size (Target Assurance of 10%)	Sample Size (90% Max Target Assurance)
EW	1	1	7	74
	1	2	8	61
	1	3	9	76
	1	Sceptical	9	79
	2	1	4	34
	2	2	3	30
	2	3	3	31
	2	Sceptical	4	32
	3	1	4	44
	3	2	4	38
	3	3	3	40
	3	Sceptical	4	43
CM	1	1	5	68
	1	2	7	70
	1	3	7	72
	1	Sceptical	7	73
	2	1	6	55
	2	2	4	51
	2	3	6	53
	2	Sceptical	6	53
	3	1	5	50
	3	2	5	48
	3	3	3	47
	3	Sceptical	5	50
BA	1	1	3	39
	1	2	5	40
	1	3	5	44
	1	Sceptical	5	44
	2	1	4	35
	2	2	3	31
	2	3	4	34
	2	Sceptical	4	34
	3	1	4	36
	3	2	4	33
	3	3	3	32
	3	Sceptical	3	32
SHELF	2	2	2	6
	2	Sceptical	2	8

optimistic view from the experts in this group.

As with the assurance calculations using Bayesian Method One, the 10% assurance level provides very small sample sizes. It is also the case for the sceptical priors in this case, as there is little difference between analysis priors. The 90% maximum assurance target, or the 50% target as used in the initial scenario, seem more appropriate choices.

The differences between the two sets of assurance calculations for a Bayesian analysis of the study demonstrates the further impact which the choice of model can have on assurance calculations.

6.6 Assurance and Power at the Maximum Feasible Sample Size

During the initial design and application for funding of the study, the designers stated that the maximum feasible number of patients who could be recruited was 120. In practice, it may not be possible to recruit more than this maximum feasible sample size. In cases such as these, it may be important to determine which effect sizes can be detected, or what assurance or power can be achieved for a set sample size.

In this section, we first investigate which effect sizes could be detected using power for a sample size of 120. We also investigate the corresponding assurance which could be achieved. Such a value may be useful to determine whether the feasible recruitment options for a study will impede its ability to be successful, and what the probability of a successful study will be in the best-case scenario.

6.6.1 Detectable Effect Sizes

For a given sample size, in this case 120, it is possible to calculate the minimum effect size to be detected which will still provide a certain level of power, here 0.9.

Under the Method One power calculations, which only uses an MCID as an input, an effect size of 0.09 or larger could be detected. Smaller effect sizes would have a lower probability of being detected given the required power. Four of the six experts suggested MCIDs larger than this value. These experts' MCIDs would be able to be detected at this sample size. The remaining two experts would consider an effect size of 0.09 to be clinically important, but would also consider smaller effects, which would not be able to be detected, to be clinically important.

Under the Method Three power calculations, which use both an MCID and expert estimates as inputs, the following effect sizes could be determined for each aggregation method, as displayed in Table 6.14.

This table demonstrates that only relatively large effect sizes could be detected when using expert estimates for the reference diagnostic test.

Table 6.14: Effect Sizes detectable for $n = 120$

Aggregation	Group	Detectable Effect
EW	1	0.24
	2	0.20
	3	0.21
CM	1	0.23
	2	0.25
	3	0.24
BA	1	0.24
	2	0.22
	3	0.23s
SHELF	2	0.14

Under each aggregation method, with the exception of SHELF, an effect size of 0.2 would not be able to be detected with the required power. Only one of the MCIDs elicited from the experts is above these effect sizes. Therefore the MCIDs of five of the six experts could not be detected with the appropriate power using this sample size.

6.6.2 Assurance and Power when $n = 120$

We consider assurance calculations when the sample size is set at 120. These assurance values represent the highest assurance that will be possible to achieve with the constraint on the number of participants.

Power

We first consider the three methods of calculating power.

The actual power provided by a sample size of 120 under Method One depends on the MCID used. As Table 6.15 demonstrates, an MCID of 0.1 is likely to provide sufficient power, while an MCID of 0.05 will not. This corresponds to the finding that an MCID of 0.09 would be appropriate given a maximum sample size of 120.

Table 6.15: Power for $n = 120$ for Method One

MCID	0.001	0.05	0.1	0.15	0.2	0.5
Power	0.06	0.69	0.94	0.99	1.00	1.00

Table 6.16 provides the power for Method Two when the sample size is 120. The power varies widely, depending on the aggregation method chosen.

Only the SHELF aggregation with Group 2 experts provides a power that would be considered sufficient by most study designs. The large discrepancy in Group 2's power estimate across different aggregation methods is due to the way individual experts are treated. The Classical Method has removed one of the experts, and it is apparent from the large drop in power that the removed expert was much more optimistic about the

Table 6.16: Power for $n = 120$ for Method Two

Aggregation	Group	Power
EW	1	0.58
	2	0.68
	3	0.13
CM	1	0.38
	2	0.07
	3	0.05
BA	1	0.41
	2	0.19
	3	0.08
SHELF	2	1.00

Table 6.17: Power for $n = 120$ for Method Three

Aggregation	Group	0.001	0.05	0.1	0.15	0.2	0.5
EW	1	0.05	0.13	0.33	0.59	0.80	1.00
	2	0.05	0.16	0.45	0.73	0.91	1.00
	3	0.05	0.15	0.40	0.69	0.88	1.00
CM	1	0.05	0.14	0.37	0.64	0.85	1.00
	2	0.05	0.12	0.32	0.57	0.79	1.00
	3	0.05	0.13	0.33	0.59	0.81	1.00
BA	1	0.05	0.13	0.33	0.60	0.81	1.00
	2	0.05	0.14	0.39	0.67	0.86	1.00
	3	0.05	0.13	0.36	0.63	0.84	1.00
SHELF	2	0.05	0.34	0.75	0.94	0.99	1.00

study than the other two experts. As such, when they received a third of the weight in the Equal Weights method, the resulting power calculation resulted in much higher power.

In this case, the Equal Weights aggregation tends to provide higher power, suggesting its inputs were more optimistic. The Classical Method resulted in the lowest power, with the Bayesian aggregation leading to power values lying between the other two mathematical aggregation methods.

We also calculate power for the Method Three power calculation, when the sample size is 120. Table 6.17 provides the resulting values.

This table demonstrates, an MCID of around 0.2 is typically required to reach an appropriate power for this method. It appears that the MCID presents a greater influence on power than the aggregation method chosen. This is expected, as the MCID represents the difference between the effectiveness of the two diagnostic tests, while the aggregation is only supplying the values for the reference test. As such, a larger difference will correspond to a larger difference between the two tests in the calculation, which in turn is easier to detect using McNemar's test.

In comparison to Table 6.16, there does not seem to be a single MCID for which the two methods are comparable. For example, the values for power using the Equal Weights aggregation in Table 6.16 are around equivalent to an MCID of 0.15, while the Bayesian aggregations' power is roughly equivalent to an MCID of between 0.1 and 0.15.

Table 6.18: Assurances for $n = 120$ for McNemar's test analysis

Aggregation Method	Design Prior Group	Method Two Assurance	Method Three, MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	1	0.89	0.05	0.41	0.82	0.97	1.00	1
EW	2	0.33	0.05	0.16	0.43	0.72	0.90	1
EW	3	0.84	0.05	0.26	0.65	0.89	0.98	1
CM	1	0.86	0.05	0.33	0.74	0.94	0.99	1
CM	2	0.29	0.05	0.15	0.40	0.69	0.90	1
CM	3	0.86	0.05	0.20	0.54	0.82	0.95	1
BA	1	0.93	0.05	0.49	0.87	0.98	1.00	1
BA	2	0.78	0.05	0.18	0.48	0.77	0.93	1
BA	3	0.87	0.05	0.28	0.68	0.91	0.98	1
SHELF	2	0.88	0.05	0.14	0.39	0.67	0.87	1.00

Assurance with Frequentist Analysis

For McNamar's test, the assurance values when $n = 120$ are provided in Table 6.18.

The majority of assurance values for Frequentist Method Two are greater than 0.8. Under this method, it would seem the study has a high probability of resulting in a statistically significant result. This is a positive sign for the researchers, as it demonstrates that their feasible sample size is a realistic and appropriate choice. Had the assurance values been low, it may have suggested that it was not appropriate to conduct the study.

In order for the Frequentist Method Three to achieve similar levels of assurance as Frequentist Method Two, an MCID of around 0.1 or 0.15 would be required.

A comparison between the power and assurance calculations can be made by comparing the power values in Tables 6.16 and 6.17 with the assurance values in Table 6.18. It appears that in the vast majority of cases, the assurance values are equal to or higher than the power for equivalent elicitation groups.

This means that the assurance calculations suggest a lower effect size could be detected with a high probability of success. For example, the power calculation suggested an effect size of around 0.2 would be the smallest which could be detected to achieve an appropriate power. The assurance calculations, however, suggest that an effect size of around 0.1 to 0.15 could be detected to achieve a 0.8 or 0.9 assurance. The additional information incorporated within the assurance calculation suggests the study may be more feasible than the power calculations suggest.

Assurance with Bayesian Analysis

We consider assurance calculations when a Bayesian analysis will be used.

Table 6.19 presents assurance values for Bayesian Method One when the sample size is 120.

Under this method, the assurance tends to drop quickly as the MCID increases. If evidence is required that the difference between tests is at least 0.1, most aggregation method design priors lead to a low assurance. The assurance using a Group 1 design prior

Table 6.19: Assurances for $n = 120$ for Bayesian Method One

Aggregation Method	Design Prior Group	Analysis Prior Group	MCID = 0.001	0.05	0.1	0.15	0.2	0.5
EW	1	1	1.00	0.23	0.03	0.00	0.00	0
EW	1	2	1.00	0.26	0.07	0.01	0.00	0.00
EW	1	3	1.00	0.23	0.06	0.01	0.00	0.00
EW	1	Sceptical	0.26	0.09	0.03	0.00	0.00	0.00
EW	2	1	1.00	0.80	0.41	0.08	0.01	0
EW	2	2	1.00	0.86	0.57	0.24	0.08	0.00
EW	2	3	1.00	0.80	0.53	0.24	0.08	0.00
EW	2	Sceptical	0.86	0.61	0.37	0.16	0.07	0.00
EW	3	1	1.00	0.45	0.20	0.09	0.04	0
EW	3	2	1.00	0.49	0.27	0.15	0.09	0.00
EW	3	3	1.00	0.49	0.29	0.15	0.09	0.00
EW	3	Sceptical	0.53	0.29	0.22	0.12	0.07	0.00
CM	1	1	1.00	0.38	0.15	0.06	0.02	0.00
CM	1	2	1.00	0.42	0.19	0.07	0.03	0.00
CM	1	3	1.00	0.38	0.17	0.08	0.03	0.00
CM	1	Sceptical	0.46	0.21	0.11	0.05	0.03	0.00
CM	2	1	1.00	0.88	0.57	0.26	0.05	0.00
CM	2	2	1.00	0.90	0.65	0.29	0.09	0.00
CM	2	3	1.00	0.88	0.61	0.29	0.11	0.00
CM	2	Sceptical	0.90	0.69	0.41	0.20	0.08	0.00
CM	3	1	1.00	0.60	0.36	0.19	0.10	0.00
CM	3	2	1.00	0.63	0.43	0.22	0.12	0.00
CM	3	3	1.00	0.60	0.40	0.24	0.13	0.00
CM	3	Sceptical	0.63	0.43	0.29	0.18	0.10	0.01
BA	1	1	1.00	0.21	0.09	0.04	0.02	0.00
BA	1	2	1.00	0.23	0.10	0.06	0.03	0.00
BA	1	3	1.00	0.23	0.10	0.05	0.02	0.00
BA	1	Sceptical	0.30	0.14	0.07	0.04	0.02	0.00
BA	2	1	1.00	0.71	0.42	0.24	0.12	0.00
BA	2	2	1.00	0.71	0.45	0.29	0.17	0.00
BA	2	3	1.00	0.71	0.42	0.25	0.16	0.00
BA	2	Sceptical	0.75	0.51	0.33	0.21	0.14	0.00
BA	3	1	1.00	0.34	0.10	0.10	0.06	0.00
BA	3	2	1.00	0.37	0.20	0.13	0.08	0.00
BA	3	3	1.00	0.37	0.20	0.11	0.07	0.00
BA	3	Sceptical	0.41	0.22	0.15	0.10	0.06	0.00
SHELF	2	2	1.00	0.99	0.84	0.17	0.00	0
SHELF	2	Sceptical	0.98	0.84	0.39	0.13	0.01	0.00

is lower than equivalent calculations for other design priors, and the sceptical analysis priors tend to lead to a lower assurance than an informative prior from any expert group.

These assurance values are quite pessimistic about the chances of a successful study overall. To find evidence in favour of at least a 10% difference between tests appears unlikely, and a lower difference is not considered clinically significant by the majority of experts.

We also consider assurance for Bayesian Method Two, when the sample size is 120. These values are provided in Table 6.20.

Under this method, the majority of variation in assurance values appears to be driven by changes in the design prior. Each analysis prior for a set design prior only differs by around 0.01. This could suggest that the analysis priors are all quite similar here, or that the sample size is large enough that the analysis prior is dominated by the observations.

Many of the resulting assurance values suggest there is a better than even chance of the study successfully finding a statistical difference between the two diagnostic tests. As seen previously, the Group 1 aggregated priors have resulted in more sceptical assurance values than the Group 2 experts.

In comparison to Table 6.19, this method provides assurance values similar to those which considered an MCID of 0.001 and 0.05. This region is not considered clinically

Table 6.20: Assurances for $n = 120$ for Bayesian Method Two

Aggregation Method	Design Prior Group	Analysis Prior Group	Assurance
EW	1	1	0.3404
EW	1	2	0.3446
EW	1	3	0.3402
EW	1	Sceptical	0.3391
EW	2	1	0.6886
EW	2	2	0.6902
EW	2	3	0.6903
EW	2	Sceptical	0.6897
EW	3	1	0.5226
EW	3	2	0.5251
EW	3	3	0.5242
EW	3	Sceptical	0.5235
CM	1	1	0.3682
CM	1	2	0.3684
CM	1	3	0.3675
CM	1	Sceptical	0.3678
CM	2	1	0.4983
CM	2	2	0.4993
CM	2	3	0.4990
CM	2	Sceptical	0.4989
CM	3	1	0.4488
CM	3	2	0.4504
CM	3	3	0.4499
CM	3	Sceptical	0.4499
BA	1	1	0.3819
BA	1	2	0.3817
BA	1	3	0.3807
BA	1	Sceptical	0.3807
BA	2	1	0.5643
BA	2	2	0.5656
BA	2	3	0.5645
BA	2	Sceptical	0.5647
BA	3	1	0.4521
BA	3	2	0.4534
BA	3	3	0.4533
BA	3	Sceptical	0.4533
SHELF	2	2	1.0000
SHELF	2	Sceptical	1.0000

important by many of the experts. This analysis, however, does not consider an MCID, instead aiming to simply compare whether a difference exists between the diagnostic tests. Should a statistically significant result be found in such an analysis, it would then be a further question as to whether the results are clinically significant.

6.7 Further Discussion

One point demonstrated throughout this chapter is the importance of the model specification. The different methods of incorporating values into the frequentist test, and the two different ways of specifying the Bayesian model, can all lead to quite different sample sizes being required. This variation demonstrates that if an inappropriate choice is made when specifying a model or statistical test, then the resulting sample size requirements can also be inappropriate.

In the case of the frequentist tests, we considered three ways of determining input values. The first simply took into account an MCID, and no additional prior information about the reference test. We suggest that for the case study considered in this thesis, the first method is the least appropriate. As the later methods demonstrate, the experts' prior distributions all suggested the reference test's effect was greater than zero, implying that this case was not realistic as to the outcome of the trial.

The second and third methods both incorporated some expert elicited information. Both seem reasonable methods to use, and so the choice between them would depend on the exact research question of interest. If a researcher is interested in detecting a statistically significant difference between the two diagnostic tests, regardless of what the difference is, then the second method would appear an appropriate choice. Alternatively, if a researcher is only interested in detecting a statistically significant MCID, then the third method may be a more appropriate choice.

Likewise in the Bayesian analysis case, the choice of model should be led by the research question. The first method looks at the improvement of the novel diagnostic test over the current test, while the second method investigates the difference between the two tests. While this distinction is subtle, the change in required sample sizes demonstrates that the choice of model clearly affects assurance calculations.

The method of incorporating an MCID into the calculations was also demonstrated to play a large effect. As shown in Section A.2.2, there is a large difference between a null hypothesis test for a difference of at least the MCID, and a null hypothesis test which tests for any difference, but is powered on an effect size of the MCID. Given there is not necessarily a standard way for assurance calculations to incorporate an MCID when conducting a Bayesian analysis, it is clearly an important consideration when using assurance.

Another consideration to be made when calculating assurance is the ownership of the beliefs represented by the prior distributions. For example, the SHELF method provides a distribution which is representative of a Rational Impartial Observer, who has considered the group's evidence. Further ownership may be assigned to the group as a whole, as they ideally would have achieved a consensus between themselves. Alternatively, the

Bayesian aggregation presented represents a decision maker's beliefs, after they have been updated to account for the elicited values from the experts. Both Equal Weight and the Classical Method, however, do not lend themselves to a single owner for the aggregated prior distribution as the individual experts will not see, or necessarily agree with, the final aggregated distribution. Instead, they are mathematical combinations of the group of experts' individual beliefs.

As the assurance calculations are dependent on the prior distributions, it follows then that prior distributions should be justifiable in order for the resulting sample sizes to be reasonable. Consideration of the ownership of the prior distributions is one way this could be considered. Given that the experts involved are experts in the relevant field, and the values they provided are their honest attempts at specifying their prior probabilities, then their individual prior distributions appear valid and justifiable to use. As the aggregated distributions are formed from these, it suggests that the information contained within them is likewise justifiable.

Whether the priors are aiming to represent the beliefs of a group of experts, or the wider view of the current knowledge within the field, is another consideration. Given a number of experts who have a current and comprehensive knowledge of the current state of their field of expertise, their prior distributions may likely take into account a wide range of expertise.

We recommend considering, and where relevant, reporting, the ownership of both design and analysis priors. Especially when these priors differ, it is important for reproducibility and clarity that the details of any informative priors used are provided. This ensures that the proper context for the design and analysis can be understood. For example, there may be different interpretations made when different prior distributions are used.

Our results tended to demonstrate that the use of sceptical analysis priors required higher sample sizes than if expert elicited priors were used. While a lower sample size is often preferable, a sceptical analysis prior does, however, ensure that there is no perception of undue influence or bias in the analysis of the study, and the results may be considered stronger or more convincing. An informative analysis prior could be viewed as influencing the final results and requiring weaker information in the observations to still be considered a significant result.

While a sceptical analysis prior may require stronger evidence and thus lead to more convincing conclusions, the informative prior still has uses. As it tended to lead to smaller requirements for sample size, in cases such as rare diseases where a large sample size may not be possible, an informative analysis prior may allow a study to be run feasibly.

We also note that the elicited prior distributions often appeared reasonably similar between Group 1 and Group 2. The small differences, however, had a noticeable impact when carried forward to assurance calculations. In many of the calculations in this chapter, the required sample sizes under Group 1 or Group 2 prior distributions varied widely. This suggests that assurance calculations can be quite sensitive to the prior distributions, which in turn outlines the importance of following proper expert elicitation guidelines. The higher the quality of the expert elicitations, the higher the quality of the assurance sample sizes.

6.8 Conclusions

In this chapter, we have demonstrated how power and assurance based sample size calculations can be used in the design of an example clinical study. We have compared assurance calculations for both a frequentist and Bayesian analysis, considering multiple different model choices. Across these, we showed that assurance calculations suggested that either smaller differences between the two diagnostic tests could be detected, or a smaller sample size could be required.

We also considered cases where different prior distributions were used for the design and analysis stages of a study, and demonstrated how this can affect the sample size requirements. We also provided example assurance and power values for a sample size of 120, which the experts designing the study felt was the maximum number of patients that could feasibly be recruited.

Part VII

Conclusions and Future Work

Chapter 7

Conclusions

7.1 Summary

Throughout this thesis, we have explored the use of subjective Bayesian methods for the design and analysis of a study investigating a novel diagnostic test. We have presented original work in the areas of elicitation and aggregation of expert judgements, and Bayesian sample size calculations.

In the context of the case study, we investigated the performance of multiple prior distribution aggregation methods. This allowed for a comparison between methods that had not previously been compared. We conducted two rounds of elicitations, involving ten experts, to cover both those directly involved in the trial and those who were not.

We suggest that assurance may be a more robust method for sample size calculation, especially when the observed results do not correspond to those used in the calculations. We have provided additional evidence about the behaviour of Bayesian assurance, often in comparison to statistical power, through simulations. We have also provided examples of how statistical power and Bayesian assurance can be used in practice through a case study.

Chapter 2 reviewed approaches to sample size calculation. We reviewed both statistical power and Bayesian assurance, and discussed issues such as effect size estimation. In the case of Bayesian assurance, we considered the use of different prior distributions in the design and analysis stages of the trial, and how this can be accounted for in the assurance calculation.

Chapter 3 reviewed elicitation of expert judgments, for use in Bayesian assurance calculations. We reviewed issues such as cognitive biases, and how they may affect the values provided by experts, as a foundation for how expert judgments should be elicited. We also considered the problem of elicitation from a group of experts, and how multiple views can be combined. We reviewed common elicitation aggregation methods, and detailed the development of two elicitations which would put these methods into practice.

Chapter 4 presented the results of the two elicitations, completed as part of the design of a study into a novel diagnostic test for Motor Neuron Disease. Results from these elicitations included judgements from experts directly involved in the development of the trial, and experts who were not involved in the development of either the diagnostic test or trial. Parameters required for both the case study sample size calculations and seed questions, to which the elicitors knew the answers, were elicited from the experts. These elicited judgments were aggregated using different aggregation methods, and the resulting aggregations used to compare common aggregation methods.

Assessing the aggregated distributions for the seed questions suggested that any method of aggregation offered an improvement over eliciting from a single expert. Across the aggregation methods, the SHELF and Classical Method both tended to perform better than other methods considered.

Chapter 5 explored statistical power and Bayesian assurance through simulation. We investigated the difference in performance between these two methods given similar inputs, and when the effect size or prior distribution mean is different to that observed when the trial is run. We also provided simulations investigating assurance, such as how the maximum assurance value can vary given the chosen design prior distribution. Additionally, we also presented simulations demonstrating how assurance varies given changes to the prior distribution to be used in the analysis, and suggested some options which may be suitable for use as analysis prior distributions.

Chapter 6 presented the sample size calculations for the case study. Using the elicited parameters and aggregated distributions, statistical power and Bayesian assurance were both used to calculate the required sample size for this study. We considered a range of different methods for specifying the inputs to the statistical power calculation, incorporating both elicited values and elicited Minimal Clinically Important Differences. We also considered two Bayesian models for the analysis of the data, with a number of different prior distributions for use in the analysis, and incorporated these into the Bayesian assurance calculations. Finally, we considered the plausible maximum number of patients who could be expected to be recruited to the study, and determined what effect size could be detected.

7.2 Future Work

While this thesis has addressed many questions, many more remain to be investigated.

While the results presented provide a comparison of popular elicitation methods, further work could seek to expand on this in a number of ways. Firstly, repeating the comparison with additional experts would provide additional evidence in comparing aggregation methods. As the SHELF method was only conducted with a single group of experts, it would be particularly useful to again conduct the SHELF method alongside the other methods to further compare between the methods with a larger group of experts.

Another area which may be of particular interest is a comparison between online and in-person elicitations. Due to the COVID-19 pandemic, in-person meetings were not an option when conducting the second round of elicitations. However, with the increase in remote working and widespread adoption of video-conference style meetings, there is an opportunity for elicitations to more easily take place without all experts and elicitors being physically present in the same location. Work comparing online survey elicitations, to online video elicitations, to in-person elicitations, could help determine the validity of each option and present recommendations for future elicitations.

The trial used for the case study presented has not, at the time of writing, resulted in the collection of any data. Should data be collected, a comparison of the trial results to the elicited prior distributions and sample size calculations would provide further insight into the methods. For example, investigating prior-data conflict would allow further comparisons between individual experts and aggregated distributions. Additionally, the sample size results could be investigated by bootstrapping the data and comparing datasets of the suggested sample sizes to determine if the assurance or power used in the sample size calculations would be replicated in practice.

There are also further assurance calculations which could be considered. Further work could consider other methods of analysis, and comment on their impact on assurance calculations. More widely, there are many other statistical tests and models for which assurance could be used to determine a sample size.

Part VIII

References

Bibliography

Daniella Acker and Nigel W. Duck. Cross-cultural overconfidence and biased self-attribution. *Journal of Socio-Economics*, 37(5):1815–1824, 2008. ISSN 10535357. doi: 10.1016/j.socec.2007.12.003.

Ziyad A Alhussain and Jeremy E Oakley. Eliciting judgements about uncertain population means and variances Prediction and Estimations View project Numerical Investigations of certain fluid flow problems using Comsol Multiphysics View project Eliciting judgements about uncertain population mean. *arXiv preprint*, 2017. URL `https://www.researchgate.net/publication/313366566`.

Ziyad A. Alhussain and Jeremy E. Oakley. Assurance for clinical trial design with normally distributed outcomes: Eliciting uncertainty about variances. *Pharmaceutical Statistics*, 19(6):827–839, 2020. ISSN 1539-1612. doi: 10.1002/PST.2040. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/pst.2040https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2040https://onlinelibrary.wiley.com/doi/10.1002/pst.2040`.

W. P. Aspinall and R. M. Cooke. Quantifying scientific uncertainty from expert judgement elicitation. In *Risk and Uncertainty Assessment for Natural Hazards*, volume 9781107006, pages 64–99. Cambridge University Press, 2011. ISBN 9781139047562. doi: 10.1017/CBO9781139047562.005.

W. P. Aspinall, R. M. Cooke, A. H. Havelaar, S. Hoffmann, and T. Hald. Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. *PLoS ONE*, 11(3), 2016. ISSN 19326203. doi: 10.1371/journal.pone.0149817.

Australian Government: Department of Health. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (Version 5.0). Technical report, 2016.

Peter Bacchetti, Charles E. McCulloch, and Mark R. Segal. Simple, Defensible Sample Sizes Based on Cost Efficiency. *Biometrics*, 64(2):577–585, 2008. ISSN 0006341X. doi: 10.1111/j.1541-0420.2008.01004_1.x. URL `http://doi.wiley.com/10.1111/j.1541-0420.2008.01004{_}1.x`.

T. Bedford, L. Walls, and Matthias Troffaes. Evaluation of elicitation methods to quantify bayes linear models. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 224(4), 2010. ISSN 1748006X. doi: 10.1243/1748006XJRR304.

James O Berger and Donald a Berry. Statistical analysis and the illusion of objectivity (Letters to editor p. 430-433). *American Scientist*, 76, 1988. ISSN 00030996.

Anja Bertsche, Frank Fleischer, Jan Beyersmann, and Gerhard Nehmiz. Bayesian Phase II optimization for time-to-event data based on historical information. *Statistical Methods in Medical Research*, 28(4):1272–1289, 2019. ISSN 14770334. doi: 10.1177/0962280217747310. URL `https://pubmed.ncbi.nlm.nih.gov/29284369/`.

Laura Bojke, Marta Soares, Aimee Fox, Dina Jankovic, Karl Claxton, Alec Morton, Linda Sharples, Christopher Jackson, Andrea Taylor, and Abigail Colson. Developing a reference protocol for expert elicitation in healthcare decision making. *Health Technology Assessment Reports*, 2019.

Fergus Bolger. The selection of experts for (probabilistic) expert knowledge elicitation. In *International Series in Operations Research and Management Science*, volume 261. 2018. doi: 10.1007/978-3-319-65052-4_16.

Fergus Bolger and Gene Rowe. The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? *Risk Analysis*, 35(1):5–11, 2015. ISSN 02724332. doi: 10.1111/risa.12272. URL `http://doi.wiley.com/10.1111/risa.12272`.

Robert F. Bordley, Keith Lehrer, and Carl Wagner. Rational Consensus in Science and Society. *Noûs*, 20(4):565, 1986. ISSN 00294624. doi: 10.2307/2214987.

Shane R. Brady. Utilizing and Adapting the Delphi Method for Use in Qualitative Research. *International Journal of Qualitative Methods*, 14(5):160940691562138, 2015. ISSN 1609-4069. doi: 10.1177/1609406915621381. URL `http://journals.sagepub.com/doi/10.1177/1609406915621381`.

Caroline Brard, Gwé Naël Le Teuff, Marie-Cécile Le Deley, and Lisa V Hampson. Bayesian survival analysis in clinical trials: What methods are used in practice? *Clinical Trials*, 14(1):78–87, 2017. doi: 10.1177/1740774516673362.

Jürgen Breckenkamp, Gabriele Berg-Beckhoff, Eva Münster, Joachim Schüz, Brigitte Schlehofer, Jürgen Wahrendorf, and Maria Blettner. Feasibility of a cohort study on health risks caused by occupational exposure to radiofrequency electromagnetic fields. *Environmental Health: A Global Access Science Source*, 8(1), 2009. ISSN 1476069X. doi: 10.1186/1476-069X-8-23. URL `https://pubmed.ncbi.nlm.nih.gov/19480652/`.

Klaus Brockhoff, A Simon, D W Smithburg, and V A Thompson. Executives' Forecasts of Earnings per Share versus Forecasts of Naive Models. In *Journal of Business*, volume 126, pages 25–52. and, Public Administration, 1970.

Matthew Browne, Vijay Rawat, Philip Newall, Stephen Begg, Matthew Rockloff, and Nerilee Hing. A framework for indirect elicitation of the public health impact of gambling problems. *BMC Public Health*, 20(1): 1–14, 2020. ISSN 14712458. doi: 10.1186/s12889-020-09813-z. URL `https://link.springer.com/articles/10.1186/s12889-020-09813-zhttps://link.springer.com/article/10.1186/s12889-020-09813-z`.

Richard H. Browne. On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940, 1995. ISSN 10970258. doi: 10.1002/sim.4780141709. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780141709https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780141709https://onlinelibrary.wiley.com/doi/10.1002/sim.4780141709`.

David V. Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2), 2015. ISSN 15265501. doi: 10.1287/mnsc.2014.1909.

Mark A. Burgman, Marissa McBride, Raquel Ashton, Andrew Speirs-Bridge, Louisa Flander, Bonnie Wintle, Fiona Fidler, Libby Rumpff, and Charles Twardy. Expert Status and Performance. *PLoS ONE*, 6(7):e22998, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022998. URL `https://dx.plos.org/10.1371/journal.pone.0022998`.

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5): 365–376, 2013. ISSN 1471-0048. doi: 10.1038/nrn3475. URL `https://www.nature.com/articles/nrn3475`.

Susan Byrne, Cathal Walsh, Catherine Lynch, Peter Bede, Marwa Elamin, Kevin Kenna, Russell McLaughlin, and Orla Hardiman. Rate of familial amyotrophic lateral sclerosis: A systematic review and meta-analysis, 2011. ISSN 00223050. URL `http://jnnp.bmj.com/`.

Gregory Campbell. FDA Regulatory Acceptance of Bayesian Statistics. In *Bayesian Methods in Pharmaceutical Research*. 2020. doi: 10.1201/9781315180212-2.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. ISSN 15487660. doi: 10.18637/jss.v076.i01.

Kevin J. Carroll. Decision Making from Phase II to Phase III and the Probability of Success: Reassured by "Assurance"? *Journal of Biopharmaceutical Statistics*, 23(5): 1188–1200, 2013. ISSN 1054-3406. doi: 10.1080/10543406.2013.813527. URL `https://www.tandfonline.com/doi/full/10.1080/10543406.2013.813527`.

David Cesarini, Örjan Sandewall, and Magnus Johannesson. Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior and Organization*, 61(3):453–470, 2006. ISSN 01672681. doi: 10.1016/j.jebo.2004.10.010.

Kathryn Chaloner, Timothy Church, Thomas A Louis, and John P Matts. Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, pages 341–353, 1993.

Winston Chang, Joe Cheng, J Allaire, Yihui Xie, and Jonathan McPherson. Shiny: web application framework for R. *R package version 0.11*, 1(4):106, 2015.

Tesfaye Getachew Charkos, Yawen Liu, and Shuman Yang. Thiazide diuretics and risk of hip fracture: A Bayesian meta-analysis of cohort studies. *Global Epidemiology*, 2: 100025, 2020. ISSN 25901133. doi: 10.1016/j.gloepi.2020.100025.

Jennifer A. Chatman and Francis J. Flynn. The Influence of Demographic Heterogeneity on the Emergence and Consequences of Cooperative Norms in Work Teams. *Academy of Management Journal*, 44(5):956–974, 2001. ISSN 0001-4273. doi: 10.5465/3069440. URL `https://journals.aom.org/doi/abs/10.5465/3069440`.

Ding-Geng Chen and Mark W Fraser. Assurance in Intervention Research: A Bayesian Perspective on Statistical Power. *Journal of the Society for Social Work and Research*, 9(1), 2018. doi: 10.1086/696239.

Samantha Low Choy, Rebecca O'Leary, and Kerrie Mengersen. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, 90(1):265–277, 2009. ISSN 0012-9658. doi: 10.1890/07-1886.1. URL `http://doi.wiley.com/10.1890/07-1886.1`.

Timothy Clark, Ursula Berger, and Ulrich Mansmann. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: Review. *BMJ (Online)*, 346(7901), 2013. ISSN 17561833. doi: 10.1136/bmj.f1135. URL `http://www.bmj.com/content/346/bmj.f1135?tab=related{#}webextra`.

Robert T. Clemen. Comment on Cooke's classical method. *Reliability Engineering & System Safety*, 93(5):760–765, 2008. ISSN 0951-8320. doi: 10.1016/J.RESS.2008.02.003. URL `https://www.sciencedirect.com/science/article/pii/S0951832008000318`.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences.* 2013. doi: 10.4324/9780203771587.

Robert J. Connor. Sample Size for Testing Differences in Proportions for the Paired-Sample Design. *Biometrics*, 43(1):207, 1987. ISSN 0006341X. doi: 10.2307/2531961.

Jonathan A. Cook, Jennifer Hislop, Temitope E. Adewuyi, Kirsten Harrild, Douglas G. Altman, Craig R. Ramsay, Cynthia Fraser, Brian Buckley, Peter Fayers, Ian Harvey, Andrew H. Briggs, John D. Norrie, Dean Fergusson, Ian Ford, and Luke D. Vale. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review, 2014. ISSN 20464924.

Jonathan A. Cook, Steven A. Julious, William Sones, Lisa V. Hampson, Catherine Hewitt, Jesse A. Berlin, Deborah Ashby, Richard Emsley, Dean A. Fergusson, Stephen J. Walters, Edward C.F. Wilson, Graeme MacLennan, Nigel Stallard, Joanne C. Rothwell, Martin Bland, Louise Brown, Craig R. Ramsay, Andrew Cook, David Armstrong, Doug Altman, and Luke D. Vale. DELTA 2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ (Online)*, 363:3750, 2018. ISSN 17561833. doi: 10.1136/bmj.k3750. URL `http://dx.doi.org/10.1136/bmj.k3750`.

Roger Cooke. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press on Demand, 1991.

Roger Cooke. Response to discussants. *Reliability Engineering and System Safety*, 93:775–777, 2008. doi: 10.1016/j.ress.2008.02.006. URL `www.elsevier.com/locate/ress`.

Roger Cooke, Max Mendel, and Wim Thijs. Calibration and information in expert resolution; a classical approach. *Automatica*, 24(1):87–93, 1988.

Andrew A. Cooper and Laren R. Conklin. Dropout from individual psychotherapy for major depression: A meta-analysis of randomized clinical trials, 2015. ISSN 18737811.

João Costa, Michael Swash, and Mamede de Carvalho. Awaji criteria for the diagnosis of amyotrophic lateral sclerosis: a systematic review. *Archives of Neurology*, 69(11):1410–1416, 2012.

Mary Kathryn Cowles and Bradley P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434), 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476956.

Adam Crisp, Sam Miller, Douglas Thompson, and Nicky Best. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharmaceutical Statistics*, 17(4):317–328, jul 2018. ISSN 1539-1612. doi: 10.1002/PST.1856. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pst.1856https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1856https://onlinelibrary.wiley.com/doi/10.1002/pst.1856.

Ioana Alina Cristea and John P. A. Ioannidis. P values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PLOS ONE*, 13 (5):e0197440, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0197440. URL https://dx.plos.org/10.1371/journal.pone.0197440.

N. Dalkey, B. Brown, and S. Cochran. Use of self-ratings to improve group estimates: Experimental evaluation of delphi procedures. *Technological Forecasting*, 1(3): 283–291, 1970. ISSN 0099-3964. doi: 10.1016/0099-3964(70)90029-3. URL https://www.sciencedirect.com/science/article/pii/0099396470900293.

Norman Dalkey and Olaf Helmer. An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science*, 9(3):458–467, 1963. ISSN 0025-1909. doi: 10.1287/mnsc.9.3.458. URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.9.3.458.

Nigel Dallow, Nicky Best, and Timothy H. Montague. Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*, 17(4):301–316, jul 2018. ISSN 1539-1612. doi: 10.1002/PST.1854. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pst.1854https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1854https://onlinelibrary.wiley.com/doi/10.1002/pst.1854.

Mamede de Carvalho, Reinhard Dengler, Andrew Eisen, John D. England, Ryuji Kaji, Jun Kimura, Kerry Mills, Hiroshi Mitsumoto, Hiroyuki Nodera, Jeremy Shefner, and Michael Swash. Electrodiagnostic criteria for diagnosis of ALS. *Clinical Neurophysiology*, 119(3):497–503, 2008. ISSN 1388-2457. doi: 10.1016/J.CLINPH.2007.09.143.

Julie De Meulemeester, Mark Fedyk, Lucas Jurkovic, Michael Reaume, Dar Dowlatshahi, Grant Stotts, and Michel Shamy. Many randomized clinical trials may not be justified: a cross-sectional analysis of the ethics and science of randomized clinical trials. *Journal of Clinical Epidemiology*, 97:20–25, 2018. ISSN 18785921. doi: 10.1016/j.jclinepi.2017.12.027.

Tim R.W. de Wilde, Femke S. Ten Velden, and Carsten K.W. De Dreu. The anchoring-bias in groups. *Journal of Experimental Social Psychology*, 76:116–126, 2018. ISSN 10960465. doi: 10.1016/j.jesp.2018.02.001.

Morris H. Degroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345), 1974. ISSN 1537274X. doi: 10.1080/01621459.1974.10480137.

Lucas J. Dixon and Jake Linardon. A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials, 2020. ISSN 16512316. URL `https://www.tandfonline.com/doi/full/10.1080/16506073.2019.1620324`.

EFSA. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *EFSA Journal*, 12(6):3734, 2014. doi: 10.2903/j.efsa.2014.3734.

Justin W Eggstaff, Thomas A Mazzuchi, and Shahram Sarkani. The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering and System Safety*, 121:72–82, 2013. doi: 10.1016/j.ress.2013.07.015. URL `http://dx.doi.org/10.1016/j.ress.2013.07.015`.

James K. Esser. Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes*, 73(2-3), 1998. ISSN 07495978. doi: 10.1006/obhd.1998.2758.

FDA. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Technical report, 2010.

Klaus Fiedler. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, 50(2):123–129, 1988. ISSN 03400727. doi: 10.1007/BF00309212. URL `https://link.springer.com/article/10.1007/BF00309212`.

Anna Filyushkina, Niels Strange, Magnus Löf, Eugene E. Ezebilo, and Mattias Boman. Applying the Delphi method to assess impacts of forest management on biodiversity and habitat preservation. *Forest Ecology and Management*, 409:179–189, 2018. ISSN 03781127. doi: 10.1016/j.foreco.2017.10.022.

Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein. Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2):330–344, 1978. ISSN 00961523. doi: 10.1037/0096-1523.4.2.330.

F. Flandoli, E. Giorgi, W.P. Aspinall, and A. Neri. Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety*, 96(10): 1292–1310, 2011. ISSN 0951-8320. doi: 10.1016/J.RESS.2011.05.012. URL `https://www.sciencedirect.com/science/article/pii/S0951832011001104`.

Craig R. Fox and Robert T. Clemen. Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, 51(9), 2005. ISSN 00251909. doi: 10.1287/mnsc.1050.0409.

B Freedman. Equipoise and the ethics of clinical research. *New England Journal of Medicine*, 1987.

Simon French. Consensus of opinion. *European Journal of Operational Research*, 7(4): 332–340, 1981. ISSN 03772217. doi: 10.1016/0377-2217(81)90090-4.

Simon French. Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1):181–206, 2011. ISSN 1578-7303. doi: 10.1007/s13398-011-0018-6. URL `http://link.springer.com/10.1007/s13398-011-0018-6`.

Lawrence M. Friedman, Curt D. Furberg, and David L. Demets. *Fundamentals of clinical trials*. Springer New York, 2010. ISBN 9781441915856. doi: 10.1007/978-1-4419-1586-3.

T Ganguly, K J Wilson, J Quigley, R M Cooke, Alessandra Babuscia, and Ming Cheung. Correspondence: Reaction to 'An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results. Technical Report 4, 2014.

Andrew Gelman. Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size. *Annals of Surgery*, 269(1):e9–e10, 2019. ISSN 0003-4932. doi: 10.1097/SLA.0000000000002908. URL `http://journals.lww.com/00000658-201901000-00046`.

Andrew Gelman and John Carlin. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, pages 1–11, 2014. doi: 10.1177/1745691614551642.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. Bayesian Data Analysis. *Bayesian Data Analysis*, 2013. doi: 10.1201/B16018. URL `https://www-taylorfrancis-com.libproxy.ncl.ac.uk/books/mono/10.1201/b16018/bayesian-data-analysis-andrew-gelman-john-carlin-hal-stern-david-dunson-aki-vehtari-`

Gerd Gigerenzer. How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases". *European Review of Social Psychology*, 2(1):83–115, 1991. ISSN 1046-3283. doi: 10.1080/14792779143000033. URL `https://www.tandfonline.com/action/journalInformation?journalCode=pers20`.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. ISSN 1467-9868.

doi: 10.1111/J.1467-9868.2007.00587.X. URL https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2007.00587.xhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2007.00587.x.

Michael Goldstein. Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3), 2006. ISSN 19360975. doi: 10.1214/06-BA116.

Ewan C. Goligher, George Tomlinson, David Hajage, Duminda N. Wijeysundera, Eddy Fan, Peter Jüni, Daniel Brodie, Arthur S. Slutsky, and Alain Combes. Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome and Posterior Probability of Mortality Benefit in a Post Hoc Bayesian Analysis of a Randomized Clinical Trial, 2018. ISSN 15383598. URL https://jamanetwork.com/.

SM Gore. Biostatistics and the medical research council. *Medical Research Council*, 1987.

John Paul Gosling. SHELF: the Sheffield elicitation framework. In *Elicitation: The Science and Art of Structuring Judgement*, pages 61–93. Springer, 2018.

Alex S. Halme, Xavier Fritel, Andrea Benedetti, Ken Eng, and Cara Tannenbaum. Implications of the minimal clinically important difference for health-related quality-of-life outcomes: A comparison of sample size requirements for an incontinence treatment trial. *Value in Health*, 18(2):292–298, 2015. ISSN 15244733. doi: 10.1016/j.jval.2014.11.004.

James K Hammitt and Yifan Zhang. Combining Experts' Judgments: Comparison of Algorithmic Methods Using Synthetic Data. *Risk Analysis*, 33(1), 2013. doi: 10.1111/j.1539-6924.2012.01833.x. URL https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2012.01833.x.

A. M. Hanea, M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. Investigate Discuss Estimate Aggregate for structured expert judgement. *International Journal of Forecasting*, 33(1):267–279, 2017. ISSN 01692070. doi: 10.1016/j.ijforecast.2016.02.008.

Anca M Hanea, Mark Burgman, and Victoria Hemming. IDEA for uncertainty quantification. In *Elicitation: The Science and Art of Structuring Judgement*, pages 95–117. Springer, 2018.

David A. Harrison, Kenneth H. Price, Joanne H. Gavin, and Anna T. Florey. Time, teams, and task performance: Changing effects of surface- and deep-level diversity on group functioning. *Academy of Management Journal*, 45(5):1029–1045, 2002. ISSN 00014273. doi: 10.5465/3069328.

David Hartley and Simon French. Elicitation and Calibration: A Bayesian Perspective. In *Elicitation: The Science and Art of Structuring Judgement*, pages 119–140. Springer, Cham, 2018. doi: 10.1007/978-3-319-65052-4_6. URL `http://link.springer.com/10.1007/978-3-319-65052-4{_}6`.

David Hartley and Simon French. A Bayesian method for calibration and aggregation of expert judgement. *International Journal of Approximate Reasoning*, 130:192–225, 2021. ISSN 0888613X. doi: 10.1016/j.ijar.2020.12.007.

Martie G. Haselton, Daniel Nettle, and Damian R. Murray. The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology*, pages 1–20. John Wiley & Sons, Inc., 2015. doi: 10.1002/9781119125563.evpsych241. URL `http://doi.wiley.com/10.1002/9781119125563.evpsych241`.

Michael Hay, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1), 2014. ISSN 10870156. doi: 10.1038/nbt.2786.

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3): e1002106, 2015. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1002106. URL `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106`.

Victoria Hemming, Terry V. Walshe, Anca M. Hanea, Fiona Fidler, and Mark A. Burgman. Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. *PLoS ONE*, 13(6):e0198468, 2018. ISSN 19326203. doi: 10.1371/journal.pone.0198468. URL `https://doi.org/10.1371/journal.pone.0198468`.

John M. Hoenig and Dennis M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55(1):19–24, 2001. ISSN 00031305. doi: 10.1198/000313001300339897.

Eva Hudlicka. Requirements elicitation with indirect knowledge elicitation techniques: comparison of three methods. In *Proceedings of the IEEE International Conference on Requirements Engineering*, 1996. doi: 10.1109/icre.1996.491424.

Mark H.B. Huisman, Sonja W. De Jong, Perry T.C. Van Doormaal, Stephanie S. Weinreich, H. Jurgen Schelhaas, Anneke J. Van Der Kooi, Marianne De Visser, Jan H. Veldink, and Leonard H. Van Den Berg. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *Journal of Neurology, Neurosurgery and Psychiatry*, 82(10):1165–1170, 2011. ISSN 1468330X. doi:

10.1136/jnnp.2011.244939. URL https://jnnp.bmj.com/content/82/10/1165https://jnnp.bmj.com/content/82/10/1165.abstract.

Susan Humphrey-Murto and Maarten De Wit. The Delphi method - more research please. *Journal of Clinical Epidemiology*, 106:136–139, 2019. doi: 10.1016/j.jclinepi.2018.06.007. URL https://doi.org/10.1016/j.jclinepi.2018.10.011.

ICH. ICH Topic E9 Statistical Principles for Clinical Trials. Technical report, 1998.

John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124, 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124. URL https://dx.plos.org/10.1371/journal.pmed.0020124.

Roman Jaeschke, Joel Singer, and Gordon H. Guyatt. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4):407–415, 1989. ISSN 01972456. doi: 10.1016/0197-2456(89)90005-6. URL http://www.contemporaryclinicaltrials.com/article/0197245689900056/fulltexthttp://www.contemporaryclinicaltrials.com/article/0197245689900056/abstracthttps://www.contemporaryclinicaltrials.com/article/0197-2456(89)90005-6/abstract.

Robert L. Winkler James E. Matheson. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096, 1976. URL https://www.jstor.org/stable/pdf/2629907.pdf.

I.L. Janis and L. Mann. *Decision making: A psychological analysis of conflict, choice, and commitment.* Free Press., 1977.

Irving Janis. Victims of groupthink, 1972. *Groupthink" is seen as a negative for such groups ...*, 36(1), 1972.

Irving Janis. Groupthink of Irving Janis. *A First Look at Communication Theory*, 18(1), 1991.

P. Joensen. Incidence of amyotrophic lateral sclerosis in the Faroe Islands. *Acta Neurologica Scandinavica*, 126(1):62–66, 2012. ISSN 00016314. doi: 10.1111/j.1600-0404.2011.01611.x. URL http://doi.wiley.com/10.1111/j.1600-0404.2011.01611.x.

Dominic D.P. Johnson and James H. Fowler. The evolution of overconfidence. *Nature*, 477(7364), 2011. ISSN 00280836. doi: 10.1038/nature10384.

Sindhu R. Johnson, George A. Tomlinson, Gillian A. Hawker, John T. Granton, Haddas A. Grosbein, and Brian M. Feldman. A valid and reliable belief elicitation method for

Bayesian priors. *Journal of Clinical Epidemiology*, 63(4):370–383, 2010. ISSN 08954356. doi: 10.1016/j.jclinepi.2009.08.005.

Mohamed N. Jouini and Robert T. Clemen. Copula Models for Aggregating Expert Opinions. *Operations Research*, 44(3):444–457, jun 1996. ISSN 0030-364X. doi: 10.1287/opre.44.3.444. URL `http://pubsonline.informs.org/doi/abs/10.1287/opre.44.3.444`.

D. Katz, J. Baptista, S. P. Azen, and M. C. Pike. Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. *Biometrics*, 34(3):469, 1978. ISSN 0006341X. doi: 10.2307/2530610.

Kenneth G Kowalski. Integration of Pharmacometric and Statistical Analyses Using Clinical Trial Simulations to Enhance Quantitative Decision Making in Clinical Drug Development. *Statistics in Biopharmaceutical Research*, 11(1):85–103, 2019. doi: 10.1080/19466315.2018.1560361. URL `https://doi.org/10.1080/19466315.2018.1560361`.

Kevin Kunzmann, Kim May Lee, Michael J. Grayling, David S. Robertson, Kaspar Rufibach, and James M.S. Wason. A review of Bayesian perspectives on sample size derivation for confirmatory trials, 2021. ISSN 23318422. URL `https://www.tandfonline.com/action/journalInformation?journalCode=utas20`.

Mary Kynn. The 'Heuristics and Biases' Bias in Expert Elicitation. *Source: Journal of the Royal Statistical Society. Series A (Statistics in Society*, 171(1):239–264, 2008. URL `https://www.jstor.org/stable/30130739?seq=1{&}cid=pdf-`.

Thomas Laage, John W. Loewy, Sandeep Menon, Eva R. Miller, Erik Pulkstenis, Natalia Kan-Dobrosky, and Christopher Coffey. Ethical considerations in adaptive design clinical trials, 2017. ISSN 21684804.

Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(NOV):863, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00863. URL `http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00863/abstract`.

Jon Landeta. Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73(5):467–482, 2006. ISSN 00401625. doi: 10.1016/j.techfore.2005.09.002.

Nathan C. Leggett, Nicole A. Thomas, Tobias Loetscher, and Michael E.R. Nicholls. The life of p: ”Just significant” results are on the rise. *Quarterly Journal of Experimental Psychology*, 66(12):2303–2309, 2013. ISSN 17470218. doi: 10.1080/17470218.2013.863371. URL `http://journals.sagepub.com/doi/10.1080/17470218.2013.863371`.

Russell V Lenth. Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, 55(3):187–193, 2001. ISSN 1537-2731. doi: 10.1198/000313001317098149. URL `https://www.tandfonline.com/action/journalInformation?journalCode=utas20`.

John A. Lewis. Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline, 1999. ISSN 02776715.

Sarah Lichtenstein and Baruch Fischhoff. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20(2):159–183, 1977. ISSN 00305073. doi: 10.1016/0030-5073(77)90001-0.

Shi-Woei Lin and Chih-Hsing Cheng. The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management*, 4(2):149–161, 2009. doi: 10.1108/17465660910973961. URL `http://www.emeraldinsight.com/doi/10.1108/17465660910973961`.

Roderick J. Little, Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molenberghs, Susan A. Murphy, James D. Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung J. Shih, Jay P. Siegel, and Hal Stern. The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012. ISSN 0028-4793. doi: 10.1056/nejmsr1203730. URL `https://www.nejm.org/doi/full/10.1056/NEJMsr1203730`.

David D. Loschelder, Johannes Stuppi, and Roman Trötschel. "€14,875?!": Precision Boosts the Anchoring Potency of First Offers. *Social Psychological and Personality Science*, 5(4):491–499, 2014. ISSN 19485514. doi: 10.1177/1948550613499942.

David D. Loschelder, Malte Friese, Michael Schaerer, and Adam D. Galinsky. The Too-Much-Precision Effect: When and Why Precise Anchors Backfire With Experts. *Psychological Science*, 27(12):1573–1587, 2016. ISSN 14679280. doi: 10.1177/0956797616666074.

David J. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 2000. ISSN 09603174. doi: 10.1023/A:1008929526011.

David Machin, Michael J. Campbell, Say Beng Tan, and Sze Huey Tan. *Sample Size Tables for Clinical Studies: Third Edition.* 2009. doi: 10.1002/9781444300710.

Albert E. Mannes, Jack B. Soll, and Richard P. Larrick. The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2), 2014. ISSN 00223514. doi: 10.1037/a0036677.

D. Marti, T.A. Mazzuchi, and R. M. Cooke. Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis. In *Expert Judgement in Risk and Decision Making*, pages 53–82. 2021.

Scott E. Maxwell, Michael Y. Lau, and George S. Howard. Is psychology suffering from a replication crisis?: What does 'failure to replicate' really mean? *American Psychologist*, 70(6):487–498, 2015. ISSN 0003066X. doi: 10.1037/a0039400. URL `/record/2015-39598-001`.

Matthew S. Mayo and Byron J. Gajewski. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Controlled Clinical Trials*, 25(2): 157–167, 2004. ISSN 01972456. doi: 10.1016/j.cct.2003.11.006.

Anna E. McGlothlin and Roger J. Lewis. Minimal clinically important difference: Defining what really matters to patients, 2014. ISSN 15383598. URL `https://jamanetwork.com/`.

Jurian V. Meijering and Hilde Tobi. The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment. *Technological Forecasting and Social Change*, 103:166–173, 2016. ISSN 00401625. doi: 10.1016/j.techfore.2015.11.008.

Curtis L. Meinert. *Clinical Trials: Design, Conduct and Analysis.* Oxford University Press, 2009. ISBN 9780199864478. doi: 10.1093/acprof:oso/9780195035681.001.0001.

Philip Meissner, Malte Schubert, and Torsten Wulf. Determinants of group-level overconfidence in teams: A quasi-experimental investigation of diversity and tenure. *Long Range Planning*, 51(6):927–936, 2018. ISSN 18731872. doi: 10.1016/j.lrp.2017.11.002.

MA Meyer and JM Booker. *Eliciting and analyzing expert judgment: a practical guide.* Society for Industrial and Applied Mathematics, 2001. URL `https://epubs.siam.org/doi/pdf/10.1137/1.9780898718485.bm`.

Petrus Mikkola, Osvaldo A Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, and Arto Klami. Prior knowledge elicitation: The past, present, and future. *arXiv preprint*, 2021. URL `https://arxiv.org/abs/2112.01380`.

Frank Miller, Sarah Zohar, Nigel Stallard, Jason Madan, Martin Posch, Siew Wan Hee, Michael Pearce, Mårten Vågerö, and Simon Day. Approaches to sample size calculation for clinical trials in rare diseases. *Pharmaceutical Statistics*, 17(3):214–230, 2018. ISSN 15391604. doi: 10.1002/pst.1848. URL `http://doi.wiley.com/10.1002/pst.1848`.

Cyr E M'lan, Lawrence Joseph, David B Wolfson, and Others. Bayesian sample size determination for binomial proportions. *Bayesian Analysis*, 3(2):269–296, 2008.

Gilberto Montibeller and Detlof von Winterfeldt. Individual and group biases in value and uncertainty judgments. In *International Series in Operations Research and Management Science*, volume 261, pages 377–392. Springer New York LLC, 2018. doi: 10.1007/978-3-319-65052-4_15. URL `https://link.springer.com/chapter/10.1007/978-3-319-65052-4{_}15`.

M. Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy, 2014. ISSN 10916490. URL `www.pnas.org/cgi/doi/10.1073/pnas.1319946111`.

Asher Mullard. Parsing clinical success rates. *Nature reviews. Drug discovery*, 15(7), 2016. ISSN 14741784. doi: 10.1038/nrd.2016.136.

K. P. Murphy. Conjugate Bayesian Analysis of the Gaussian Distribution. *Def*, 1(7), 2007. ISSN <null>.

Radford M. Neal. MCMC using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. 2011. doi: 10.1201/b10905-6.

Gregory F. Nemet, Laura Diaz Anadon, and Elena Verdolini. Quantifying the Effects of Expert Selection and Elicitation Design on Experts' Confidence in Their Judgments About Future Energy Technologies. *Risk Analysis*, 37(2), 2017. ISSN 15396924. doi: 10.1111/risa.12604.

NICE. *Methods for the development of NICE public health guidance*. Number September. 2012.

NICE. How NICE measures value for money in relation to public health interventions. Technical report, 2013.

Frederick Novomestky and Saralees Nadarajah. Package truncdist, 2016. URL `https://cran.r-project.org/web/packages/truncdist/index.html`.

Jeremy E Oakley and Anthony O'Hagan. Uncertainty in prior elicitations: a nonparametric approach. *Biometrika*, 94(2):427–441, 2007.

Jeremy E Oakley and Anthony O'Hagan. SHELF: the Sheffield Elicitation Framework (version 3.0). *School of Mathematics and Statistics, University of Sheffield*, 2016.

Joy Ogden. QALYs and their role in the NICE decision-making process. *Prescriber*, 28 (4):41–43, 2017. ISSN 09596682. doi: 10.1002/psb.1562. URL `http://doi.wiley.com/10.1002/psb.1562`.

Anthony. O'Hagan. *Uncertain judgements : eliciting experts' probabilities*. Wiley, 2006. ISBN 0470033304.

Anthony O'Hagan. Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 2019(S1):69–81, 2019. ISSN 1537-2731. doi: 10.1080/00031305.2018.1518265. URL `https://www.tandfonline.com/action/journalInformation?journalCode=utas20`.

Anthony O'Hagan and Jeremy E. Oakley. Probability is perfect, but we can't elicit it perfectly. In *Reliability Engineering and System Safety*, volume 85, 2004. doi: 10.1016/j.ress.2004.03.014.

Anthony O'Hagan and John W Stevens. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230, 2001.

Anthony O'Hagan, John W Stevens, and Michael J Campbell. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201, 2005.

Chitu Okoli and Suzanne D. Pawlowski. The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1):15–29, 2004. ISSN 0378-7206. doi: 10.1016/J.IM.2003.11.002. URL `https://www.sciencedirect.com/science/article/pii/S0378720603001794`.

Claudia Pedroza, Weilu Han, Van Thi Thanh Truong, Charles Green, and Jon E. Tyson. Performance of informative priors skeptical of large treatment effects in clinical trials: A simulation study. *Statistical Methods in Medical Research*, 27(1):79–96, 2018. ISSN 14770334. doi: 10.1177/0962280215620828. URL `http://journals.sagepub.com/doi/10.1177/0962280215620828`.

Scott Plous. A Comparison of Strategies for Reducing Interval Overconfidence in Group Judgments. *Article in Journal of Applied Psychology*, 1995. doi: 10.1037/0021-9010.80.4.443. URL `https://www.researchgate.net/publication/232537444`.

Martyn Plummer. rjags: Bayesian graphical models using MCMC, 2016.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. {CODA}: Convergence Diagnosis and Output Analysis for {MCMC}. *R News*, 6(1), 2006.

Matthew A Psioda and Joseph G Ibrahim. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20(3):400–415, 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy009. URL `https://academic.oup.com/biostatistics/article/20/3/400/4935054`.

John Quigley, Abigail Colson, Willy Aspinall, and Roger M Cooke. Elicitation in the classical model. In *Elicitation: The Science and Art of Structuring Judgement*, pages 15–36. Springer, 2018.

Shijie Ren and Jeremy E Oakley. Assurance calculations for planning clinical trials with time-to-event outcomes. *Statistics in Medicine*, 33(1):31–45, 2014.

A. Ring, B. Lang, C. Kazaroho, D. Labes, R. Schall, and H. Schütz. Sample size determination in bioequivalence studies using statistical assurance. *British Journal of Clinical Pharmacology*, 85(10):2369–2377, 2019. ISSN 0306-5251. doi: 10.1111/bcp.14055. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/bcp.14055`.

Paul J Roback and Geof H Givens. Supra-Bayesian pooling of priors linked by a deterministic simulation model. *Communications in Statistics-Simulation and Computation*, 30(3):447–476, 2001. doi: 10.1081/SAC-100105073org/10.1081/SAC-100105073. URL `http://www.tandfonline.com/action/journalInformation?journalCode=lssp20www.dekker.com`.

Gary L. Rosner. Bayesian Methods in Regulatory Science. *Statistics in Biopharmaceutical Research*, 12(2):130–136, 2020. ISSN 1946-6315. doi: 10.1080/19466315.2019.1668843. URL `https://www.tandfonline.com/doi/full/10.1080/19466315.2019.1668843`.

Vivekananda Roy. Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020. doi: 10.1146/ANNUREV-STATISTICS-031219-041300. URL `https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-031219-041300`.

David D. Rutstein. The Ethical Design of Human Experiments. *Daedalus*, 98:523–541, 1969.

Benedetta Santarlasci, Andrea Messori, Corinne Zara Yahni, and Denyse Demers. Clinical trial response and dropout rates with olanzapine versus risperidone. *Annals of Pharmacotherapy*, 37(4):556–563, 2003. ISSN 10600280. doi: 10.1345/aph.1C291. URL `http://journals.sagepub.com/doi/10.1345/aph.1C291`.

Jonathon P. Schuldt, Christopher F. Chabris, Anita Williams Woolley, and J. Richard Hackman. Confidence in Dyadic Decision Making: The Role of Individual Differences.

Journal of Behavioral Decision Making, 30(2):168–180, 2017. ISSN 08943257. doi: 10.1002/bdm.1927. URL `http://doi.wiley.com/10.1002/bdm.1927`.

Adil E. Shamoo. The myth of equipoise in phase 1 clinical trials. *Medscape General Medicine*, 10(11):254, 2008. ISSN 15310132. URL `/pmc/articles/PMC2605120/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605120/`.

Brian J. Smith. boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11), 2007. ISSN 15487660. doi: 10.18637/jss.v021.i11.

Jack B. Soll and Joshua Klayman. Overconfidence in Interval Estimates, 2004. ISSN 02787393. URL `/record/2004-11031-001`.

Andrew Speirs-Bridge, Fiona Fidler, Marissa McBride, Louisa Flander, Geoff Cumming, and Mark Burgman. Reducing Overconfidence in the Interval Judgments of Experts. *Risk Analysis*, 30(3):512–523, 2010. ISSN 02724332. doi: 10.1111/j.1539-6924.2009.01337.x. URL `http://doi.wiley.com/10.1111/j.1539-6924.2009.01337.x`.

D J Spiegelhalter and L S Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13, 1986.

David J. Spiegelhalter, Laurence S. Freedman, and Mahesh K. B. Parmar. Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):357, 1994. ISSN 09641998. doi: 10.2307/2983527. URL `https://www.jstor.org/stable/10.2307/2983527?origin=crossref`.

David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons, 2004.

Akira Takazawa and Satoshi Morita. Optimal Decision Criteria for the Study Design and Sample Size of a Biomarker-Driven Phase III Trial. *Therapeutic Innovation and Regulatory Science*, 54(5):1018–1034, 2020. ISSN 21684804. doi: 10.1007/s43441-020-00119-1. URL `https://pubmed.ncbi.nlm.nih.gov/31989540/`.

Say Beng Tan, Y. F.Alexander Chung, Bee Choo Tai, Yin Bun Cheung, and David Machin. Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. *Controlled Clinical Trials*, 24(2): 110–121, 2003. ISSN 01972456. doi: 10.1016/S0197-2456(02)00318-5.

Karl Halvor Teigen and Magne JØrgensen. When 90% confidence intervals are 50% certain: on the credibility of credible intervals. *Applied Cognitive Psychology*, 19(4):455–475,

2005. ISSN 0888-4080. doi: 10.1002/acp.1085. URL http://doi.wiley.com/10.1002/acp.1085.

D. W. Thomas, J. Burns, J. Audette, A. Carroll, C. Dow-Hygelund, and M. Hay. Clinical development success rates 2006–2015. *BIO Industry Analysis*, 2016.

Andrew Timm, Andrew Gleman, and John Carlin. CRAN - Package retrodesign, 2019. URL https://cran.r-project.org/web/packages/retrodesign/index.html.

André M. Travessa, Filipe B. Rodrigues, Tiago A. Mestre, and Joaquim J. Ferreira. Fifteen years of clinical trials in Huntington's disease: A very low clinical drug development success rate. *Journal of Huntington's Disease*, 6(2):157–163, 2017. ISSN 18796400. doi: 10.3233/JHD-170245.

B. J. Traynor, M. B. Codd, B. Corr, C. Forde, E. Frost, and Orla Hardiman. Incidence and prevalence of ALS in Ireland, 1995-1997 a population- based study. *Neurology*, 52(3):504–509, 1999. ISSN 00283878. doi: 10.1212/wnl.52.3.504. URL https://n.neurology.org/content/52/3/504https://n.neurology.org/content/52/3/504.abstract.

Phuong N. Truong, Gerard B.M. Heuvelink, and John Paul Gosling. Web-based tool for expert elicitation of the variogram. *Computers & Geosciences*, 51:390–399, 2013. ISSN 0098-3004. doi: 10.1016/J.CAGEO.2012.08.010. URL https://www.sciencedirect.com/science/article/pii/S0098300412002890.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 1973. ISSN 00100285. doi: 10.1016/0010-0285(73)90033-9.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. ISSN 00368075. doi: 10.1126/science.185.4157.1124.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. ISSN 00368075. doi: 10.1126/science.7455683. URL http://science.sciencemag.org/.

J. H. van der Lee, J. Wesseling, M. W.T. Tanck, and M. Offringa. Efficient ways exist to obtain the optimal sample size in clinical trials in rare diseases, 2008. ISSN 08954356.

Joost van Rosmalen, David Dejardin, Yvette van Norden, Bob Löwenberg, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27(10):3167–3182, 2018. ISSN 14770334. doi: 10.1177/0962280217694506.

Carl Van Walraven, Jeffrey L. Mahon, David Moher, Clara Bohm, and Andreas Laupacis. Surveying physicians to determine the minimal important difference: Implications for sample-size calculation. *Journal of Clinical Epidemiology*, 52(8):717–723, 1999. ISSN 08954356. doi: 10.1016/S0895-4356(99)00050-5.

Shravan Vasishth and Andrew Gelman. The statistical significance filter leads to overconfident expectations of replicability. *Journal of Memory and Language*, 103:151–175, 2017. URL `http://arxiv.org/abs/1702.00556`.

Shravan Vasishth, Daniela Mertzen, Lena A. Jäger, and Andrew Gelman. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175, 2018. ISSN 0749-596X. doi: 10.1016/J.JML.2018.07.004.

M C Vázquez, C Ketzoián, C Legnani, I Rega, N Sánchez, A Perna, M Penela, X Aguirrezábal, M Druet-Cabanac, and M Medici. Incidence and Prevalence of Amyotrophic Lateral Sclerosis in Uruguay: A Population-Based Study. *Neuroepidemiology*, 30:105–111, 2008. doi: 10.1159/000120023. URL `www.karger.com`.

Elena Verdolini, Laura Diaz Anadon, Jiaqi Lu, and Gregory F. Nemet. The effects of expert selection, elicitation design, and R&D assumptions on experts' estimates of the future costs of photovoltaics. *Energy Policy*, 80, 2015. ISSN 03014215. doi: 10.1016/j.enpol.2015.01.006.

Kristian Wahlbeck, Arja Tuunainen, Antti Ahokas, and Stefan Leucht. Dropout rates in randomised antipsychotic drug trials. *Psychopharmacology*, 155(3):230–233, 2001. ISSN 00333158. doi: 10.1007/s002130100711. URL `https://link.springer.com/article/10.1007/s002130100711`.

Rosalind J Walley, Claire L Smith, Jeremy D Gale, and Phil Woodward. Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. *Pharmaceutical statistics*, 14(3):205–215, 2015.

Fei Wang and Alan E Gelfand. A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. Technical Report 2, 2002.

Mike West. Bayesian Aggregation. *Journal of the Royal Statistical Society. Series A (General)*, 147(4):600, 1984. ISSN 00359238. doi: 10.2307/2981847. URL `https://www.jstor.org/stable/10.2307/2981847?origin=crossref`.

Richard F. West and Keith E. Stanovich. The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin* *and Review*, 4(3):387–392, 1997. ISSN 10699384. doi: 10.3758/BF03210798. URL https://link.springer.com/article/10.3758/BF03210798.

Cameron J. Williams, Kevin J. Wilson, and Nina Wilson. A comparison of prior elicitation aggregation using the classical method and SHELF. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, page rssa.12691, 2021. ISSN 0964-1998. doi: 10.1111/rssa.12691. URL https://onlinelibrary.wiley.com/doi/10.1111/rssa.12691.

Kevin J. Wilson. An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, 33(1):325–336, 2017. ISSN 01692070. doi: 10.1016/j.ijforecast.2015.11.014.

Kevin J Wilson and Malcolm Farrow. Combining judgements from correlated experts. In *Elicitation: The Science and Art of Structuring Judgement*, pages 211–240. Springer, 2018.

Kevin J. Wilson, S. Faye Williamson, A. Joy Allen, Cameron J. Williams, Thomas P. Hellyer, and B. Clare Lendrem. Bayesian sample size determination for diagnostic accuracy studies. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2108.08594v1.

R L Winkler. Scoring Rules and the Evaluation of Probabilities. Technical Report 1, 1996. URL https://link.springer.com/content/pdf/10.1007/BF02562681.pdf.

Robert L. Winkler. Combining Probability Distributions from Dependent Information Sources. *Management Science*, 27(4):479–488, 1981. ISSN 0025-1909. doi: 10.1287/mnsc.27.4.479. URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.27.4.479.

Bram Wisse, Tim Bedford, and John Quigley. Expert judgement combination using moment methods. *Reliability Engineering & System Safety*, 93(5):675–686, 2008. ISSN 0951-8320. doi: 10.1016/J.RESS.2007.03.003. URL https://www.sciencedirect.com/science/article/pii/S0951832007000956.

WMA. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects, 2001. ISSN 00429686.

Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxx069. URL https://academic.oup.com/biostatistics/article/20/2/273/4817524.

Thomas Wutzler. CRAN - Package logitnorm, 2018. URL https://cran.r-project.org/web/packages/logitnorm/index.html.

Jingjing Ye, Gregory Reaman, R. Angelo De Claro, and Rajeshwari Sridhara. A Bayesian approach in design and analysis of pediatric cancer clinical trials. *Pharmaceutical Statistics*, 19(6):814–826, 2020. ISSN 1539-1604. doi: 10.1002/pst.2039. URL `https://onlinelibrary.wiley.com/doi/10.1002/pst.2039`.

Appendix A

Appendix

A.1 Elicitation Supplementary Materials

The following pages contain the information provided to the experts alongside both elicitations. This supplementary material included both further background about the novel diagnostic test, questions about the experts' background, and the seed questions, followed by further guidance for using the elicitation application.

Elicitation Evidence

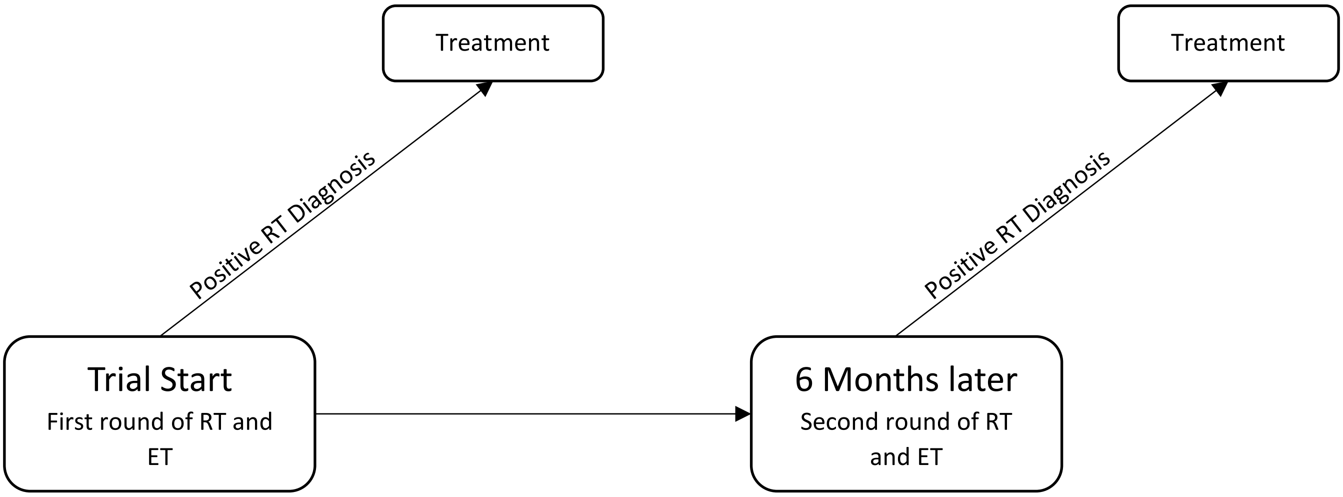
The following document will provide some background information and definitions.

Definitions

MND	Motor Neurone Disease, which involves the progressive degeneration of motor neurones in the cerebral cortex, brainstem and spinal cord.
RT	Reference Test
Positive RT Diagnosis	A diagnosis using the Reference Test leading to a patient being assigned treatment for MND
ET	Experimental Test
Positive ET Diagnosis	A diagnosis using the Experimental Test and Reference Test leading to a patient being assigned treatment for MND
Median	The value where an outcome is equally likely to occur above or below.
Best Estimate	The median.
Lower 25% Quartile	Assuming the outcome will occur below the median, this quartile is the value where an outcome is equally likely to occur above or below.
Upper 25% Quartile	Assuming the outcome will occur above the median, this quartile is the value where an outcome is equally likely to occur above or below.

Trial Summary

The participants in this trial are patients suspected of having MND. They will be tested using both the RT and ET at an initial time point, and after a further 6 months. Patients that receive a positive diagnosis from the RT at the first time point will leave the trial to begin treatment. ET results will not affect a patient's position in the trial.



The anticipated improvement from ET will be an earlier diagnosis than that from using the RT alone.

Additional information was provided to the experts. Results from previous trials using the ET were given to assist the experts. The full version also contained additional definitions for the specific experimental and reference tests.

Elicitation Records

Clinical Trial Name

Name:	
Date:	
Job:	
Declaration of interest:	

Experience and prior knowledge

Please provide some details into your background.

What is your background in researching MND?	
How long have you been involved in MND research?	
What sources of information is your knowledge of MND based on?	
What are your strengths and weaknesses regarding this topic?	
Please rate your knowledge on Motor Neuron Disease from 1 (least) to 5 (most).	
Please rate your knowledge on the reference test from 1 (least) to 5 (most).	
Please rate your knowledge on the experimental test from 1 (least) to 5 (most).	
Please list any sources of quantitative information about experimental test you are aware of. This information will be shared with other participants	

Practice Questions

These questions are designed to provide practice for specifying uncertainty around estimates. Information from these questions will be analysed as part of the PhD project.

For each question, please provide your best estimate, as well as a range of possible values in which you believe there is a 50% probability that the true value would fall.

For example, consider the incidence rate for MND in the UK. Out of 100,000 people, a researcher may consider the absolute minimum number of new cases of MND in the UK to be 500, and the absolute maximum to be 3000. From here, they may believe the true rate is equally likely to lie between 500 and 600, 600 and 700, 700 and 1000 and 1000 and 3000.

Their best estimate (median) would be 700, as they believe there is an equal chance that the incidence rate will be above or below this value.

	Minimum	Lower 25%	Best Estimate	Upper 75%	Maximum
For the years 2006 to 2009, what was the incidence rate per 100,000 people of ALS in the Netherlands?					
For the years 2006 to 2009, what was the prevalence rate per 100,000 people of ALS in the Netherlands?					
For the years 1995 to 1997, what percentage of people with ALS in Ireland were male?					
For the years 2002 to 2003, what percentage of people with ALS in Uruguay were male?					
For the years 1995 to 1997, what percentage of new ALS diagnoses in Ireland were male?					
For the years 2002 to 2003, what percentage of new ALS diagnoses in Uruguay were male?					
For the years 1985 to 2006, what percentage of ALS diagnoses in New					

Zealand were familial?					
For the years 1989 to 1992, what percentage of ALS diagnoses in Hong Kong were familial?					
For the years 1987 to 2009, what was the incidence rate per 100,000 people of ALS in the Faroe Islands?					
For the years 1987 to 2009, what was the prevalence rate per 100,000 people of ALS in the Faroe Islands?					

Minimal Clinically Important Difference

What is the smallest percentage increase in correct positive diagnoses from using experimental test you would need to see to implement it in diagnoses?	

Elicitation

We wish to elicit from you a number of probability distributions. These distributions will represent your uncertainty about what will happen during the trial. Constructing these will involve specifying minimum and maximum values, median (or best estimate) values and upper and lower quartiles.

You will need access to the internet to complete this elicitation stage. The application can be accessed here: https://cwilliams.shinyapps.io/shinyelicitation/

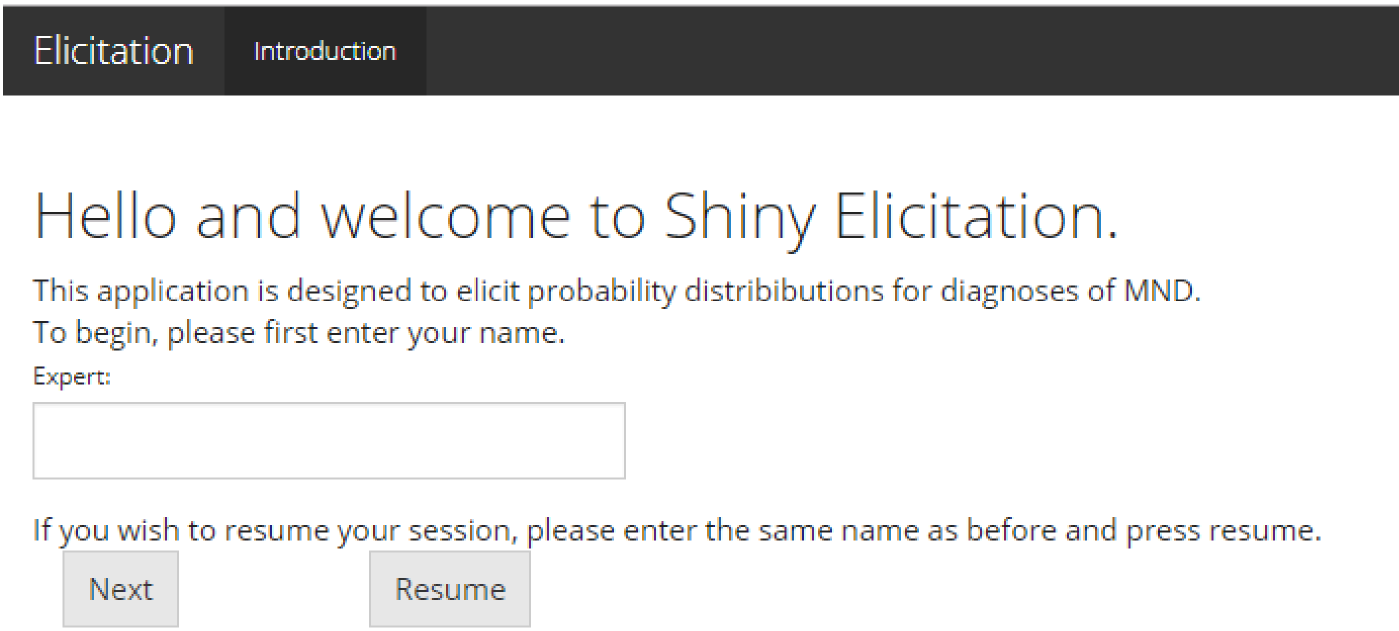
The following example will provide a walkthrough of how to use the application, and what different options mean.

We are interested in determining your knowledge and uncertainty about the trial, so please refrain from using external sources when completing the elicitation.

Example

Introduction

On the Introduction page, you will need to enter your name. If you do not finish the elicitation in a single session, or need to reload the application, you will need to enter the same name as you did previously. This will include the same spaces and capital letters.



Hello and welcome to Shiny Elicitation.

This application is designed to elicit probability distribibutions for diagnoses of MND.
To begin, please first enter your name.

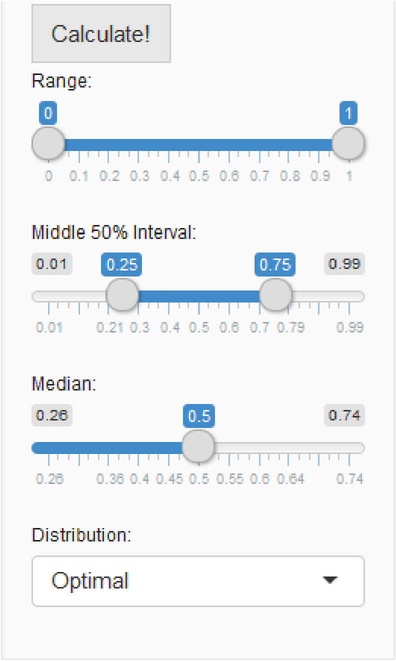
Expert:

If you wish to resume your session, please enter the same name as before and press resume.

After clicking Next or Resume, the application may take a few seconds to load.

Inputting Values

On the Elicitation pages, there is a panel on the left where you can enter your values. To begin, select the lowest and highest possible values you think this parameter could feasibly take.



Range: 0 – 1

Middle 50% Interval: 0.25 – 0.75

Median: 0.5

Distribution: Optimal

Next, input the upper and lower quantiles, represented here as the ends of a Middle 50% Probability Interval. Finally, input your best estimate of the parameter of interest.

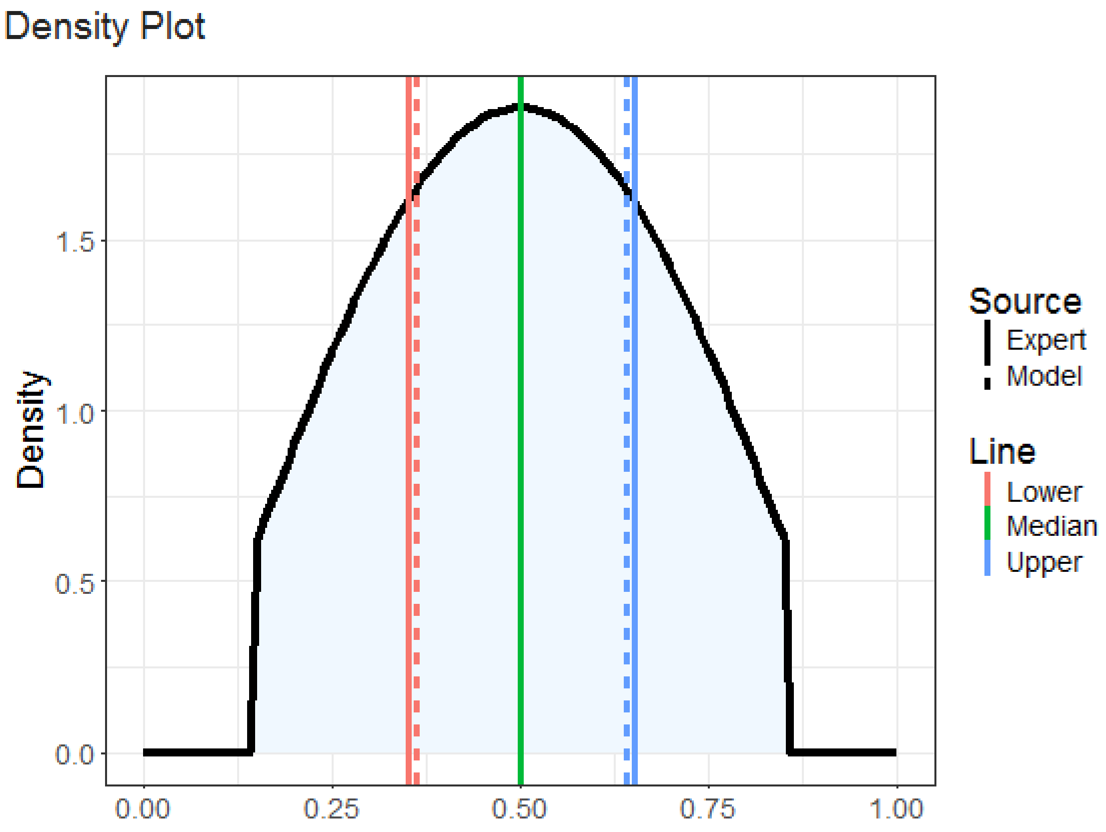
Below is an option to change the distribution fitted to your values. It is recommended to start with the Optimal option, to allow the application to select the distribution that provides the best fit. If you are not happy with the suggested fit, you can adjust the distribution fit to a number of set options.

Density Plot

The application will return a density plot and histogram of the fitted distribution. Overlaid are two sets of lines. The solid lines represent the values you have inputted as your median and lower and upper quartiles. The dashed lines represent the equivalent median and lower and upper quartiles of the fitted distribution. These two sets of lines will likely not be exactly the same, as it is unlikely your values will perfectly match a distribution. If only one line for a colour is showing, it suggests the distribution value and your own were the same.

Summary statistics are provided at the bottom of the page.

Consider the shape of the distribution and the location of the dashed lines. If you think these represent your beliefs about the parameter, clicking Save will move you on to the next parameter. If not, try changing your inputted values to adjust the distribution.

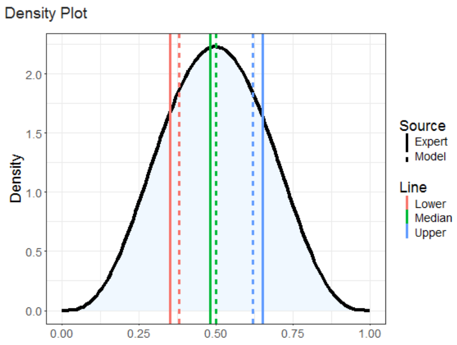


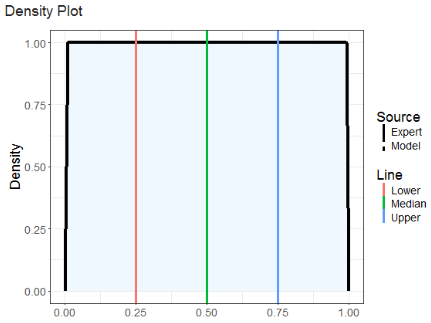
Distributions

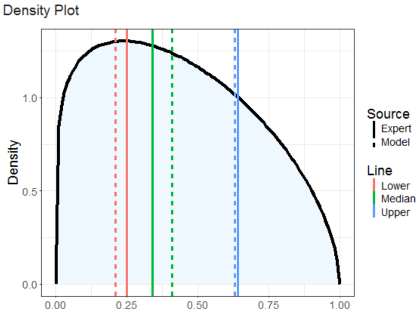
If adjusting the inputted values does not provide an appropriate distribution, you can manually set the type of distribution that will be fit to ensure a particular shape.

Beta

The Beta distribution allows for many different shapes of distribution. It can provide distributions that are flat, skewed or symmetrical.

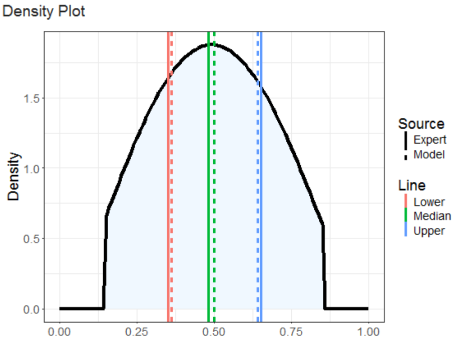






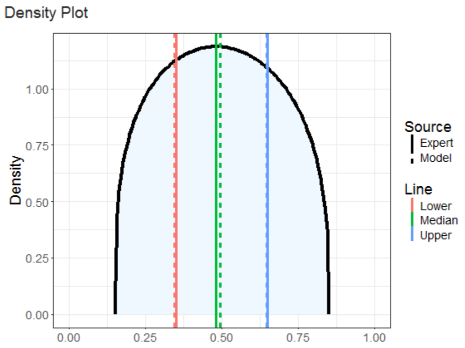
Beta Truncated

This is a variant of the Beta distribution in which the minimum and maximum values are strictly enforced.



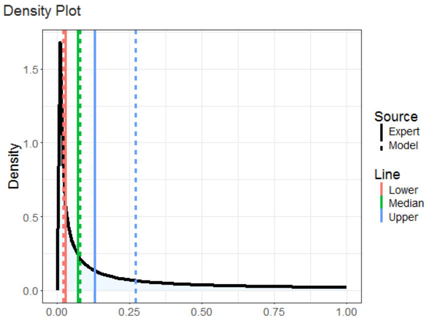
Beta Rescaled

This is a variant of the Beta distribution which has been rescaled to ensure the distribution lies between the minimum and maximum values.



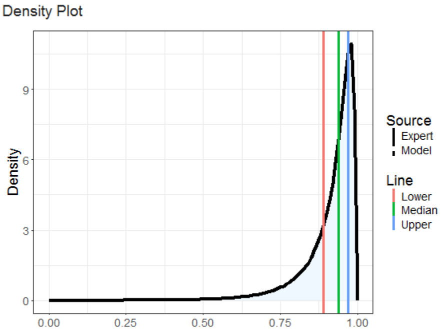
Log Norm

The Log Normal distribution is heavily skewed, allowing for cases where the parameter is most likely to have values close to zero.



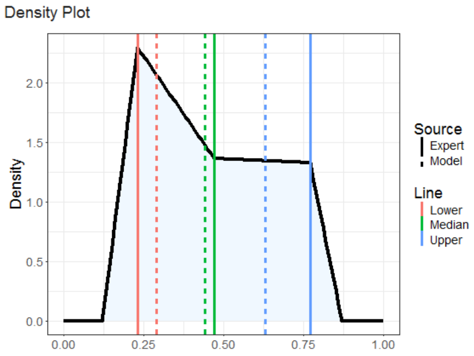
Logit Norm

The Logit Normal distribution allows for skew, and will work well for cases where the parameter is most likely to have values close to one.



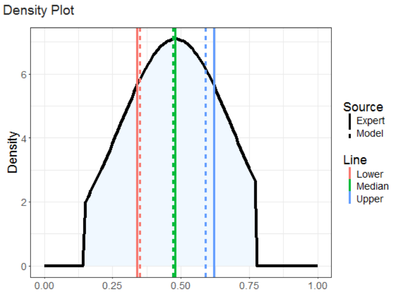
Polygon

The polygon distribution will provide a distribution when the quantiles do not naturally fit any of the other available distributions. It is likely to present sharp points, which may not be reflective of beliefs.



Normal Truncated

The Normal truncated distribution will fit a symmetrical distribution, with cut-offs determined by the minimum and maximum values.



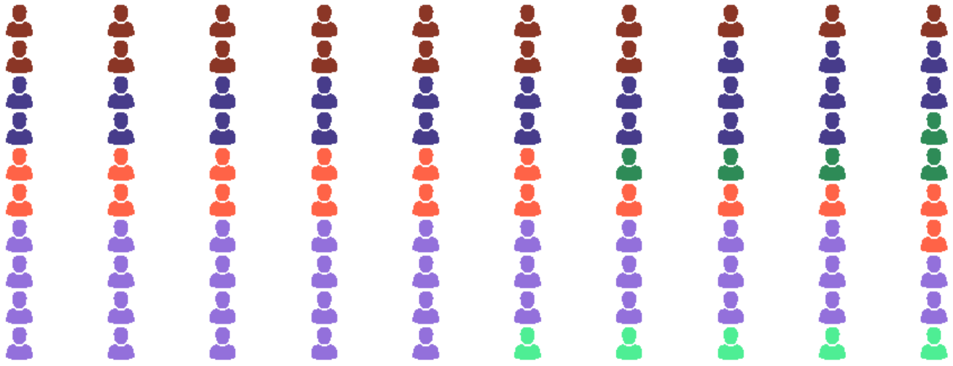
Checks

In order to ensure the values you have provided make sense, there are two checks you can use to verify your results.

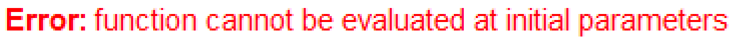
The first calculates an estimate of the probability of a positive experimental test diagnosis for a randomly selected patient from the trial, regardless of their reference test results. A 90% probability interval has been included for this value.

If you are satisfied with these values, continue on. If not, you can return to previous pages to modify your previous results. Possible options to modify the experimental test rate are listed in the application.

The final image displays the suggested results for 100 simulated patients who would be similar to those in the trial. Dark colours represent patients with positive experimental test results, and lighter colours represent negative experimental test results. The red people received a positive reference test result at the first time point, purple a positive reference test at the second time point and green a negative reference test at both time points. The proportions of each are provided in a table.



In some rare cases, inputted values may return the following warning.



Error: function cannot be evaluated at initial parameters

This should only occur when inputted values are very close together, and means the algorithm is unable to fit a distribution for the provided values. If this error occurs, try leaving more space between values.

A.2 Further simulations

The following sections provide results from further simulations not included in Chapter 5.

A.2.1 Assurance for ANOVA

Assurance and power can both be calculated for more complex methods, such as ANOVA.

We consider a one factor ANVOA, with three factor levels. Four combinations of factor levels have been chosen, to represent potential relationships between factor levels. Figure A.1 provides the four cases, labelled A, B, C, and D. For each, the design prior distributions have a standard deviation of 0.1, and the 95% probability interval for each has been plotted. The first, combination A, is where the prior beliefs are that two factor levels are equal, and the third is different. In this case, we would expect to find Factor Level 3 statistically significantly different from the other two levels, which themselves would not be different from each other.

The second, combination B, is where the prior beliefs are that all three factor levels are different, and there is no overlap in the 95% interval ranges in the design priors. In this case, we would expect to find all three factor levels statistically significantly different from each other.

The third case, combination C, is where the prior beliefs are that the three distinct factor levels do have overlap in their interval ranges, but the mean of each design prior is not with the 95% interval of another factor level. In this case, we would expect a result showing each factor level to be statistically significantly different from each other, though such an effect would be harder to detect than combination B due to the increase in overlap between the design priors.

The final case, combination D, is one where the prior beliefs are that the three factor levels have overlap in their 95% interval range, and the mean of one prior lies within the intervals of both other factor levels. In this case, we would expect to find Factor Level 1 statistically significantly different from Factor Level 3, while Factor Level 2 is not statistically significantly different from either other level.

Figure A.1 then provides assurance curves for the four combinations of factor levels.

The more overlap present between design prior distributions of the different factor levels, the lower the assurance is for a given n. This is a reasonable finding, as more distinct groups will be easier to identity in smaller datasets. Those combinations with overlapping design priors mean it is harder to determine which are statistically different from each other, if any are at all.

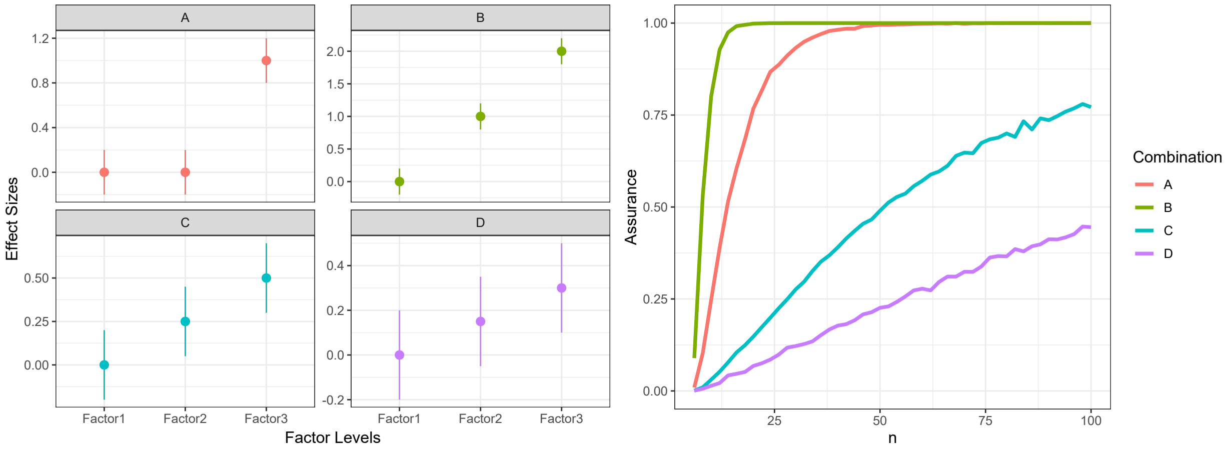


Figure A.1: Four different level combinations for a one factor ANOVA, and the corresponding assurance curves.

A.2.2 Minimal Clinically Important Differences

An MCID can be incorporated into the design and analysis of a trial in a number of different ways.

Firstly, it could be incorporated in the design stage. In this case, it would be used as an effect size in a power calculation, to ensure that the trial could detect an effect size of at least the MCID with a certain level of power. In a Bayesian setting, the MCID could be used as a best estimate for the design prior. However, this use is less natural, as the MCID does not represent the most plausible value for a parameter, but rather one the researchers feel is necessary for the treatment to have a positive impact. Additionally, the design prior places probabilities on either side of the best estimate. In the case of a distribution which is symmetrical such as a normal distribution, this prior would suggest that there is a reasonable probability that the true effect size is being less than the MCID.

The other stage in which a MCID could be incorporated is in the analysis. A Frequentist Z-test, for example, will often have a null hypothesis that the mean effect size is equal to zero, $\theta_0 = 0$. Likewise, a Bayesian analysis can use the posterior probability $P(\theta > 0 \mid X)$ in order to make a judgement. However, if only values equal to or greater than the MCID are of interest, then the posterior probability $P(\theta \geq \text{MCID} \mid X)$ may be more appropriate.

The effect of a change in MCID varies depending on whether it is incorporated with the design or analysis. Figure A.2 demonstrates an example of this, where the assurance curves have opposite slopes as the MCID increases.

This example has a set sample size of 50 for an exact binomial test, with a null hypothesis of $\theta_0 = 0.01$ for the case where the MCID is incorporated into the design, and a null hypothesis of $\theta_0 = MCID$ for the case where the MCID is incorporated into

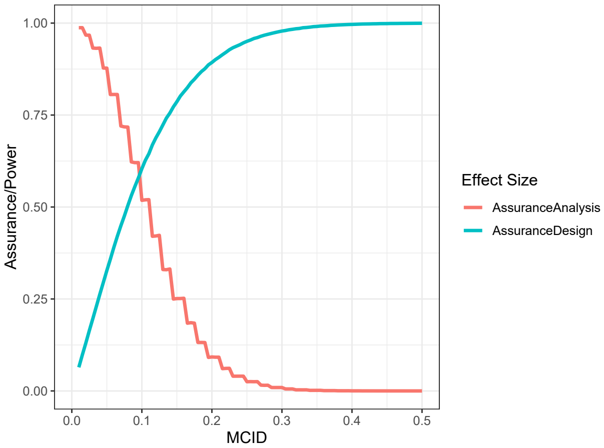


Figure A.2: Example assurance curves for a varying MCID. The blue curve incorporates the MCID into the design prior, and the red line incorporates the MCID into the analysis through the null hypothesis.

the analysis. When the MCID is accounted for in the design, the design prior was given as $\theta \sim Beta(\gamma MCID, \gamma(1 - MCID))$ such that the mean of the distribution is given to be $MCID$, and the γ term chosen to affect the variance of the distribution. In the alternative case, where the MCID is accounted for in the design, the design prior was given as $\theta \sim Beta(10, 40)$.

As the MCID increases, the two assurance curves change in opposite directions.

When the MCID is used in the design, the assurance increases as MCID increases. As the sample size and standard deviation are constants, the increasingly larger effect expected to be present is then expected to be easier to detect. This means it is increasingly likely that a significant result will be found, and as such the assurance increases.

For the assurance curve incorporating the MCID into the analysis, an increase in the MCID results in a decrease in assurance. Increasing the MCID, in this case, means increasing the required test statistic for a significant results to be found. As the design prior is constant, this means the probability of observing the required test statistic decreases as the MCID increases.

For cases such as with binomial data, which has discrete steps, this may not be true locally as there may still be a step pattern which changes the underlying distribution as additional observations of successes are required. Overall, however, the assurance in this case will still decrease as the MCID increases.

Ultimately, the reason for these differences is in the way the MCID has been incorporated into the calculations. It is not unexpected that incorporating the value in a different way will change how the power or assurance is affected. When the MCID is included into the analysis, to be used as a value against which to compare such as in a null hypothesis, it will effect the calculations differently to when it is used as the estimate of the effect, such as that used as an alternative hypothesis.

A.2.3 Replication using previous effect sizes

As mentioned in the previous section, and in Chapter 2, using previous estimates as an input to power calculations can lead to underpowered trials. This is due to the fact that an estimated effect is unlikely to be exactly the same as the population-level effect. Vasishth et al. (2018) demonstrates how for low powered studies leads to overestimates for statistically significant effects, and that relying on statistical significance can result in non-reproducible study results.

We will demonstrate how using assurance instead of power helps to account for this disparity between estimated and population effects.

A.2.4 Normal observations

For a Z-test, with an effect size of 0.5, a power calculation reveals we require a sample size of 42. We then simulate 100,000 datasets of size 42, from a $N(0.5, 1)$ distribution. Figure A.3 shows a histogram of the estimated effect sizes from each of the 100,000 datasets, coloured by statistical significance. As would be expected, these estimates form a normal distribution with a mean centred on 0.5, and a standard deviation of $\frac{1}{\sqrt{42}}$.

We then use these estimated effect sizes as the inputs for a second round of power calculations. Figure A.4 shows a histogram of the required sample sizes based on the previous estimates.

Including all results, significant or not, 50% of the sample sizes are below the required 42. When we only consider the statistically significant results, 55% of the sample sizes are

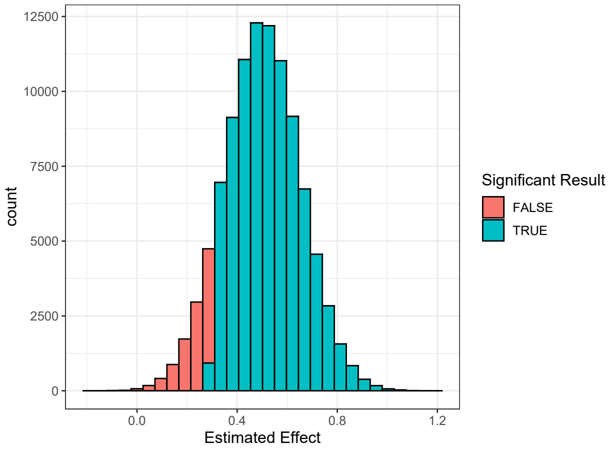


Figure A.3: Estimates of the effect size using a Z-test on datasets simulated from a $N(0.5, 1)$ distribution.

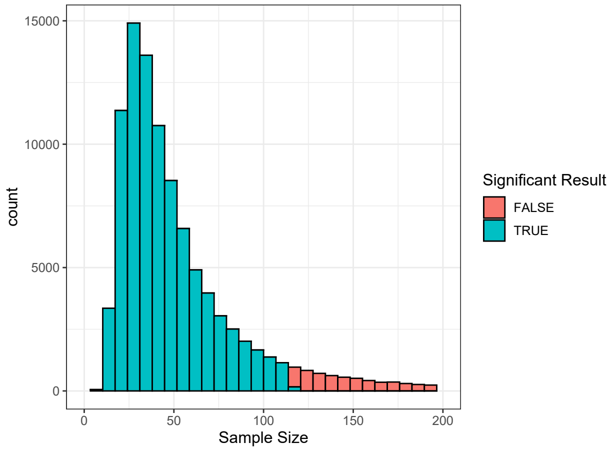


Figure A.4: Estimates of the sample size calculated using the effect sizes from a Z-test on datasets simulated from a $N(0.5, 1)$ distribution.

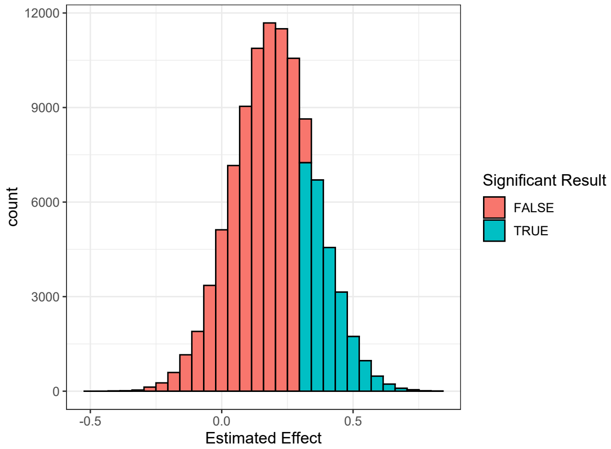


Figure A.5: Estimates of the effect size using a Z-test on datasets simulated from a $N(0.3, 1)$ distribution.

below 42. As a sample size of 42 is required for the trial to be properly powered at the 90% level, any sample size less than this value will lead to an underpowered trial.

While these results show that there is a slight increase in underpowered trials when only considering significant results, there is an assumption that the initial trials were correctly powered.

We consider the case where a true effect size of 0.25 is present, but the trial is powered on an effect size of 0.5. As before, the required sample size from the power calculation would be 42 with this incorrect estimate. If the correct effect size had been used, the required sample size would be 169.

Figure A.5 shows the estimated effect sizes from such a trial. Due to the smaller than required sample size, the majority of the results found are not statistically significant.

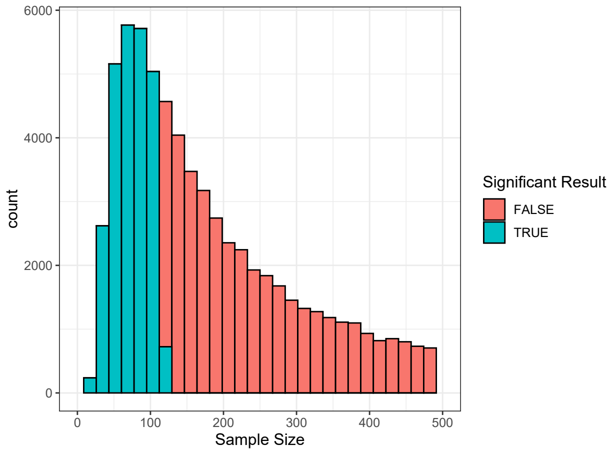


Figure A.6: Estimates of the sample size calculated using the effect sizes from a mis-powered Z-test on datasets simulated from a $N(0.5, 1)$ distribution.

We then use these estimated effect sizes as the inputs for a second round of power calculations. Figure A.6 shows a histogram of the required sample sizes based on the previous estimates.

As the plot shows, the sample sizes required in future trials are typically very large. In this case, only 2% of all sample sizes, or 10% of the sample sizes from the hypothetical significant trials, would lead to future trials being under powered. However, if these estimates were used to power future trials, the vast majority of the trials will instead be over-powered. Of the total estimates, 84% would lead to sample size calculations requiring twice as many samples as actually needed.

This demonstrates the issue with using previous trial's estimates. Even if the previous trial had not misspecified the inputs to its power calculation, there is still a reasonable probability of future sample size calculations under-powering a trial. If the original trial was itself underpowered, then the new trial is also likely to be mis-powered and will often produce a higher sample size than necessary. This can lead to increased costs and difficulties in running the trial.

Although attention has been drawn to the problems associated with only publishing significant results, the over-respresenation of significant results still remains (Head et al., 2015).

As such, the instances where a significant result has been used limits the effect sizes to those that happen to be larger, and in turn limit the potential sample sizes to those that are smaller.

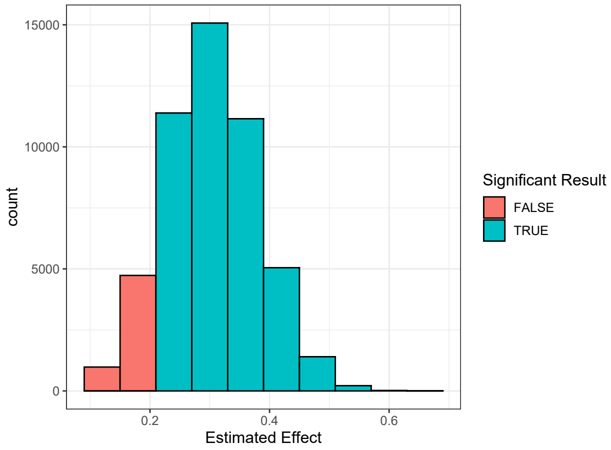


Figure A.7: Estimates of the effect size using an exact binomial test on datasets simulated from a $Binomial(32, 0.3)$ distribution.

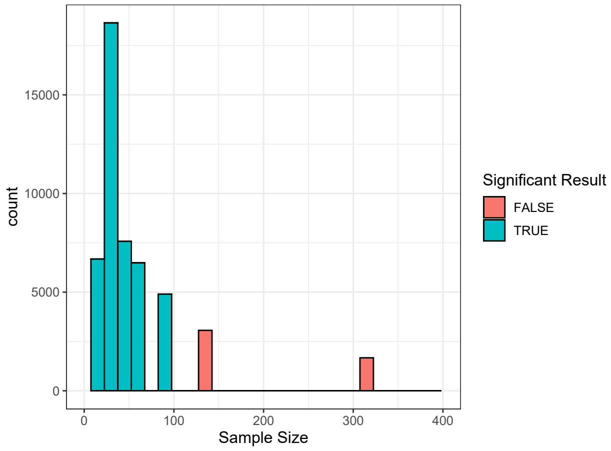


Figure A.8: Estimates of sample sizes using effect sizes from an exact binomial test.

A.2.5 Binomial observations

We can consider a similar case for binomial data as well. As in the previous section, we first simulate trial estimates and then power a future trial based on the initial results. First, we simulate trials of size 32, with a probability of success of 0.3. Figure A.7 shows the resulting estimates of the probability of success. Estimates ranging from 0.22 and 0.69 were found to be significantly different than the null hypothesis $\theta_a = 0.1$.

We then use these estimated effect sizes as the inputs for a second round of power calculations. Figure A.8 shows a histogram of the required sample sizes based on the previous estimates.

In this case, 51% of the sample sizes for replications would have less than the required 32 samples in order to achieve a power of 0.9. If only significant results were used to power future tests, then 57% of the trials would have insufficient sample sizes.

This demonstrates the importance of selecting an accurate effect size, and that such a selection may not be an easy task. Incorporating multiple studies through meta-analysis or eliciting expert opinions may allow for a wider range of knowledge from the field of study to be included in calculations.

A.2.6 Aggregating Priors, Assurances, Powers, and Sample Sizes

In order to gain a broader and more nuanced view of the field, it may make sense to elicit priors from multiple experts. In Chapter 3, we review aggregation methods for prior distributions in this context.

A potential question of interest is at which stage the aggregation should take place. While we have reviewed the aggregation of priors into a single group prior, it is also possible to aggregate assurances or sample sizes instead.

We consider three cases. Firstly, where expert priors are aggregated first, and then an assurance calculation and sample size decision. Secondly, we consider calculating assurance values for each of the expert priors, and then aggregating the assurance values in order to determine a sample size. Finally, we calculate individual assurance curves for each expert prior, and aggregate the selected sample size from each.

We consider a trial in which we will make a binomial observation, where a probability of greater than 0.1 is considered clinically significant. We will run a Bayesian analysis on the data, with an analysis prior of a $Beta(1, 19)$ distribution. Such a prior places approximately 13.5% of its probability above the clinically significant value.

We consider three fictional experts, with varying opinions on the parameter of interest. We give them prior distributions $Beta(1, 1)$, $Beta(2, 1)$ and $Beta(3, 1)$, to represent increasing levels of optimism about the trial. These correspond to prior means for the probability of 0.5, 0.67 and 0.75.

Figure A.9 provides assurance curves based on the three individual experts. As would be expected, the expert with the most optimistic design prior had the highest assurance values. The maximum assurance for each expert is, respectively, 97.20%, 99.95% and 99.99%.

In a trial design utilising the knowledge of these three experts, there are a number of different ways in which their priors could be aggregated. We will use an equal weighting method as outlined in Chapter 3. While we only consider the case of an equal weighting scheme, alternative weights such as those from the Classical Method, could also be used in a similar manner.

Under this method, the aggregated prior is the average of the densities for each prior. Figure A.10 shows the design priors of the three experts, and the aggregated prior in bold. All prior distributions are optimisitic about the results, as the majority of their area is above the clinically significant value.

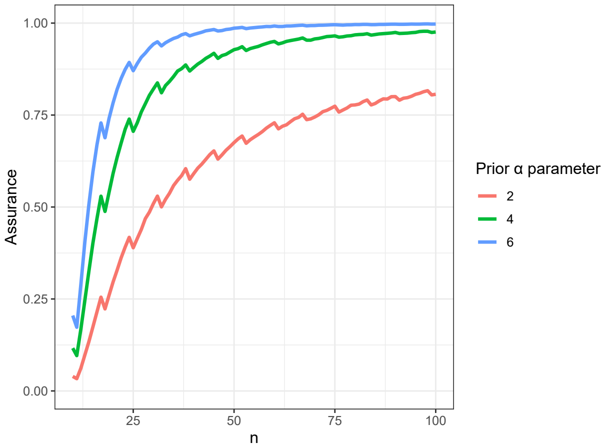


Figure A.9: Assurance curves for three experts, with design priors $Beta(1, 1)$, $Beta(2, 1)$ and $Beta(3, 1)$.

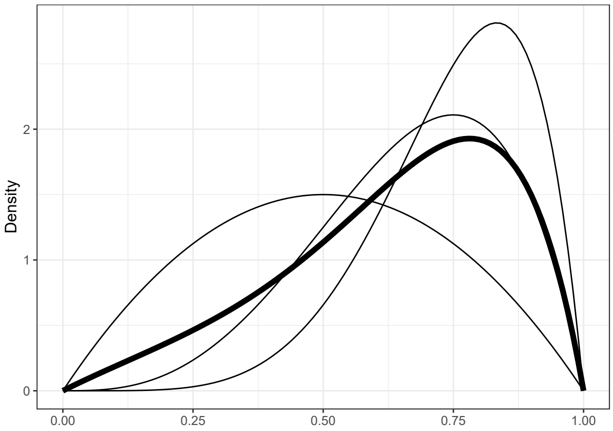


Figure A.10: Design priors of the three experts, and the equal weights aggregation prior, in bold.

We investigate whether aggregating the results before or after calculating the assurance changes the outcome. Figure A.11 provides an assurance curve for the aggregated design prior distribution in black, and compares it to the average of the three experts' individual assurance curves, plotted in red.

As this demonstrates, the aggregation of individual assurance values is not the same as the assurance for the aggregate prior.

Another approach would be to determine sample sizes before aggregating. Table A.1 provides the required sample size in order to reach certain levels of assurance, based on the aggregated prior, individual priors and the average sample size from the three experts.

As this table shows, aggregating the priors first leads to a required sample size of 19 and 37 respectively, for assurances of 50% and 80%. If we averaged the experts' assurance curves, the required sample sizes would be 22 and 50, and if we averaged the experts'

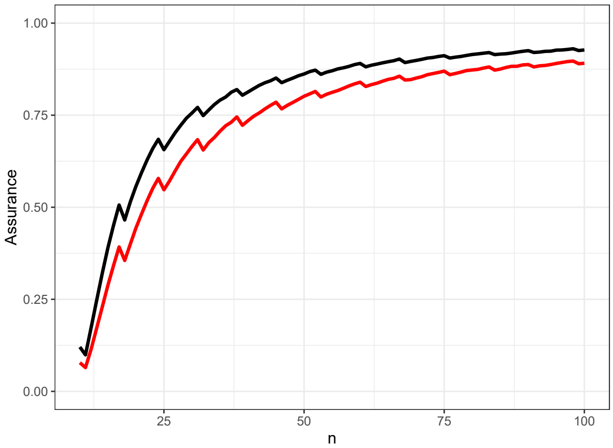


Figure A.11: Assurance curves for an equal weighted design for the three experts (black) and the average assurance value for each of their assurance curves (red).

Table A.1: Approximate sample sizes required to reach levels of assurance

Required Assurance	Aggregated Prior Sample Size	Average Assurance	Expert 1 Sample Size	Expert 2 Sample Size	Expert 3 Sample Size	Average Sample Size for Experts
0.8	37	50	90	29	27	49
0.5	19	22	30	17	14	21

required sample sizes, we would require sample sizes of 21 and 49.

There are clear differences between the results for the three methods. As such, they can not be used interchangeably.

We suggest that aggregation at the prior level is the more sensible option. Such an aggregation leads to a single prior distribution, taking into account each of the group members' views. This allows the interpretation of the assurance to be simpler, and incorporates the multiple sources of prior information at a stage where it is natural to do so. Aggregating at the sample size stage instead, for example, ignores the underlying mechanisms involved in the prior distribution and assurance calculation.

Notably, by aggregating the priors together. smaller sample sizes are required. This may suggest a better incorporation of the prior information.

One further aspect that can be considered is that aggregating at the prior level allows for a behavioural aggregation method to be used. In these cases, there are no mathematical weighting equivalents and so there is not an option to aggregate the assurance curves or final sample sizes. The benefits of a behavioural aggregation are only possible when aggregating the priors.

The behaviour of these aggregations for a normal model is similar.

We consider a trial in which interest lies in the difference between two treatments which we assume follows a a $N(\mu, 1)$ distribution, where the priors for μ are provided by three

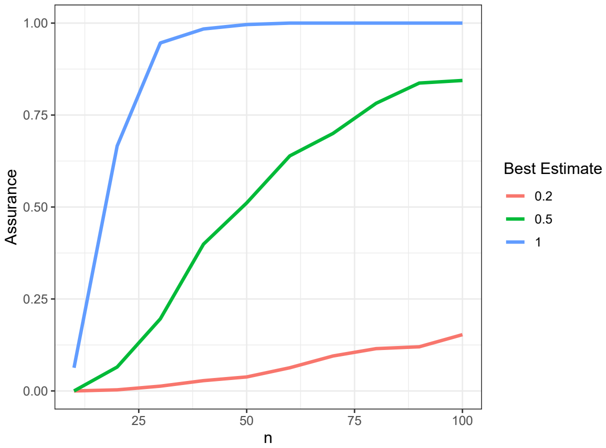


Figure A.12: Assurance curves for three experts, with design priors $N(0.2, 0.1)$, $N(0.5, 0.1)$ and $N(1, 0.1)$.

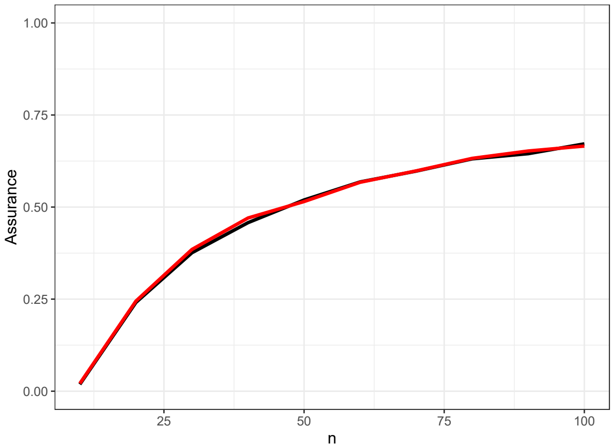


Figure A.13: Assurance curves for an equal weighted design for the three experts (black) and the average assurance value for each of their assurance curves (red).

fictional experts. Their individual priors have been chosen to be $N(0.2, 0.1)$, $N(0.5, 0.1)$ and $N(1, 0.1)$, to represent a group which agrees there is a positive effect, but has differing opinions as to how large it is.

We take a difference of 0.1 to be required for a clinically significant difference, and so the posterior distributions will be compared to this value. This will be done using a Bayesian analysis, in which a sceptical analysis prior of $N(0, 0.2)$ is used. Figure A.12 provides the three assurance curves.

In this case, the three curves provide more varying differences in the probability of success.

Figure A.13 provides the aggregate prior's assurance and the averaged assurance curve. As the plot demonstrates, in this scenario it does not matter whether the priors are aggregated before calculating assurance, or whether the assurances are aggregated at each value of n.

While it is possible to average over assurances in this case, it is not always possible to average over the sample sizes. The curve with a best estimate of 0.2 in Figure A.13 has a maximum assurance lower than the other curves. If the chosen assurance cutoff is above this maximum, for example at a value of 0.8, then the average cannot be calculated as there will not be a sample size to be calculated for this curve.

This chosen set of prior distributions further demonstrates how aggregating the prior distributions in the calculation is a more sensible option. Furthermore, it can be the case that some assurance curves will not reach the required assurance to calculate a sample size at all, and thus individual sample sizes could not be aggregated.

While we have shown the benefit of aggregating the prior, rather than the assurance value, it is also important to consider the use of an individual expert's assurance curve. The individual assurance curves should contain the aggregated assurance curve between them, and so they could be used as a boundaries. As the curve represent what different expert's believe, and assurance value within their range for a given value of n could be considered more reasonable than those values outside it.