Bayesian design of pilot trials for complex interventions

Abstract

. . .

1 Confirmatory trial designs

Prior to considering how to design the pilot, we need to know how we will use pilot data to design the confirmatory trial. Consider only hybrid methods, i.e. those which assume a frequentist analysis in the main trial but make use of probability distributions describing uncertainty in the nuisance parameters, the treatment effect, or both. Note that a hybrid approach may be a special case of a more general fully Bayesian approach with some specific analysis priors, so include fully Bayesian methods in the review.

Example problems include: uncertainty in a single ICC in a fully nested design; uncertainty in multiple variance components in a partially nested design; uncertainty in feasibility aspects such as recruitment rates, cluster sizes, data collection rates, adherence etc.; uncertainty in effect sizes and their correlations in the case of multiple endpoints.

[**?**] - Assurance in clinical trial design Defines assurance for any meaningful outcome of the trial, not just the outcome "null hypothesis is rejected". For example, can also apply to "null rejected and intervention is truly superior". The idea of calculating the unconditional probability of the outcome, with respect to some prior, remains the same. In the discussion, talks about more complex processes than a single trial, e.g. a phase II leading to a phase III, and how these can be dealt with providing the process can be modelled e.g. the relationship between the phase II and III endpoints. Focus is on using these assurance quantities to help guide the design process, as opposed to automating it through some optimality criteria

Citing [**?**]:

[**?**] - Sample size and the probability of a successful trial Just a practitioner-focussed illustration and discussion of the methods in [**?**].

[**?**] - Evidence-based sample size calculations based upon updated meta-analysis Outlines a simulation based approach for SSD based on an existing meta-analysis. hybrid approach considering assurance, where the prior comes from the meta-analysis and the outcome is rejecting the null in a test based on updating the meta-analysis. Focusses on using a prior for the treatment effect, but notes that the method can easily incorporate priors for nuisance parameters. Considers fixed and random effects meta-analysis models. For random effects, they use MCMC in the meta-analysis to generate priors for both the treatment effect and the between-trial heterogeneity, noting that uncertainty in the latter will likely be significant since only a few studies will be available. Sampling a new treatment effect then involves sampling from N(theta, tau) where theta and tau have a joint distribution. Refers to the computational burden of a two stage nested MCMC approach in the case of the random effects model - the 2nd MCMC coming from having to update the meta-analysis for each hypothetical trial, which requires MCMC (compared with the fixed effect model which does not). Consider a few outcomes: rejecting the null (so average or unconditional power); the width of a confidence interval (or more generally, the variance in the estimate); and 'limits of equivalence'.

[?] - Simple, Defensible Sample Sizes Based on Cost Efficiency Argues against the typical approach where first a power constraint is decided upon, then the sample size determined, in favour of explicitly considering the trade-off between the two characteristics. They avoid having to quantify the projected value of a study. They ask for a function f(n) to be specified such that v(n*)/f(n*) ¿= v(n)/f(n) for all n ¿= n* - i.e. as we increase n, the ratio between the true value and the chosen f decreases. Then we use this f in place of v, and choose the n which minimises c(n)/f(n), i.e. that which would be most cost effective (for this surrogate f). They suggest two fs, each leading to a suggested sample size which we can be sure is more cost effective than any larger sample size. Not focussed on designing using priors, although they do mention this in an example to show such an approach can agree with theirs.

[?] - Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation A review

[?] - Bayesian sample size for exploratory clinical trials incorporating historical data Fully Bayesian, suggesting outcomes such as having a high posterior probability that the treatment effect is better than the control. Asks for a trial which will convince us that either the intervention is better than control, or that the difference is less than some hypothesised value. Argues that the fully Bayesian approach is appropriate for early phase trials. "Misspecification of utilities can lead to seriously sub-optimal designs . . . implementation [of decision-theoretic Bayesian approaches] in clinical trial design has been and probably will remain limited".

[?] - Optimal Cost-Effective Go-No Go Decisions in Late-Stage Oncology Drug Development Designing trials to maximise cost-effectiveness, requiring specification of all costs and all benefits.

[?] - Power for studies with random group sizes Motivated by observational studies where we don't know in advance how many will be in each 'arm', but we need to choose the overall sample size. Considers uncertainty in p, the probability of being in one of the two groups. Given p, the average power can be calculated. To acknowledge uncertainty in p, integrate over a prior for it. Generalises to multiple groups, as we would see in e.g. a 2-by-2 table. Note that here, the group membership and the effect size are entwined. Say that the method is not appropriate for RCTs since the group size is controlled, but this isn't true in practice, particularly for our complex intervention trials.

[?] - Bayesian Design of Noninferiority Trials for Medical Devices Using Historical Data Fully Bayesian. Uses the [?] idea of a 'sampling' prior for the design stage, and a 'fitting' prior for use in the analysis. They define a Bayesian power quantity as the probability (under the sampling prior) that the posterior probability (under the fitting prior and conditional on the data) that the treatment effect reaches the required threshold is high (where high is specified exactly). They define two sampling priors, a null type with support on the intersection of the null and alternative parameter spaces, and an alternative type with support only within the alternative parameter space. Then they define Bayesian type I error to be the power under the former, and power to be under the latter. How are we to interpret these two sampling priors? In their simple example, they use two point mass priors. They discuss how to incorporate historical data from previous trials (only relating to the control) using i) hierarchical priors, and ii) power priors. - Uses point priors to try and mimic frequentist hypothesis testing and error rates, which seems to defy the point of a fully Bayesian approach. - Bayesian methods for incorporating historical data, e.g. power priors and comensurate priors, may help bridge the gap between the settings of the pilot and the main trial. That is, we want to use pilot data on e.g. recruitment rates, but know these are likely to be different in the main trial.

[?] - A Quantitative Approach for Making Go/No-Go Decisions in Drug Development Nothing of interest, discusses the analogy between screening/diagnostic tests and hypothesis testing in trials and describes assurance.

[?] - Optimal Sample Sizes and Go/No-Go Decisions for Phase II/III Development Programs Based on Probability of Success Considers trial design and SSD simultaneously across a phase II and III trial program. The two trials are identical other than their sample size, 2 arm trials with a continuous normal endpoint with known variance. They give an 'empirical Bayes' approach - they suggest a uniform prior for the treatment effect at phase II, which gives an updated distribution (in closed form) for the phase III prior, from which POS can be calculated. Then they consider simultaneous SSD. Prior to observing the phase II data, they consider POS as a random variable (since it is a function of the phase II sample estimate). They suggest we should consider the median of this POS distribution when doing SSD, but their notation is not clear - they need to specify some treatment effect, in fact, the expected value of the phase II sample estimate, but its not clear where this comes from (having previously said they have a uniform prior on the true effect). So, actually this is all conditional on some supposed effect size. They suggest making a stop/go decision after the phase II data based on the POS, and suggest some possible cut-off values. Note that these can be converted back to decision rules based on the phase II sample estimates. - Considers POS as a random variable, prior to phase II data being observed. - Limits themselves to simple conjugate cases where posterior distributions can be derived in closed form.

[?] - The perils with the misuse of predictive power

[?] - Planning future studies based on the conditional power of a meta-analysis Builds on suggestions that we should design trials with respect to existing meta-analyses. Frequestist approach, using conditional power - given the data (from past studies) we have already observed, what is the power to detect a pre-specified effect size. Discusses issues around multiple testing, and essentially frames the problem as an adaptive design. Not Bayesian, a frequestist alternative to [?].

[?] - Predictive power to assist phase 3 go/no go decision based on phase 2 data on a different endpoint Focusses on oncology, where the phase III endpoint will be OS and the phase II will usually e PFS (or response rate). Generalises to any situation where the two parameters (effect sizes for each endpoint) are asymptotically (?) bivariate normal. Phase III stat is the estimates log OS hazard ratio. They use a bivariate normal for the prior of the effects conditional on the correlation between them, with that correlation in turn having a prior distribution. They note this is a conjugate prior, since its estimate is known to be asymptotically bivariate normal. They derive expressions for predictive power (i.e. assurance) over the phase II and III trials, both when we only use PFS at phase II and when we also incorporate the OS phase II information. Need to properly go through the appendix to understand the derivations of these predictive probabilities. When considering SSD for the phase II trial, they calculate the predictive power you would get from different combinations of phase II sample size and the phase II data (i.e. the observed OS HR). Seem to suggest what we just look at the resulting contour plot and try to pick a Phase II sample size at which the increasing predictive power starts to level off. Doesn't formally consider the sampling distribution of the phase II statistic. - In the predictive probability expressions we need to say how many events we will see in the phase II and the phase III trials. Analogous to sample size, but also clearly unknown at the design stage, so should be integrated out also?

[?] - Discounting phase 2 results when planning phase 3 clinical trials In a Bayesian framework and using the average power type of assurance measure. Focus is on dealing with biased phase II results. Builds upon / contrasts with [?]. Mentions 'dual criteria' and gives some references - here, we have two hypothesised treatment effects, a minimal effect and a desired effect, and we can ask for different probabilities that the effect exceeds these before moving to a phase III trial (so implying a Bayesian approach to phase II analysis, although they go on to say you could use confidence intervals, so not clear). They use dual criteria as part of the decision rule - the 90% CI should exclude one threshold, while the 75% CI should exclude another, for example. Compares 5 methods which jointly specify a stop/go decision rule from moving from II to III, and a SSD method for the phase III trial. The main contribution is suggesting that for the latter, we can multiply the phase II estimate by some factor and use this in the sample size calculation. This seems strange - why should we want to look for an effect smaller than the MCID? Why should our target be derived from a phase II estimate? Connects with choosing the effect size to detect based on its plausibility rather then its clinical relevance. In comparing the methods in a number of scenarios, they calculate assurance, where the prior on the treatment effect is somehow based on the phase II data - not clear exactly how. They cite [?] and looking back there, it seems they are proposing that the prior should be normal, with mean equal to the estimated treatment effect, and sd equal to the sample standard error. This may be what you would get if you started with a uniform prior? They use a simple formula for assurance which takes as arguments estimated treatment effects and variances, possibly adjusted according to which method is used. They say plugging in these estimates gives you a 'estimated' assurance, while plugging in the 'true' values gives 'theoretical' assurance. Not convinced these have meaningful interpretations. As a way to discount phase II data, the methods are simple but do not appear to have a strong theoretical foundation - contrast with alternatives like commensurate priors. - Follow the references on 'dual criteria', from a frequentist perspective this suggests having more than one power constraint, similar to out thinking of power in a multi-dimension parameter space. - An example where the interpretation of the assurance metric isn't clear, particularly in terms of the justification for the prior distribution which is fed into it.

[?] - Evaluating and utilizing probability of study success in clinical development Assurance type approach. The prior on the treatment effect they use is really a posterior, based on existing data. From the start, they state that we will generally need to use simulation to do the assurance calculations, as noted in [?]. Note that the MC routine involves sampling from the joint distribution on p(x, theta). Doesn't offer any new methodology, just described actual applications in a pharma company. Examples are as interested in the historical / existing / phase II data side as they are on the assurance side. - Assumes that CTS will be needed when doing the assurance integration. - Example of sampling from p(x, theta) in the MC estimation scheme, confirming that moving from simulating conditional power to average power is trivial. - Notes another benefit of running CTS's is that we can get distributions on other OCs of interest, such as study duration.

[?] - On the Limiting Behavior of the Probability of Claiming Superiority in a Bayesian Context Proves some results along the lines of average power converging to the prior probability of rejecting the null, as the sample size increases to infinity. Does this for Bayesian success criteria, which generalises to the usual frequentist success criteria.

[?] - Joint probability of statistical success of multiple phase III trials Argues that a regulator will generally require 2 phase III 'successes' to accept a new drug, so we should calculate this joint probability of success when designing our trials. Considers the prior for the treatment effect as being a result of a phase II trial and an initial uninformative prior. Notes that the joint probability is not just the product of two separate probabilities, because the two test statistics will be correlated. Focusses on the usual normal scenario, and derives analytic results for the joint predictive power. Extends to the case where we have more than two trials to design, and where success means that we reject in null in some pre-specified number of these.

[?] - A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach Phase II, 2 stage, binary outcome, randomised setting. Proposes that we use a Bayesian predictive probability criteria when optimising designs, rather than the usual type I and II rates. So, fully Bayesian. They propose a Bayesian power type approach, where we calculate the probability of obtaining a large posterior probability of effectiveness. Makes use of analysis and design priors.

[?] - Assurance calculations for planning clinical trials with time-to-event outcomes First considers an exponential model for survival. They use a large sample expression describing the normally distributed test statistic. Gives some details on how they elicited the two parameters in the model, which is done indirectly. The calculation of assurance is computationally simple using MCMC. Goes on to consider a Weibull model. Similarly to the exponential case, they use the fact that the test statistic will have some sampling distribution (in this case, t). The Weibull model requires a further two parameters, making the elicitation more challenging, again they describe how the joint distribution can be constructed by eliciting on observable quantities. Going on to using non-parametric models, they look at how pilot data can be incorporated with expert opinion when deriving the necessary prior distributions. - example of eliciting relatively complex model parameters - example of applying the general assurance approach to a specific situation, similar to what we are planning; if we want to talk about eliciting ICCs and related quantities, this type of paper could be the result; note that in our case an MC approach to calculations would perhaps be more involved that sampling a test statistic directly from its sampling distribution

[?] - Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study Not Bayesian design, but does consider how to use pilot data. Considers using pilot data to estimate both an SD and an event rate in a control arm. Uses simulation to describe the various sampling distributions (despite the fact that analytic descriptions are available), and tries to quantify the trade-off between the sample size of the pilot and the resulting precision of the nuisance parameter estimate. From this, gives some rough guidance based on how large the pilot should be, based on diminishing returns in precision. Shows that in many cases, inflating an estimate (e.g. up to an upper CI limit) to guard against its imprecision will lead to unnecessarily large main trials and overall it would be more efficient to increase the pilot sample size.

[?] - The statistical interpretation of pilot trials: should significance thresholds be reconsidered? Suggests a Bayesian analysis of pilot data and a decision rule which asks for a certain posterior probability (having observed the pilot data) that the true effect s greater than the MCID. Does not consider how we should design the pilot to facilitate this.

[?] - A Quantitative Process for Enhancing End of Phase 2 Decisions From a pharma perspective, outlines a general strategy to guide moving from phase II to phase III. Two stages - first synthesising the evidence, then using models to make decisions. They propose to use meta-regression to build a model which relates the different outcomes at phase II and III - an example result is a model which predicts the true phase III effect as a multiple of the estimated phase

II effect. They propose using a random effects model for the meta-regression, under a Bayesian framework with non-informative hyperparameters. In the modelling stage, they assume that the phase II estimate is normally distributed, and define 3 default (noninfomative, enthusiastic, sceptical) on the true phase II effect. Corresponding posteriors are constructed given the phase II estimate (all conjugate). The three posteriors are combined with the meta-regression model to give predictive distributions of the phase III effect. These are in turn used to simulate the planned phase III trial and calculate the probability of success (however this is defined). Uses MC (although refers to it as MCMC?). Does not consider optimal design of either the phase II or phase III, just the evaluation of some proposed phase III design. - Is the proposed mete-regression approach the right way to quantify surrogacy between phase II and III outcomes? - Suggests a benchmark PoS will be study specific, but could be taken to be the current industry average

[**?**] - Bayesian probability of success for clinical trials using historical data

[**?**] - Sample size planning for phase II trials based on success probabilities for phase III Considers a setting where a phase II and III trial are to be done in the same population with the same outcome, with unbiased estimates of the treatment effect. Time to event endpoint, assuming normally distributed log hazard ratio estimates. They show that under certain conditions (inc a non-informative normal prior on the treatment effect), a frequentist stop/go decision rule at phase II can be equivalent to a Bayesian rule asking for a certain probability that the treatment effect is beyond a certain threshold. Uses the phase II estimate to design the phase III trial, so introducing randomness in that sense rather than through a prior. Proposes measures of the probability of a successful phase II and III program, and the probability of a successful phase III trial given that the decision was made to do one, both conditional on an assumed treatment effect. They suggest we should take two hypothesised effects as we do in frequentist testing, but design the trials around these probabilities rather than the usual error rates (although these are still used to design the phase III trial). They go on to look at unconditional probabilities, using a mixture of an effective prior and an ineffective prior. Not clear why. Considers the prior as a distribution of a large set of treatment effects, avoiding a Bayesian interpretation. - Asks for even more constraints to be set in order to define an optimal phase II design. - Another example of using a phase II estimate as the effect to detect in phase III, which seems unprincipled - Not convinced that the Bayesian and frequentist views are properly separated - Gives some discussion about the implications of assuming an infinite number of treatments are ready to be evaluated, as in Stallard's paper; connected I think to the 'long run' frequentist view compared with the 'short run' Bayesian view.

[**?**] - Utility-based optimization of phase II/III programs Same group as [**?**]. Same setting, but starts here goes straight to putting a prior on the unknown treatment effect and then defining the probability of going to phase III and getting a significant result there. Still assumes the phase II estimate will be used to choose the phase III n. Describes a utility function. Costs are per patient and set up costs, for both phase II and III. Benefits are discrete - they set three levels, and say that the appropriate level is determined by the upper 95% CI of the *estimated* treatment effect. Their utility is the sum of these, so they split the expected utility into expected benefit and expected costs. - How can we define benefit in terms of an estimated effect as opposed to the true effect? - Uses a number of fixed cost and benefit parameters, doesn't look at (probabilistic) sensitivity analysis - Optimises over phase II sample size and decision rule

Subsequently found:

[**?**] - Accounting for Variability in Sample Size Estimation with Applications to Nonadherence and Estimation of Variance and Effect Size Related to our CI problems in particular, as considers the effect of uncertainty in the number of non-compliers. Referes to [**?**] for a discussion of the usual methods to correct, which assume the parameter is known. They a non-parametric model, where those who get the treatment follow one distribution and those who don't follow another, with these distributions defined only by their mean and variance. They derive the expectation and variance of the sample treatment difference, followed by a sample size formula. Actually, doesn't consider the adherance parameters as random - their improvement over [**?**] is through considering only the population parameters, as opposed to being required to know in advance exactly what proportions we will see in the study. Then goes on to consider variability from plugging in random parameter estimates into a sample size formula. All frequentist, assumes that estimates will be plugged into a sample size formula, suggests that to maintain high expected power the nominal power level should be increased. - Avoids any discussion of a Bayesian view, instead claiming that uncertainty comes only from sampling distributions of previous estimates, and uses 'fiducial' arguments

[**?**] - Sample size calculations for clinical studies allowing for uncertainty about the variance Considers sample size formula for a t test of two means, where the usual plug-in estimate of the true variance is available from a previous study. Gives a formula, but the proofs in the appendix are not clear or convincing.

[**?**] - Bayesian sample size determination for binomial proportions Concerned with choosing sample size to satisfy a range of Bayesian criteria such as the Average Length Criteria, which are all reviewed and extended to the binomial case. For those cases where Monte Carlo methods are needed to evaluate sample size, they suggest fitting linear regression models to the (one dimensional) data to help find the optimal n.

[**?**] - Using historical data for Bayesian sample size determination Fully Bayesian approach, designing to give some kind of bound on an expectation or probability with respect to some quality of the posterior. Uses design and analysis priors. Uses power priors to incorporate historical data into the design prior. This introduces a parameter describing the weight given to historical evidence, which must be chosen when doing SSD. Suggest to plot this against the corresponding sample size requirement. Notes that some power priors have a simple form, but that in general we may have to use simulation in the computations. - Uses simulation to evaluate a given sample size; involves calculating a measure based on the posterior in each sample

[?] - A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models Early mention of using design and analysis priors, and of a general simulation approach to estimating performance criteria. Simulation routine is intensive, again requiring sampling from the posterior for each sample of the data.

[?] - Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study Suggests using two constraints on posterior probabilities that the effect is greater than the MCID, and that it is greater than zero. Argues that PPV (i.e. the pre-posterior probability of a meaningful effect given we have a positive test result) and NPV are better than unconditional power. They use a fixed estimate of sigma rather than a prior, noting that using a prior typically does not affect the assurance because the latter is approximately linear over the range of sigma, and the prior tends to be symmetric. But variance priors may not be symmetric, particularly if they could be close to zero. They say that this behaviour does not reflect an asymmetric utility in power, where a loss in power has a greater effect than the same gain in power. Note that in our utility model this isn't always true, with there being an optimal power (both conditional and unconditional) for a fixed alpha.

[?] - Adapting the sample size planning of a phase III trial based on phase II data Frequentist model where phase II effect estimate is used to plan phase III, and they suggest using a lower CI limit of that estimate to guard against the possibility that the true effect in the phase III setting and population is lower than that in phase II.

[?] - Optimal sample sizes for phase II clinical trials and pilot studies Claims to approach the problem from a Bayesian decision-theoretic perspective, but takes a long-run view of a program of phase II and subsequent phase III trials. Aim is to minimise the number of patients required per successful phase III, although as we see later this leads to recommendations of lots of tiny trials. Bayesian since they propose a prior on each treatment effect, which is assumed to be identical for each treatment - so doesn't interpret as a distribution of actual treatment effects. Optimisation is only over the OCs of the phase II trials, with phase III OCs assumed to be fixed. In the example of a comparison of normal means, they suggest that alpha should be around 0.2 and a power of around 0.4, and they suggest that a phase II should generally be around 0.03 times the size of the phase III.

[?] - A mathematical model for maximizing the value of phase 3 drug development portfolios incorporating budget constraints and risk Develops a complex model for optimising a portfolio of drugs and trials, making use of a Bayesian view when calculating the probability of a trials' success.

[?] - Designing phase II studies in the context of a programme of clinical research For a fixed number of patients, shows how to choose the number of treatments which should be tested, assuming their effectiveness follows a beta prior distribution. Also mentions expected gain, using a similar function to ours.

[?] - Sample sizes for phase II and phase III clinical trials: An integrated approach

[?] - On an Approach to Bayesian Sample Sizing in Clinical Trials

[?] - Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations

[?] - Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study

[?] - A Bayesian predictive sample size selection design for single-arm exploratory clinical trials

[?] - Bayesian methods for setting sample sizes and choosing allocation ratios in phase II clinical trials with time-to-event endpoints

[?] - The what, why and how of Bayesian clinical trials monitoring

[?] - The chart trials: Bayesian design and monitoring in practice

[?] - Implications for trials in progress of publication of positive results

[?] - Meta-analysis in the design and monitoring of clinical trials

[?] - Bayesian interim analysis of phase II cancer clinical trials

[?] - Prospective application of Bayesian monitoring and analysis in an ‘open’ randomized clinical trial

[?] - A Bayesian re-assessment of two phase II trials of gemcitabine in metastatic nasopharyngeal cancer

[?] - Early stopping of a clinical trial when there is evidence of no treatment benefit: Protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project

[?] - A framework for prospectively defining progression rules for internal pilot studies monitoring recruitment

Notes

The most general way to calculate assurance is through trial simulation. This involves specifying a simulation program which takes all parameters of interest as input, along with the planned design, and outputs whether or not the event occurred. The problem of optimal design in this context is exactly what we are dealing with in WP4. We can apply the methods directly by simulating from prior distributions all the parameters which we currently condition on, and adjusting the power constraint to reflect that it is now unconditional. This is done in, for example, [?].

To illustrate and evaluate different approaches, we need to define priors. We can use the default priors set out in Speigelhalters book for this, particularly sceptical and enthusiastic priors for treatment effects. Would be good to have some actual examples to draw on and re-design. Look at the Turner papers for examples of ICC priors built from historic data sets

---

Themes to pick out / summarise

The range of events which are considered a ‘success’; e.g. rejecting the null; rejecting the null when the alternative is true; getting a significant result in an updated meta-analysis; rejecting the null in a number of trials

Guidance on how the measures proposed should be used in optimal trial design; e.g. Stallards criteria of minimising number of patients required for each phase III success

Approaches to doing to computations; e.g. focussing on conjugate scenarios; using simulation

How priors are developed; e.g. expect elicitation; updating an uninformative prior using phase II data; using historical trial data, potentially from a meta-analysis; using default priors; design and analysis priors; power priors to discount historical data

Difficulties in applying the methods to our types of problems. Any conjugacy approaches will generally not work due to multilevel models, but can use simulation when that is proposed. Elicitation of our type of parameters, e.g. complex variance components, and feasibility parameters e.g. recruitment rates. Defining success in terms of multiple endpoints. Advantage is that the same endpoint is used in both pilot and phase III, although setting and population may differ.

Try to identify one or two key approaches / papers which we can focus on extending to the complex intervention setting.

Scope of the problem:

Fully Bayesian (reducing to hybrid in a special case), or hybrid? - Aim for a general fully Bayesian method, and apply to a hybrid example.

Use historical data when describing the priors? Discount using power priors? - Assume that the correct priors are available, possibly from pilot data.

Conjugate or more general (with implications for computations)? General, including ways to deal with computational burden.

Parameters to include - treatment effects, variance components, adherence rates, recruitment rates, follow-up rates? - All.

Our main motivation is to provide a framework which methods to design pilot trials can slot into, so we need to accommodate the features particular to them - i.e. information about likely treatment effects (possibly across multiple endpoints), variance components (in the continuous case), and adherence, recruitment and data collection rates (following the paper on progression criteria for pilot trials). So our starting point assumes that we have prior distributions for all of these, probably obtained following a pilot trial.

We may need to choose a primary endpoint, which should account for our beliefs about effect sizes, MCIDs, and relative clinical importance. Note that we will typically only have two or three options, so we may wish to look at optimal design for all and then choose between these.

Recruitment models are seemingly mostly concentrating on making predictions whilst the trial is ongoing, using accumulating accrual data. We are only interested in a prior prediction of the final sample size. Can we assume that we are limited in the number of potential recruits? In practice, trials fail to recruit on time and have to require an extension. When we design a trial, what are we specifying - a recruitment deadline, or a sample size, or a number to be screened? Choose the sample size, then estimate how long it will take, then look at how uncertainty in recruitment rate affects uncertainty in study duration.

For follow-up / data collection, if we randomly drop a proportion of observations we would be modelling a MCAR process, but would want to analyse under MAR assumption. Could model a MAR mechanism, as in example 1 of [?]. Does the process in the data generating affect the power / precision of the trial, or does that come through only the method of analysis?

For adherence, what we've seen so far in [?] and [?] models non-compliance by substituting the treatment effect for that in the other arm, assuming an ITT analysis at the end. In that case, they give simple formula for inflating usual sample sizes. We could go further, as in the third example of [?], and plan a causal analysis.

Beyond parameter uncertainty, we may also consider structural uncertainty. For example, the mechanism of missing data or the model for the effects of non-compliance / adherence.

The general method is simple enough, but the implementation may be less than straightforward - in particular, defining the model which allows for the above features, and doing the necessary computations.

Given a set of assurance measures and constraints defining optimal confirmatory trial design, we can extend these to encompass the larger process of both the pilot and phase III trial if we assume a Bayesian updating after the prior. The challenge here is computational, and we should look at the methods for EVSI computation to see if and how they could be applied.

---

2 Multiple endpoints and choosing a primary

How do we currently choose a primary endpoint? According to draft FDA guidance on trials with multiple endpoints: "Endpoints are frequently ordered by clinical importance, with the most important being designated as primary (e.g., mortality or irreversible morbidity). This is not always done, however, for a variety of reasons. The most common reasons not to order endpoints by clinical importance are if there are likely to be too few of the more clinically important endpoint events to provide adequate power for the study, or if the effect on a clinically less important endpoint is expected to be larger. In these cases, endpoints are often ordered by the likelihood of demonstrating an effect" [?].

[?] - How do you design randomised trials for smaller populations? A framework "Statistically speaking, the best primary outcome to use is the one with the greatest information content; that is, the one which minimises the variance of the treatment effect relative to the magnitude of the treatment effect. In terms of information content, there is generally a hierarchy for outcome measures with continuous outcomes tending to hold most information, followed by, in order, time-to-event, ordinal and finally binary outcome measures. From a statistical perspective, it is, thus, sensible to use the most information-rich primary outcome available. It is always costly in terms of sample size to split continuous or ordered outcome data into two categories. Clearly the primary outcome measure must be important from the perspective of both patients and treating clinicians: the practical convenience of needing fewer patients should not determine the choice of outcome unless candidate outcome measures are considered relevant for decision-making for all interested parties, including patients, clinicians, relevant health authorities and, potentially, regulators."

But "likelihood of demonstrating an effect" doesn't have a clear meaning. Does this mean highest conditional power at the MCID? Or at the expected effect size? Or does it mean the likelihood of rejecting the null, unconditional on the true parameter values? Or the joint probability of rejecting the null and the true effect being meaningful? If we only consider conditional power, only the estimated sd of the endpoints will influence our decision, so we could easily end up choosing an endpoint with a much lower expected effect because it had a slightly lower sd.

As a start, we can look at a variety of hypothetical endpoints described by probability distributions on their treatment effect and their sd (which we assume are independent). These distributions can be thought of as resulting from a pilot trial. An initial question to ask is, will a Bayesian perspective ever lead us to choosing a different endpoint than the usual frequentist one would? The freq approach would essentially only look at a point estimate of the sd (which we take to be the mean of the distribution), and then looks at conditional power. So, all other information about the likely effect size, our uncertainty around it, or our uncertainty around the sd, is ignored. The Bayesian view accounts for all this. We use an assurance measure where success is defined as rejecting the null while the true treatment effect is greater than the MCID. It is not hard to find a pair of endpoints where the two approaches lead to different choices of endpoints, with the implications in terms of necessary sample size being quite significant (e.g. reducing 460 per arm to 320).

Some things to note: - The assurance measure is quite flat over the range of n. Suggests that a constrained optimisation approach is not sensible, since we may be able to dramatically reduce sample size at only a small cost to quality. Want to plot the costs and benefits to help decision making. - The measure will always benefit from greater uncertainty in the treatment effect since this will lead to greater prior probability of a meaningful effect. We can check the probability of going forward and the effect being ¡ 0, and we see in our example that it is always very low (¡ 0.001). - There will be clear interactions between the uncertainty in the parameters after the pilot, and the amount of missing data, which may be quite different for different outcomes. So we may want to focus on assessing missing data in the pilot, incorporating this into our framework by constructing distributions for missingness parameters. - Defining a measure to use to choose an endpoint will automatically extend to choosing the sample size of the main trial. So we can include multilevel aspects, i.e. learning about variance components in the pilot to help choose the best main trial design. - Can optimal outcome choice ever depend on n?

Consider two candidate endpoints which could be used in the main trial as the primary. Both are assumed to be normally distributed outcomes in our population, with mean μ_i and variance σ_i^2 for $i = 1, 2$. We expect there to be missing data for each outcome, with the number of complete data items following binomial distributions with parameters p_i. Following the results of a pilot study, our belief regarding the true values of these parameters is expressed through probability distributions, with our belief regarding one parameter of an endpoint being independent of the values of the other parameters relating to the same endpoint. Specifically, $\mu_i \sim N(m_i, s_i^2), \sigma_i^2 \sim Gamma^{-1}(f_i, g_i)$, and $p_i \sim Beta(a_i, b_i)$ (where the inverse gamma distribution has a shape f and rate g parametrisation).

We take the minimal clinically important difference to be μ^* in each case. The statistics used to measure each endpoint are denoted S_i. To estimate classical power, we calculate the probability of rejecting the null in a trial of effective sample size $p_i n$ where the treatment effect is $\mu_i = \mu^*$, the variance of the outcome is $\sigma_i^2 = \frac{g_i}{f_i - 1}$ (i.e. the mean of the prior distribution), and the proportion of data retained is $p_i = \frac{a_i}{a_i + b_i}$ (again, the mean of the prior distribution). Denote this power function as $\beta(n, \mu_i, \sigma_i^2)$ (note that power depends on p_i only through n).

Using only these point estimates and conditioning the treatment effect at the alternative hypothesis, the endpoint which gives the larger power could be selected. An alternative metric which accounts for uncertainty in the true parameter values is the joint probability of rejecting the null hypothesis and the true treatment effect being meaningful (i.e. $\mu_i \geq \mu^*$) [**?**]:

$$\int \int \int \beta(n, \mu_i, \sigma_i^2) p(\mu_i, \sigma_i^2, n) d\mu_i d\sigma_i^2 dn. \tag{1}$$

We compute a Monte Carlo estimate of this integral by taking N samples from the joint distribution $p(\mu, \sigma^2, p, n)$ (specifically we sample $\mu \sim N(m, s^2), \sigma^2 \sim Gamma^{-1}(f, g), p \sim Beta(a, b), n \sim Bin(n, p)$) and for the jth sample computing $y^j = \beta(n, \mu, \sigma^2) \times I[\mu \geq \mu^*]$ where $I[.]$ is the indicator function. The Monte Carlo estimate is then $\frac{1}{N} \sum_{j=1}^{N} y^j$.

3 Bayesian hybrid designs for pilot trials

Suppose that the parameter space Φ has been partitioned into three subsets:

- Φ_{-1}: We don't want to proceed to a phase III trial, even after modification to the trial or intervention design (decision a_{-1});
- Φ_0: We want to modify the trial or intervention design and the proceed to phase III (decision a_0);
- Φ_1: We want to proceed straight to phase III without any modifications (decision a_1).

Assume we have also defined a prior distribution $p(\phi)$. An example partitioning of a two-dimensional parameter space is given in Figure **??**, where each point has been sampled from the prior distribution.

We consider a decision-making process following the trial based on posterior probabilities. Specifically, we consider posterior probabilities $p_i = Pr[\phi \in \Phi_i \mid x]$ for $i = -1, 0, 1$. The vector $p = (p_{-1}, p_0, p_1)$ will lie in the space $P = \{p_i \in$

$[0, 1] \mid p_{-1} + p_0 + p_1 = 1\}$. Prior to the trial, we partition the set P into three subsets P_i, and define the decision rule as taking action a_i iff $p \in P_i$. An example partitioning of P is given in Figure **??**.

For state i and action j, $i, j \in \{-10, 1\}$, we can calculate

$$
\begin{aligned}
Pr[\phi \in \Phi_i, x \in X_i \mid D] &= Pr[\theta \in \Theta_{i,j}] && (2)\\
&= \int I[\theta \in \Theta_{i,j}] p(\theta) d\theta && (3)\\
&\approx \frac{1}{N} \sum_{k=1}^{N} I[\theta_k \in \Theta_{i,j}], && (4)
\end{aligned}
$$

where $\theta_k \sim p(\theta)$ are samples of the joint parameter $\theta = (\phi, x)$, which we can generate easily through simulation (first sampling ϕ, then $x \mid \phi$).

Three states and three actions gives nine probabilities. We already have the marginal probabilities $Pr[\phi \in \Phi_i]$ from the prior. How are we to process these attributes and use them to compare designs? Can we specify thresholds to specific probabilities or combinations of them, or weight them according to their relative importance?

Computationally, for any given sample size we can simulate a single sample of θ's, to which different decision rules can be applied with minimal extra computational effort. So, focus first on defining optimality criteria for a decision rule for a given fixed sample size. Then we can use efficient optimisation methods to find the optimal sample size with such a decision rule.

On posterior decision rules, [**?**] consider a Bayesian binary regression problem and use the posterior prob. of the effect being greater than 0 as their success criteria, with cut-off *alpha* "generally 0.1, 0.05, or 0.025, depending on the nature of the problem", and then asking for a probability of that being satisfied of $1 - \beta$, "often between 0.8 and 0.9". So they seem to make a link with frequentist operating characteristics, but not explicitly or with discussion. They cite [**?**].

[**?**] - Robust Bayesian sample size determination in clinical trials

[**?**] - Bayesian clinical trials in action

[**?**] - Bayesian Hypothesis Testing in Two-Arm Trials with Dichotomous Outcomes States that the Bayesian equivalent of a hypothesis test is calculating the posterior prob of the null and accepting the treatment if this prob is below α. Compares the frequentist and Bayesian tests to show when they are equivalent,

[**?**] - A Bayesian analysis of classical hypothesis testing Focusses on the case where one hypothesis is simple and the other composite, but notes at the start that when both are simple, or both composite, there is not much of a problem.

4 Frequentist designs

We have a multi-dimensional parameter space, some dimensions corresponding to nuisance parameters. We want to evaluate and optimise a design, which is a sample size and a decision rule, with three possible decisions. Evaluation should focus on error rates, of which we can define 3 types and elicit constraints. Ideally we want to calculate the maximum of each error rate, but as we can't know a priori where these will be attained, we would be left with an optimisation problem. Error rates cannot be determined analytically since there won't generally be a closed form expression for the multivariate sampling distribution of the statistics. Finding maximum error rates is therefore infeasible, so we instead elicit a number of points where we would like them to be controlled, inc nuisance parameters. Considering only a set of hypotheses means we should avoid flexible parametric decision rules, since these may behave strangely in areas away from the point hypotheses, so we stick with the rectangular rules currently used. Evaluation can then sample under the hypotheses using some kind of bandit algorithm to determine if there are any violations. Using the same sample, we can adjust the decision rule are re-evaluate. The evaluation output can be a combined error rate violation measure, e.g. a mean square, and we can then optimise sample size as in WP4.

Questions: - When can and cant we know a priori where error rates will be maximised? - When can and cant we get a closed form sampling distribution? (In this case we need at least a counter example to motivate things, showing that the true sampling distribution can be quite different to something constructed from marginals)

Challenges: - How to sample over different hypotheses most efficiently - How to optimise a decision rule using the same sample - How to combine error rates into a single measure and use this in an expensive optimisation problem

Evaluation: - How does the number of hypotheses affect performance? - How far off are actual error rates to those from combining univariate tests? Connection to multiple testing problems? -

5 Bayesian designs

A natural extension to asking people for a set of point hypotheses is to ask them for a partition of the parameter space. As above, working with the whole space under a conditional (i.e. frequentist) framework is difficult, as we need to look for maximums. If we take a Bayesian view, we only have to consider the prior and can evaluate a design with respect to this alone. Decision making after a Bayesian analysis is based on posterior probabilities, which we can reduce down to the probabilities of each hypothesis. Decision rules can therefore be defined in terms of partitions of this probability space. A design can be evaluated similarly to the above, but now we have unconditional probabilities of each cell and therefore of the errors we have defined. There is a computational challenge from taking expectations over the pilot data and within that over the posteriors. We can try to deal with this using the nonparametric regression technique in Strong.

6 Decision-theoretic designs

The Bayesian hypothesis testing approach defines a kind of latent utility function in order to do the necessary computations. If we define the utility function ourselves, both analysis and design come straight out of the Bayesian machine - we do not have to consider things like error probabilities. Defining a utility function on our three decisions and full parameter space lets us specify more fully the preferences and trade-offs which are hinted at in the elicited hypothesis partition. We can apply value of information computation methods again, now somewhat more directly.