Designing Simulations for Estimating the Operating Characteristics of a Bayesian Adaptive Randomized Controlled Trial

Yang Lu

Department of Mathematics and Statistics

McGill University, Montreal

June 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Contents

Abstract

Background: Randomized controlled trials (RCTs) are commonly used in evaluating the efficacy of medical interventions. However, RCTs are not easy to conduct due to constraints of cost and feasibility. Bayesian adaptive designs can be used to help improve the efficiency of RCTs while maximizing the chance that participants receive a more advantageous intervention. Designing a Bayesian adaptive trial requires extensive simulations to ensure that the operating characteristics (i.e. the Type I and Type II error) of the trial satisfy pre-determined criteria. Planning such trials is a time-consuming process that can be daunting. The literature on examples illustrating how such simulations are designed is still limited.

Objective: To provide a checklist of items to be considered when designing simulations for a Bayesian adaptive trial and illustrate its implementation in the context of the DEFINE randomized controlled trial of flu vaccines in patients with rheumatoid arthritis.

Methods: Based on our experience in designing simulations for the DEFINE RCT, it was determined that study design features of a trial which influence how it can be adapted include: number of intervention arms, number of interim analyses and the timing of interim analyses. Aspects of the simulation design that need to be specified include a guess value of the sample size, number of trials simulated and number of posterior samples obtained from each trial. For the DEFINE trial, an informative mixture prior was employed, reflecting the level of confidence in the credibility of prior information. The number of arms was set to 3, the number of interim analyses was set to 1, the number of simulated trials was set to 1000 to estimate operating characteristics and the number of posterior samples drawn in each trial was set to 10000.

Results: We defined a checklist with 8 points covering aspects of the study design and simulation design. For the DEFINE trial, we found that there were 15 different possible adaptations. The sample size needed to achieve the trial's objectives using a frequentist sample size calculation was 1683 to demonstrate the superiority and noninferiority of ADJ with 5% probability of making a type I error and 95% power, given that the minimum clinical significant differences between ADJ and SD and ADJ and HD were assumed to be 11.2% and 2%, respectively. Our simulations showed that with a sample size of 800 it was possible to demonstrate the superiority of ADJ vs SD with Type I and Type II errors of 0.05 in a Bayesian adaptive design. However, to demonstrate non-inferiority of ADJ vs HD, a Type II error of 0.1 had to be tolerated if the maximum sample size is constrained to 1000. Further, with the sample size of 800 and the probability that the trial would be adapted due to superiority being detected in Year 1 was 98.4%.

Conclusions: Availability of a checklist can facilitate the planning of simulations necessary to evaluate a Bayesian adaptive trial design. In the case of the DEFINE trial, the simulations allowed us to evaluate its feasibility under different scenarios. In particular, we were able to calculate the probability of having a certain adaptation under various settings.

Résumé

Contexte: Les essais contrôlés randomisés (ECR) sont couramment utilisés pour évaluer l'efficacité des interventions médicales. Cependant, les ECR ne sont pas faciles à réaliser en raison de contraintes de coût et de faisabilité. Les dévis adaptatives bayésiennes peuvent être utilisées pour aider à améliorer l'efficacité des ECR tout en maximisant les chances que les participants reçoivent une intervention plus avantageuse. Le plan d'expérience d'un essai adaptatif bayésien nécessite des simulations approfondies pour s'assurer que les caractéristiques de fonctionnement (c'est-à-dire les erreurs de type I et de type II) de l'essai satisfont à des critères prédéterminés. La planification de tels essais est un processus qui prend du temps, qui peut être intimidant et qui est limité par une littérature sur les exemples illustrant comment de telles simulations sont conçues, un peu mince.

Objectif : Produire des listes de verification pour concevoir des simulations pour un essai adaptatif bayésien et illustrer sa mise en œuvre dans le contexte de l'essai contrôlé randomisé DEFINE de vaccins contre la grippe chez des patients atteints de polyarthrite rhumatoïde.

Méthodes : Sur la base de notre expérience dans la conception de simulations pour l'ECR DEFINE, il a été déterminé que les caractéristiques du plan d'expérience d'un essai qui influencent la façon dont il peut être adapté comprennent : le nom-

bre de bras d'intervention, le nombre d'analyses intermédiaires et le calendrier de analyses intermédiaires. Les aspects du plan d'expérience de la simulation qui doivent être spécifiés comprennent une valeur approximative de la taille de l'échantillon, le nombre d'essais simulés et le nombre d'échantillons a posteriori obtenus à partir de chaque essai. Pour l'essai DEFINE, un mélange informatif a priori a été utilisé, reflétant le niveau de confiance dans la crédibilité des informations a priori. Le nombre de bras a été fixé à 3, le nombre d'analyses intermédiaires a été fixé à 1, le nombre d'essais simulés a été fixé à 1000 pour estimer les caractéristiques opératoires et le nombre d'échantillons à postériori prélevés dans chaque essai a été fixé à 10000.

Résultats : Nous avons défini une liste de contrôle avec 8 points couvrant les aspects de la conception de l'étude et de la conception de la simulation. Pour l'essai DEFINE, nous avons constaté qu'il y avait 15 adaptations différentes possibles. La taille de l'échantillon nécessaire pour atteindre les objectifs de l'essai en utilisant un calcul fréquentiste de la taille de l'échantillon était de 1683 pour démontrer la supériorité et la non-infériorité de l'ADJ avec une probabilité de 5% de commettre une erreur de type I et une puissance de 95%, étant donné que les différences cliniques significatives minimales entre ADJ et SD et ADJ et HD ont été supposés être respectivement de 11.2% et 2%. Nos simulations ont montré qu'avec une taille d'échantillon de 800, il était possible de démontrer la supériorité de l'ADJ par rapport à la SD avec des erreurs de type I et de type II de 0.05 dans une conception adaptative bayésienne. Cependant, pour démontrer la non-infériorité de l'ADJ par rapport à la HD, une erreur de type II de 0.1 devait être tolérée si nous limitons la taille de l'échantillon à une maximum de 1000. De plus, avec la taille de l'échantillon de 800 et la probabilité que l'essai soit adapté

en raison de la supériorité détectée au cours de l'année 1 était de 98.4%.

Conclusions : La disponibilité d'une liste de contrôle peut faciliter la planification des simulations nécessaires pour évaluer un plan d'essai adaptatif bayésien. Dans le cas de l'essai DEFINE, les simulations nous ont permis d'évaluer sa faisabilité sous différents scénarios. En particulier, nous avons pu calculer la probabilité d'avoir une certaine adaptation dans divers contextes.

Acknowledgements

I would like to express my sincere appreciation by acknowledging my supervisors Dr. Nandini Dendukuri and Dr. Russell Steele, who had given me invaluable guidance, encouragement, and support throughout the span of this research. Throughout this journey, Dr. Dendukuri not only spent a considerable amount of time taught me academic knowledge but also enlightened me on how to approach challenges in life and how to keep going whenever I stumble. Meanwhile, since the first day I entered the program, Dr. Steele not only taught me academically, but also helped me greatly and constantly in planning graduate school life for me and immersing me in the department's culture and philosophy.

I would also like to acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) awarded to Dr. Dendukuri, as well as the award from the Department of Mathematics and Statistics at McGill University that supported me financially during a significant period of my research. It would not be possible for me to complete this study without your generous support.

I would also like to thank Dr. James A. Hanley and Erica Moodie from the Department of Epidemiology and Biostatistics at McGill University, who taught me and provided me with opportunities to practice the statistical tools I learned in real-world problems. Additionally, I also wish to thank Dr. James A. Hanley for

agreeing to serve as the committee member.

I would also like to thank all the group members in the DEFINE trial. In particular, I would like to thank Dr. Inés Colmegna and Dr. Nandini Dendukuri for getting me involved in this trial, from which I get the inspiration. It would not be possible for me to complete my study without your help.

I wish to thank all my friends that support me. In particular, I would like to thank Peiyuan Huang for supporting me with many helpful discussions.

Personally speaking, I wish to give my biggest thank to my parents and my partner Xintong Jiang who gives me boundless love and support through this journey. Thank you for always being the most supportive people around me. It is your encouragement that supports me to the destination of this long journey.

Chapter 1

Introduction

1.1 Background

Randomized control trials (RCT) are comparative studies used to evaluate the efficacy of one experimental intervention compared to another. The classical RCT consists of a control group and an intervention group, where the participants are equally likely to be assigned to either group. RCTs are considered superior to non-randomized trials because, first, the randomized allocation eliminates potential confounding bias introduced by the selective allocation of participants. This could happen, for example, when the investigator or participants may prefer receiving one of the interventions over the other based on several factors which may also be related to the outcome under study. Randomization ensures that the participants with different prognostic characteristics and demographic characteristics, will be evenly allocated in the intervention and control groups. Secondly, the procedure of randomization guarantees the validity of statistical tests of significance testing. For example, Student's t-test for comparing the difference in means between two groups is valid and does not require any further adjustments for the balance of covariates.

Despite the advantage of randomization, RCTs are not easy to conduct due to constraints of cost and feasibility. Another disadvantage of RCTs is the lack of generalizability as the participants are not typically representative of the population of interest. Also, obtaining informed consent from individuals can be complicated especially if the intervention is perceived as inconvenient or harmful or if participation in the trial does not offer a reasonable chance of receiving a promising treatment. (Sanson-Fisher et al., 2007).

One solution that can mitigate some of the cons of the classic RCT design is the adaptive design. In a sequential adaptive design, multiple interim analyses are carried out within the trial to determine whether the evidence is sufficiently strong to conclude in favour of one of the interventions, and reconsider the allocation of patients or stop the trial early. The adaptive design thus provides a more flexible and cost-effective way to conduct a traditional RCT study while allowing the possibility to reduce the duration and the cost of the trial and increase the probability that participants receive a beneficial intervention.

The sequential design can be implemented not only using the frequentist approach but also under a Bayesian framework. The Bayesian approach for conducting RCT design and analysis has been used and accepted increasingly. This is because the Bayesian framework offers a formal way to specify and update previous information using prior distributions, thus further strengthening the cost-effective usage of the available knowledge about an intervention. Also, the inferences based on posterior distributions can be expressed as probability statements and are easy to interpret. Besides, calculations of some important frequentist characteristics, such as Type I and Type II error, can also be justified even under the Bayesian framework (Rubin, 1984; Berry, 2010).

A distinctive feature of planning an adaptive trial is the usage of simulations. Simulations are necessary because the final design of the study is not known at the time of planning such that the estimates of the operating characteristics may not be obtained analytically. Different versions of the design are plausible based on the number of intervention groups, number of interim analyses and possible modifications at each interim analysis. Simulating these different versions allows researchers to determine various aspects of planning the trial, including the sample size. The literature on the topic of planning simulations is still young. Trial designers need more examples to guide them.

1.2 Outline

The purpose of this thesis is to propose a checklist that will outline a step-by-step process for designing simulations. The simulation approach is employed to determine aspects of the trial design that can be controlled, e.g. the sample size, to obtain the desired strength of evidence under various settings that cannot be controlled, e.g. anticipated efficacy of the intervention. In Chapter 2, I will review basic concepts in Bayesian inference as well as different aspects that must be considered when designing a Bayesian adaptive trial. In Chapter 3, I will propose a checklist for the key steps necessary to plan the simulations of an adaptive trial. I will also demonstrate the application of this checklist in the context of a motivating example of a 3-arm randomized controlled trial of influenza vaccines. In

Chapter 4, I will discuss the findings.

Chapter 2

Literature Review

2.1 Bayesian Inference

Bayesian inference is a paradigm for statistical inference. It differs from the conventional frequentist approach in the sense that the inclusion of information about the unknown parameters, which is external to the research study at hand, can be done formally using prior distributions. In this section we briefly introduce the basic concepts and notation used in the Bayesian framework.

Probability Notation

Let a be an event, and H be the context where a might occur. Let $p(a|H)$ denote a conditional probability density of event a conditioning on the context H. Let $p(a)$ denote the marginal density of event a, i.e., the density function obtained by aggregating $p(a|H)$ across all values of the context H.

2.1.1 Bayes' Theorem

Let $p(y,\theta)$ denote the joint probability distribution function for the observed data vector y and an unknown parameter vector θ. The joint distribution function can be written as the product of a marginal distribution function and a conditional distribution function:

$$p(y,\theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y). \qquad (\textit{Eq. 2.1.1.1})$$

Bayes' theorem states that:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \qquad (\textit{Eq. 2.1.1.2})$$

where $p(\theta|y)$ is referred to as the posterior distribution, $p(\theta)$ is referred to as the prior distribution, $p(y|\theta)$ is referred to as the likelihood or the sampling distribution and $p(y) = \Sigma_\theta p(y|\theta)p(\theta)$ $(or \int_\theta p(y|\theta)p(\theta)$ if θ is continuous), the overall probability of y across all possible values of θ. The term $p(y)$ is a constant which does not depend on θ. It is typically very difficult to obtain $p(y)$ in an analytical form. Therefore the unnormalized posterior density is given by

$$p(\theta|y) \propto p(y|\theta)p(\theta) \qquad (\textit{Eq. 2.1.1.3})$$

This expression is often exploited by computational methods as it allows us to make inferences about θ while knowing only the likelihood and prior distribution.

2.1.2 Basic Components in the Bayesian Framework

Thus, there are three components involved in any Bayesian analysis no matter how simple or complex. The first component that needs to be specified in Bayesian analysis is the likelihood, which relates the data from the current trial (internal information) to the unknown parameters of interest. The second is the prior distribution that allows formal inclusion of external information about the unknown parameters and is required to be specified explicitly. The last component is the posterior distribution, which arises from updating the prior distribution with the information in the likelihood using Bayes theorem. The posterior distribution is the final result of the Bayesian analysis summarizing the probability distributions of the parameters of interest. From this distribution, we can obtain point and interval estimates and graphical summaries of the parameters of interest. The following sub-sections give more details on each of these key components in a Bayesian analysis.

2.1.2.1 Likelihood

The likelihood tells us how likely different values of the unknown parameters are given the sample data. It is expressed as a function of the unknown parameters and has the same functional form as the joint probability density function of the observed data $p(y|\theta)$. Suppose that we have n observations $y_1,...y_n$, then the likelihood, $p(y|\theta)$, is given by

$$p(y|\theta) = p(y_1,...,y_n|\theta), \qquad (\textit{Eq. 2.1.2.1.4})$$

where $\theta = (\theta_1, ...\theta_m) \in \mathbb{R}^m$ denotes the model parameters. If we further assume that those observations are independent of each other, the likelihood becomes

$$p(y|\theta) = p(y_1, ..., y_n|\theta) = \Pi_{i=1}^{n} p(y_i|\theta). \qquad (\textit{Eq. 2.1.2.1.5})$$

The likelihood contains the information in the observed data. If two probability models $p(y|\theta)$ have the same likelihood functions, then they would lead to the same inference on θ.

Example: Likelihood for Binomial Random Variable

Let Y denote the binomial random variable, that is the number of successes in n independent Bernoulli trials, where each individual trial has a dichtomous outcome, such as 1 or 0, yes or no, success or failure, with a probability of p or $1-p$, respectively. Let y denote the observed number of successes. The likelihood of p given y, and n is given by

$$p(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad (\textit{Eq. 2.1.2.1.6})$$

2.1.2.2 Prior Distribution

The prior distribution represents our knowledge about the unknown parameter, θ, that is external to the observed data. Priors can be informative or non-informative and in some cases hybrid priors are also used. A non-informative prior provides no specific information about a parameter. Informative prior distributions can be obtained by expressing previous study results or subjective opinions in the form of a probability distribution. Non-informative priors are often used to maintain the objectivity of the analysis. They may be preferred to informative priors due to the unavailability or unreliability of the external information, or simply because the researchers want the statistical analyses to be conducted solely based on the data from the current research study.

The literature on elicitation of prior information is well-developed. Several methods such as matching means and standard deviations (Lee, 2012) or matching percentiles of probability intervals (Press, 1989) to a pre-specified family of distributions have been proposed to convert the elicited prior information into a prior distribution.

Example: Non-informative and Informative Beta Priors for Binomial Probability

A $Beta[a,b]$ is suitable for the Binomial probability θ in the previous example as it spans (0,1). A commonly used non-informative prior can be specified with $a = b = 1$. The $a = b = 1$ prior is considered non-informative because it allows equal weight for all values of θ. In contrast, a informative prior can be specified with, for instance, $a = 1, b = 9$, and the mean of such a prior is $\frac{a}{a+b} = 0.1$. The greater the value of $a+b$, the smaller variability in the prior distribution.

2.1.2.3 Posterior Distribution

Since the posterior distribution $p(\theta|y)$ is proportional to the product of the likelihood and the prior distribution, it is a compromise between these two pieces of information about θ. It follows that the conclusions drawn based on the posterior are often more reliable than those drawn based on the observed data only as long as the prior information is valid.

2.1.2.4 Estimation of Bayesian Conjugate Models

The likelihood and prior distributions are called conjugate and the Bayesian model is called a conjugate model if the posterior distribution and prior distribution are from the same probability distribution family. It is computationally beneficial to conduct conjugate analysis in the sense that it is possible to derive a posterior distribution analytically and the posterior distribution has a closed-form expression.

2.1.2.4.1 Example: Beta-Binomial Conjugate model

This example, illustrates a Bayesian conjugate model. Let $y_1,...,y_n$ denotes the results of n trials each of which only takes values either 0 or 1, i.e., a sequence of Bernoulli trials. Then it follows that:

$$p(y|\theta) \propto \theta^y(1-\theta)^{n-y},$$

where $y = \sum_{i=1}^{n} y_i$ denotes the number of successes out of n trials. Suppose we have a Beta prior distribution $Beta(a,b)$ for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, 0 \leq \theta \leq 1, a, b > 0.$$

Applying Bayes theorem to combine the likelihood and the prior distribution yields:

$$\begin{aligned}\text{Posterior} &\propto \text{Likelihood} \times \text{Prior}\\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^y(1-\theta)^{n-y}\\ &\propto \theta^{a+y-1}(1-\theta)^{b-1+n-y}\\ &\propto Beta[a+y,\ b+n-y],\end{aligned}$$

i.e., a kernel of the $Beta[a+y,\ b+n-y]$ distribution, which belongs to the same family of distributions as the prior.

2.1.3 Posterior Predictive Distribution

Another fundamental goal of statistical modeling is to make predictive inferences on unobserved but observable quantities. We can distinguish between making predictions based on the prior distribution or based on the posterior distribution. An example of a problem where we need to make predictions based on the prior distribution is when planning a study and no data have been observed yet. Predictions based on the posterior distribution would be relevant after completing the data analysis.

Let y denote the observed data and $\tilde{y}$ denote some future observations. The distribution of $\tilde{y}$ conditioning on y is given by:

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)f(\theta)d\theta, \qquad (\textit{Eq. 2.1.3.7})$$

where $f(\theta)$ is the distribution of unknown parameter θ. If we further assume that $\tilde{y}$ and y are independent conditional on θ, the equation becomes:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)f(\theta)d\theta. \qquad (\textit{Eq. 2.1.3.8})$$

If we replace $f(\theta)$ in $Eq.$ 2.1.3.8 with $p(\theta)$, a prior distribution for θ, then $p(\tilde{y}|y)$ is called the prior predictive distribution. In contrast, when we replace $f(\theta)$ by $p(\theta|y)$, the posterior distribution for θ, we have the posterior predictive distribution $p(\tilde{y}|y)$. It is posterior since it depends on y, and it is predictive since it gives the probability distribution of unknown $\tilde{y}$.

2.1.4 Hybrid Prior Distribution

The posterior distribution serves to update our prior knowledge about the unknown parameter θ with the observed data. Thus it can reflect variability or agreement between previous studies and the current study. In the event the current study is in agreement with previous observations, then the posterior distribution results in more precise knowledge about θ. On the other hand, disagreement between the prior distribution and the observed data would result in greater uncertainty in the posterior distribution than in the prior distribution. Excluding relevant prior information simply on account of a conflict between the observed data and the prior distribution will lead to a potentially biased posterior distribution that underestimates the uncertanty.

When planning a new study, we are faced with a situation where we do not know if the future data will be in agreement with the prior distribution or not. If, as explained in the previous section, we rely on the prior predictive distribution to generate observable data, then we are more likely to create situations where the prior and data are in agreement. This may lead to an artificially precise posterior distribution and underestimation of the sample size necessary. Early in the study of a technology, available prior information may be based on small studies and potentially unreliable. Therefore, it has been proposed to use a hybrid prior distribution, recognizing that our prior knowledge is based in part on some observed data but also in part on the limitations of this observed data. A robustified prior distribution achieves a compromise between what we know and what we do not by using a mixture of informative and non-informative prior distributions.

2.1.4.1 Robust Priors in Clinical Trials with Historical Control Information

Bayesian methods are perfectly suited in combining and accumulating information from current trial data with relevant historical information. In particular, many adaptive clinical trials use information about the placebo control group to reduce the number of patients allocated in the control group (Berry, 2006). However, the quality of the available external information may not be ideal for various reasons. In this case the robustified mixture prior can help mitigate concerns about the over dependence on prior information in a context where prior-data conflict cannot be ruled out due to the novelty of the intervention under study and offer appropriate ways to combine the prior information under a meta-analytic framework (Schmidli 2014).

The mixture prior (or robust prior) has two components. The first component is a meta-analytic-predictive (MAP) prior (Spiegelhalter et al., 2004) that incorporates and expresses the information among previous trials. The second component is a non-informative prior. The MAP prior now expresses aggregated information about previous trials, which is already better than a simple conjugate informative prior based on a single trial in that it accounts for information from multiple trials and takes into account the heterogeneity between them. In addition, after being diluted with a non-informative prior, the impact of including such informative priors is further downplayed. The ratio of mixing should be specified based on the level of confidence we have in the credibility of the historical trials we wish to include.

2.1.4.2 Example: Construct Robust Prior Using Beta Conjugate Prior

Let $Beta[\alpha,\beta]$ denote the conjugate prior distribution for an unknown parameter elicited from the historical trial. Then the robust prior for the parameter is given by

$$\hat{p}_{robust} = (1-w_r) * Beta(\alpha,\beta) + w_r * Beta[1,1] \qquad (\textit{Eq. 2.1.4.2.9})$$

where the weight w_r can be chosen based on the credibility of the simple conjugate prior.

If multiple prior distributions were included, then a robust MAP prior can be approximated as:

$$\hat{p}_{robust\ MAP} = (1-w_r)\sum_{k=1}^{K} w_k * Beta[\alpha_k,\beta_k] + w_r * Beta[1,1], \quad (\textit{Eq. 2.1.4.2.10})$$

where the hyperparameters w_k, α_k, and β_k can be obtained as maximum likelihood estimates (Schmidli et al., 2014).

2.2 Clinical Trials

In this section, I will introduce some concepts involved in evaluating interventions within the context of clinical trials.

2.2.1 Types and Phases of Clinical Trials

Clinical trials are used to study many different types of health interventions including medications, vaccines and medical devices. Clinical trials can be classified into four phases corresponding to four different stages covering different objectives related to the intervention under study, such as proving safety, superiority, noninferiority, etc. The four main phases of clinical trials will be discussed in the following sub-sections.

2.2.1.1 Phase I Studies

In phase I of a drug development study, clinical trials are usually designed to estimate the tolerability of the drug and identify the pharmacokinetics and pharmacodynamics (Friedman et al., 1998). The maximally tolerated dose is often evaluated during this stage. To determine the maximally tolerated dose, the researchers often start with a low dose and escalate the dose until a certain level of toxicity is reached. The starting point is usually determined by extrapolating the model fitted using animal data, by assuming the same relationship holds in humans (Le Tourneau et al., 2009). In general, the number of participants enrolled in phase I is usually small, and the maximal tolerable dose is often determined if one or several unacceptable toxicities are observed from these participants.

2.2.1.2 Phase II Studies

After evaluating the tolerable range of dosage, the next objective in evaluating the intervention in the phase II study is to demonstrate the efficacy of the health intervention with an increased number of participants. The number of additional participants depends on the particular settings and the desired level of precision. The evaluation can be conducted by comparing each intervention arm or comparing concurrent intervention arms to historical data. A pre-post comparison can also be made to demonstrate the treatment effect. Multiple dosages may be studied in different intervention arms if the optimal dosage is not yet available. The credibility of the phase II study depends on that of the phase I study, and it affects the quality of the following phase III study.

2.2.1.3 Phase III and Phase IV Studies

Once conclusions obtained from the previous phase II study appear to be promising, the phase III studies commence, and the main goal is to refine and validate the evaluation of safety and the efficacy conclusions from previous phases on a larger group of participants than before. The investigators often focus on the efficacy of the intervention(s). The superiority trial design, equivalence trial design and noninferiority trial design are often used to capture and quantify the differences between treatments using efficacy measures described in Section 2.2.2.3 later in this chapter. In addition, phase III trials often have a follow-up period for future evaluations related to the safety or efficacy, ranging from months up to a lifetime, depending on the nature and purpose of the intervention. For example, in one study, long-term follow-up surveillance of patients who received preoperative chemoradiotherapy resulted in a median follow-up of 134 months. The original study showed an improved local control rate, but the follow-up study did not show any survival benefits (Sauer et al., 2012). Therefore, follow-up surveillance is often necessary after successfully conducting the phase III study. Still, uncertainty may remain about the balance between the benefit and side effects until the phase IV study is conducted on a larger population.

2.2.2 Efficacy criteria evaluated in clinical trials

Randomized controlled trials can be categorized into three types: the superiority trial, the equivalence trial, and the noninferiority trial based on the hypothesis that is being tested. The parameters involved in the hypothesis depend on the health outcome being studied. When the health outcome is dichotomous, the hypothesis is often framed in terms of parameters like the risk difference or risk ratio which we will use for illustration in the sections below.

2.2.2.1 Risk Difference and Risk Ratio

The risk difference (RD) is defined as the difference in proportion of subjects with the outcome between the two groups being compared. For example, suppose that we are interested in the efficacy of a new vaccine. Subjects in the treatment group are assigned the new vaccine, and patients in the control group are assigned a placebo. The outcome of interest could be the number of infections prevented. The difference between the two groups in the percentage of subjects with the outcome is given by:

$$RD = R_A - R_B, \qquad (\textit{Eq. 2.2.2.1.11})$$

where R_A and R_B are the risk of infection in the treatment and control groups, respectively.

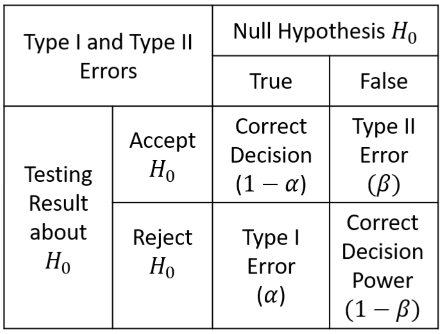
Similarly, the risk ratio (RR) is defined using R_A and R_B as follows:

$$RR = \frac{R_A}{R_B} \qquad (\textit{Eq. 2.2.2.1.12})$$

2.2.2.2 Probability of Errors in Decision Making

Analyses of randomized controlled trials are expected to lead to decisions regarding the intervention under study. At the time of the planning of the trial, it is of interest to determine the probability of an incorrect decision as a function of design parameters such as the sample size and decision criteria. This helps in planning the trial as it makes optimal use of the available resources. A poorly planned trial could be a waste of money if it has an unacceptably high risk of resulting in a wrong decision. In order to determine the probability of a wrong decision, we typically work with important operating characteristics such as type I error rate and type II error rate in the context of a frequentist inferential framework. In the Bayesian framework, though we do not use hypothesis testing, it is still possible to define these operating characteristics and use them to plan the trial.

The frequentist hypothesis testing setup requires us to specify the null hypothesis, the hypothesis that we will test, and its complement, the alternative hypothesis. When the goal of the research study is to compare two groups, the null hypothesis typically assumes that there is no difference between the two groups. A Type I error occurs if one rejects the null hypothesis H_0 when it is true. The probability of making a Type I error, i.e., P(Type I Error), is called the level of significance of the test, denoted by α. A Type II error occurs if one does not reject the null hypothesis when the alternative hypothesis H_a is true. The probability of making a Type II error, i.e., P(Type II Error), is denoted as β, where $1-\beta$ is often referred to as the power of the test.



Type I and Type II Errors		Null Hypothesis H_0	
		True	False
Testing Result about H_0	Accept H_0	Correct Decision $(1-\alpha)$	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decision Power $(1-\beta)$

Figure 2.2.2.2.1: Table of Type I and Type II Errors.

To illustrate, consider the specification of hypothesis test for two groups A and

B in terms of risks R_A and R_B as follows:

$$H_0 : R_A = R_B$$
$$H_a : R_A \neq R_B$$

Then the probability of making a Type I error and a Type II error is given by

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is True}) = P(\text{Reject } H_0 | R_A = R_B)$$
$$\beta = P(\text{Do not reject } H_0 | H_a \text{ is True}) = P(\text{Do not reject } H_0 | R_A \neq R_B).$$

The power of a test, denoted as $1 - \beta$, is the probability of concluding the alternative hypothesis while it is true. The power depends on the significance level of the test as well as the particular value of the unknown parameter being tested under the alternative hypothesis. The larger the absolute value of δ is, the higher power the test would have. Ideally, we want to minimize the probability of having both types of error rates simultaneously. However, the trade-off between the type I error and type II error should also be considered: the lower the significance level of the test, the lower probability that the test correctly rejects the null hypothesis when the alternative is true.

2.2.2.3 Superiority Trials

A superiority trial is designed to capture the superiority of one treatment over others. Consider once again the example where we are interested in the efficacy of a new vaccine. If one expects a positive difference in response rates between the new vaccine and placebo, then the criterion to conclude efficacy of the new vaccine can be formulated as (Committee for Proprietary Medicinal Products, 2008)

- The two-sided 95% confidence interval for the difference in proportions between the means lies entirely above 0.

- The two means are statistically significantly different at the α = 5% level (p-value< 0.05),

under the frequentist framework for statistical inference.

Within the context of a Bayesian analysis, these statements above can be converted into a one-sided probabilistic statement in terms of the posterior probability or predictive posterior probability. In this thesis, I will stick with the posterior probability as follows:

$$P(R_A - R_B > 0|\text{ data}) \geq 97.5\%, \qquad (\textit{Eq. 2.2.2.3.13})$$

where R_A and R_B are the proportions of outcomes in the group receiving treatment A and the group receiving treatment B, respectively. This statement can be generalized into

$$P(R_A - R_B > \Delta|\text{ data}) \geq 97.5\%, \qquad (\textit{Eq. 2.2.2.3.14})$$

where Δ is the clinically meaningful difference between R_A and R_B.

In practice, we have many choices for specifying the probabilistic statement defining superiority in terms of the cut-off value above which we have a clinically meaningful difference, and even the confidence level. For example, one can use the risk ratio as the statistic instead of the risk difference, and 2 as the value for the margin of difference at the 90% confidence level. The new statement is now

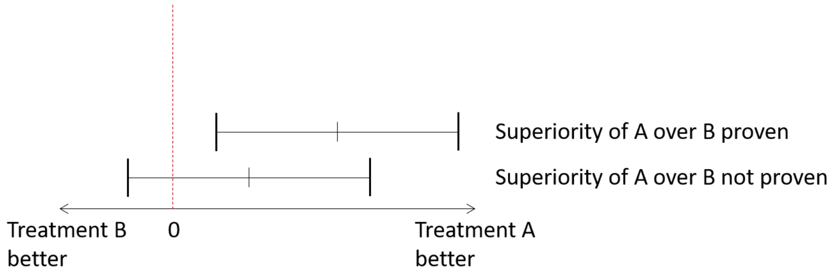

Superiority of A over B proven
Superiority of A over B not proven
Treatment B better
0
Treatment A better

Figure 2.2.2.3.1: Possible Scenarios in Superiority Trials.

given by:

$$P(\frac{R_A}{R_B} > 2|\text{ data}) \geq 95\% \qquad (\textit{Eq. 2.2.2.3.15})$$

2.2.2.4 Equivalence Trial

An equivalence trial is designed to prove that the treatments are clinically indifferent. Unlike the superiority trial, the equivalence trial requires two margins of clinical equivalence. That is, the one-sided probabilistic statement of the chosen statistics can be written as:

$$P(-\Delta_1 < R_A - R_B < \Delta_2|\text{ data}) \geq 97.5\%, \qquad (\textit{Eq. 2.2.2.4.16})$$

where the margins $-\Delta_1$ and Δ_2 can also be chosen symmetrically, and the risk difference is chosen to represent the difference between treatment A and treatment B.

2.2.2.5 Noninferiority Trial

A noninferiority trial is designed to confirm that one of the interventions is not inferior to the others. For example, we consider a new treatment A is non-inferior to an existing treatment B when treatment A is more effective than or as effective

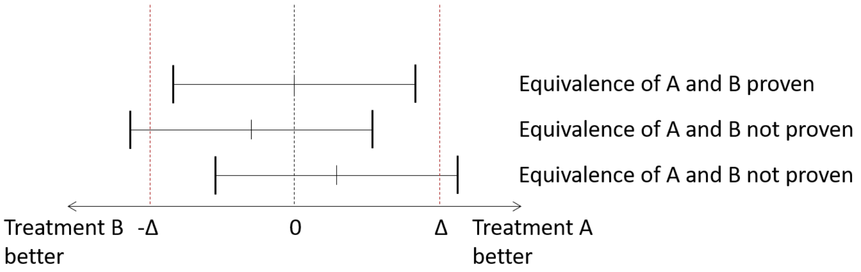

Equivalence of A and B proven
Equivalence of A and B not proven
Equivalence of A and B not proven
Treatment B better
-Δ
0
Δ
Treatment A better

Figure 2.2.2.4.1: Possible Scenarios in Equivalence Trials.

as treatment B. Suppose that we use risk difference as the statistic for decision making, $-\Delta$ as the value for the acceptable margin of difference and a 97.5% as the desired probability of achieving the criterion, the statement above can also be formulated mathematically using a one-sided probabilistic statement as follows:

$$P(R_A - R_B > -\Delta|\text{ data}) \geq 97.5\% \qquad (\textit{Eq. 2.2.2.5.17})$$

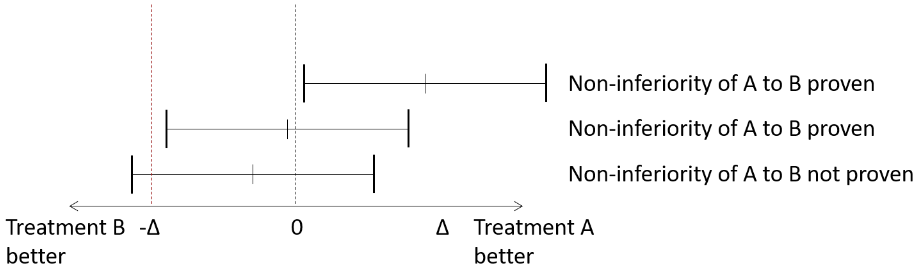

Non-inferiority of A to B proven
Non-inferiority of A to B proven
Non-inferiority of A to B not proven
Treatment B better
-Δ
0
Δ
Treatment A better

Figure 2.2.2.5.1: Possible Scenarios in Noninferiority Trials.

2.2.2.6 Computational Issue and Monte Carlo Methods

When making Bayesian inference, it is usually difficult to obtain the marginal density of y, $p(y)$, when one goes beyond simple problems of the type described above involving conjugate distributions. Moreover, when specifying a Bayesian model, we can include as many parameters as demanded, but usually, only a few of them are of primary interest, and inferences are made on the marginal posterior densities for these parameters. Again, it is challenging to obtain the marginal posterior densities analytically. This is the case even when the number of unknown parameters is small, e.g., the problem of estimating the posterior distribution of the difference between two proportions is mathematically challenging. To address these challenges, different simulation-based methods have been proposed. A detailed discussion of these methods can be found in reference texts (Thisted, 2000; Brooks, 2011). These methods are used to obtain a random sample from the joint posterior distribution of the parameters of interest. This sample may be used to estimate statistics such as the posterior mean and quantiles for each parameter of interest. The availability of fast computers has made it easy to draw a sample of a large size and obtain the desired precision in these estimates.

For the purposes of this thesis, we only require the method of Monte Carlo simulation as illustrated in the following example. Let us consider the problem of comparing the risk of an outcome in two arms A and B of a randomized controlled trial using risk difference. In each arm, we observe a Binomial variable, i.e., y_A successes out of n_A subjects and y_B successes out of n_B subjects. We can use a conjugate prior or robust prior for R_A and R_B to obtain their posterior distributions as described in Example 2.1.2.4.1, respectively. In order to obtain the posterior of the risk difference, $R_A - R_B$, we can simply use the sample obtained from the posterior distributions of the risk in each group separately and then take the difference. Probabilistic metrics like $P(R_A - R_B > 0)$ can also be calculated by drawing multiple samples from the posterior distribution.

2.2.3 Adaptive Design in Clinical Trials

Adaptive designs of clinical trials allow for the study design to be modified during the course of the trial based on evidence examined at interim analyses. Possible adaptations include changing the randomization ratio to favour a more promising intervention and halting an intervention arm that is shown to be less efficacious. In traditional clinical trial designing, the sample size is calculated to achieve a certain power level based on the prespecified treatment effect δ and type I error rate α. An adaptive design can reduce the overall sample size by stopping recruiting patients into the arm that shows a strong tendency towards inferiority and reallocating them into the groups where the treatment appears to be more beneficial (Paul et al., 2006; Berry, 2011). This technique can also be combined with using a preliminary randomization ratio calculated from reliable prior information such that more patients will benefit from being assigned with treatments that appear to be more efficacious and safer.

2.2.3.1 The Fully Bayesian Approach and Bayes as a Frequentist Tool

There are two common strategies for conducting Bayesian analyses in medical intervention development, one is a fully Bayesian approach which is often used in the context of decision making, and the other is a mixed approach that employs Bayes rule as a tool under the frequentist framework (Berry, 2010). The fully Bayesian analysis uses the likelihood function and the prior distribution together with a utility function. The utility function is equivalent to a loss function that measures the value of a certain outcome, e.g. in terms of the cost, and it is very popular in many fields in the decision making context (Berry, 2010; Körding and Wolpert, 2006a; Grünwald and Dawid, 2004; Yuille and Bülthoff, 1996; Körding and Wolpert, 2006b). However, there are some limitations that often arise in practice when trying to find a suitable utility function in the context of clinical trials. For example, in the context of adaptive trials, determining the cost of each interim decision may be time consuming as it may depend on several variables that are difficult to quantify. Therefore, the fully Bayesian approach does not work out well in clinical trials.

Still, we can use Bayesian inference as a tool under the frequentist framework to plan and design clinical trials (Berry, 2010). As the modern clinical trial become more and more complicated, the expected sample size calculated using a frequentist approach is often too conservative. We can use Bayesian inference to design a trial that is capable of achieving the same goal with a smaller sample size. Although it is often hard to obtain the operating characteristics analytically, we can still get good estimates using simulation methods and use them as the measures for evaluating the feasibility of a trial.

2.2.3.2 Simulations in Bayesian Adaptive Trials

Simulations are thus vital and will be used to evaluate the influence that a clinical trial's design can have on its operating characteristics. An operating characteristic is a function of several trial design parameters including the rule for early stopping, the cutoff values used for evaluating the superiority and noninferiority, the sample size and some other trial design parameters, such as the trial duration and number of interim analyses. These configurations have a significant impact on the potential and the ability of the trial to capture the hypothetical differences between treatments. A large number of trials can be simulated from the prior predictive distribution of the model parameters. The behaviour of operating characteristic functions across these trials is observed. In particular, the probability that the function meets pre-determined criteria for making a decision (e.g. to stop and conclude the intervention is efficacious) are examined to determine the Type I error (i.e. the probability of wrongly concluding the intervention is beneficial even though it is not). By looking at results simulated using general settings, researchers can see the big picture and empirically anticipate how a trial design may unfold and adjust the trial design accordingly.

2.2.3.3 Sample Size in Bayesian Adaptive Trials

The sample size of a Bayesian Adaptive Trial cannot be pre-determined with certainty as the trial design involves the possibility that the trial may stop or be modified at multiple points. One can use the frequentist sample size calculation for a fixed design trial as a starting point. It may be attractive to try and make a clear decision with a smaller sample size than that of a fixed design trial. However, it may be possible that the sample size required for a Bayesian adaptive trial is higher as it takes into account information as it accrues and the assumptions made at the start of the trial may not hold as it proceeds.

2.2.3.4 Methods for Trial Simulations

A few pedagogical articles have provided guidance for conducting adaptive trials and designing the corresponding simulation studies. In the articles by Thorlund et al. (2018) and Satlin et al. (2016), the authors elaborated the various steps that need to be considered by clinician researchers who want to conduct adaptive trials. Hansen et al. (2018) published an article that focused more specifically on simulations. Their interest was in providing guidance for programming using WinBUGS within the SAS interface to implement the pre-trial simulation study. In the article by Pallmann et al. (2018), the authors discussed the principles and potential obstacles researchers are likely to be confronted with in conducting simulation studies for adaptive trials, with more emphasis placed on Frequentist approaches. In the article by Thorlund et al. (2019), the authors developed an open-source application that allow clinical trial investigators to conduct simulation studies to explore different scenarios under a numbers of different settings.

None of the articles published so far provides an overview of the different steps researchers must follow to design their own simulations. In the next chapter of this thesis, I will propose a checklist that addresses this gap.

Chapter 3

A Checklist for Designing Simulations for Adaptive Trials: Description and Application to a Trial of Influenza Vaccines

The planning of every adaptive randomized controlled trial raises unique challenges. However, certain common elements are involved in designing simulations for all trials. In this chapter, I will propose a checklist to help statisticians design simulations and illustrate its use by application to the planning of a vaccine trial. In Section 3.1, I will enumerate the elements of the checklist. In Section 3.2, I will use an influenza vaccine trial as an example to illustrate how the checklist can be implemented. I will discuss the results from simulations of the vaccine trial under different settings in Section 3.4.

3.1 Checklist for Designing Simulations to Evaluate the Feasilibity of an Adaptive Trial Design

Box 3.1 below presents a checklist containing some generic elements that need to be considered and clearly defined when we want to evaluate the feasibility of a trial design via simulations. In the following sub-sections I provide more explanation on each of these elements.

Box 3.1 - A Checklist for Planning Simulations to Evaluate the Feasibility of a Bayesian Adaptive Trial Design

1. Identify the interventions and outcomes of interest.
2. Define the criteria to be evaluated to answer the objectives of the trial.
3. Specify the number of interim analyses and the decision rules to be used in each interim analysis and the final analysis.
4. Encode possible outcomes evaluated at each interim analysis and the final analysis.
5. Determine the prior distributions for each unknown parameter.
6. Determine the sample size under a frequentist analysis as well as the minimum and maximum sample size.
7. Define the simulation settings.
8. Specify the desired type I and type II errors for each outcome studied.

3.1.1 Identify the Interventions and Outcomes of Interest

A clinical trial can have multiple interventions and outcomes. To conduct a simulation study, we first need to identify the number of interventions to be compared, which determines the number of arms in the study. The outcomes of interest should also be identified. More than one outcome, e.g. efficacy and safety, may be evaluated at each interim analysis and the final analysis. The criteria for decision making, defined later on, depend on the outcomes of interest identified at this stage.

3.1.2 Define the Criteria to be Evaluated to Answer the Objectives of the Trial

When we want to evaluate whether a certain intervention results in a benefit, the criteria used to demonstrate a benefit need to be defined. For example, in order to evaluate the safety of a particular intervention, we may define a criterion in terms of the risk of adverse events being "comparable" to another intervention which has been demonstrated as safe. The word "comparable" can be expressed by defining constraints on the risk difference, such as $R_A - R_B < 0.02$, where R_A denotes the risk of adverse events associated with the intervention of interest (A) and R_B denotes the risk of adverse events associated with the intervention B which has been demonstrated as being safe. This criterion can be further extended into a probabilistic statement, and the confidence level can be chosen as desired, e.g., $P(R_A - R_B < 0.02) > 97.5\%$.

3.1.3 Specify the Number of Interim Analyses and the Decision Rules to be Used in Each Interim Analysis and the Final Analysis

An interim analysis is an analysis of current data from an ongoing trial before the trial is completed. Each interim analysis can result in a decision to adapt the trial via actions such as modifying allocation ratio, dropping treatment arms early or stopping the trial early. In order to carry out simulations, we need to determine the number of interim analyses and the possible decisions we could have at each interim analysis. For example, in a Bayesian adaptive multi-arm trial, we are allowed to stop any of the arms early if we find any predefined signal that suggests safety issues in any of the interim analyses.

Decisions at each interim analysis can rely on the use of the posterior distribution or the posterior predictive distribution of the relevant parameters. The posterior distribution is often used for confirmatory decision making when sufficient data has accrued. However, when it comes to stopping early for futility, the posterior predictive distribution which provides information on whether the trial will succeed eventually may be more relevant (Berry, 2010). The predictive distribution is often used for expected futility when there are several planned interim analyses and the trial will stop if the predictive probability of success at the final analysis is sufficiently low.

3.1.4 Encode Possible Outcomes Evaluated at Each Interim Analysis and the Final Analysis

As some clinical trial designs can become fairly complicated with numerous adaptations, it is necessary to identify a way to summarize the results we could obtain at each decision point to get a bird's eye view. One possible solution could be identifying and encoding the possible outcome that a trial could end up having. A flow diagram illustrating the different possible interim analyses and decisions leading to different possible endpoints for the study can be very useful. The frequency of each outcome of a trial can be obtained to estimate the probability that the trial ends up having certain results.

3.1.5 Determine the Prior Distributions for Each Unknown Parameter

Incorporating prior information from earlier studies is a key advantage of the Bayesian approach. Accordingly, the choice of prior distribution greatly impacts the design of a Bayesian adaptive trial. In order to arrive at reliable inferences, prior distributions with high credibility need to be elicited from historical trials that can be updated with the information from the trial that is being planned. Prior information elicited objectively from previous trials that had the same objectives in the same population is often sensible. The parameters for a simple prior distribution can be obtained by matching means and standard deviations. A MAP prior can be used to incorporate prior information from multiple trials and further diluted with a non-informative prior to form a robust MAP prior.

3.1.6 Determine the Sample Size under a Frequentist Analysis as well as the Minimum and Maximum Sample Size

In planning the Bayesian adaptive trial, it is helpful to commence with the frequentist sample size necessary to achieve these error rates as an estimate of the maximum sample size. However, the final result obtained from the simulation may not necessarily be smaller than the sample size suggested by the frequentist approach depending on the trial designs and the historical information we use in the Bayesian model, especially when we have a prior-data conflict. A minimum and maximum value of the sample size to be considered in the simulations should be specified so the variation in the Type I and Type II errors across these sample sizes can be examined. The maximum sample size may be determined by feasibility constraints.

3.1.7 Specify the Desired Type I and Type II Errors for Each Outcome Studied

Regulatory agencies expect that randomized controlled trials are designed such that the risk of type I and type II errors are controlled at predefined, acceptable levels. The level may differ for efficacy and safety outcomes and also by the nature of the intervention and the medical condition being addressed. For example, for safety outcomes, interest may focus on the type II error of not detecting an important safety violation when it exists. For efficacy outcomes, type I error is typically more important as it is preferable to avoid the erroneous conclusion that an intervention is beneficial when in fact it is not.

3.1.8 Define the Simulation Settings

We need to specify the number of simulated trials that will be drawn and the number of posterior samples that will be drawn in each trial. We can also specify other relevant statistics that can be calculated from the simulations such as the expected sample size or the expected duration of the trial. In addition, it is necessary to specify the assumed values of the unknown parameters. The Type I error is estimated under the null hypothesis, i.e. the assumption that there is no difference between the two groups being compared. The Type II error is measured under the alternative hypothesis that there is a difference between the two groups. The assumed value of this difference needs to be specified, and a range of values should be considered which cover the minimum clinically meaningful difference.

3.2 Description of the DEFINE Vaccine Trial

3.2.1 Background

In this section I will describe the planning of the DEFINE (a**D**juvantEd in**F**luenza vacc**I**ne i**N** rh**E**umatoid arthritis) randomized controlled trial, which serves as a motivating example for the methods developed in this thesis. DEFINE is being planned as a 3-arm randomized controlled trial of influenza vaccines for Rheumatoid Arthritis (RA) patients. In RA, IIV3-SD (SD) and IIV3-HD (HD) are two types of inactivated vaccines recommended for protection against seasonal influenza. An earlier study has shown that patients who received the high dose (HD) were 2-3 times more likely to seroconvert than those who received the standard dose (SD) (Colmegna et al., 2020). However, the HD is not widely used due to its cost. A cost-effective alternative to the HD could be the adjuvanted IIV3 (ADJ) which has been shown as having a better efficacy than the SD in the elderly, while being relatively more affordable than the HD. Still, its safety and efficacy are unclear in RA patients. The DEFINE trial will focus on evaluating the safety and the efficacy of the ADJ compared with the SD and the HD in RA patients.

When designing the trial, in lieu of conducting a traditional fixed vaccine trial, conducting a Bayesian adaptive trial can not only make use of historical trial data to potentially reduce the sample size required to draw conclusions, but also allow subjects a greater chance to be randomized to a more promising treatment by permitting stopping of a less efficacious treatment at the interim analysis.

3.2.2 Objective of the Trial

The objectives of this experiment are

- To evaluate the safety of the ADJ arm compared to the HD and SD arms.
- To evaluate the superiority of the ADJ arm to the SD arm after successfully meeting the safety criterion.
- To evaluate the noninferiority of the adjuvanted dose arm to the HD arm after successfully meeting the safety criterion.

3.2.3 Recruitment and Schedule of Interim Analyses

The study will take place during the fall influenza season over a 2-year period. The recruitment of patients will commence on Oct 1st, and end on Jan 1st for both study years (see Figure. 3.2.3.1).

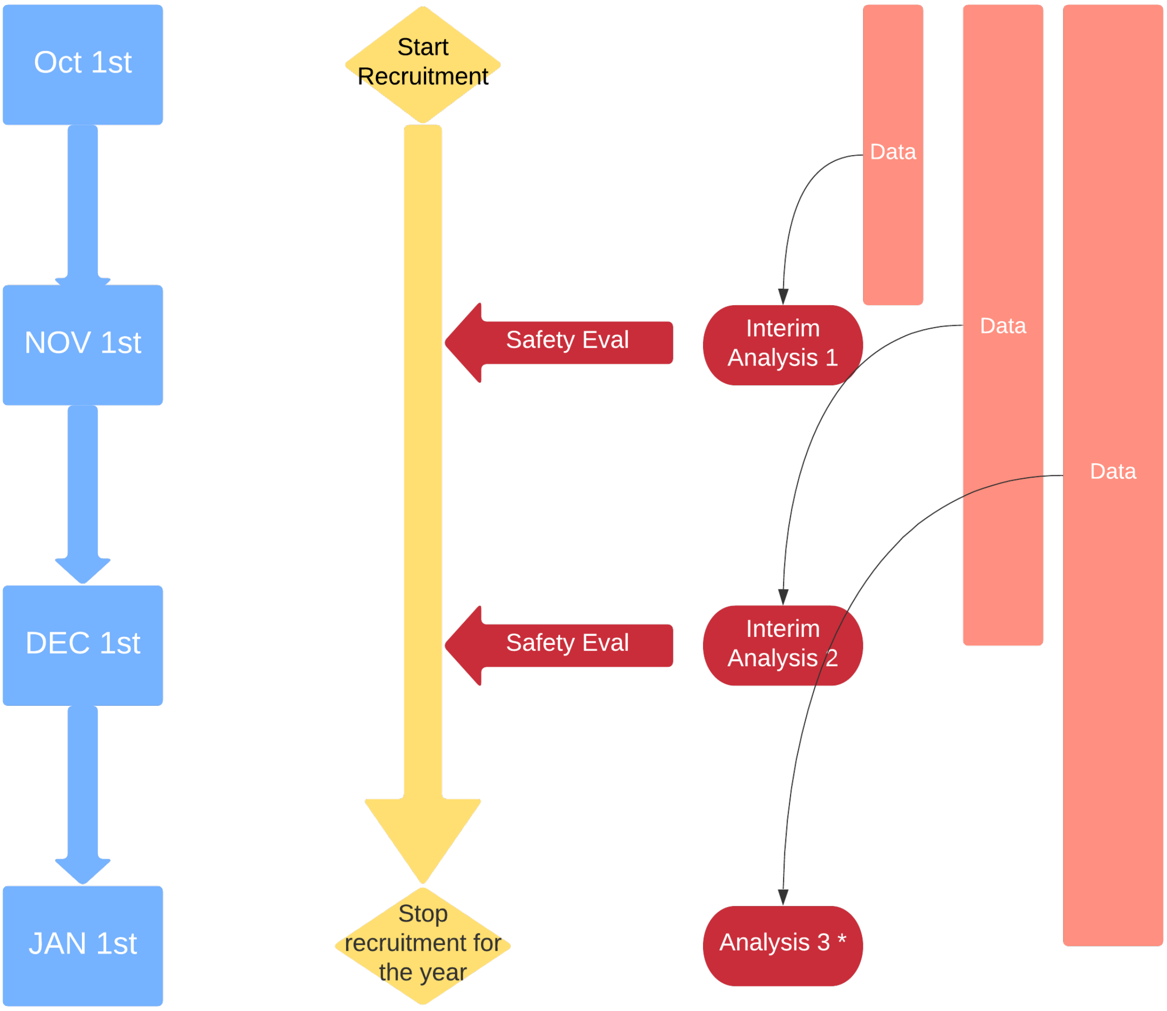


Figure 3.2.3.1: Illustration of subject recruitment and data analysis plan for the DEFINE trial in each year.
*In year 1, Analysis 3 is an interim analysis, while in year 2 it is the final analysis.

3.3 Application of the Checklist for Designing Simulations

3.3.1 Identify the Interventions and Outcomes of Interest

3.3.1.1 Identify the Interventions of Interest

In the DEFINE trial, there are three types of influenza vaccines and therefore three treatment arms involved: the SD arm, the ADJ arm, and the HD arm.

3.3.1.2 Identify the outcomes of interest

Safety

The primary endpoint for safety is the risk of significant flares, i.e., worsened RA symptoms.

Immunogenicity

The primary endpoint for efficacy can be defined in terms of the seroconversion rate (SCR) per vaccine strain, defined as the percentage of RA patients who seroconverted in terms of the geometric mean titre (GMT) (Center for Biologics Evaluation and Research, 2007). For simplicity, in this thesis, I will assume only one strain is used at a time and only use SCR as the primary endpoint.

3.3.2 Define the Criteria to be Evaluated to Answer the Objectives of the Trial

At each interim analysis a decision must be made whether the trial can proceed. This decision will be based on a Bayesian analysis of the data gathered at that point. The criteria used in these interim analyses for demonstrating success, futility and inconclusive results in safety and efficacy based on Bayesian posterior probabilities will be elaborated in this section. Given only one interim analysis of efficacy was planned, it was decided not to employ the posterior predictive distribution for determining futility. Each criterion corresponds to one of the objectives in Section 3.2.2. A summary of all criteria, including the timing for each evaluation, can be found in Table. 3.3.3.1.

3.3.2.1 Criteria for Evaluating Safety of ADJ vs HD or SD

Criteria for Continuation of the Trial:

The safety of the ADJ arm compared to the SD arm (or HD arm) can be concluded if the upper bound of the one-sided 97.5% posterior credible interval for the risk difference of significant flares of ADJ arm vs. SD arm (or HD arm) is less than or equal to 0.2. These criteria can be written as

$$P(R_{ADJ,\ flare} - R_{SD,\ flare} \geq 0.2) < 2.5\%, \qquad (\textit{Eq. 3.3.2.1.1})$$

and

$$P(R_{ADJ,\ flare} - R_{HD,\ flare} \geq 0.2) < 2.5\%, \qquad (\textit{Eq. 3.3.2.1.2})$$

where $R_{SD,\ flare}$, $R_{HD,\ flare}$ and $R_{ADJ,\ flare}$ are risk of flares in the SD arm, HD arm and the ADJ arm, respectively.

Criterion for Stopping the Trial Early:

The ADJ arm will be considered as less safe than the SD arm (HD arm) if the lower bound of the one-sided 50% predictive credible interval at the final analysis for the risk difference of significant flares of ADJ arm over SD arm (HD arm) is greater than or equal to 0.2. These criteria can be written as

$$P(R_{ADJ,\ flare} - R_{SD,\ flare} \geq 0.2) > 50\%, \qquad (\textit{Eq. 3.3.2.1.3})$$

and

$$P(R_{ADJ,\ flare} - R_{HD,\ flare} \geq 0.2) > 50\%, \qquad (\textit{Eq. 3.3.2.1.4})$$

The ADJ arm will be dropped from the trial if any of these criteria above is met.

Criterion for Concluding Safety:

The safety of the ADJ arm will be concluded only if the criteria 3.3.2.1.1 and 3.3.2.1.2 are satisfied when the complete dataset is available at the end of the trial.

Criterion for Inconclusive Results:

If neither of the criteria for demonstrating success and futility is satisfied at the final analyses, the safety of the ADJ arm compared to the SD arm or the HD arm cannot be concluded even at the end of the trial.

3.3.2.2 Criteria for Evaluating Efficacy

Criteria for Evaluating superiority of ADJ vs SD

Criteria for Concluding Success:

The superiority of the ADJ arm over the SD arm can be concluded if the lower bound of the one-sided 97.5% posterior credible interval for the difference in SCR of ADJ arm compared to SD arm is greater than or equal to 0. This is equivalent to

$$P(R_{ADJ,\, SCR} - R_{SD,\, SCR} \geq 0) > 97.5\%, \qquad (\textit{Eq. 3.3.2.2.5})$$

where $R_{ADJ,\, SCR}$ and $R_{SD,\, SCR}$ are risks in seroconversion in the ADJ arm and the SD arm, respectively.

Criterion for Concluding Futility:

The ADJ arm will be considered as not superior over the SD arm if the upper bound of the one-sided 50% predictive credible interval for the difference in SCR of ADJ arm and SD arm is less than or equal to 0. This criterion can be written as

$$P(R_{ADJ,\, SCR} - R_{SD,\, SCR} \leq 0) > 50\%, \qquad (\textit{Eq. 3.3.2.2.6})$$

Criterion for Inconclusive Results:

If neither of these criteria for demonstrating success or futility is satisfied, the superiority of the ADJ arm over the SD arm or the HD arm cannot be concluded even at the end of the trial.

Criteria for Evaluating Noninferiority of ADJ vs HD

Criteria for Concluding Success:

The noninferiority of the ADJ arm to the HD arm can be concluded if the lower bound of the one-sided 97.5% posterior credible interval for the difference in SCR of the ADJ arm compared to the SD arm (HD arm) is greater than or equal to -10%. This is equivalent to

$$P(R_{ADJ,\,SCR} - R_{HD,\,SCR} \geq -0.1) > 97.5\%. \qquad (\textit{Eq. 3.3.2.2.7})$$

Criterion for Concluding Futility:

The ADJ arm will be considered as being noninferior to the HD arm if the upper bound of the one-sided 50% predictive credible interval for the difference in SCR of ADJ arm and the HD arm is less than or equal to -10%. This criterion can be written as

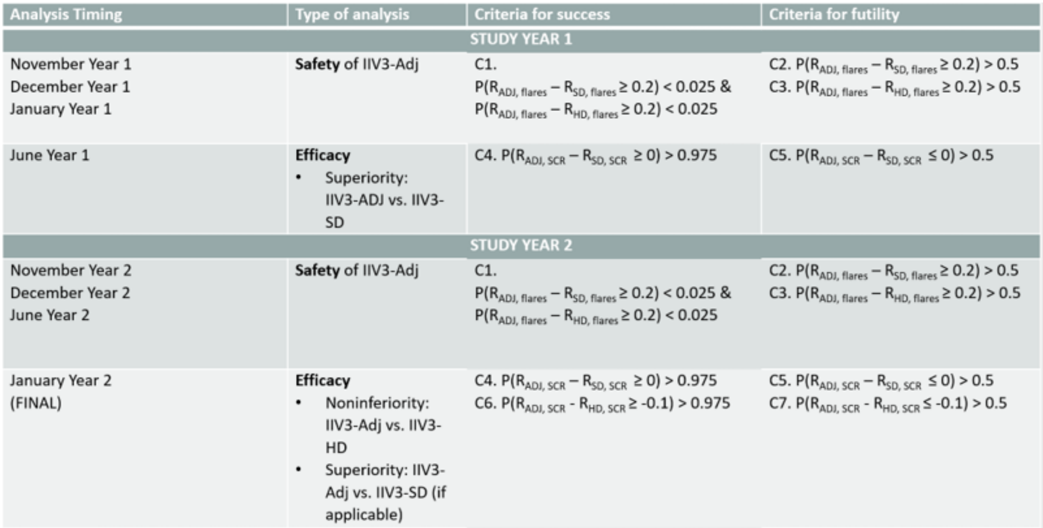
$$P(R_{ADJ,\,SCR} - R_{HD,\,SCR} \leq -0.1) > 50\% \qquad (\textit{Eq. 3.3.2.2.8})$$

Criterion for Inconclusive Results:

If neither of these criteria for demonstrating success or futility is satisfied, the noninferiority of ADJ arm to the HD arm cannot be concluded.

3.3.3 Specify the Number of Interim Analyses and the Decision Rules to be Used in Each Interim Analysis and the Final Analysis

Table 3.3.3.1: Schedule of IAs and Criterion to Be Evaluated.



Analysis Timing	Type of analysis	Criteria for success	Criteria for futility
STUDY YEAR 1			
November Year 1 December Year 1 January Year 1	**Safety** of IIV3-Adj	C1. $P(R_{ADJ,\ flares} - R_{SD,\ flares} \geq 0.2) < 0.025$ & $P(R_{ADJ,\ flares} - R_{HD,\ flares} \geq 0.2) < 0.025$	C2. $P(R_{ADJ,\ flares} - R_{SD,\ flares} \geq 0.2) > 0.5$ C3. $P(R_{ADJ,\ flares} - R_{HD,\ flares} \geq 0.2) > 0.5$
June Year 1	**Efficacy** • Superiority: IIV3-ADJ vs. IIV3-SD	C4. $P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \geq 0) > 0.975$	C5. $P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \leq 0) > 0.5$
STUDY YEAR 2			
November Year 2 December Year 2 June Year 2	**Safety** of IIV3-Adj	C1. $P(R_{ADJ,\ flares} - R_{SD,\ flares} \geq 0.2) < 0.025$ & $P(R_{ADJ,\ flares} - R_{HD,\ flares} \geq 0.2) < 0.025$	C2. $P(R_{ADJ,\ flares} - R_{SD,\ flares} \geq 0.2) > 0.5$ C3. $P(R_{ADJ,\ flares} - R_{HD,\ flares} \geq 0.2) > 0.5$
January Year 2 (FINAL)	**Efficacy** • Noninferiority: IIV3-Adj vs. IIV3-HD • Superiority: IIV3-Adj vs. IIV3-SD (if applicable)	C4. $P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \geq 0) > 0.975$ C6. $P(R_{ADJ,\ SCR} - R_{HD,\ SCR} \geq -0.1) > 0.975$	C5. $P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \leq 0) > 0.5$ C7. $P(R_{ADJ,\ SCR} - R_{HD,\ SCR} \leq -0.1) > 0.5$

C1: the safety criteria for continuation of the trial based on posterior probabilities. The success in safety is demonstrated if C1 is satisfied in the final analysis.
C2&C3: the criteria for demonstrating futility in the safety of the ADJ arm based on predictive probabilities at the final analysis.
C4: the criterion for demonstrating success in superiority of the ADJ arm to the SD arm based on posterior probabilities.
C5: the criterion for demonstrating futility in the superiority of the ADJ arm to the SD arm based on predictive probabilities at the final analysis.
C6: the criterion for demonstrating success in the noninferiority based on posterior probabilities.
C7: the criterion for demonstrating futility in the noninferiority based on predictive probabilities at the final analysis

At the end of each month of the trial, the researcher conducts a safety assessment for the ADJ arm that determines whether we should drop the ADJ arm and terminate the trial due to serious safety reasons (e.g., severe flares due to receiving the ADJ vaccine).

In the event that the superiority of ADJ over SD is proven, at the end of the first study year, the trial design allows a potential early dropout of the SD arm so that more patients can benefit from being reallocated to the other two arms.

At the end of the second study year, the trial design allows a second chance for demonstrating the superiority of the ADJ arm over the SD arm if the SD arm is not dropped at the end of Year 1. At the end of Year 2, the noninferiority of the ADJ arm to the HD arm will be evaluated in addition.

3.3.4 Encode Possible Outcomes Evaluated at Each Interim Analysis and the Final Analysis

Since demonstrating that the ADJ arm is not unsafe is a pre-requisite for the continuation of the trial, the safety criteria are always evaluated prior to the efficacy (i.e., superiority and noninferiority) criteria. The structure introduced by the ordering of evaluations diversifies the possible outcomes. In order to plan the simulations it is helpful to list all possible combinations of outcomes that may occur at each interim analysis.

Consider a 5-D outcome vector defined as follows:

$$V_{outcome} = \begin{pmatrix} \text{Safety}_{\text{Y1}} \\ \text{Superiority}_{\text{Y1}} \\ \text{Safety}_{\text{Y2}} \\ \text{Superiority}_{\text{Y2}} \\ \text{Noninferiority} \end{pmatrix} \qquad (\textit{Eq. 3.3.4.9})$$

where each element of the vector represents different results from the evaluation in chronological order and the subscription denotes the study year of the evaluation. Each element of the vector can take 4 different values: 0, 1, 2 and 9 that corresponds to “Futility”, “Success”, “Inconclusive Result”, and “Not Evaluated”. For example,

$$V_{outcome} = (1,2,1,1,0)^T = \begin{pmatrix} \text{Safety}_{\text{Y1}} = \text{Success} \\ \text{Superiority}_{\text{Y1}} = \text{Inconclusive} \\ \text{Safety}_{\text{Y2}} = \text{Success} \\ \text{Superiority}_{\text{Y2}} = \text{Success} \\ \text{Noninferiority} = \text{Futility.} \end{pmatrix}$$

The decision rules and the possible adaptations for each interim analysis are elaborated in Table 3.1. A table that lists all possible scenarios allowed by the trial in a compact manner can be found in Table 3.3.4.1. Another table that elaborates all possible results with more details can be found in Table 3.3.4.2. Some outcomes are impossible because of the trial design. For example, consider the first cell of Table 3.3.4.2, which is struck out, where all of the three arms are dropped. In particular, if the HD arm is dropped, then it implies that the trial has reached the noninferiority evaluation near the end of the trial, and C6 must be satisfied. In this case, the trial is ended, and there is no chance to drop the ADJ arm.

The flow diagram in Figure. 3.3.4.3 demonstrates the order of the key decisions made at the interim and final analyses for safety and immunogenicity. The different possible outcomes that a trial can end up with depending on the adaptation are also demonstrated in this graph. The probability of each outcome in each scenario will be calculated.

Table 3.1: Decision Rules.

Analysis	Encoding	Decision Rule (Possible Adaptation)
Safety Analysis 1-5	1 or 0	Continue or Not Safe (Stop ADJ)
Safety Analysis 6	1 or 0 or 2	Safe or Not Safe or Inconclusive
Efficacy analysis 1 (ADJ vs SD)	1 or 0 or 2 or 9	Efficacious (Stop SD) or Not Efficacious (Stop ADJ) or Inconclusive (Continue) or Not Evaluated
Efficacy analysis 2 (ADJ vs SD)	1 or 0 or 2 or 9	Efficacious or Not Efficacious or Inconclusive or Not Evaluated
Efficacy analysis 3 (ADJ vs HD)	1 or 0 or 2 or 9	Efficacious or Not Efficacious or Inconclusive or Not Evaluated

Table 3.3.4.1: Possible Outcomes (Compact Version).

Success 1
Futility 0
Inconclusive 2
Not evaluated 9

Possible outcomes:
11191, 12111, 11192, 12112,
11190, 12110, 11099, 12121,
12099, 09999, 10999, 12120,
12100, 12101, 12102, 12122

Final Status of the Three Arms		SD Dropped			
		Yes		No	
		HD Dropped		HD Dropped	
		Yes	No	Yes	No
ADJ Dropped	Yes		11099 11190 12110		12099 09999 10999 12120 12100 12101 12102
	No	11191 12111	11192 12112	12121	12122

Each cell represents the possible outcome $V_{outcome}$ corresponding to different V_{Status}.
Encoding rules: 0- Futility; 1- Success/Continuation of the trial; 2- Inconclusive; 9- Not Evaluated.
Outcomes represented by the cell that is strikethrough can not appear due to the trial design.

Table 3.3.4.2: Possible outcomes encoded.

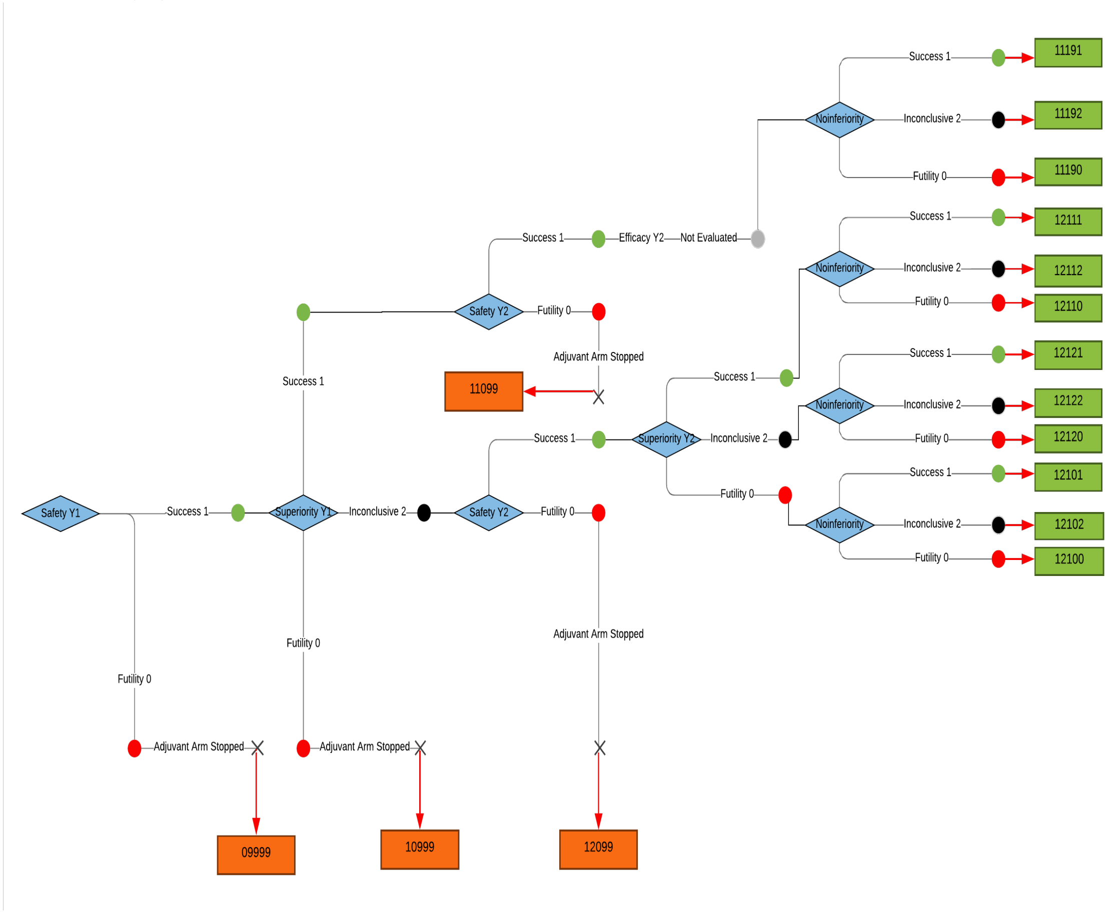
Final Status of the Three Arms		SD Dropped			
		Yes		No	
		HD Dropped		HD Dropped	
		Yes	No	Yes	No
ADJ Dropped	Yes		Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 11099		Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12099
					Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 09999
			Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 11190		Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 10999
					Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12100
			Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12110		Safety Y1, Efficacy, Safety Y2, Superiority Y2, Noninferiority 12101
					Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12102
					Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12120
	No	Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 11191	Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 11192	Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12121	Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12122
		Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12111	Safety Y1, Superiority Y1, Safety Y2, Superiority Y2, Noninferiority 12112		

Possible outcomes:
11191, 12111, 11192, 12112, 11190, 12110, 11099, 12121, 12099, 09999, 10999, 12120, 12100, 12101, 12102, 12122

Success 1
Futility 0
Inconclusive 2
Not evaluated 9

Each cell represents the possible outcome $V_{outcome}$ corresponding to different V_{Status}.
Encoding rules: 0- Futility; 1- Success/Continuation of the Trial; 2- Inconclusive; 9- Not Evaluated

Figure 3.3.4.3: Flow diagram illustrating possible adaptations and outcomes in DEFINE.



Each blue diamond represents a decision point for safety and immunogenicity at each interim analysis.
The orange rectangles indicate different situations where the ADJ arm is dropped.
The green rectangles indicate the possible combination of outcomes emerging when the study is completed in Y2.

3.3.5 Determine the Prior Distribution for Each Unknown Parameter

For the DEFINE trial, only one historical trial is available providing information on safety and efficacy in the HD and SD groups. The prior information from the earlier study will be used in the analysis and a mixture prior allows us to do so in a conservative manner that does not give that single small study too much importance.

3.3.5.1 Derive the Beta Prior

The parameters a and b of a beta prior $Beta[a,b]$ can be obtained by matching its mean and standard deviation (sd) with those from the previous study. For example, for the DEFINE trial, the normal approxiamted 95% confidence interval for the risk of seroconversion in the SD arm from the previous study is [4.6%, 14.9%]. Then, by matching [mean-1.96*sd, mean + 1.96*sd], we get

$$mean = 0.0975,\ sd = 0.0268.$$

Solving the equations

$$mean = \frac{a}{a+b},\ sd = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}},$$

with respect to a and b, we get $a = 12.3$ and $b = 114.1$. It is worth mentioning that there is no unique solution and alternatively, the prior parameter values can also be obtained by matching any two quantiles of the Beta distribution. This calculation can be done using the function $beta.parms.from.quantiles()$ implemented in R

(Belisle, 2017). The information from the previous study which we used for the risk of flare and SCR, and the corresponding Beta prior distributions obtained via quantile matching are listed in Table 3.2.

Table 3.2: Beta Prior Distributions for Flare and Seroconversion Rate.

Arms	SD*	HD*	ADJ**
95% confidence interval of risk of significant flares	[0, 5%]	[0, 5%]	NA
Prior distribution over risk of significant flares	$Beta[2,\ 98]$	$Beta[2,\ 98]$	$Beta[1,\ 1]$
95% confidence interval of risk of seroconversion	[4.6%, 14.9%]	[15.8%,30.3%]	NA
Prior distribution over risk of seroconversion	$Beta[10.6, 106.2]$	$Beta[28.6, 97.7]$	$Beta[1, 1]$

* Informative priors from historical trials used for the SD arm and the HD arm are obtained via quantile matching.
** Non-informative Beta priors are used for the ADJ arm.

3.3.5.2 Constructing a Robust Mixture Prior

The mixing proportion to construct a robust prior distributions is specified to be 0.5 to give equal weight to the non-informative prior and the informative prior based on the single previous study. The formula for constructing robustified mixture prior distribution is given by

$$P(\theta) = 0.5 * \underbrace{Beta(\alpha,\beta)}_{\text{Informative}} + 0.5 * \underbrace{Beta(1,1)}_{\text{Noninformative}} . \qquad (Eq.\ 3.3.5.2.10)$$

The robustified mixuture prior distributions can be constructed using the formula from 3.3.5.2.10. For example, the mixture prior distribution for the risk of seroconversion in the SD arm is given by

$$P(R_{SD,\ SCR}) = 0.5 * Beta[12.3, 114.1] + 0.5 * Beta[1, 1]$$

3.3.6 Determine the Sample Size under a Frequentist Analysis as well as the Minimum and Maximum Sample Size

Suppose that the risk of seroconversion in each arm is given by $R_{SD,\ SCR}$, $R_{ADJ,\ SCR}$ and $R_{HD,\ SCR}$, respectively, along with the corresponding observed proportion of seroconversion $p_{SD,\ SCR}$, $p_{ADJ,\ SCR}$ and $p_{HD,\ SCR}$. Then the typical frequentist estimate of sample size required to achieve a significance level of $(1-\alpha)\%$ and a power of $(1-\beta)\%$ for a superiority trial of the ADJ arm is given by (Blackwelder, 1982)

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 * [R_{ADJ,\ SCR} * (1 - R_{ADJ,\ SCR}) + R_{SD,\ SCR} * (1 - R_{SD,\ SCR})]}{(R_{ADJ,\ SCR} - R_{SD,\ SCR})^2}, \qquad (Eq.\ 3.3.6.11)$$

Concerning the noninferiority trial of the ADJ arm with an allowed difference of $\Delta = 0.1$, the frequentist estimated sample size is given by

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 * [R_{ADJ,\ SCR} * (1 - R_{ADJ,\ SCR}) + R_{SD,\ SCR} * (1 - R_{SD,\ SCR})]}{(R_{ADJ,\ SCR} - R_{SD,\ SCR} + 0.1)^2}, \qquad (Eq.\ 3.3.6.12)$$

As the population risk cannot be observed, a guess value for the observed proportion of seroconversion can be plugged in to obtain the estimated sample size. Using the formula 3.3.6.11 and 3.3.6.12, we can now calculate the frequentist sample size under different settings. Under the first setting of the efficacy evaluation, we aim to arrive at a positive conclusion about the superiority of the ADJ arm and a negative conclusion about the noninferiority of the ADJ arm with $\alpha = 0.05$

and $\beta = 0.05$. The frequentist calculation results in a sample size of 585 for each group. Under the second setting, we aim to arrive at a positive conclusion in both the superiority and the noninferiority of the ADJ. Using the frequentist sample size formula, the sample size required for the superiority trial obtained is 207 and that for the noninferiority trial obtained is 561 for each group. Therefore, the frequentist calculation results in a sample size of 561 for each group.

The minimum and maximum sample sizes can be set according to feasibility constraints. The DEFINE study was planned as a multi-centre study with a maximum of 100 subjects available per centre per year. Therefore we defined the minimum sample size as that which would be available in the 1st year if only one centre participated, i.e. 100. For the maximum sample size, we selected 1000 as this was the highest number of subjects permissible under budget constraints.

3.3.7 Specify the Desired Type I and Type II Errors for Each Outcome Studied

In this section, I will define the null and alternative hypotheses corresponding to the safety and immunogenicity outcomes and the definitions for the Type I and Type II errors corresponding to the decision criteria presented in the previous section.

3.3.7.1 Type I and Type II Errors Corresponding to Decision Criteria for Safety of the ADJ Arm

Formulate Hypothesis

Recall the criteria for evaluating the safety of the ADJ arm given by Equation

3.3.2.1.1 and 3.3.2.1.2. The corresponding hypothesis can be defined as

$$H_0 : \text{the ADJ arm is at least as safe as the SD arm and the HD arm,}$$
$$H_a : \text{the ADJ arm is less safe than the SD arm or the HD arm.}$$

It follows that the probability of making type I error rate and type II error rate can be defined as

$$\begin{aligned} P(\text{Type I Error}) &= \alpha = P(\text{C2 holds} \vee \text{C3 holds} | H_0 \text{ is true}), \\ P(\text{Type II Error}) &= \beta = P(\text{C1 holds} | H_a \text{ is true}), \end{aligned}$$

respectively, where C1 (criterion #1 from 3.3.2.1.1 and 3.3.2.1.2) is given by

$$\begin{aligned} \text{C1} : &P(R_{ADJ,\, flare} - R_{SD,\, flare} \geq 0.2 | \text{data}) < 2.5\% \\ &\& \; P(R_{ADJ,\, flare} - R_{HD,\, flare} \geq 0.2 | \text{data}) < 2.5\%, \end{aligned}$$

and C2 (criterion #2 from E.q. 3.3.2.1.3) and C3 (criterion #3 from E.q. 3.3.2.1.4) are given by

$$\begin{aligned} \text{C2} :&P(R_{ADJ,\, flare} - R_{SD,\, flare} \geq 0.2) > 50\%, \\ \text{C3} :&P(R_{ADJ,\, flare} - R_{HD,\, flare} \geq 0.2) > 50\%. \end{aligned}$$

To calculate the type I and type II error rates for evaluating safety, the null hypothesis and alternative hypothesis settings for simulation need to be specified.

Null Hypothesis Setting for Safety Evaluation

Under the null hypothesis, the risk of flare in the SD arm and HD arm are assumed to be 0.1, and the ADJ arm is assumed to be equally safe. Concerning the efficacy, the ADJ arm is expected to have intermediate efficacy between the SD and HD arms. However, in order to prevent the ADJ from being dropped early due to the efficacy evaluations, the ADJ arm is set to be as efficacious as the HD arm. The full setting can be found in Table.3.3

Alternative Hypothesis Setting for Safety Evaluation

Under the alternative hypothesis setting for safety evaluation, the ADJ arm is set to be less safe than the SD arm and the HD arm. In particular, a value of 0.3 is chosen for the risk of flare in the ADJ arm based on the safety evaluation criterion. The full setting can be found in Table.3.4

Table 3.3: Settings for Safety Evaluation under Null Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.1	0.22**
HD Arm	0.1*	0.22

* Under the null hypothesis setting, the ADJ is as safe as the other two arms. This value may be replaced by any value ≤ 0.025.
** The ADJ arm is assumed to be as efficacious as the HD arm to avoid being dropped due to efficacy reasons.

Table 3.4: Settings for Safety Evaluation under Alternative Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.3 *	0.22 **
HD Arm	0.1	0.22

* Under the alternative hypothesis setting, the ADJ is not as safe as the other two arms. The value of 0.3 may be replaced by any value ≥ 0.3.
** The ADJ arm is assumed to be as efficacious as the HD arm to avoid being dropped due to efficacy reasons. The rest of the setting for efficacy does not matter when the focus is safety.

3.3.7.2 Type I and Type II Errors Corresponding to Decision Criteria for Efficacy of the ADJ Arm

3.3.7.2.1 Superiority

Formulate Hypothesis

The criteria for demonstrating superiority of the ADJ arm given by Equation 3.3.2.2.5 can be formulated in a hypothesis testing context. The corresponding hypothesis can be defined as

$$\begin{aligned} &H_0 \text{ :the ADJ arm is at most as efficacious as the SD arm,} \\ &H_a \text{ :the ADJ arm is more efficacious to the SD arm} \end{aligned}$$

It follows that the probability of making corresponding type I and type II errors can be defined as

$$\begin{aligned} P(\text{Type I Error}) = \alpha = P(\text{C5 holds}|H_0 \text{ is true}), \\ P(\text{Type II Error}) = \beta = P(\text{C4 holds}|H_a \text{ is true}), \end{aligned}$$

respectively, where C4 (criterion #4 from E.q. 3.3.2.2.5) and C5 (criterion #5 from E.q. 3.3.2.2.6) are given by

$$\begin{aligned} \text{C4} : P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \geq 0|\text{data}) > 97.5 \\ \text{C5} : P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \leq 0|\text{data}) > 50\% \end{aligned}$$

Null Hypothesis Setting for Superiority Evaluation

Under the null hypothesis setting for the superiority trial, the ADJ arm is assumed to be as efficacious as the SD arm. In particular, a value of 0.088 is chosen for SCR in the ADJ arm. In addition, to prevent the ADJ arm from being dropped early due to the safety evaluations, the ADJ arm is set to be as safe as the other two arms. The SD and HD settings are taken from the results for the Michigan strain in the previous trial. The full setting can be found in Table.3.5.

Table 3.5: Settings for Superiority Evaluation under Null Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.1*	0.088**
HD Arm	0.1	0.22

* Under the null hypothesis setting for the superiority evaluation, the ADJ is assumed to be as safe as the other two arms to avoid being dropped due to the safety evaluations.
** The ADJ arm is assumed to be just as efficacious as the SD arm.

Alternative Hypothesis Setting for Superiority Evaluation

Under the alternative hypothesis setting for the superiority trial, the ADJ arm is assumed to be more efficacious than the SD arm. In particular, a value of 0.15 is chosen for SCR in the ADJ arm. In addition, to prevent the ADJ arm from being dropped early due to the safety evaluations, the ADJ arm is set to be as safe as the other two arms. The full setting can be found in Table.3.6.

Table 3.6: Settings for Superiority Evaluation under Alternative Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.1*	0.15**
HD Arm	0.1	0.22

* Under the alternative hypothesis setting for the superiority evaluation, the ADJ is assumed to be as safe as the other two arms to avoid being dropped due to the safety evaluations.
** The ADJ arm is assumed to be more efficacious than the SD arm.

3.3.7.2.2 Noninferiority

Formulate Hypothesis

Similarly, the criteria for demonstrating noninferiority of the ADJ arm given by Equation 3.3.2.2.7 can also be formulated in a hypothesis testing context. The corresponding hypothesis can be defined as

$$\begin{aligned}&H_0 : \text{the ADJ arm is less efficacious than the HD arm,}\\&H_a : \text{the ADJ arm is at least as efficacious as the HD arm}\end{aligned}$$

It follows that the probability of making the corresponding type I and type II errors can be defined as

$$\begin{aligned}P(\text{Type I Error}) = \alpha = P(\text{C7 holds}|H_0 \text{ is true}),\\P(\text{Type II Error}) = \beta = P(\text{C6 holds}|H_a \text{ is true}),\end{aligned}$$

respectively, where C6 (criterion #6 from E.q. 3.3.2.2.7) and C7 (criterion #7 from E.q. 3.3.2.2.8) are given by

$$\begin{aligned}&\text{C6 :}P(R_{ADJ,\,SCR} - R_{HD,\,SCR} \geq -0.1|\text{data}) > 97.5\%,\\&\text{C7 :}P(R_{ADJ,\,SCR} - R_{HD,\,SCR} \leq -0.1|\text{data}) > 50\%.\end{aligned}$$

Null Hypothesis Setting for Noninferiority Evaluation

Under the null hypothesis setting for the noninferiority trial, the ADJ arm is assumed to be inferior to the HD arm. In order to avoid the possible impact of early dropout of the ADJ arm due to the superiority evaluation with the SD arm, the value of $R_{ADJ,\ SCR}$ should lie between 0.088 and 0.12. In particular, a value of 0.12 can be specified for $R_{ADJ,\ SCR}$. The full setting can be found in Table.3.7.

Alternative Hypothesis Setting for Noninferiority Evaluation

Under the alternative hypothesis setting for the noninferiority trial, the ADJ arm is assumed to be more efficacious than the SD arm. In particular, a value of 0.20 is specified for $R_{ADJ,\ SCR}$. The full setting can be found in Table.3.8.

Table 3.7: Settings for Noninferiority Evaluation under Null Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.1*	0.12
HD Arm	0.1	0.22

* Under the null hypothesis setting for the noninferiority evaluation, the ADJ is assumed to be as safe as the other two arms to avoid being dropped due to the safety evaluations.
** The ADJ arm is assumed to be much less efficacious than the HD arm.

Table 3.8: Settings for Noinferiority Evaluation under Alternative Hypothesis.

Arms	Risk of Flare	SCR
SD Arm	0.1	0.088
ADJ Arm	0.1*	0.20**
HD Arm	0.1	0.22

* Under the alternative hypothesis setting for the noninferiority evaluation, the ADJ is assumed to be as safe as the other two arms to avoid being dropped due to the safety evaluations.
** The ADJ arm is assumed to be noninferior to the HD arm.

3.3.8 Define the Simulation Settings

The number of simulated trials is specified as N_s=1000, where posterior samples for six parameters (safety endpoint and immunogenicity endpoint for each treatment arm) will be monitored in each trial, and N_p=10000 posterior samples will be drawn in each trial to estimate the probability of each decision criterion. Finally, the Type I and Type II errors are estimated as the probability that the decision criterion meets the pre-determined cut-off across the 1000 simulated trials. For example, concerning the superiority evaluation of the ADJ arm, the corresponding type I and type II error rates are given by

- Probability of making a type I Error:

  $$P(\text{Type I Error}) = \frac{\text{\# of posterior samples such that } C_4 \text{ is met}}{N_s},$$

  given the settings that the ADJ arm is equal to the SD arm, i.e., $R_{ADJ,\ SCR} = R_{SD,\ SCR}$.

- Probability of making a type II Error:

  $$P(\text{Type II Error}) = \frac{\text{\# of posterior samples such that } C_4 \text{ is not met}}{N_s},$$

  given the settings that the ADJ arm is superior to the SD arm. See Section.3.3.8.1.2 for more details.

The criterion C4

$$P(R_{ADJ,\ SCR} - R_{SD,\ SCR} \geq 0) > 0.975$$

is estimated and evaluated as

$$\frac{\text{\# of } (R_{ADJ,\ SCR} - R_{SD,\ SCR}) \geq 0}{N_p} > 0.975$$

3.3.8.1 Specify the Assumed Values of the Unknown Parameters

3.3.8.1.1 Settings for Safety Evaluations

To study the type II error rates in the safety evaluations, the assumed value for $R_{ADJ,\ flare}$ and $R_{ADJ,\ SCR}$ need to be specified. According to Section.3.3.7, the following setting will be used:

- $R_{SD,\ flare} = 0.1$, $R_{ADJ,\ flare} = 0.3$, $R_{HD,\ flare} = 0.1$
- $R_{SD,\ SCR} = 0.088$, $R_{ADJ,\ SCR} = 0.22$, $R_{HD,\ SCR} = 0.22$.

The assumed value for $R_{ADJ,\ SCR}$ is set to be 0.22 to prevent the ADJ arm from being dropped due to the efficacy evaluations. and the efficacy evaluations.

3.3.8.1.2 Settings for Efficacy Evaluations

To study the type I error rates in the efficacy evaluation (especially for the superiority evaluations), the ADJ arm is assumed to be not efficacious than the SD arm, while being comparably safe to the SD arm and the HD arm. In particular, $R_{ADJ,\ SCR}$ will be set equall to $R_{SD,\ SCR}$ i.e.,

- $R_{SD,\ flare} = 0.1$, $R_{ADJ,\ flare} = 0.1$, $R_{HD,\ flare} = 0.1$
- $R_{SD,\ SCR} = 0.088$, $R_{ADJ,\ SCR} = 0.088$, $R_{HD,\ SCR} = 0.22$.

To study the type II error rates in the efficacy evaluations, the ADJ arm is assumed to be comparably safe and the assumed value for $R_{ADJ,\ flare}$ will be shared across the superiority and noninferiority evaluation. i.e., $R_{SD,\ flare} = 0.1, R_{ADJ,\ flare} = 0.1, R_{HD,\ flare} = 0.1$. Still, different values for $R_{ADJ,\ SCR}$ need to be specified for each evaluation, respectively. Since the value of $R_{SCR,ADJ}$ is expected to lie in between $R_{SCR,SD}$ and $R_{SCR,HD}$, we will specify two potential values, 0.15 and 0.20, for $R_{SCR,ADJ}$ under the alternative hypothesis for both the superiority and noninferiority evaluation.

1. $R_{SCR,SD} = 0.088$, $R_{SCR,ADJ} = 0.15$, $R_{SCR,HD} = 0.22$;

2. $R_{SCR,SD} = 0.088$, $R_{SCR,ADJ} = 0.20$, $R_{SCR,HD} = 0.22$.

Again, the SD and HD settings are taken from the results for the Michigan strain in the previous trial. Under the first setting, ADJ is in the middle between SD and HD. Under the second setting, ADJ is closer to HD.

3.4 Simulation and Results

In this section, simulations carried out under various settings and their corresponding results will be discussed. All the simulations are programmed, conducted, and analyzed using R software (R Core Team, 2019). The main goals of the simulations are to illustrate the impact of sample size on evaluating safety and efficacy of the ADJ arm under different scenarios. Settings from Section.3.3.8.1 are used for the simulation study.

3.4.1 Results of the Simulations for the Safety Outcomes

The probability of making a type I error and that of making a type II error is our main interest. However, the importance of the type I error rate and type II error rate are not always the same. For example, a failure in not rejecting that the ADJ arm is safe while it is unsafe (i.e., making a type II error) is much more disastrous than a failure in rejecting that the ADJ arm is safe while it is safe (i.e., making a type I error). Therefore, in this thesis, I will mainly focus on evaluating the impact of the sample size on the type I error rate in concluding safety, and the type I and type II error rates in concluding the superiority and the noninferiority of the ADJ arm.

The safety outcomes from across all settings with various sample sizes ranged from 100 to 1000 are presented. The relevant statistics, and the probability of making a type II error in demonstrating the safety of the ADJ arm are plotted in Figure 3.4.1.1 and Figure 3.4.1.2. This figure shows that the trial design can capture the signal that suggests the ADJ is unsafe at all sample sizes, eliminating any possible type II error, which is quite important.

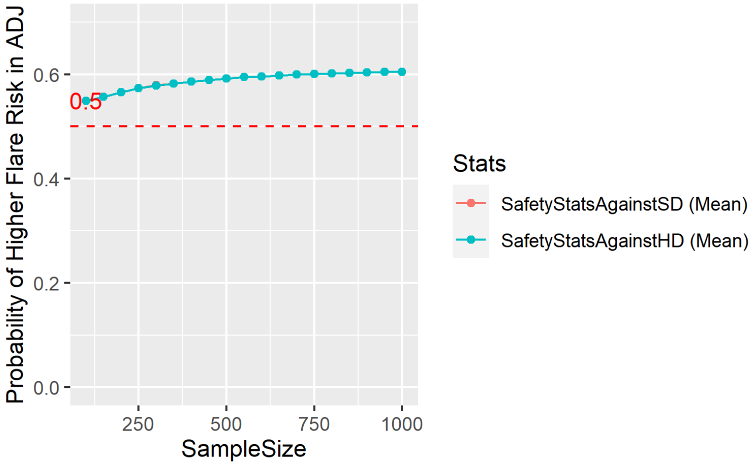


Figure 3.4.1.1: Mean Probability of Detecting Higher Flare Risk in ADJ When ADJ is Unsafe.

The mean probability of comparable flare risk in the ADJ arm against the SD arm (orange line) and against the HD arm (blue line) is computed using the mean values of the corresponding statistics from C2 and C3 (i.e., left hand side of the inequality) from Table 3.3.3.1 among 1000 simulations, respectively. The orange line and blue line overlap and both are far above the cutoff line so that both lines are correctly suggesting that the ADJ is unsafe under the alternative hypothesis setting at all sample sizes.

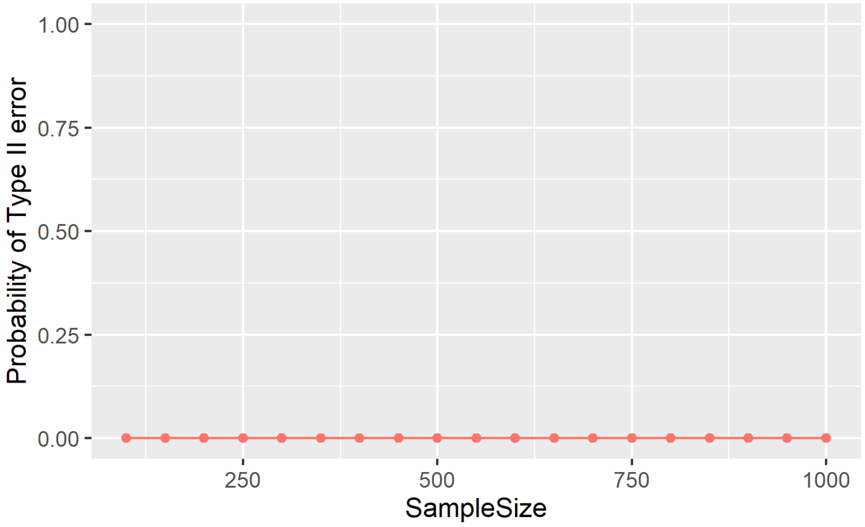


Figure 3.4.1.2: Probability of Type II Error in Demonstrating Safety.

The probability of type II error in demonstrating futility of safety of the ADJ arm is plotted against the sample sizes.

3.4.2 Efficacy Outcomes

The two efficacy outcomes of superiority of ADJ vs SD and non-inferiority of ADJ vs HD are related and will be discussed together since they both rely on the SCR of the ADJ arm.

3.4.2.1 Under the Null Hypothesis Setting

Superiority. Under the null hypothesis setting, the ADJ arm is assumed to be not only as efficacious as the SD arm, but also inferior to the HD arm. The impact of the sample size on demonstrating the superiority of the ADJ arm under the null hypothesis is shown in Figure 3.4.2.1.1 and Figure 3.4.2.1.2. Accordingly, Figure 3.4.2.1.1 shows that the criteria for demonstrating the superiority can not be met, neither at the end of the first study year nor before the end of the trial at the criterion in Table 3.3.3.1. It follows that the simulated trials are unlikely to produce type I error across all sample sizes as Figure 3.4.2.1.2 reflected.

Noninferiority. The impact of the sample size on demonstrating the noninferiority of the ADJ arm under the null hypothesis setting is shown in Figure.3.4.2.1.3 and Figure.3.4.2.1.4. Basically, the simulated trials are very unlikely to demonstrate the noninferiority of the ADJ arm when it is inferior under the null hypothesis across all sample sizes.

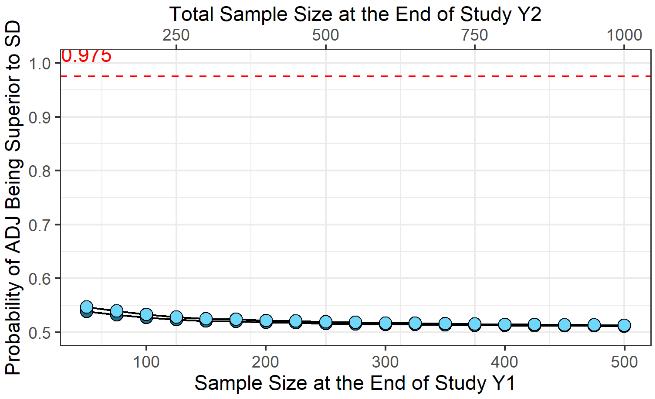


Figure 3.4.2.1.1: Mean Probability of Superior ADJ, $R_{ADJ,\ SCR} = 0.088$

The mean probability of demonstrating the superiority of the ADJ arm to the SD arm under the null hypothesis is plotted against the sample size at the end of the first study year and that at the end of the second study year (i.e., total sample size). The light blue dots show the impact of the sample size at the end of the first study year on the probability of demonstrating the superiority of the ADJ arm to the SD arm at that time (i.e., early dropout of the SD arm). The dark blue dots show the impact of the sample size on the probability of demonstrating the superiority of the ADJ arm to the SD arm before the trial ends, i.e., either at the end of the first study year or that at the end of the second study year. The mean values calculated across 1000 simulations of the probability of the ADJ arm being superior to the SD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,SD} > 0)$ from the left-hand side of C4 in Table 3.3.3.1) are very unlikely to cross over the 0.975 boundary with a sample size up to 1000 (or 500 at the end of the first study year).

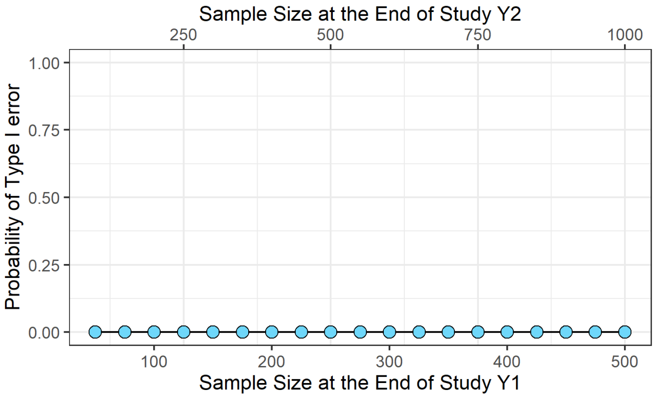


Figure 3.4.2.1.2: Probability of Type I Error in Demonstrating Superiority, $R_{ADJ,\ SCR} = 0.088$.

The simulated trials do not reject the null hypothesis in demonstrating the superiority of the ADJ arm to the SD arm when it is true across all sample sizes.

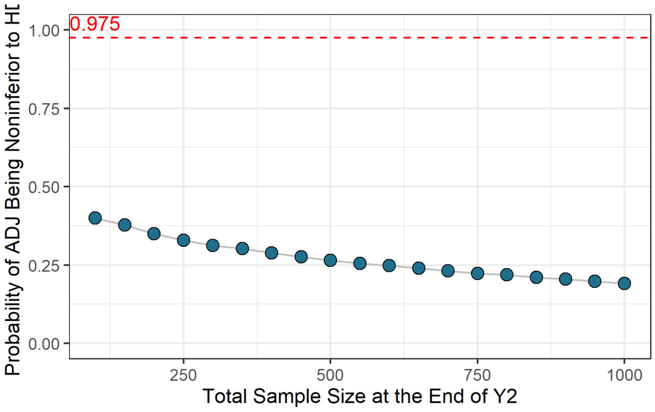


Figure 3.4.2.1.3: Mean Probability of Noninferior ADJ, $R_{ADJ,\ SCR} = 0.088$

The mean probability of demonstrating the noninferiority of the ADJ arm to the HD arm is plotted against the total sample size under the null hypothesis. The mean values of the probability calculated across 1000 simulations of the ADJ arm being noninferior to the HD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,HD} > -0.1)$ from the left-hand side of C6 in Table 3.3.3.1) are very unlikely to cross over the 0.975 boundary to demonstrate the noninferiority of the ADJ arm given current SCR setting across all sample sizes.

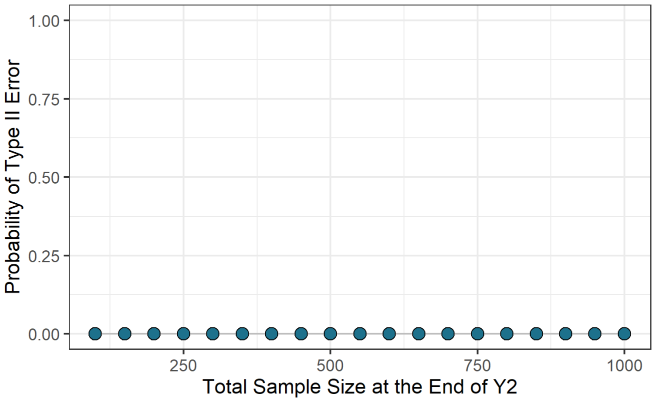


Figure 3.4.2.1.4: Probability of Type I Error in Demonstrating Noninferiority, $R_{ADJ,\ SCR} = 0.088$.

The simulated trials do not demonstrate the noninferiority of the ADJ arm to the HD arm under the null hypothesis with the largest sample size of 1000.

3.4.2.2 Under Alternative Hypothesis Setting

3.4.2.2.1 Simulations Under the First SCR Setting for the Alternative Hypothesis Setting

The impact of the sample size on demonstrating the superiority of the ADJ arm under the first alternative hypothesis setting is shown in Figure 3.4.2.2.1 and Figure 3.4.2.2.2. Accordingly, Figure 3.4.2.2.1 shows that the criteria for demonstrating the superiority can be met neither at the end of the first study year nor before the end of the trial at the criterion in Table 3.3.3.1. It follows that the simulated trials always fail to demonstrate the superiority of the ADJ arm when it is true and produces an unacceptably high type II error of 100% as Figure 3.4.2.2.2 suggests. In other words, it is impossible to demonstrate the superiority unless we bring down the strictness of the criteria, e.g., use a higher setting for $R_{ADJ,\ SCR}$ than 0.15 or a lower coverage percentage than 95%.

The impact of the sample size on demonstrating the noninferiority of the ADJ arm is presented in Figure 3.4.2.2.3 and Figure 3.4.2.2.4. Similar to the above situation, the noninferiority of the ADJ arm is barely demonstrated with a sample of size 1000 with current setting of $R_{ADJ,\ SCR}$.

All in all, despite the appropriateness of the plans in setting up the trial, under current alternative settings, the trial is unlikely to successfully demonstrate the superiority and the noninferiority of the ADJ arm. In particular, the criteria for demonstrating superiority and noninferiority are not met even with the highest feasible sample size of 1000. Therefore, in order to demonstrate the immunogenicity of the ADJ arm accurately, it is necessary to consider using higher alternative values for $R_{ADJ,\ SCR}$ that are reasonable or compromising by relaxing criteria's cutoff values accordingly.

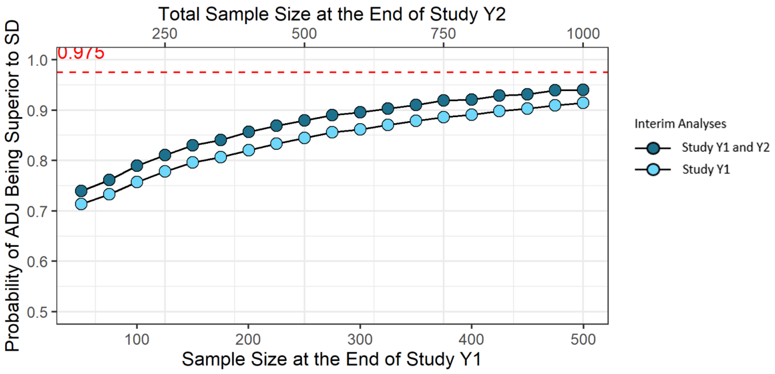


Figure 3.4.2.2.1: Mean Probability of Superior ADJ, $R_{ADJ,\ SCR} = 0.15$

The mean probability of demonstrating the superiority of the ADJ arm to the SD arm is plotted against the sample size at the end of the first study year and that at the end of the second study year (i.e., total sample size). The light blue dots show the impact of the sample size at the end of the first study year on the probability of demonstrating the superiority of the ADJ arm to the SD arm at that time (i.e., early dropout of the SD arm). In contrast, the dark blue dots show the impact of the sample size on the probability of demonstrating the superiority of the ADJ arm to the SD arm before the trial ends, i.e., either at the end of the first study year or that at the end of the second study year. The mean values calculated across 1000 simulations of the probability of the ADJ arm being superior to the SD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,SD} > 0)$ from the left-hand side of C4 in Table 3.3.3.1) are not high enough in both curves to cross over the 0.975 boundary with a sample size up to 1000 (or 500 at the end of the first study year).

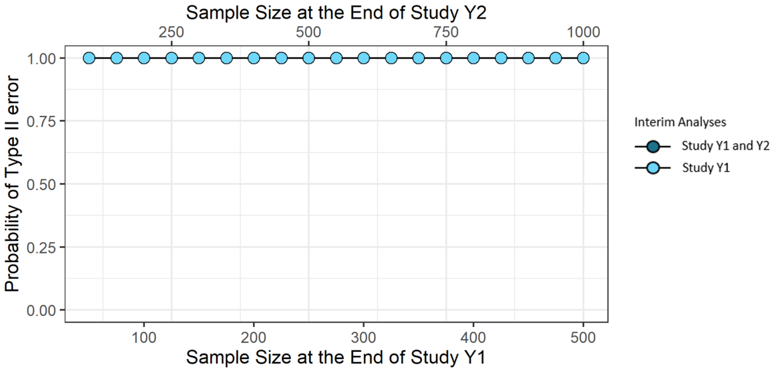


Figure 3.4.2.2.2: Probability of Type II Error in Demonstrating Superiority, $R_{ADJ,\ SCR} = 0.15$.

The simulated trials fail to demonstrate the superiority of the ADJ arm to the SD arm when it is true across all sample sizes.

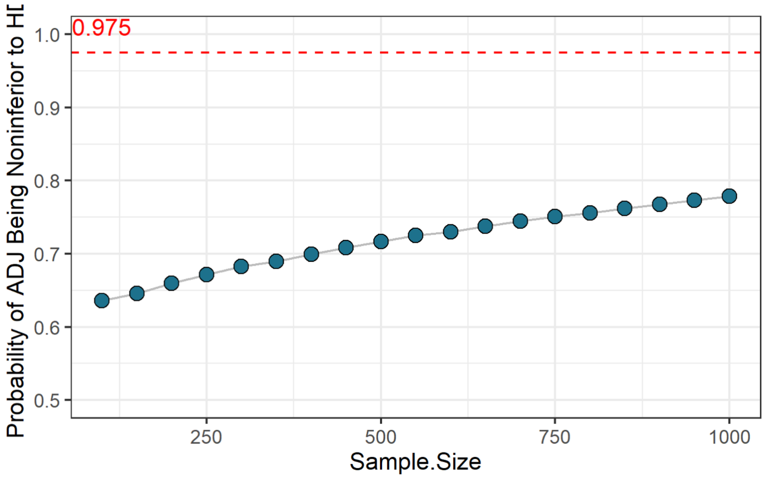


Figure 3.4.2.2.3: Mean Probability of Noninferior ADJ, $R_{ADJ,\ SCR} = 0.15$

The mean probability of demonstrating the noninferiority of the ADJ arm to the HD arm is plotted against the total sample size. The mean values of the probability calculated across 1000 simulations of the ADJ arm being noninferior to the HD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,HD} > -0.1)$ from the left-hand side of C6 in Table 3.3.3.1) are still not high enough to cross over the 0.975 boundary to demonstrate the noninferiority of the ADJ arm given current SCR setting with a sample size of 1000.

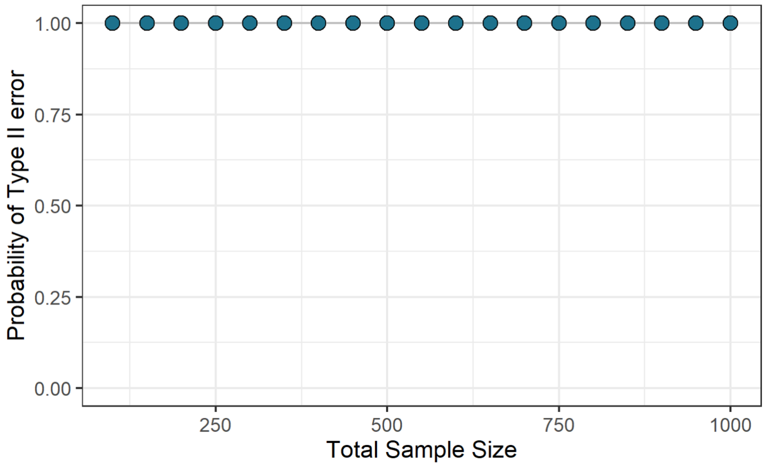


Figure 3.4.2.2.4: Probability of Type II Error in Demonstrating Noninferiority, $R_{ADJ,\ SCR} = 0.15$.

The simulated trials fail to demonstrate the noninferiority of the ADJ arm to the HD arm when it is true with the largest sample size of 1000.

3.4.2.2.2 Simulations Under the Second SCR Setting and Criterion Cutoff Value for the Alternative Hypothesis Setting

In this subsection, simulations are carried out with a higher value for SCR in the ADJ arm than the original setting, i.e.,

$$R_{ADJ,\ SCR} = 0.20,$$

while the remaining configurations remain unchanged.

In addition, for the noninferiority outcome, a lower cutoff value for demonstrating success in the noninferiority of the ADJ arm is lowered to 0.95. i.e., the criterion C6 becomes

$$P(R_{ADJ,\ SCR} - R_{HD,\ SCR} > -0.1) > 0.95.$$

The noninferiority outcome using the original and the relaxed cutoff values will be compared and the differences will be discussed.

Superiority. The impact of the sample size on demonstrating the superiority of the ADJ arm at the end of the first study year and that before the end of the trial are demonstrated in Figure 3.4.2.2.5 and Figure 3.4.2.2.6. In Figure 3.4.2.2.5, the mean value for the relevant statistics calculated at the end of the first study year, $P(R_{ADJ,\ SCR} - R_{SD,\ SCR} > 0)$, exceed 0.975 when the sample size is larger than 375 at that moment (i.e., the light blue dots crossover the 0.975 boundary at sample size = 375), whereas the mean probability calculated from the interim analyses at the end of the first study year and the end of the second study year exceed 0.975 when the sample size surpasses 400 (i.e., the dark blue dots cross over the 0.975 boundary at sample size = 400). This is because it is much easier for the trial to demonstrate the superiority after an increment in the hypothesized difference in SCR between the ADJ arm and the SD arm than before. Compared with the results obtained from Section.3.4.2.2.1, the results obtained under the second SCR setting suggest that with a larger difference in the hypothesized SCRs between these two arms (i.e., $R_{ADJ,\ SCR} - R_{SD,\ SCR} = 0.112$ under the second SCR setting rather than $R_{ADJ,\ SCR} - R_{SD,\ SCR} = 0.07$ under the first SCR setting), the trial is more likely to demonstrate the superiority of the ADJ arm correctly with a feasible sample size. Figure 3.4.2.2.6 shows that the type II error rates in demonstrating the superiority at the end of the first study year reduce to an acceptably low level if the sample size surpasses 400 (or 800 at the end of the trials) at the end of the first study year by the light blue curve, while the type II error rates in demonstrating the superiority before the end of the trial reduce tremendously if the sample size at the end of the trial surpasses 400 by the dark blue curve.

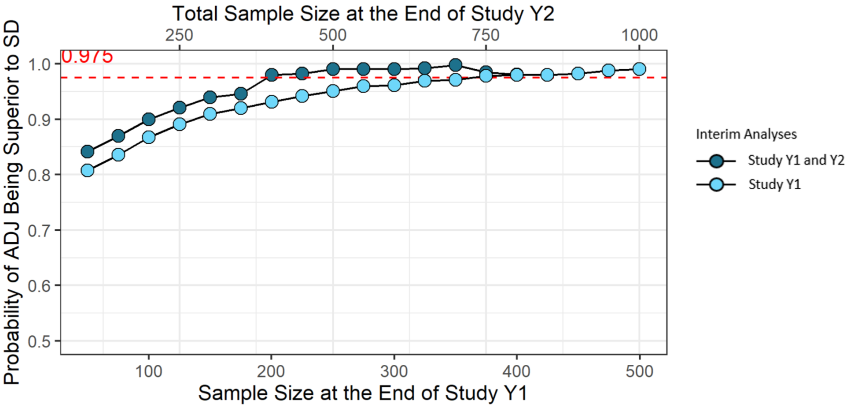


Figure 3.4.2.2.5: Mean Probability of Superior ADJ, $R_{ADJ,\ SCR} = 0.20$.

The mean probability of demonstrating the superiority of the ADJ arm over the SD arm is plotted against the sample size at the end of the first study year and that at the end of the second study year (i.e., total sample size). The light blue curve shows the impact of the sample size at the end of the first study year on the probability of demonstrating the superiority of the ADJ arm to the SD arm at that time (i.e., early dropout of the SD arm). In contrast, the dark blue curve shows the impact of the sample size on the probability of demonstrating the superiority of the ADJ arm to the SD arm before the trial ends, i.e., either at the end of the first study year or that at the end of the second study year. The mean values calculated across 1000 simulations of the probability of the ADJ arm being superior to the SD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,SD} > 0)$ from the left-hand side of C4 in Table 3.3.3.1) exceed 0.975 once certain sample sizes are achieved. The probability of demonstrating the superiority at the end of the first study year exceeds 0.975 if the sample size at that time surpasses 375. In contrast, the probability of demonstrating the superiority of the ADJ by the end of the trial exceeds 0.975 if the sample size surpasses 400.

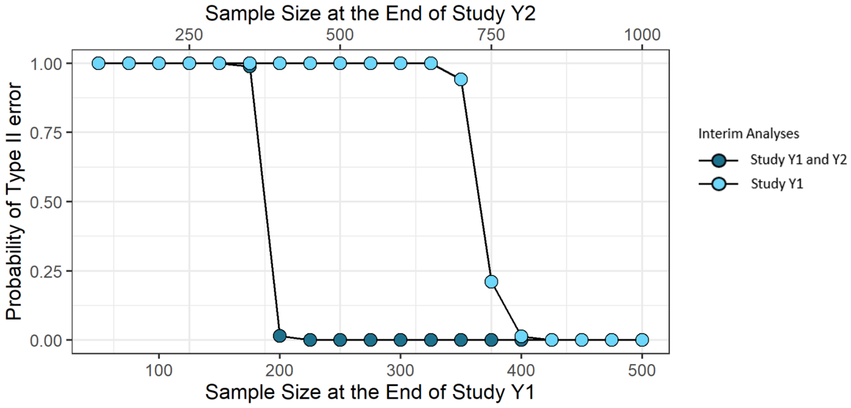


Figure 3.4.2.2.6: Probability of Type II Error in Demonstrating Superiority, $R_{ADJ,\ SCR} = 0.20$.

The simulated trials under current settings will start demonstrating the superiority of the ADJ arm over the SD arm once certain sample sizes are achieved. The probability of making a type II error in demonstrating the superiority at the end of the first study year can be reduced to an acceptable level if the sample size at that time exceeds 400. In contrast, the probability of making a type II error in demonstrating the superiority before the end of the trial can be reduced to an acceptable level if the total sample size surpasses 450.

Noninferiority. The impact of the sample size on demonstrating noninferiority is illustrated in Figure 3.4.2.2.7 and Figure 3.4.2.2.8 when different cutoff values for decision making were employed. More specifically, Figure 3.4.2.2.7 shows that a high enough mean values of the relevant metric, $P(R_{SCR,ADJ} - R_{SCR,HD} > -0.1)$, can be achieved if the sample size at the end of the study year 2 is larger than 800 when the cutoff value is set to be 0.975. In contrast, the relevant metric for demonstrating the noninferiority of the ADJ arm to the HD arm exceeds 0.95 as the sample size surpasses 700. This impact of lowering the decision cutoff value is also reflected in Figure 3.4.2.2.8: the probability of making a type II error drop to a fairly low level when the sample size surpasses 950 if the cutoff value is set to be 0.975, whereas that can be achieved when the sample size surpasses 750 if the cutoff value is set to be 0.95.

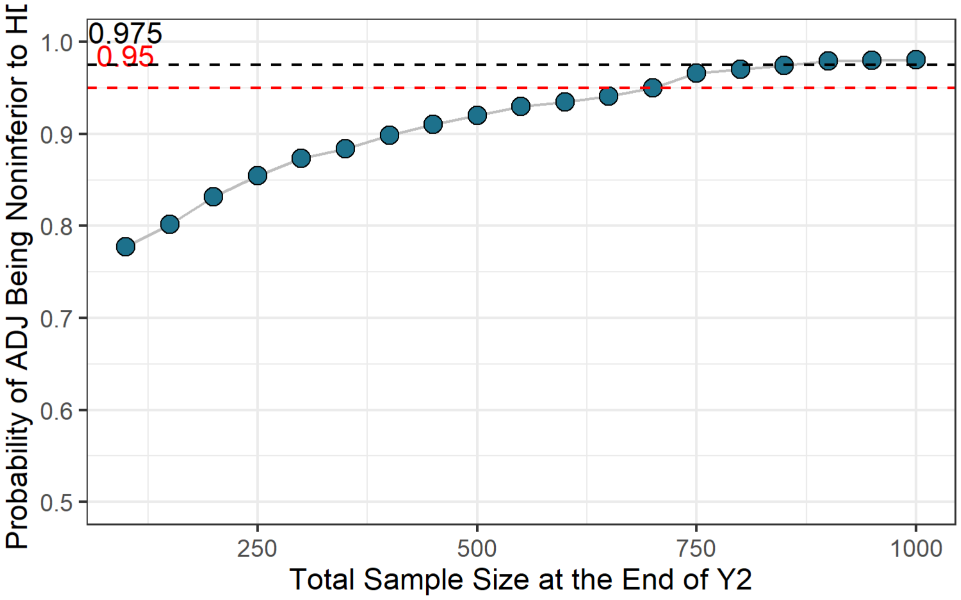


Figure 3.4.2.2.7: Mean Probability of Noninferior ADJ, $R_{ADJ,\ SCR} = 0.20$.

The mean probability of demonstrating the noninferiority of the ADJ arm to the HD arm is plotted against the total sample size. The black dashed line represents the original cutoff value for C6 and the red dashed line represents the lowered cutoff value for C6. When the cutoff value is set to be 0.975, the mean values of the probability calculated across 950 simulations of the ADJ arm being noninferior to the HD arm (i.e., $P(R_{SCR,ADJ} - R_{SCR,HD} > -0.1)$ from the left-hand side of C6 in Table 3.3.3.1) cross over the 0.975 boundary, while, when the cutoff value is lowered to 0.95, the same result is achieved when the sample size surpass 750.

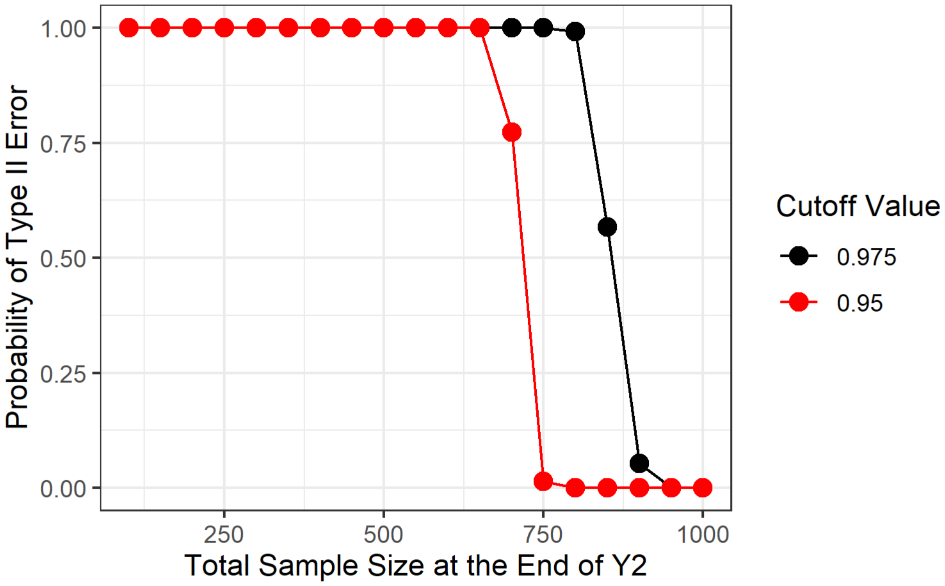


Figure 3.4.2.2.8: Probability of Type II Error in Demonstrating Noninferiority, $R_{ADJ} = 0.20$.

The black line presents the relationship between the probability of making a type II error when the cutoff boundary is set to be 0.975 and the red line presents that probability when the cutoff boundary is lowered to 0.95. Under the original setting with 0.975 as the cutoff value, the simulated trials start demonstrating the noninferiority of the ADJ arm to the HD arm once the sample size exceeds 800, and the probability of making a corresponding type II error eventually reduces to an acceptable level once the sample size reaches 950. In contrast, when the lower cutoff value of 0.95 was used, the simulated trials start demonstrating the noninferiority of the ADJ arm to the HD arm once the sample size exceeds 650, and the probability of making a corresponding type II error eventually reduces to an acceptable level once the sample size surpasses 800.

Possible Scenarios. Another interesting inference from simulations is regarding the likelihood of different combinations of outcomes under each simulation setting. The possible scenarios that arise over 1000 simulations under the modified setting with lowered cutoff value are demonstrated in Figure.3.4.2.2.9. Under the relaxed setting, 98.4% of the trials end up with the scenario encoded as 11191 (Safety Y1 = Success, Superiority Y1 = Success, Safety Y2 = Success, Superiority Y2 = Not Evaluated, Noninferiority = Success) and 1.6% of the trials end up with the scenario encoded as 12111 (Safety Y1 = Success, Superiority Y1 = Inconclusive, Safety Y2 = Success, Superiority Y2 = Success, Noninferiority = Success) when the sample size is 800.

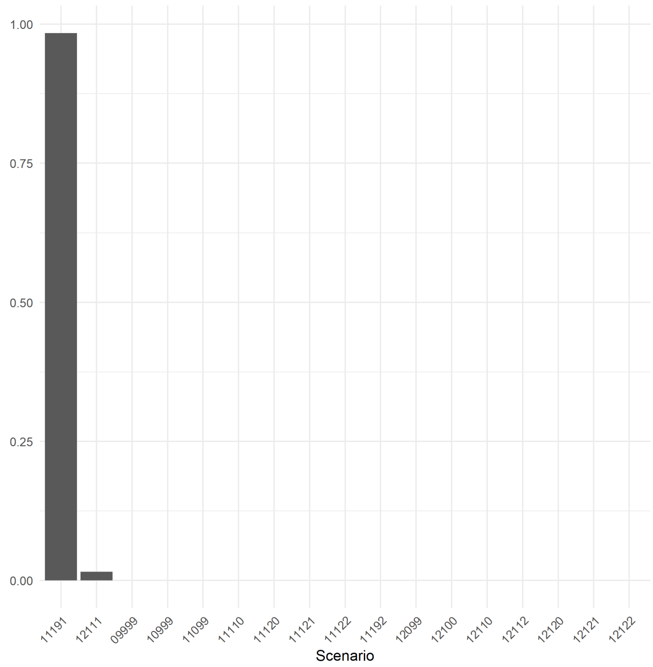


Figure 3.4.2.2.9: Possible Scenario.

The probability that each scenario arising is calculated with the following settings $R_{ADJ} = 0.20$, $P(R_{SCR,ADJ} - R_{SCR,HD}) > -0.1) > 0.95$, and a sample size of 800. 98.4% of the trials end up with the scenario encoded as 11191 (Safety Y1 = Success, Superiority Y1 = Success, Safety Y2 = Success, Superiority Y2 = Not Evaluated, Noninferiority = Success) and 1.6% of the trials end up with the scenario encoded as 12111 (Safety Y1 = Success, Superiority Y1 = Inconclusive, Safety Y2 = Success, Superiority Y2 = Success, Noninferiority = Success).

Chapter 4

Discussion

4.1 Summary

We have illustrated how the aid of a simple checklist can help statisticians to go through the different steps necessary in planning an adaptive trial design. In previous sections, using the DEFINE RCT as an example, we have seen that the simulations can be used to evaluate the feasibility of the trial in different scenarios by computing user-defined metrics such as type I and type II error rates. For example, when we have the seroconversion rate setting $R_{SCR,SD} = 0.088$, $R_{SCR,ADJ} = 0.15$, and $R_{SCR,HD} = 0.22$, the simulation results suggest that the trials with sample size smaller than 1000 are not likely to meet the criterion for adaptation and early dropping of the SD arm, whereas, when we have the seroconversion rate setting $R_{SCR,SD} = 0.088$, $R_{SCR,ADJ} = 0.2$, and $R_{SCR,HD} = 0.22$, we have seen that the superiority of the ADJ arm over the SD arm could be demonstrated at the end of the first study year if the sample size surpasses 400. The DEFINE trial investigators are advised to proceed with a minimum sample size of 1000 and a relaxation of the decision cutoff value for the noninferiority criterion from 97.5% to 95%.

We can also use the simulation to estimate the probability that the trial design will be adapted. To illustrate, when we have the SCR setting of $R_{SCR,SD} = 0.088$, $R_{SCR,ADJ} = 0.2$, and $R_{SCR,HD} = 0.22$ and the relaxed cutoff value of 0.95, the probability that the trial design is adapted was 98.4%. In addition, the simulation can also be used to study the relationship between the sample size and the probability to achieve the adaptation criteria under different scenarios. This permits researchers to fine-tune the trial design to achieve the same feasibility at a lower cost by adjusting the decision rules, such as cut-off values of the criterion. For example, we achieved the same error rates in demonstrating the noninferiority of the ADJ arm to the HD arm at a smaller sample size by decreasing the cut-off value in C6. This goal can possibly also be achieved by increasing the number of interim analyses, adjusting the time of interim analyses, etc., though we did not study these designs. Alternatively, if researchers believe the relaxed simulation settings are not realistic, they can be advised to seek a higher budget or collaborations with more centres in order to design a study that has adequate power to detect the associations of interest.

4.2 Limitation and Future Work

Some limitations of the work in this thesis should be noted. First, the weight of the historical prior in the robust prior is not unique. The weight we used in the DEFINE trial is 0.5, while the impact of other weights could also be explored. Different weights can be used depending on the credibility of the historical prior included in different trials. Secondly, the number of trials or posterior samples drawn can be increased to evaluate the feasibility of a trial design more precisely.

Thirdly, I did not explore other possible adaptations as they are not feasible in our trial. For example, we were not able to add more interim analyses as the study was planned for two years, and the efficacy can only be estimated once per year.

Work on this thesis has identified several gaps in current knowledge and practice that could be addressed as follows:

- **Developing an app**. Existing apps such as HECT Simulator (Thorlund et al., 2019) have been designed to support those unfamiliar with programming simulations in planning a trial. It would be interesting to investigate the possibility of developing an app leading from the checklist proposed in this thesis, that is geared towards statisticians or those at ease with conducting simulations. This could greatly facilitate the trial design feasibility evaluation along with parameter/hyperparameter tuning.

- **Application to complex trial designs** - Apply the checklist to a more complex design, e.g., a design for trials that span for a long period that allows more interim analyses, or the trials that have more treatment arms, and check if it can be further improved based on simulation studies.

- **Application to more complete scenarios** - Apply the checklist to a denser list of simulation settings. Appropriate methods can be applied to model the relationship between simulation settings and the output operating characteristics to fill in the blank between settings.

Bibliography

Belisle, P. (2017).

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.

Berry, D. A. (2010). *Bayesian adaptive methods for clinical trials*. CRC.

Berry, D. A. (2011). Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology*, 9(4):199–207.

Blackwelder, W. C. (1982). "proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, 3(4):345–353.

Brooks, S. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall CRC.

Center for Biologics Evaluation and Research (2007). Clinical data to support licensure of seasonal inactivated flu vaccine.

Colmegna, I., Useche, M. L., Rodriguez, K., McCormack, D., Alfonso, G., Patel, A., Ramanakumar, A. V., Rahme, E., Bernatsky, S., Hudson, M., and Ward, B. J. (2020). Immunogenicity and safety of high-dose versus standard-dose inactivated influenza vaccine in rheumatoid arthritis patients: a randomised, double-blind, active-comparator trial. *The Lancet Rheumatology*, 2(1):e14–e23.

Committee for Proprietary Medicinal Products (2008). Points to consider on switching between superiority and non-inferiority. *British Journal of Clinical Pharmacology*, 52(3):223–228.

Friedman, L., Furberg, C., and DeMets, D. (1998). *Fundamentals of Clinical Trials*. Springer.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433.

Hansen, C. H., Warner, P., Walker, A., Parker, R. A., Whitaker, L., Critchley, H. O., and Weir, C. J. (2018). A practical guide to pre-trial simulations for bayesian adaptive trials using sas and bugs. *Pharmaceutical Statistics*, 17(6):854–865.

Körding, K. P. and Wolpert, D. M. (2006a). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7):319–326.

Körding, K. P. and Wolpert, D. M. (2006b). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7):319–326.

Le Tourneau, C., Lee, J. J., and Siu, L. L. (2009). Dose Escalation Methods in Phase I Cancer Clinical Trials. *JNCI: Journal of the National Cancer Institute*, 101(10):708–720.

Lee, P. M. (2012). *Bayesian statistics: an introduction*. Wiley.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., and et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1).

Paul, G., Christy, C.-S., Vladimir, D., Brenda, G., Michael, K., and Jose, P. (2006). Adaptive designs in clinical drug development-an executive summary of the phrma working group. *Journal of Biopharmaceutical Statistics*, 16(3):275–283. PMID: 16724485.

Press, S. J. (1989). *Bayesian statistics Principles, models and applications*. John Wiley & Sons, Inc.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., and D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine*, 33(2):155–161.

Satlin, A., Wang, J., Logovinsky, V., Berry, S., Swanson, C., Dhadda, S., and Berry, D. A. (2016). Design of a bayesian adaptive phase 2 proof-of-concept trial for ban2401, a putative disease-modifying monoclonal antibody for the treatment of alzheimer's disease. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 2(1):1–12.

Sauer, R., Liersch, T., Merkel, S., Fietkau, R., Hohenberger, W., Hess, C., Becker, H., Raab, H.-R., Villanueva, M.-T., Witzigmann, H., and et al. (2012). Preoperative versus postoperative chemoradiotherapy for locally advanced rectal cancer: Results of the german cao/aro/aio-94 randomized phase iii trial after a median follow-up of 11 years. *Journal of Clinical Oncology*, 30(16):1926–1933.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health care evaluation*. Wiley.

Thisted, R. A. (2000). *Elements of statistical computing: Numerical computation*. Chapman & Hall/CRC.

Thorlund, K., Golchi, S., Haggstrom, J., and Mills, E. (2019). Highly efficient clinical trials simulator (hect): Software application for planning and simulating platform adaptive trials. *Gates Open Research*, 3:780.

Thorlund, K., Haggstrom, J., Park, J. J., and Mills, E. J. (2018). Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ*, 360.

Yuille, A. and Bülthoff, H. (1996). Bayesian decision theory and psychophysics. *Perception as Bayesian Inference*, page 123–162.