BLOST: Bayesian Longitudinally Ordinal Sequential Trial Design for Evaluating Respiratory Disease Treatments

by

Yulia Kozhevnikova

M.A., New Economic School, 2023
B.Sc., Moscow State University, 2021

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2025

Declaration of Committee

Name:	**Yulia Kozhevnikova**
Degree:	**Master of Science**
Thesis title:	**BLOST: Bayesian Longitudinally Ordinal Sequential Trial Design for Evaluating Respiratory Disease Treatments**
Committee:	**Chair:** Wei (Becky) Lin Lecturer, Statistics and Actuarial Science **Haolun Shi** Supervisor Assistant Professor, Statistics and Actuarial Science **Himchan Jeong** Committee Member Assistant Professor, Statistics and Actuarial Science **Liangliang Wang** Examiner Associate Professor, Statistics and Actuarial Science

Abstract

In clinical studies, traditional clinical trial designs may be inadequate for the rapid development of effective treatments during acute respiratory outbreaks. To address the drawbacks, we propose a Bayesian Longitudinally Ordinal Sequential Trial (BLOST) framework to optimize drug development processes and enhance resource efficiency in response in such settings. The design is based on a Bayesian model for longitudinally observed ordinal outcomes, accounting for patient heterogeneity and enabling information borrowing across time points. Our sequential framework is designed to compare the experimental treatment with a standard one. We consider three analytical approaches: a standard Bayesian method based on Hamiltonian Monte Carlo (BLOST), an enhanced version applying Bayesian model selection (BLOST-BMS), and a conventional frequentist approach. Simulation studies demonstrate that the proposed Bayesian approaches are more efficient than a conventional frequentist approach across different scenarios. Moreover, our approach requires fewer patients and less time to reach trial conclusions.

Keywords: Bayesian Design; Drug Efficacy; Ordinal Outcomes; Phase III Clinical Trial

Acknowledgements

I would like to express my deepest gratitude to my supervisor Haolun Shi. I truly appreciate his guidance, expertise, support, and patience throughout my graduate journey at SFU. Working on this paper was challenging, but it was also an incredibly valuable experience for me. I am grateful for the opportunity to work on this project and for his valuable advice and encouragement along the way.

I am also very grateful to the professors who have contributed to my learning experience through engaging courses and valuable feedback: Ian Bercovitz, Richard Lockhart, Rachel Altman, Liangliang Wang, and Lin Zhang. Additionally, I would like to extend my thanks to Wei (Becky) Lin — it was a great pleasure to work as a Teaching Assistant for her courses. A special thanks to the Department of Statistics and Actuarial Science for their continuous support, especially to Caitlin, Charlene, and Anna, who were always there to help with any questions.

I am thankful for my fellow graduate students, whose collaboration and knowledge-sharing made this journey especially enriching. Special thanks to So Yeon Park, who became my first friend in Canada. I sincerely thank Yueyang Han and Shifan Jia, who were an incredible support in many ways.

I deeply appreciate my family and friends in Moscow, who have always been just a phone call away. A special thanks to my mom — her strength and love have shaped my path to success and happiness. I am also grateful to my family and friends in Vancouver for their support and love — it means the world to me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The emergence of novel respiratory diseases emphasizes the critical need for efficient clinical trial designs. Respiratory diseases, such as those caused by SARS-CoV-2, have profound global health implications, with millions of confirmed infections and high mortality rates around the world. As of September 27, 2023, the COVID-19 pandemic had resulted in approximately 770.9 million confirmed cases and 6.9 million deaths worldwide (Manirambona et al. 2024 [1]). Rapid development and evaluation of effective treatments are vital, particularly in severe cases, where timely therapeutic intervention can save lives.

Traditional clinical trial models face significant challenges in addressing the urgent demands of respiratory diseases. Conventional drug development cycles, spanning three to ten years, are often prolonged by administrative delays and rigid protocols. However, such timelines are insufficient during pandemics or acute respiratory outbreaks, where the rapid escalation of cases demands accelerated processes. In addition, trial frameworks often face inefficiencies, such as small sample sizes, lack of randomization, and inadequate control groups, which limit the generation of reliable evidence. A comprehensive approach to data borrowing and trial design is essential to address these limitations and identify promising treatments efficiently.

Rapidly evolving health crises present unique challenges that require specific considerations for trial design. For chronic diseases or oncology trials, the population of patients remains stable. However, respiratory diseases involve rapidly growing and changing patient cohorts. This increase in patient numbers demands innovative trial designs capable of leveraging the rapid accumulation of data in real time. Furthermore, efficacy endpoints in respiratory disease cycles are often longitudinal ordinal, providing a detailed representation of symptom progression over time. For example, the World Health Organization (WHO) recommends a 10-level ordinal scale to track the severity of the disease, ranging from mild ambulatory conditions to hospitalization or death. Incorporating such detailed and dynamic outcomes into trial designs can enhance the evaluation of treatment efficacy and provide more nuanced insights.

The application of ordinal outcomes in clinical trials has become a popular topic in recent years, and many studies have explored their advantages. Whitehead and Horby (2017) [2] presented a sequential design for emerging pandemics, using odds ratios as the basis for hypothesis testing. They only used a frequentist approach and highlighted the importance of applying ordinal outcomes. D'Amico et al. (2020) [3] demonstrated the statistical power of using ordinal outcomes compared to binary outcomes in studies on cirrhosis decompensation. Biswas et al. (2018) [4] showed a frequentist approach for response adaptive designs in Phase III clinical trials with a focus on ordinal outcomes. Furthermore, various papers relied on group sequential trials that use a Bayesian framework. Ryan et al. (2019) [5] illustrated the use of Bayesian adaptive designs in respiratory trials, focusing on improving efficiency with early stopping for efficacy or futility. However, they did not use ordinal outcomes in this research. Casey et al. (2020) [6] applied a Bayesian sequential design with interim analyses to evaluate Hydroxychloroquine's efficacy for COVID-19, leveraging an ordinal scale for patient outcomes. Similarly, Ryan et al. (2022) [7] discussed Bayesian adaptive clinical trial designs for respiratory medicine, emphasizing flexibility in accommodating dynamic and evolving patient cohorts. Their approaches included multi-arm trials based on response adaptive randomization and arm dropping. The authors used a binary outcome, defined as death within 7 days. Moreover, some of the recent papers suggested other Bayesian approaches to work with longitudinally ordinal data in clinical trials. Rohde et al. (2024) [8] introduced the Bayesian ordinal transition models for ordinal longitudinal outcomes. Wu et al. (2024) [9] used a Bayesian multivariate hierarchical model to model mixed types of outcomes, including ordinal outcomes. Researchers focused on improving estimation efficiency by building the connection between correlated mixed types of outcomes. The last two papers were not group sequential trials but provided a valuable Bayesian framework for working with ordinal outcomes.

Current methods in clinical trial design consider longitudinal ordinal outcomes, Bayesian frameworks, and group sequential trial designs, which enable early stopping decisions and improve trial adaptability. While these approaches have been explored in existing research, certain gaps remain. Specifically, existing methods often fail to fully account for patient heterogeneity or discuss strategies to capture individual-level variability, despite its significant impact on trial outcomes. Moreover, many models do not explicitly account for the effect of time, which limits their ability to comprehensively model disease dynamics. Incorporating assumptions about drug efficacy progression and introducing corresponding features could enhance their ability to effectively borrow information across time points.

To address these limitations, we propose the Bayesian Longitudinally Ordinal Sequential Trial (BLOST) design. This design incorporates several key features, including the use of longitudinal ordinal outcomes, the group sequential trial structure, and the Bayesian framework. Moreover, BLOST introduces a time progression function to capture detailed information about patient recovery trajectories and drug efficacy over time. Additionally, the proposed method incorporates random effects to account for patient-individual heterogeneity. Finally, we extend the BLOST design with a Bayesian Model Selection (BMS) framework to compare models with different disease progression functions. These contributions make BLOST more efficient and adaptable than traditional frequentist methods, particularly in scenarios requiring dynamic and adaptive trial designs. By addressing these research gaps, BLOST advances trial design with a flexible framework for evaluating experimental treatments.

The rest of this article is organized as follows. In Chapter 2, we describe the methodology, including the data structure, model specification, and estimation approach for longitudinal ordinal outcomes. Additionally, we demonstrate the extension of the method through Bayesian model selection. Chapter 3 focuses on the design of the study and presents the illustration of the proposed approach. Chapter 4 describes the simulation study and outlines the findings. Chapter 5 provides a discussion of the results and the areas for improvement.

Chapter 2

Methodology

2.1 Data Structure

In this study, we consider a two-arm trial: a treatment arm with one experimental dose of the drug and a control arm with a standard drug or a placebo. In each of the arms we get an ordinal efficacy outcome over a time period T. The ordinal levels of efficacy are denoted as $1, \ldots, K$, where K is the total number of ordinal levels for efficacy. For each patient, we obtain a progression of the efficacy outcomes for time points $t = 1, 2, \ldots, T$. Let $Y^{(t)}$ denote an observed efficacy outcome at a timepoint t. For example, $Y^{(t)} = 1$ may represent death, $Y^{(t)} = 2$ severe symptoms, $Y^{(t)} = 3$ mild symptoms, $Y^{(t)} = 4$ no symptoms. The efficacy outcome for patients in the standard arm is denoted by $W^{(t)}$.

We denote a vector of efficacy probabilities for the ordinal outcomes at time t as $\boldsymbol{\pi}_t^E = (\pi_{t,1}^E, \ldots, \pi_{t,K}^E)^\top$, where $\pi_{t,k}^E = \Pr(Y^{(t)} = k)$ for a treatment arm with an experimental dose. A vector of efficacy probabilities for a standard arm is $\boldsymbol{\pi}_t^S = (\pi_{t,1}^S, \ldots, \pi_{t,K}^S)^\top$, where $\pi_{t,k}^S = \Pr(W^{(t)} = k)$. The superscripts E and S stand for experimental and standard arm, respectively.

The objective of the study is to determine whether the experimental treatment is more effective than a standard one. To measure the effect we need to introduce a metric on which we rely to make a decision at each timepoint t. This summary metric is based on the efficacy probabilities $\left\{\boldsymbol{\pi}_t^E : t = 1, \ldots, T\right\}$ and is computed as follows:

$$p^E = \frac{\sum_{t=1}^T w(t) \boldsymbol{b}^\top \boldsymbol{\pi}_t^E}{\sum_{t=1}^T w(t)},$$

where $w(t)$ is the weight of the observation in a time window, $\boldsymbol{b}^\top$ is a vector of weightings for the four ordinal levels. The weighted-averaged probability p is defined over the whole trial period $[1, T]$. The $w(t)$ is defined as 1 for $t = T$ and 0 otherwise in most cases, since the main interest lies in the final efficacy probability within the observed period. The $\boldsymbol{b}^\top$ is a predefined vector of ordinal levels. Its specific form is typically determined in consultation with clinicians or physicians. If let $\boldsymbol{b}^\top = (0, 0, 1, 1)$, $\boldsymbol{b}^\top \boldsymbol{\pi}_t^E$ implies that strong and medium efficacy outcomes have positive weights, while mild and low efficacy outcomes are excluded.

For the standard arm, the analogous metric is determined:

$$p^S = \frac{\sum_{t=1}^T w(t) \boldsymbol{b}^\top \boldsymbol{\pi}_t^S}{\sum_{t=1}^T w(t)}.$$

Thus, the hypothesis test of this study is formulated as:

$$H_0 : p^E \leq p^S \quad \text{versus} \quad H_a : p^E > p^S.$$

2.2 Model

Let i denote the index for an individual. The outcomes $Y_i^{(t)}$ in the experimental arm and $W_i^{(t)}$ in the standard arm suggest the use of a latent variable model. The ordinal outcomes arise from a categorization process of continuous latent variables $\tilde{Y}_i^{(t)}$ and $\tilde{W}_i^{(t)}$. The categorization is provided through predefined $K-1$ thresholds that break down outcomes into K categories. We denote the thresholds as $\{\tau_{y,k}\}$ and assume they are the same for all individuals. The thresholds $\{\tau_{y,k}\}$ are predefined in increasing order and used directly in ordinal categorization, ensuring their correct ordering.

If we assume that $\tilde{Y}_i^{(t)}$ is a normally distributed variable and its cumulative distribution is denoted by $\Phi(\cdot)$, we show that the probability of the outcome $Y_i^{(t)}$ to be in a category k is:

$$\Pr(Y_i^{(t)} = k) = \Phi(\tau_{y,k+1}) - \Phi(\tau_{y,k}),$$

where $k = 1, \ldots, K$, $\tau_{y,1} = -\infty$ and $\tau_{y,K+1} = \infty$. The probability of the outcome W_i is defined similarly:

$$\Pr(W_i^{(t)} = k) = \Phi(\tau_{w,k+1}) - \Phi(\tau_{w,k}),$$

where $k = 1, \ldots, K$, $\tau_{w,1} = -\infty$ and $\tau_{w,K+1} = \infty$.

To account for individual variability, we model the latent variable as a function of disease progression over time, including both patient-specific random effects and an error term. The latent variable for the experimental arm $\tilde{Y}_i^{(t)}$ is modeled as:

$$\tilde{Y}_i^{(t)} = \beta_0 g(t) + u_i + \epsilon_y,$$

$$u_i \sim N(0, \sigma_{u,y}^2),$$

$$\epsilon_y \sim N(0, 1),$$

where u_i is subject-specific random effect, ϵ_y is an error term, β_0 is a regression coefficient and $g(t)$ is a function displaying a progression of disease. The shape of $g(t)$ can be estimated from preliminary data, but for simplicity, we assume $g(t) = \log t$ in the baseline BLOST model.

The latent variable for the control arm $\tilde{W}_i^{(t)}$ is modeled as:

$$\tilde{W}_i^{(t)} = \gamma_0 h(t) + u_i + \epsilon_w,$$

$$u_i \sim N(0, \sigma_{u,w}^2),$$

$$\epsilon_w \sim N(0, 1),$$

where u_i is subject-specific random effect, ϵ_w is an error term, γ_0 is a regression coefficient, $h(t)$ is a function displaying a progression of disease, the form of which, for simplicity, will be the same as that of $g(t)$.

2.3 Estimation

We denote $\mathbf{\Omega}_E = (\beta_0, \sigma_{u,y}^2, \tau_{y,1}, \dots, \tau_{y,K})$ as the set of unknown parameters of the experimental treatment and $\mathbf{\Omega}_S = (\gamma_0, \sigma_{u,w}^2, \tau_{w,1}, \dots, \tau_{w,K})$ as the parameters related to the standard treatment. Let i denote the index for a patient, and let T_i denote the most recent observed time point for patient i. The likelihood for the experimental arm is:

$$L(\mathbf{\Omega}_E|\ \mathcal{D}) = \prod_{i=1}^{n} \phi\left(\frac{u_i}{\sigma_{u,y}}\right) \prod_{t=1}^{T_i} \left\{ \Phi\left(\tau_{y,Y_i^{(t)}+1} - \beta_0 g(t) - u_i\right) - \Phi\left(\tau_{y,Y_i^{(t)}} - \beta_0 g(t) - u_i\right) \right\}.$$

The likelihood for the standard arm is:

$$L(\mathbf{\Omega}_S|\ \mathcal{D}) = \prod_{i=1}^{n} \phi\left(\frac{u_i}{\sigma_{u,w}}\right) \prod_{t=1}^{T_i} \left\{ \Phi\left(\tau_{w,W_i^{(t)}+1} - \gamma_0 g(t) - u_i\right) - \Phi\left(\tau_{w,W_i^{(t)}} - \gamma_0 g(t) - u_i\right) \right\}.$$

$\Phi(\cdot)$ and $\phi(\cdot)$ are cumulative distribution function and probability density function of the standard normal distribution, respectively. We implement the next vague priors to the model parameters:

$$\begin{aligned} \tau_{y,k}, \tau_{w,k} &\sim N(0, \nu), \qquad\qquad k = 2, \dots, K \\ \beta_0, \gamma_0 &\sim N(0, \nu), \\ \sigma_{u,y}^2, \sigma_{u,w}^2 &\sim IG(a_u, b_u), \end{aligned}$$

where $IG(a_u, b_u)$ represents the inverse gamma distribution with a shape parameter a_u and a scale parameter b_u. We set $\nu = 10^4$ and $a_u = b_u = 10^{-3}$ to result in a vague prior. Hamiltonian Monte Carlo (HMC) is used to sample from the posterior distributions, which are obtained separately for the experimental and standard arms. The models are fitted using the `brms` package, which implements HMC via Stan.

2.4 Test Criterion

The efficacy test criterion $\Pr(p^E > p^S \mid \mathcal{D})$ is based on the values obtained from the posterior samples of $\boldsymbol{\Omega}_E$ and $\boldsymbol{\Omega}_S$. For the experimental treatment, the probability of the efficacy outcome being in the category k is:

$$\pi^E_{t,k} = \Pr(Y^{(t)} = k\ ) = \Phi\left(\frac{\tau_{y,k+1} - \beta_0 g(t)}{\sqrt{1+\sigma^2_{u,y}}}\right) - \Phi\left(\frac{\tau_{y,k} - \beta_0 g(t)}{\sqrt{1+\sigma^2_{u,y}}}\right),$$

for $k = 1, \ldots, K$. The terms inside the brackets result from standardizing the latent variable, centering it by its mean and scaling by its standard deviation to use the standard normal CDF.

For the standard treatment, the probability of the efficacy outcome is:

$$\pi^S_{t,k} = \Pr(W^{(t)} = k\ ) = \Phi\left(\frac{\tau_{w,k+1} - \gamma_0 g(t)}{\sqrt{1+\sigma^2_{u,w}}}\right) - \Phi\left(\frac{\tau_{w,k} - \gamma_0 g(t)}{\sqrt{1+\sigma^2_{u,w}}}\right),$$

for $k = 1, \ldots, K$.

The posterior estimate of p^E and p^S can be expressed as

$$\widehat{p}^E = \frac{\sum_{t=1}^T w(t)\boldsymbol{b}^\top \widehat{\boldsymbol{\pi}}^E_t}{\sum_{t=1}^T w(t)},$$

$$\widehat{p}^S = \frac{\sum_{t=1}^T w(t)\boldsymbol{b}^\top \widehat{\boldsymbol{\pi}}^S_t}{\sum_{t=1}^T w(t)},$$

where $\widehat{\boldsymbol{\pi}}^E_t = (\widehat{\pi}^E_{t,1}, \ldots, \widehat{\pi}^E_{t,K})^\top$ and $\widehat{\boldsymbol{\pi}}^S_t = (\widehat{\pi}^S_{t,1}, \ldots, \widehat{\pi}^S_{t,K})^\top$ are computed from averages of the posterior samples.
The test criterion $\Pr(p^E > p^S \mid \mathcal{D})$ is based on the comparison of the posterior samples of $\widehat{p}^E$ and $\widehat{p}^S$. It is estimated as the proportion of posterior samples that satisfy $\widehat{p}^E > \widehat{p}^S$:

$$\Pr(p^E > p^S \mid \mathcal{D}) = \frac{\text{Number of posterior samples where } \hat{p}^E > \hat{p}^S}{\text{Total number of posterior samples}}.$$

This probability is evaluated at multiple interim time points during the study. It is then compared to predefined thresholds based on the chosen alpha spending function. If the probability exceeds the threshold at any interim stage, the trial may stop early due to strong evidence that the experimental treatment is more effective.

2.5 Bayesian model selection

To consider a comprehensive range of specifications, we propose the Bayesian Longitudinally Ordinal Sequential Trial design with Bayesian model selection (BLOST-BMS). In this framework, we consider multiple models to characterize the treatment-control relationship and to identify the most probable model. We consider a range of different models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. Each of them might involve a specific time progression function $g(t)$, such as $\mathcal{M}_1 : g(t) = t$, $\mathcal{M}_2 : g(t) = \log t$, and $\mathcal{M}_3 : g(t) = \sqrt{t}$.

We further focus our derivation on models related to the experimental treatment. The corresponding model selection procedure for the standard treatment can be derived similarly. We denote $f(\mathbf{\Omega}_E)$ as the joint prior distribution of the parameters $\mathbf{\Omega}_E$. Let $\mathbf{u} = (u_1, \ldots, u_n)$ denote the vector of subject-specific random effects and $f(\mathbf{u} \mid \mathbf{\Omega}_E) = \prod_i^n \phi(u_i|\sigma_{u,y})$ is the probability density function conditional on the model parameters. We can compute the marginal likelihood of model $\mathcal{M}_k$ integrating out $\mathbf{\Omega}_E$ and $\mathbf{u}$ with respect to the corresponding prior distribution,

$$P(\mathcal{D} \mid \mathcal{M}_k) = \int P(\mathcal{D} \mid \mathbf{\Omega}_E, \mathcal{M}_k) f(\mathbf{\Omega}_E) f(\mathbf{u} \mid \mathbf{\Omega}_E) \mathrm{d}\mathbf{\Omega}_E \mathrm{d}\mathbf{u},$$

and the posterior model probability for $\mathcal{M}_k$ is

$$P(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k) P(\mathcal{M}_k)}{\sum_{i=1}^{M} P(\mathcal{D} \mid \mathcal{M}_i) P(\mathcal{M}_i)},$$

where $P(\mathcal{M}_k)$ is the prior probability of model $\mathcal{M}_k$.

In case there is no prior preference for any model, we may take a discrete uniform prior model probability, such as $P(\mathcal{M}_k) = 1/K$. In each interim analysis, the model with the largest posterior model probability $P(\mathcal{M}_k \mid \mathcal{D})$ is chosen to derive the test statistics for decision making.

Chapter 3

Design

3.1 Design Specification

The trial proceeds in one stage with several interim analyses. Relying on a group sequential testing framework, we compare the experimental treatment with the standard treatment. The trial includes five interim evaluations, with decisions made every $0.5T$ days starting from T (the initial evaluation period) up to $3T$ days. Let α denote the desired type I error rate, and the desired power is $1-\beta$. Let c_j denote the posterior probability cutoff for the efficacy endpoint, i.e., if $\Pr(p^E > p^S \mid \mathcal{D}) \geq c_j$ at the j-th interim analysis, we will stop the trial and declare the superiority of the experimental treatment, where $j = 1, \ldots, J$.

From $t = T$ to $t = 3T$, we use a group sequential design to compare the treatment arm with the standard arm. At each interim, we do a hypothesis test of $H_0 : p^E \leq p^S$ versus $H_1 : p^E > p^S$ using the criterion $\Pr(p^E > p^S \mid \mathcal{D})$. The test criterion is repeatedly applied under a group sequential framework. If $\Pr(p^E > p^S \mid \mathcal{D}) \geq c_j$, we can stop the trial and recommend the experimental treatment. Otherwise, we continue enrolling the next cohort of patients, or if none of tests succeeds at the end of the trial, we conclude that the trial fails to declare superiority of the experimental treatment.

3.2 Type I Error Rate and Power

While conducting the study we control the frequentist type I error rate. The type I error is defined as the probability that the trial is concluded with a positive results given the experimental treatment being exactly the same as the control treatment. Let α_E denote the target type I error rate. The control of the efficacy type I error rate towards the target α_E is achieved through calibrating the value of c_j. At each interim analysis in the group sequential design, we control for the spending of the type I error rate, i.e. the probability of exceeding the critical boundaries at the kth interim, so that the overall type I error is controlled. Based on the numerical search approach proposed by Shi and Yin (2019) [10], we calibrate the values of c_j with the initial guess being $c_j = \Phi(a_j)$, where a_j are the critical constant under the frequentist group sequential design with the type I error rate of α_E. We consider both Pocock and O'Brien-Fleming type spending function. For Pocock's design, the type I error rate spending can be approximated as $\alpha \log\{1+(e-1)j\}$. The type I error rate spending for the O'Brien-Fleming design can be approximated as $2-2\Phi(z_{\alpha/2}/\sqrt{j})$.

To calibrate the type II error rate, we use an alternative scenario, which is defined as $p^E > p^S$. Let the desired power be denoted as $1-\beta_E$. Achieving this power involves ensuring that the trial detects a true treatment effect with high probability when the experimental treatment is superior to the control. We iteratively adjust the sample size using a line search approach until the desired power is reached, as power generally increases with sample size. At each iteration, the sample size is evaluated under the alternative hypothesis to ensure that the calculated power matches the target $1-\beta_E$. By combining the spending function approach for type I error control with this iterative procedure, we maintain rigorous control over both error rates while ensuring the trial's robustness and efficiency.

3.3 Illustration of the Trial Design

We illustrate the conduct of our proposed trial design using a hypothetical clinical trial with a cohort size of 30. Suppose we aim to evaluate the efficacy of an experimental treatment compared to a standard treatment. The trial proceeds in a single stage with five interim analyses conducted at regular intervals. Specifically, the evaluations occur every 5 days, starting from day 10 ($T = 10$) and continuing until day 30 ($3T = 30$).

At each interim analysis, a cohort of 30 patients is enrolled and their outcomes are observed. Using the Bayesian framework, we compare the experimental treatment's efficacy (p^E) against the standard treatment (p^S) by evaluating the posterior probability $\Pr(p^E > p^S \mid \mathcal{D})$. If this probability exceeds the threshold c_j at any interim analysis, the trial stops early, declaring the experimental treatment superior. Otherwise, the trial continues to the next interim analysis. For example, at the first interim analysis on day 10, the posterior probability might be below the threshold c_1, prompting continuation. By the third interim analysis on day 20, the probability could exceed c_3, resulting in early termination and recommendation of the experimental treatment. If none of the thresholds are reached, the trial concludes on day 30, and the final decision is made based on all accumulated data.

This design ensures efficient use of resources while maintaining rigorous control of type I and type II error rates. By dynamically assessing the data at interim points, the trial maximizes the likelihood of identifying an effective treatment while minimizing patient exposure to potentially ineffective therapies.

Chapter 4

Simulation

4.1 Setup

We conduct 1000 independent simulations to evaluate the performance of the BLOST design and the BLOST-BMS design with three models for $g(t)$, i.e., $\mathcal{M}_1 : g(t) = t$, $\mathcal{M}_2 : g(t) = \log t$, and $\mathcal{M}_3 : g(t) = t^{0.5}$. The first phase of the trial, leading up to the initial interim analysis, lasts 10 days ($L_1 = L = 10$ days). The second phase lasts $0.5L$ days, so the time point for the second interim decision is at $L_2 = 1.5L$. Subsequent interim decisions are made every $0.5L$ days. The total trial duration is $L_5 = 3L = 30$ days. We consider two cohort sizes: 10 and 30. On each day during the trial, patients of cohort sizes of 10 or 30 are equally assigned to control arm and the experimental arm, constituting a total of 20 or 60 patients per day.

In terms of the definition of the summary metric, we set $\boldsymbol{b}^\top = (0, 0, 1, 1)$, i.e., the third and fourth ordinal efficacy levels are of interest. Moreover, we set the time weighting function $w(t)$ to be 1 at $t = 10$ and 0 otherwise. The coefficient parameters of the underlying regression model, β_0 and γ_0, as well as the threshold value for the ordinal outcomes, $\{\tau_{y,k}\}$ and $\{\tau_{w,k}\}$, are calibrated towards the assumed values of p^E and p^S in the simulation scenarios. To simulate the ordinal outcomes for efficacy, we first generate the latent variable from the normal distribution, and then calculate the ordinal outcomes through categorization. The standard deviation of the ϵ_y and ϵ_w is 1. The standard deviation of the random effect u_i is 0.1.

We compare the BLOST design and BLOST-BMS design with a frequentist design based upon the sample proportion of the binary outcomes that equals 1 if at $t = 10$, the observed efficacy outcomes falls into the third and fourth ordinal levels, and equals 0 otherwise. The counterpart frequentist group sequential design on the binary sample proportion is adopted for the interim and final decisions.

We consider various scenarios for efficacy probabilities, with a total of 10 scenarios. Three of these scenarios have equal values for p^E and p^S, specifically 0.2, 0.3, and 0.4. The remaining seven scenarios represent cases where p^E is greater than p^S, with p^S fixed at either 0.2 or 0.4. All 10 scenarios are run for cohort sizes of both 10 and 30. We apply both O'Briend-Fleming and Pocock alpha spending functions.

4.2 Simulation Results

To evaluate the design's operating characteristics in terms of time and resources, we compute various design metrics averaged from $1,000$ independent simulations. The results are shown in the Table 4.1. This table has dimensions of $1,000 \times 5$, representing $1,000$ simulations and 5 interim analysis points for scenario 6, where $p^S = 0.4$ and $p^E = 0.6$, under the BLOST-BMS model, with a cohort size of 10. Each value in the table corresponds to the test criterion value $\Pr(p^E > p^S \mid \mathcal{D})$, which represents the posterior probability that the efficacy parameter of the experimental treatment exceeds that of the standard treatment at a given interim analysis point. These probabilities are subsequently compared to the threshold c_j, where $j = 1, \ldots, 5$, to determine whether the null hypothesis can be rejected.

Table 4.1: Posterior probabilities computed at each interim analysis ($t = 1, \ldots, 5$) across 1,000 simulation runs. Each value represents the proportion of cases where the efficacy parameter for the experimental treatment exceeds that of the standard treatment. We select 10 representative rows out of 1,000 for illustration.

Interim1	Interim2	Interim3	Interim4	Interim5
0.979	0.998	0.998	0.999	1.000
0.964	0.993	0.975	0.993	0.999
0.978	0.997	0.997	1.000	1.000
0.973	0.991	0.998	0.997	0.996
0.862	0.714	0.987	0.993	1.000
0.987	1.000	0.998	1.000	1.000
1.000	1.000	1.000	1.000	1.000
0.936	0.968	0.977	0.999	0.999
0.762	0.939	0.975	0.990	0.932
0.882	0.781	0.867	0.970	0.979

Table 4.2 is derived by comparing the posterior probabilities to thresholds c_j that are determined for each interim analysis. These thresholds depend on the chosen alpha spending function, alpha and beta level. A rejection is recorded for a given simulation if any posterior probability exceeds its corresponding threshold at any interim analysis point. In other words, this table shows, for each of the three approaches, the probability of rejecting the null hypothesis $H_0 : p^E \leq p^S$. Specifically, it reports the percentage of 1,000 simulated cases in which the null hypothesis was rejected. The power $1 - \beta$ equals 0.8 for all calculated scenarios. The power is defined as the probability that the experimental treatment is correctly identified and the trial is concluded with a positive results when the experimental treatment is superior to the standard one. Table 4.2 presents the results for the O'Brien-Fleming type spending function with a cohort size of 10. First 7 scenarios consider the cases

Table 4.2: The percentage of cases where the null hypothesis ($H_0 : p^E \leq p^S$) is rejected under the O'Brien-Fleming alpha spending function with a cohort size of 10, $\beta = 0.2$ and $\alpha = \{0.1, 0.05\}$. Scenarios vary by efficacy probabilities (p^E, p^S) for experimental and standard treatments.

p^E	p^S	**Scenario**	**BLOST**		**BLOST-BMS**		**Frequentist**	
			$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$
0.6	0.2	1	1.000	1.000	1.000	1.000	0.989	0.978
0.5	0.2	2	1.000	1.000	1.000	1.000	0.917	0.852
0.4	0.2	3	0.981	0.960	0.982	0.950	0.696	0.561
0.8	0.4	4	1.000	1.000	1.000	1.000	0.989	0.976
0.7	0.4	5	1.000	1.000	1.000	1.000	0.900	0.841
0.6	0.4	6	0.982	0.951	0.974	0.950	0.657	0.548
0.5	0.4	7	0.621	0.466	0.635	0.509	0.336	0.226
0.2	0.2	8	0.080	0.035	0.092	0.046	0.104	0.062
0.4	0.4	9	0.079	0.035	0.100	0.051	0.107	0.065
0.3	0.3	10	0.079	0.036	0.099	0.046	0.113	0.055

where $p^E > p^S$, for the last 3 scenarios $p^E = p^S$. Based on the first 7 scenarios, we can conclude whether the design maintains the power. All scenarios where p^E is greater than p^S by more than 0.1 confirm that the power is maintained in the design. However, cases with p^E greater than p^S by only 0.1 (Scenario 7) fail to reject the null hypothesis with a probability exceeding 80%. Thus, if the treatment effect of the experimental drug is minimal and nearly indistinguishable from the control, the design is unlikely to capture these effects. The last 3 scenarios help to assess whether the type I error rate is maintained. We observe that for both BLOST and BLOST-BMS approaches the proportion of cases, where the null hypothesis is rejected, is less than the predefined alpha level of 5 or 10%. Similarly, for the frequentist approach, the numbers can be slightly higher than the predefined alpha, but remain relatively close. Comparing the three approaches, we can conclude that the Bayesian methods tend to be more effective in correctly rejecting the null hypothesis for almost all scenarios, indicating a higher percentage of correct rejections compared to the frequentist approach.

Table 4.3 represents the results for the O'Brien-Fleming type spending function with a cohort size of 30. We see similar patterns in the results, indicating that Bayesian methods appear to be more efficient in correctly rejecting the null hypothesis. Moreover, in this case for Scenario 7, where $p^E = 0.5$ and $p^S = 0.4$, we observe a much higher percentage of correct rejection for BLOST and BLOST-BMS, compared to those in Table 4.2. The values for this scenario in the frequentist approach have also increased but have still not exceeded 80%. In general, increasing the cohort size from 10 to 30 leads to a higher percentage of correct H_0 rejections across all three methods but does not result in significant differences compared to the previous table.

Table 4.3: The percentage of cases where the null hypothesis ($H_0 : p^E \leq p^S$) is rejected under the O'Brien-Fleming alpha spending function with a cohort size of 30, $\beta = 0.2$ and $\alpha = \{0.1, 0.05\}$. Scenarios vary by efficacy probabilities (p^E, p^S) for experimental and standard treatments.

p^E	p^S	**Scenario**	**BLOST**		**BLOST-BMS**		**Frequentist**	
			$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$
0.6	0.2	1	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.2	2	1.000	1.000	1.000	1.000	0.999	0.996
0.4	0.2	3	1.000	1.000	1.000	1.000	0.969	0.932
0.8	0.4	4	1.000	1.000	1.000	1.000	1.000	1.000
0.7	0.4	5	1.000	1.000	1.000	1.000	0.995	0.993
0.6	0.4	6	1.000	1.000	1.000	1.000	0.936	0.893
0.5	0.4	7	0.946	0.895	0.935	0.878	0.539	0.422
0.2	0.2	8	0.073	0.029	0.121	0.067	0.096	0.047
0.4	0.4	9	0.085	0.045	0.149	0.095	0.101	0.053
0.3	0.3	10	0.085	0.045	0.138	0.085	0.099	0.058

Table 4.4: The percentage of cases where the null hypothesis ($H_0 : p^E \leq p^S$) is rejected under the Pocock alpha spending function with a cohort size of 10, $\beta = 0.2$ and $\alpha = \{0.1, 0.05\}$. Scenarios vary by efficacy probabilities (p^E, p^S) for experimental and standard treatments.

p^E	p^S	**Scenario**	**BLOST**		**BLOST-BMS**		**Frequentist**	
			$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$
0.6	0.2	1	1.000	1.000	1.000	1.000	0.980	0.947
0.5	0.2	2	1.000	1.000	1.000	1.000	0.880	0.794
0.4	0.2	3	0.950	0.930	0.963	0.934	0.608	0.470
0.8	0.4	4	1.000	1.000	1.000	1.000	0.982	0.955
0.7	0.4	5	1.000	1.000	1.000	1.000	0.860	0.778
0.6	0.4	6	0.962	0.916	0.962	0.913	0.574	0.444
0.5	0.4	7	0.526	0.374	0.560	0.422	0.276	0.177
0.2	0.2	8	0.069	0.030	0.080	0.038	0.085	0.048
0.4	0.4	9	0.061	0.029	0.094	0.037	0.101	0.060
0.3	0.3	10	0.076	0.031	0.091	0.039	0.093	0.060

The next tables demonstrate the results for the Pocock alpha spending function. Table 4.4 presents results for a cohort size of 10, while Table 4.5 displays results for a cohort size of 30. The overall dynamics of the results are quite similar to those observed in the first two tables, where Bayesian approaches consistently demonstrate greater efficiency in correctly rejecting the H_0 compared to the frequentist approach. Similarly, we notice that for scenario 7, where p^E is greater than p^S by only 0.1, all methods struggle to detect the effect, particularly when the values of p^S and p^E are relatively small. The situation improves if we increase the cohort size and the values are not small. Overall, the findings in the last

Table 4.5: The percentage of cases where the null hypothesis ($H_0 : p^E \leq p^S$) is rejected under the Pocock alpha spending function with a cohort size of 30, $\beta = 0.2$ and $\alpha = \{0.1, 0.05\}$. Scenarios vary by efficacy probabilities (p^E, p^S) for experimental and standard treatments.

p^E	p^S	**Scenario**	**BLOST**		**BLOST-BMS**		**Frequentist**	
			$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$	$\alpha = .1$	$\alpha = .05$
0.6	0.2	1	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.2	2	1.000	1.000	1.000	1.000	0.996	0.992
0.4	0.2	3	1.000	1.000	1.000	1.000	0.942	0.886
0.8	0.4	4	1.000	1.000	1.000	1.000	1.000	1.000
0.7	0.4	5	1.000	1.000	1.000	1.000	0.993	0.988
0.6	0.4	6	1.000	1.000	1.000	1.000	0.913	0.825
0.5	0.4	7	0.904	0.839	0.903	0.837	0.477	0.325
0.2	0.2	8	0.061	0.022	0.090	0.051	0.067	0.038
0.4	0.4	9	0.075	0.042	0.141	0.097	0.084	0.050
0.3	0.3	10	0.071	0.029	0.118	0.070	0.081	0.041

two tables consistently show similar results to those obtained using the O'Brien-Fleming alpha spending function.

We also consider the average trial duration (in days) and the average total sample size. As shown in Figures 4.1a, 4.1b, 4.3a, and 4.3b, the frequentist approach consistently requires a larger sample size compared to both Bayesian approaches. The only exception is in scenarios where $p^E = p^S$: all methods require the maximum possible number of patients. Furthermore, as shown in Figures 4.2a, 4.2b, 4.4a, and 4.4b, we observe that the Bayesian design generally results in a shorter trial duration in most scenarios. This indicates that the borrowing of information across the longitudinally ordinal outcomes greatly accelerates the progress of the trial.

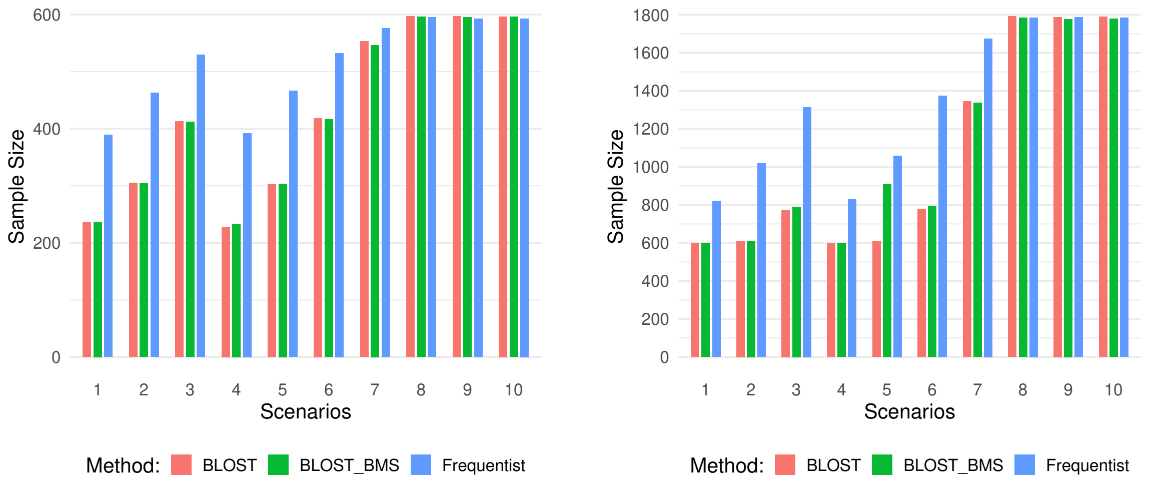

Figure 1a

Figure 1b

Figure 4.1: Average sample size required to determine if the experimental treatment is more effective than the standard treatment, based on the O'Brien-Fleming spending function with a cohort size of 10 (Figure 1a) and a cohort size of 30 (Figure 1b). The red, green, and blue bars represent the BLOST method, the BLOST-BMS method, and the frequentist approach, respectively.

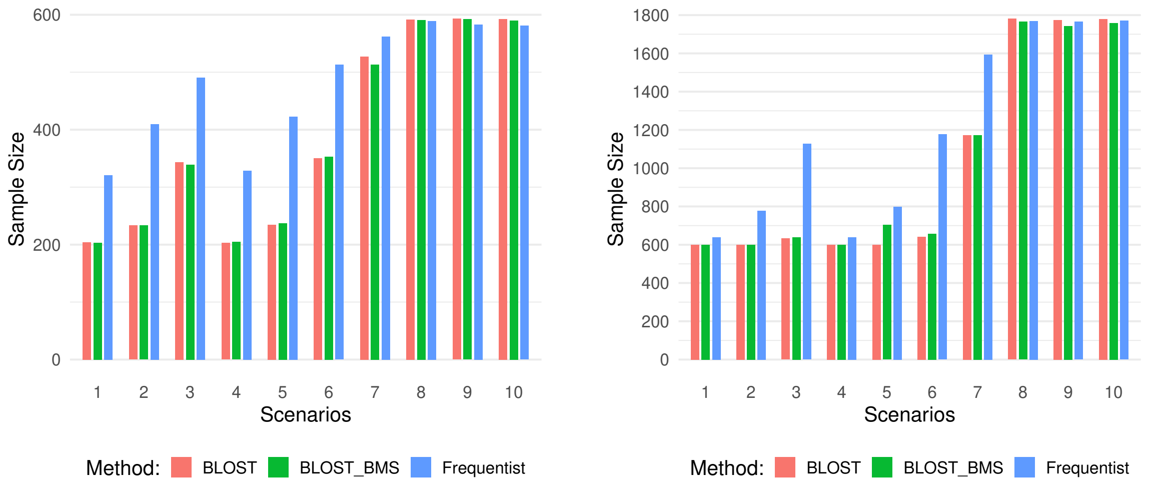

Figure 3a

Figure 3b

Figure 4.3: Average sample size required to determine if the experimental treatment is more effective than the standard treatment, based on the Pocock spending function with a cohort size of 10 (Figure 3a) and a cohort size of 30 (Figure 3b). The red, green, and blue bars represent the BLOST method, the BLOST-BMS method, and the frequentist approach, respectively.

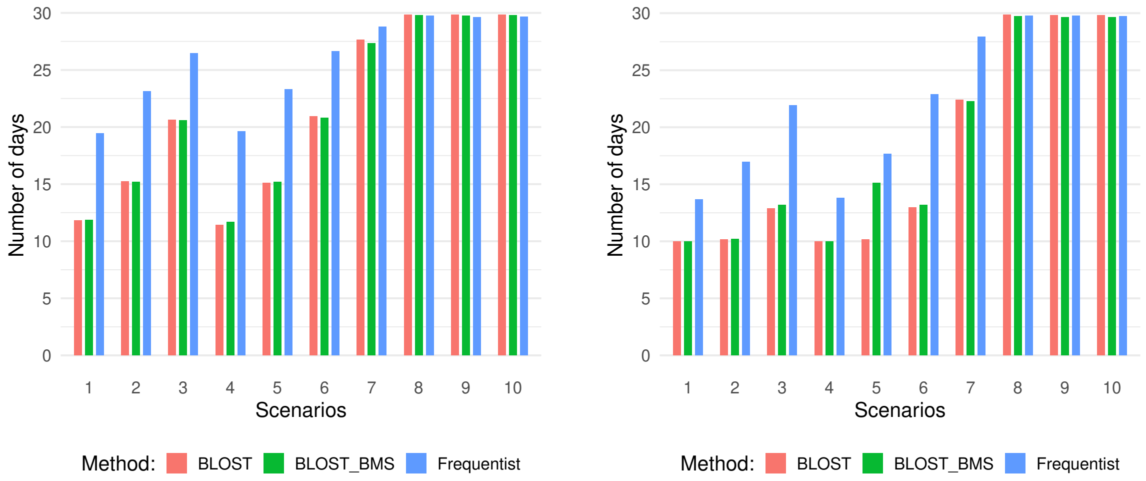

Figure 2a

Figure 2b

Figure 4.2: Average trial duration required to determine if the experimental treatment is more effective than the standard treatment, based on the O'Brien-Fleming spending function with a cohort size of 10 (Figure 2a) and a cohort size of 30 (Figure 2b). The red, green, and blue bars represent the BLOST method, the BLOST-BMS method, and the frequentist approach, respectively.

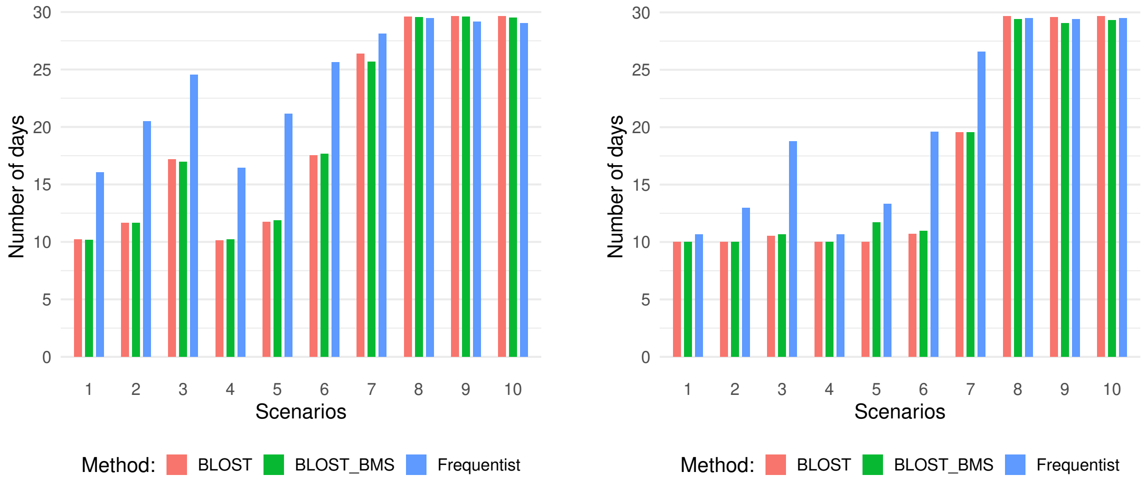

Figure 4a

Figure 4b

Figure 4.4: Average trial duration required to determine if the experimental treatment is more effective than the standard treatment, based on the Pocock spending function with a cohort size of 10 (Figure 4a) and a cohort size of 30 (Figure 4b). The red, green, and blue bars represent the BLOST method, the BLOST-BMS method, and the frequentist approach, respectively.

Chapter 5

Discussion

We propose a Bayesian Longitudinally Ordinal Sequential Trial design that compares the experimental treatment with the standard one. Our design effectively proposes Bayesian modeling for the longitudinally ordinal outcomes and is well suited for clinical trials of respiratory diseases. We demonstrate two methods based on Bayesian approaches: a standard one using Hamiltonian Monte Carlo, and an enhanced version incorporating Bayesian model selection (BMS). Additionally, we implement the same framework using the conventional frequentist approach to provide a comprehensive comparison of the results. From the simulation results, the proposed Bayesian methods are more efficient and accurate than the frequentist approach.

One way to enhance the trial design could be to consider multiple doses of the treatment. In extensive clinical trials, several doses are often analyzed simultaneously if they have been identified as the most promising based on Phase II trial results. Thus, by covering more possibilities, the study would become more applicable to real-life circumstances. However, it would also become significantly more complex due to the additional comparisons among the analyzed doses.

An important consideration for treatment trials during a pandemic is that the completion of a Phase III trial with stringent type I error rate control might not always be necessary before approving a drug for use. The reason is based on the urgent need for any potentially effective treatment against the respiratory disease. Compared to the traditional large-scale Phase III frequentist design, our design may carry a higher risk of false positive errors due to possible model misspecification. However, our proposed design significantly enhances the efficiency of identifying potentially useful treatments. This accelerates the treatment development process, which is a crucial advantage in responding to respiratory diseases. While this approach does sacrifice some control over the type I error rate, the benefits of increased design power and trial efficiency provide substantial compensation. To mitigate the risk of model misspecification, a conservative strategy could involve setting α to, for example, two-thirds of the desired type I error rate. Although this adjustment might result in an excess conservation of the type I error rate and a consequent loss of power in case the model is correctly specified, such a precaution is essential to prevent false positive errors.

Moreover, another valuable addition to our design could be the implementation of a more complex approach to modeling the progression function $g(t)$. In clinical practice, disease progression is commonly observed to vary among individuals. For example, some patients are likely to experience a rapid recovery, while others may endure severe deterioration over time. To address this variability, integrating a random effects term to model the impact of $g(t)$ or employing a more sophisticated modeling technique, such as a spline-based approach, could substantially enhance the model's flexibility. These improvements can be incorporated within the proposed Bayesian ordinal regression framework, providing a robust tool to capture the nuanced dynamics of disease progression.

Bibliography

[1] E. Manirambona et al. “Evolution and implications of SARS-CoV-2 variants in the post-pandemic era”. In: *Discover Public Health* 21.1 (2024), p. 16.

[2] J. Whitehead and P. Horby. “GOST: A generic ordinal sequential trial design for a treatment trial in an emerging pandemic”. In: *PLoS Neglected Tropical Diseases* 11 (2017), e0005439.

[3] G. D’Amico et al. “Ordinal outcomes are superior to binary outcomes for designing and evaluating clinical trials in compensated cirrhosis”. In: *Hepatology* 72.3 (2020), pp. 1029–1042.

[4] A. Biswas, R. Bhattacharya, and S. Das. “A response adaptive design for ordinal categorical responses”. In: *Journal of Biopharmaceutical Statistics* 28 (2018), pp. 1169–1181.

[5] E. G. Ryan et al. “Using Bayesian adaptive designs to improve phase III trials: a respiratory care example”. In: *BMC Medical Research Methodology* 19 (2019), pp. 1–10.

[6] J. D. Casey et al. “Rationale and design of ORCHID: a randomized placebo-controlled clinical trial of hydroxychloroquine for adults hospitalized with COVID-19”. In: *Annals of the American Thoracic Society* 17.9 (2020), pp. 1144–1153.

[7] E. G. Ryan, D. L. Couturier, and S. Heritier. “Bayesian adaptive clinical trial designs for respiratory medicine”. In: *Respirology* 27.10 (2022), pp. 834–843.

[8] M. D. Rohde et al. “Bayesian transition models for ordinal longitudinal outcomes”. In: *Statistics in Medicine* (2024).

[9] D. Wu et al. “A Bayesian multivariate hierarchical model for developing a treatment benefit index using mixed types of outcomes”. In: *BMC Medical Research Methodology* 24.1 (2024), p. 218.

[10] H. Shi and G. Yin. “Control of Type I Error Rates in Bayesian Sequential Designs”. In: *Bayesian Analysis* 14.2 (2019), pp. 399–425.