Bayesian Analysis (2023) **18**, Number 2, pp. 519–546

A Multi-Armed Bayesian Ordinal Outcome Utility-Based Sequential Trial with a Pairwise Null Clustering Prior*

Andrew Chapple†, Yussef Bennani‡, and Meredith Clement§

Abstract. A multi-armed trial based on ordinal outcomes is proposed that leverages a flexible non-proportional odds cumulative logit model and numerical utility scores for each outcome to determine treatment optimality. This trial design uses a Bayesian clustering prior on the treatment effects that encourages the pairwise null hypothesis of no differences between treatments. A group sequential design is proposed to determine which treatments are clinically different with an adaptive decision boundary that becomes more aggressive as the sample size or clinical significance grows, or the number of active treatments decreases. A simulation study is conducted for 3 and 5 treatment arms, which shows that the design has superior operating characteristics (family wise error rate, generalized power, average sample size) compared to utility designs that do not allow clustering, a frequentist proportional odds model, or a permutation test based on empirical mean utilities.

Keywords: multi-armed trials, group sequential trials, Bayesian clustering.

1 Introduction

COVID-19 quickly became a global pandemic in early 2020, with extensive spread seen in Europe, South America, Asia, and the United States. Early on in the pandemic, there were no known effective therapies for the disease. Several therapies were thought to be potentially beneficial based on in vitro studies, including Hydroxychloroquine (Yao et al. (2020), Liu et al. (2020)) and chloroquine, with or without Azithromycin, and Lopinavir/Ritonavir (Chu et al. (2020)). Due to COVID-19's devastating clinical consequences for patients as well as the tremendous strain on the healthcare system, quickly determining an effective treatment was imperative. In March, 2020, researchers at the University Medical Center (UMC) in New Orleans, LA, wanted to test two readily available therapies: (1) Hydroxychloroquine and (2) Hydroxychloroquine + Azithromycin versus (3) standard supportive care to determine if either of these treatments could provide a first step towards global recovery from the pandemic.

The trial (NCT # 04344444) was planned as a large-scale phase III trial with a maximum of 600 patients (ClinicalTrials.gov (2020)). Inclusion criteria included a positive

*This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science (Pordes (2007), Sfiligoi et al. (2009)).

†Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA, achapp@lsuhsc.edu

‡Department of Medicine, School of Medicine, LSU Health Sciences Center, New Orleans, LA

§Department of Medicine, School of Medicine, LSU Health Sciences Center, New Orleans, LA

 https://doi.org/10.1214/22-BA1316

COVID-19 test, onset of symptoms less than 7 days of trial enrollment, oxygen saturation above 94 %. Patients who were pregnant, lactating, or under 18 were excluded from the trial. Patients who used Hydroxychloroquine or Azithromycin were also excluded, in addition to patients who needed ICU care, and those who could not take oral medication or provide consent. Patients receiving the Hydroxychloroquine only treatment would receive 400 mg orally on day 1 of enrollment followed by 200 mg per day on days 2 through 5. Patients receiving Hydroxychloroquine and Azithromycin would receive 500 mg of Hydroxychloroquine on day 1 followed by 250 mg of Azithromycin on days 2 through 5. Patients receiving placebo would only receive standard therapy for COVID-19 patients at that time. To optimize benefit to study subjects, the researchers at UMC wanted to quickly eliminate the placebo arm during the course of enrollment if either of the treatments showed benefit in patients.

The researchers at UMC explained the various potential patient ordinal outcomes for COVID-19 patients after treatment. These potential day 14 outcomes included death, need for mechanical ventilation, continued hospitalization with or without supplemental oxygen, and discharge from care with or without supplemental oxygen. This outcome is ordinal in nature, since at day 14 of COVID-19 course the patients potential outcomes in terms of optimality are ordered as death $<$ on mechanical ventilation $<$ hospitalized with supplemental oxygen $<$ discharged with supplemental oxygen $<$ hospitalized without supplemental oxygen $<$ discharge without supplemental oxygen. The research clinicians felt that a patient requiring supplemental oxygen at day 14 may indicate that their condition could worsen in the future, so a hospitalized patient not on oxygen was preferred over a discharged patient on oxygen. This ordinal outcome was the primary outcome of the trial.

They wanted to account for this ordinal outcome in making treatment decisions, which was used as the primary outcome in the proposed trial (no secondary outcomes, including safety outcomes, were considered in adaptive decision making). Ordinal outcomes provide greater power to determine a treatment difference compared to binary ones, due to a greater level of granularity. For example, if we only examine whether or not a patient dies from COVID-19, we would be missing out on determining if a treatment also increases the probability of a hospital discharge. This ordinal scale is similar to the 8-point WHO clinical progression scale, which was published after this trial was designed (Marshall and et al (2020)). In order to determine treatment optimality from this ordinal outcome, numerical utilities on the range of $[0, 100]$ were elicited from clinicians for each potential outcome. These ordinal outcomes were scored based on clinical benefit to the patient. Without the use of utilities, a non-proportional odds model will not produce statistics summarizing overall treatment benefit as the parameters characterize distributions on each ordinal outcome category but not overall clinical benefit (Murray et al. (2018)). Utility (optimality) scores have been used in various trials since being introduced by Thall and Cook (2004) via the Efftox design. While they are not the most popular approach, real-world trials have been conducting using several trial approaches (Hoftsetter (2020), Amsbaugh and et al (2019), Shah et al. (2015), Brock et al. (2017), Murray et al. (2016)).

Using utilities, a treatment improvement can be seen solely through a decrease in the probabilities of poor events or even an increase in discharge probabilities, regardless of

whether all outcome probabilities are shifted in a single direction. For example, if fewer patients are placed on ventilators for a given treatment arm in the 14 days following hospital admission, this is a treatment benefit - regardless of change in day 14 death probabilities. This is a highlighted benefit of using numerical utilities in categorical/ordinal outcomes, as it may pick up on on-proportional shifts in outcome probabilities. These trial goals motivated a multi-armed utility-based group sequential trial based on ordinal outcomes that could repeatedly remove ineffective treatment arms at various interim analyses until trial completion or a set of treatments was declared optimal.

Multi-armed multi-staged trials (MAMS) are especially useful in quickly sorting through new potential therapies for COVID-19 or other pandemics. Multi-armed trials have been used for over half a century, with Dunnett (1955) providing test statistic corrections to give desired type I errors. Thall et al. (1989) extended multi-armed trials of experimental agents against a control to a two-staged design, where at the first stage the trial is stopped for futility or an optimal experimental therapy is selected.

In the second stage, the chosen optimal therapeutic is compared against historical results from a control. If the therapeutic is deemed effective, a randomized comparison of the chosen treatment and control is conducted in phase III. Since then there have been a myriad of MAMS trials that extends the two stage approach to multiple sequential looks. Stallard and Todd (2003) discuss a design strategy that is similar to Thall et al. (1989) and discuss design considerations for these trials. Our design differs from the MAMS trials described above in that all treatment arms including the control, can be dropped at various interim looks. Likewise a set of treatments, potentially including the control, can be declared equally optimal and graduated to larger-scaled trials. This was the desire of the research team in order to give needed treatments to COVID-19 patients as soon as possible. However, the design could be adjusted to either only allow dropping a control at later interim looks or to keep the control arm enrolling throughout the trial, similar to other MAMS trials. In general, this design can be used in settings where all treatments are experimental so being able to drop all treatments will allow the trial to better explore promising treatments based on early data.

Lastly, we introduce a Bayesian clustering prior on the treatment specific marginal outcome probabilities that *a priori* favors pairwise null hypotheses. Simulation studies show this novel clustering prior reduces family-wise error rates (by .25-.5), improves the generalized power (by .05-.20), and reduces the needed sample size (by 100-200 patients) for the trial compared to a typical exchangeable prior. Compared to a frequentist proportional odds model, the proposed clustering prior had a better controlled family wise error rate and average sample sizes by 200-300. When a true proportional odds relationship held, the clustering design had a slight decrease in generalized power (0 and .04) compared to the proportional odds model. However, when the proportional odds truth did not hold, the clustering design had an increase in generalized power of .09 and .23 for 3 and 5 treatments, respectively, which represents a drastic improvement while still maintaining lower sample sizes. The proposed design shows general superiority to a permutation based method based on empirical mean utilities. These results are shown in Figures 3 and 4, and in Tables 2, $S1$.

These posterior mean treatment specific marginal probabilities of each outcome are then used to compute mean utility scores for each treatment, which is the weighted sum

of each outcome's marginal probabilities times it's numerical utility score. The posterior distributions of the mean utility score for each treatment are used to make decisions to (1) drop a treatment for inferiority or (2) declare a set of treatments equally optimal. We propose a multi-armed randomized group sequential trial that monitors ordinal categorical outcomes for each patient, stops inferior treatment arms and stops the trial to declare an optimal treatment or set of optimal treatments when one is found.

To facilitate trial conduct, covariate adjustment was preferred over stratification based on patient covariates. The proposed method uses Bayesian methods similar to Harrell and Lindsell (2020) to adjust for the effect of age, Charleston Comorbidity Index (CCI) (Charlson et al. (1987)), and initial condition when estimating the probabilities of each outcome for a given treatment. Our approach differs in that we assume a non-proportional odds model with clustering induced on marginal treatment effects after covariate adjustment. We investigate whether this covariate adjustment is necessary via simulation study. The assumed structure produces a Multi-Armed Bayesian Ordinal Utility Sequential Trial, which we call MABOUST, that can be used to effectively weed out inferior treatment arms and determine an optimal set of treatments.

The remainder of this manuscript proceeds as follows. Section 2 discusses the statistical model used and the priors placed on the parameters, while Section 3 explains the decision making process used for stopping rules in the trial. Section 4 includes a simulation study of the design for 3 and 5 treatment arms for a factorial of designs with levels related to whether covariates were adjusted for, and whether pairwise null-clustering was used. The simulation study conducted with 3 treatments was actually used in planning the proposed trial. In addition, we simulate the proposed design under a frequentist proportional odds model and a permutation test based on empirical mean utilities for comparison. Functions needed to implement or simulate the trial design are provided in the R package *MABOUST*, which is discussed in detail in the supplemental materials (Chapple et al., 2022). Section 5 ends with a discussion of the method and its other possible uses.

2 Models, Priors, and Posterior Sampling

In this section we outline the statistical model assumed for the ordinal outcome variable and treatment/covariate relationship. We discuss how marginal quantities of interest are determined after covariate adjustment which are then used in trial decision making. We discuss the prior distributions assumed on the treatment vectors, which have a so-called *House-Party-Prior* that favors pairwise null hypotheses between all considered treatments.

2.1 Statistical Model

Consider an categorical outcome Y that takes on values $j = 1, \ldots, J$ and a randomized patient treatment assignment T_i which takes values $1, \ldots, K$. The goal is to determine which of the K treatments, or sets of K treatments, is optimal in terms of the J patient outcomes. To determine which of the K treatments are optimal, we potentially adjust

the treatment effect estimates by a vector of additional patient covariates, $\boldsymbol{X}_i$, that are known to affect the outcome probabilities. For trials involving COVID-19, these factors include Charleston Comorbidity Index (CCI), age group in decades, and an initial categorical condition upon arrival into the hospital. These were factors that the trial physicians wanted adjustment for, with age being dichotomized into decades to reflect the way that outcome data was reported in Louisiana (Whitfield and Swenson (2020)). We assume a cumulative logistic regression model:

$$logit(P[Y_i \leq j | T_i = k, \boldsymbol{X}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\beta}]) = \theta_{k,j} + \boldsymbol{X}_i\boldsymbol{\beta} \tag{1}$$

where $\theta_{k,j}$ is the kth treatment effect on the cumulative probability of observing $Y \leq j$. This implies that the vectors $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,J-1})$ represent the marginal effects of treatment k, after adjusting for additional patient covariates $\boldsymbol{X}_i$. $\boldsymbol{X}_i$ should be specified so that increased values are associated with poorer patient outcomes, and we enforce this into the model by assuming that *a priori* each entry of the vector $\boldsymbol{\beta} < 0$, which forces probabilities of $Y = j$ to be higher for lower values j and increased $\boldsymbol{X}_i$.

While we model $logit(P[Y_i \leq j | T_i, \boldsymbol{X}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\beta}])$ and obtain posterior samples for $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and $\boldsymbol{\beta}$ our interest lies in the quantity $logit[P[Y_i \leq j | T_i, \{\boldsymbol{\theta}_k\}_{k=1}^K]$ which is the probability of each outcome for a given treatment, marginalizing over the covariate space $\boldsymbol{X}_i$. We estimate this marginal probability via the quantity $P[Y_i = j | T_i = k] = logit^{-1}(\theta_{k,j}) - logit^{-1}(\theta_{k,j-1})$. When $j = J$, $logit^{-1}(\theta_{k,j})$ is replaced by 1, and when $j = 1$ $logit^{-1}(\theta_{k,j-1})$ is replaced by 0.

2.2 Prior Distributions

Since we are going to test whether a given treatment arm is inferior repeatedly throughout the trial design, we will constrain the prior distribution on each $\boldsymbol{\theta}_k$ vector to favor the null hypothesis of no treatment difference, using a Bayesian clustering prior similar to that described in Chapple and Thall (2018) and Chapple (2021). This introduces discrete random latent parameters $\zeta_1, \ldots, \zeta_K$ that take on values of $1, \ldots, K$ where $\zeta_k = \zeta_l$ for some $l \neq k$ indicates that treatments k and l are identical in terms of the ordinal outcome Y. Let $I[\zeta_k = l]$ denote the indicator that treatments k and l are equivalent. We assume a *spike-and-slab* type prior of the form:

$$\boldsymbol{\theta}_k | \zeta_k \sim I[\zeta_k = k]\pi_k(\boldsymbol{\theta}_k) + \sum_{l \neq k} I[\zeta_k = l]\delta_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}_k), \tag{2}$$

where $\delta_{\boldsymbol{\theta}_l}(\cdot)$ represents the dirac probability measure at $\boldsymbol{\theta}_l$ and $\pi_k(\cdot)$ is an unspecified prior distribution on treatment k. This implies that when $\zeta_k = l$, the parameter vectors of treatments k and l are clustered together, i.e. $\boldsymbol{\theta}_k = \boldsymbol{\theta}_l$, with $\boldsymbol{\theta}_l$ able to move freely in posterior sampling and $\boldsymbol{\theta}_k$ moving along with it. For the purposes of this study, we assume that the priors are equal for all K treatments, i.e. $\pi_k(\cdot) = \pi(\cdot) \forall k$, and place flat priors on the vector $\boldsymbol{\theta}_k$ with the restriction that $\theta_{k,j} < \theta_{k,j+1}$ for all k and j. This implies that the conditional prior distribution on $\boldsymbol{\theta}_k$ given $\zeta_k = k$

$$\boldsymbol{\theta}_k | \zeta_k = k \propto \prod_{j=1}^{J-1} I[\theta_{k,j} < \theta_{k,j+1}].$$

Other priors could be used in place of the non-informative flat prior, so long as this ordinal restriction holds, because without this it is possible that there will be estimated probabilities of $Y = j$ that are negative. Similarly, treatment-specific priors can be used if treatments target the disease in different manner. This could be used to guide the trial early on based on clinical or published experience with each treatment. We note that this idea is similar to that discussed for prognostic patient subgroups (instead of treatment groups) in Chapple and Thall (2018) and Chapple (2021). For the March 2020 trial planned at UMC, there was hesitancy to impose differential priors on the treatment groups to avoid influencing adaptive decision making. To ensure that *a priori* this configuration favors the pairwise null hypothesis of no treatment difference, we assume that

$$P[\zeta_k \neq k] = p_\zeta = 1 - P[\zeta_k = k],$$

for all treatments. Like Chapple and Thall (2018), we define a set $S = \{l : \zeta_l = l\}$ which is used to define the conditional distribution of $\zeta_k|\zeta_k \neq k$. The probability that $\zeta_k = l \in S$ is equal to one divided by the cardinality of S for any element in the set. This implies that the prior probability that all treatments are equal is p_ζ^{K-1} since for example, this indicates that treatments 2,..., K are clustered on treatment 1. Since all prior distributions for each treatment are equal, this logic can be applied similarly to any other treatment to cluster on. When $K = 5$ and $p_\zeta = .9$, this probability is about .65 indicating that the prior that at least 2 treatments are different is .35. This favors the global null. Marginally, decreasing the probability p_γ should increase the familywise type I error rate and also the power to drop the most inferior treatments. We explore decreasing p_γ to values of .75 and .5 (equally likely that treatments are equal or unequal in utility) via simulation in terms of these operating characteristics. In general, our simulation results favor using $p_\gamma = .9$ to reduce familywise type I error rates.

2.3 Prior Features and Differences From Other Clustering Priors

Here we briefly discuss some features of the prior distribution described above, which we fondly call a *house-party-prior*. This name fittingly characterizes how the prior acts in sampling. Consider a neighborhood with K houses, each with one family. If a family k is home, then $\zeta_k = k$, and they would be welcome to company. But if that family is at another house, i.e. $\zeta_k \neq k$, that family would not invite other families to their home since they are away. Consider another hypothetical scenario for this analogy, where several families have gone to family m's home, i.e. $\zeta_m = m$ (family m is home) and $\zeta_k = m$ for several choices of k which also implies that $\boldsymbol{\theta}_k = \boldsymbol{\theta}_m$ for those families. If at a given iteration of MCMC, $\zeta_l = l$ and we propose setting $\zeta_l \neq l$, we note that the set S is now smaller than $K - 1$ since several families are over at house m. This in turn makes it more likely that family l will also go to house m and join their cluster, setting $\boldsymbol{\theta}_l = \boldsymbol{\theta}_m$. This latter feature favors one global cluster if evidence is starting to mount that the number of clusters is small in the MCMC. The priors on $(\zeta_1, \ldots, \zeta_K)$ and hence $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ behave exactly like this real-life example.

In previous uses of this prior distribution, each set of parameters indexed by k had their own unique prior distribution (Chapple and Thall (2018) Chapple (2021)). In the

house-party-prior analogy, different homes might have different amenities (TVs, snacks, dogs, etc) so if two families are clustered together in house k instead of house m, their posterior experience would differ based on the prior settings of each house. MABOUST uses the same prior for each treatment group, so this aspect of the prior configuration does not come into play for this manuscript.

One might ask why this *house-party-prior* is favorable compared to assuming a finite mixture model or Dirichlet prior. A finite mixture model with K latent classes would produce the clusters of treatment effects, but these effects would not be exactly equal and would vary according to the prior for that latent cluster. It is also not possible to set up dirac measures in the finite mixture where parameters in the same latent distribution are clustered together. It would be possible to make components of the mixture distribution dirac measures, but these would need to be set on specific numerical values rather than other treatment based parameter vectors. The proposed pairwise null clustering method allows for fully continuous posterior samples of each θ_k while also favoring clustering truly pairwise null treatment effects together, and also favoring a global null when only one cluster exists.

2.4 Posterior Sampling

We obtain the posterior distribution parameter vector $\left(\{\boldsymbol{\theta}_k, \zeta_k\}_{k=1}^K, \boldsymbol{\beta}\right)$ through Markov Chain Monte Carlo (MCMC) sampling with Metropolis-Hastings Steps. For each iteration of the MCMC, we propose adjusting the clustering structure $\{\boldsymbol{\theta}_k, \zeta_k\}_{k=1}^K$ by randomly choosing a value of k updating each $\boldsymbol{\theta}_k|\zeta_k$. When $\zeta_k = k$, we propose clustering $\boldsymbol{\theta}_k$ with some randomly chosen m in the random set S and setting $\zeta_k = m$. When $\zeta_k = m$, $\boldsymbol{\theta}_k = \boldsymbol{\theta}_m$, and we propose setting $\zeta_k = k$ and unclustering $\boldsymbol{\theta}_k$ by randomly adjusting one element, under the constraint that $\theta_{k,j} < \theta_{k,j+1}$ for all j. These moves are accepted with probability proportional to the likelihood ratio times the prior ratio for ζ_k, since the prior on $\boldsymbol{\theta}_k$ is flat. Within each iteration, we also sample $\boldsymbol{\theta}_k|\zeta_k = k$ and $\boldsymbol{\beta}$ using adaptive Metropolis-Hastings, centered around the previous values, where the entry-specific proposal variance parameters are doubled (halved) if the acceptance rate for every 100 iterations is above .2 (below .6). This is done until half of the MCMC iterations are completed. For $\boldsymbol{\theta}_k$, normal proposal distributions are used, while for $\boldsymbol{\beta}$, we generate our proposal from a log-normal distribution with mean $\log(-\boldsymbol{\beta})$. We appropriately adjust the acceptance probability for entries of $\boldsymbol{\beta}$ using the proposal ratio, due to it's non-symmetry.

This sampling structure guarantees that the model adheres to the cumulative logit model assumptions, while borrowing strength across treatment groups for better estimation of these probabilities - through adaptive clustering and the covariate effects $\boldsymbol{\beta}$. When two treatment vectors are not clustered, borrowing can occur through adjustment for additional covariates, but we do not explore methods where treatment effects borrow hierarchically instead of via adaptive clustering. Code is produced in c++ to encourage computational speed. 2,000 iterations of the MCMC produced satisfactory trace plots of the parameter vector entries, indicating convergence, and were able to accurately approximate the true probabilities of each event for simulated data of size 200.

3 Decision Making

Rather than assuming proportional odds treatment effects, like several other authors including Harrell and Lindsell (2020), we flexibly estimate the probabilities of each outcome j and use a numerical utility score for each outcome j to determine a mean score of each treatment. We do this because a utility score makes it easy to compare different treatments numerically under a non-proportional odds model. Formally, let U_j denote elicited utility scores obtained from clinicians for outcome j, with $U_1 = 0$ and $U_J = 100$. For all other outcomes, it should be the case that $U_{j-1} < U_j < U_{j+1}$, i.e. that the utility score increases with j - which reflects better outcomes.

Eliciting the utility scores $U_1, \ldots, U_J$ was done by first meeting with team clinicians and determining the ordinal outcome structure to be used for the trial. While eliciting these outcomes, the team statistician encouraged clinicians to think about how they would score these ordinal outcomes in terms of relative clinical benefit, under the restriction that $U_{j-1} < U_j$. The clinicians met as a group until they agreed on a scoring system.

Two of the authors, Dr. Bennani and Dr. Clement took the lead on establishing these scores based on their clinical experience as infectious disease physicians and experience treating COVID-19 patients early on in the pandemic. After establishing their own scores, they had a series of meetings discussing their personal scoring system until they reached a consensus. If there was a firm difference in opinions for a given U_j the score was set as the average score between their two scoring systems. Afterwards, this scoring system was presented to all other team clinicians for final approval. This process took under 4 days in total to establish a final scoring system.

Given estimated marginal probabilities $P[Y_i = j|T_i = k, \{\boldsymbol{\theta}_k\}_{k=1}^K]$ for a single group trial, we can obtain a mean utility score for treatment k by computing:

$$\bar{U}(k) = \sum_{j=1}^{J} U_j P[Y_i = j|T_i = k, \{\boldsymbol{\theta}_k\}_{k=1}^K]. \tag{3}$$

Where $P[Y_i = j|T_i = k, \{\boldsymbol{\theta}_k\}_{k=1}^K]$ are determined using the estimated cumulative regression model for $P[Y_i \leq j|T_i = k, \{\boldsymbol{\theta}_k\}_{k=1}^K]$. Since we are using a Bayesian approach, we can obtain a posterior distribution of the mean utility scores $\bar{U}(k)|\mathcal{D}_n$ for each $k = 1, \ldots, K$, which we will use in decision making throughout the trial. Here $\mathcal{D}_n = \{Y_i, T_i, \boldsymbol{X}_i\}_{i=1}^n$ is the dataset of patients after n patients have been enrolled in the trial.

We will make M interim looks throughout the trial, after $n = n_1, \ldots, n_M$ patients have had their outcomes evaluated. For our proposed design, this sequence of interim looks was predetermined by clinicians, but it could have been optimized based on design operating characteristics (OCs). Our goal is to stop a treatment k if any other treatment l has a clinically meaningful improvement in terms of mean utility, arbitrarily denote $\Delta > 0$, over treatment k. We will stop treatment k after n patients are enrolled into the study if:

$$\max_{l \neq k} P[\bar{U}(l) > \bar{U}(k) + \Delta|\mathcal{D}_n] > c(n, \Delta, J, \mathcal{A}_n),$$

where $\mathcal{A}_n$ denotes the set of active treatments after enrolling n patients are enrolled and $c(n, \Delta, J, \mathcal{A}_n) > 0$ is a decreasing boundary function as n and Δ increase and as the number of active treatments (denoted $|\mathcal{A}_n|$) decreases. If the posterior probability that any treatment $l \neq k$ has a utility higher than k by at least $\Delta > 0$ is sufficiently high, we will stop treatment arm k. Under this framework, It's possible to stop multiple treatment arms at each interim analysis - if several treatments are showing futility compared to a promising therapy. If all but one treatment has been stopped, the trial ends and the last remaining treatment is declared superior. For the cutoff function, we assume that:

$$c(n, \Delta, J, \mathcal{A}_n) = \gamma_0 + exp\left(-\gamma_1 \Delta - \gamma_2 \frac{n}{J|\mathcal{A}_n|}\right),$$

where $\gamma_0, \gamma_1, \gamma_2 > 0$ are design parameters that are calibrated to obtain desired type I error probabilities and power under a wide variety of simulation scenarios. This function approaches γ_0 as Δ, n increase and $|\mathcal{A}_n|, J$ decrease. This decreasing decision boundary is necessary due to the conservative nature of the pairwise null clustering prior on the treatment effect vectors. By clustering the treatment specific parameters that govern the multinomial distribution, this prior also clusters the utilities which makes stopping for a utility difference less likely, which happens with a non-negligible posterior probability when two treatments have true utilities that are close to each other but not equal.

This gives us additional flexibility to test multiple increasing values of Δ, since the evidence required to declare a treatment inferior will decrease as Δ increases. A vector $\boldsymbol{\Delta}$ of increasing clinical significance could be constructed to facilitate operating characteristics. In the trial at UMC, $\boldsymbol{\Delta} = (2.5, 5, 10, 15, 20)$ was chosen. The decreasing nature of the function $c(\cdot)$ is constructed to require less evidence for stopping a treatment for a utility improvement of 10 compared to a minor improvement of 2.5. These choices represent small, moderate, and large utility increases on the $[0, 100]$ range of plausible values. Additionally, we impose a stopping rule for utility equivalency (i.e. for futility) between the active treatments $\mathcal{A}_n$ in a trial after the first interim look based on the criterion

$$\min_{l,k \in \mathcal{A}_n} P[|\bar{U}(l) - \bar{U}(k)| < \min(\boldsymbol{\Delta}) | \mathcal{D}_n] > c(n, min(\boldsymbol{\Delta}), J, \mathcal{A}_n)$$

This implies that if the posterior probability that all active treatments have utilities within $\min(\boldsymbol{\Delta})$ of each other is high, that the trial will be stopped early and the treatments $\mathcal{A}_n$ will be declared equally optimal in terms of their effect on the ordinal patient outcome. Figure 1 displays three possible trial replicates under the proposed stopping rules with $M = 5$ interim looks, $K = 3$ treatments, and a single value of $\boldsymbol{\Delta}$ to test (i.e. a 1-vector). Here P_k denoting the posterior probability that treatment k is inferior to the best performing treatment, and P_{Fut} denoting the posterior probability that all treatments in $\mathcal{A}_n$ have mean utilities within $min(\boldsymbol{\Delta})$. C_Δ denotes $c(n, \Delta, J, \mathcal{A}_n)$ for a given J, n, and $\mathcal{A}_n$ at that point in the trial.

In the first trial replication, treatments 2 and 3 are declared optimal and the trial is stopped at look 4. In the second trial replication, all treatments are continued at the first interim look and at the second interim look, both treatments 1 and 2 are declared inferior to treatment 3, ending the trial. In trial replication 3, all treatments are continued until

the final interim decision, when treatment 1 is stopped due to inferiority. The probability P_{Fut} is then computed only for treatments 2 and 3. Since this value is $< c_\Delta$, we do not declare treatments 2-3 to be different in terms of mean utility.

m	c_Δ	P_1	P_2	P_3	P_{Fut}	Comment
1	.99	1.00	0	.85	–	Drop treatment 1
2	.97	–	0	.75	.2	Continue with treatments 2,3
3	.87	–	0	.2	.5	Continue with treatments 2,3
4	.75	–	0	.05	.95	Stop Trial, Declare treatments 2-3 optimal

m	c_Δ	P_1	P_2	P_3	P_{Fut}	Comment
1	.99	.90	.75	0	–	Continue all treatments
2	.97	.99	.98	0	0	Stop trial and declare treatment 3 to be optimal

m	c_Δ	P_1	P_2	P_3	P_{Fut}	Comment
1	.99	.7	0	.85	–	Continue all treatments
2	.97	.65	0	.75	0	Continue all treatments
3	.87	.69	0	.70	0	Continue all treatments
4	.75	.5	0	.45	0	Continue all treatments
5	.65	.7	0	.33	.55	Treatment 1 is stopped P_{Fut} is calculated based on $\|\mathcal{A}_n\| = 2$, trial ends.

Figure 1: Three possible trial results for $M = 5$ interim decisions and $K = 3$ treatments, and a single value of $\mathbf{\Delta}$ to test. P_k is the posterior probability that treatment k is $min(\mathbf{\Delta})$ inferior to the most superior treatment and P_{Fut} is the posterior probability that all active treatments are $min(\mathbf{\Delta})$ optimal.

The MABOUST design requires the trial statistician to work with the trial team to specify 5 different sets of parameters.

- $n_1, .., n_M$: The sample sizes for each interim look.
- $\mathbf{\Delta}$: The set of clinical improvements to try.
- $(\gamma_0, \gamma_1, \gamma_2)$: the parameter vector used to determine the cutoff function $c(\cdot)$.
- p_ζ: The prior probability of a pairwise null hypothesis. Setting $p_\zeta = 0$ indicates that no clustering will take place.
- Whether or not covariates will be adjusted for in the analysis.

We recommend setting $p_\zeta = .9$, $\gamma_0 = .5$, and $\gamma_1, \gamma_2 < .2$. These parameter settings decrease the likelihood of a type I error. We explore how different choices of γ_1, γ_2 affect operating characteristics in the simulation study. Simulation results show that covariate adjustment does not affect trial accuracy, but does lead to trials with smaller sample sizes. This reduces the decision-making from clinicians to only consider choices for $\boldsymbol{\Delta}$ and $n_1, \ldots, n_M$ in most cases.

The Bayesian approach in particular makes it easier to control family-wise error rates by using Bayesian clustering. This is shown in our simulation study where we compare the proposed clustering approach to a frequentist proportional odds approach, a Bayesian model without clustering, and a model that uses permutation tests and empirical outcome frequencies. MABOUST with pairwise-null clustering reduces the probability of type I errors across these multiple pairwise comparisons.

4 Simulation Study

To show the potential benefit of various aspects of MABOUST, we investigate the operating characteristics of 4 modeling schemes: (1) clustering and covariate adjustment, (2) no clustering and covariate adjustment, (3) clustering and no covariate adjustment, and (4) no clustering or covariate adjustment. We also compare each of the 4 explored methods to a frequentist approach that assumes a proportional odds model and a permutation test based on empirical mean utilities. We set $p_\zeta = .9$ as described in Section 2 to encourage the pairwise null hypothesis between each set of two treatments. When we do not allow clustering treatment effects, we set $p_\zeta = 0$ and compare operating characteristics to when clustering is allowed.

4.1 Simulation Parameters and Operating Characteristics of Interest

We simulate trial replications for 1,000 randomly generated scenarios for 3 and 5 treatments with truly proportional odds and non-proportional odds treatment-outcome relationships. For the 3-treatment trial, we make sequential decisions after 100, 200, 400, and 600 patients are enrolled. This was the proposed group sequential decision structure of the trial to be conducted at UMC. For 5 treatment groups, we make interim decisions after 200, 400, 600, 800, 1000 patients are enrolled in the trial. We planned our maximum sample sizes for the two trials by allotting 200 patients per treatment, which we felt could adequately explore the 6 ordinal outcomes, particularly for events that might have true probabilities of occurrence that are $< .05$. As stated in Section 3, we use a utility improvement grid of $\boldsymbol{\Delta} = (2.5, 5, 10, 15, 20)$ and set $(\gamma_0, \gamma_1, \gamma_2) = (.5, .05, .05)$. In the bottom row of Figure 2, display the cutoff functions over the sample sizes from 0 to 1000 for the chosen $\boldsymbol{\Delta}$ values (including 0 for declaring an optimal set), and 3 and 5 treatments.

The two cutoff decision boundary plots are shown on the scale of $n = 0$ to $n = 1000$, which reflects the group sequential boundaries for the planned 5-armed MABOUST trial. We show this size for both, because the set of active treatments $\mathcal{A}_n$ are used in the

decision boundary function $c(n, \Delta, J, \mathcal{A}_n)$. If a 5-armed trial is reduced to 3 treatments through stopping two arms for inferiority, the boundaries will thereafter resemble the decision cutoffs for 3 treatments - unless another treatment is stopped thereafter. By setting $\gamma_0 = .5$, we must have that $c(n, \Delta, J, \mathcal{A}_n) \geq .5$ for all function values. However, note that the decision boundaries for the 3-armed design converge to .5 much quicker than those for 5 treatments. This aspect of the trial is a novel proposal in multi-armed designs and allows for more aggressive decision making as the trial hones in on optimal therapies. We argue that this boundary decrease is justified, because we have already ruled out several treatments as optimal by this time. We also like to note that with increasing Δ, the curves shift downwards. With $\Delta = 0$, which is used in determining whether the trial should stop and declare a set of treatments equally optimal, the decision boundary is 1 (permitting no stopping) until $n = 300$ for a 3-armed trial and $n \approx 500$ for a 5-armed trial. After treatments are dropped from the design this decision becomes easier due to decreased futility boundaries.

We compare our proposed Bayesian method to a frequentist proportional odds (PO) cumulative logistic regression model. For the PO model, we perform pairwise hypothesis tests for each pair of treatments, testing whether the proportional odds effect is at least 1.05. This reflects an increase of 5%, which is similar to $min(\boldsymbol{\Delta}) = 2.5$ for an average utility value of 50. We drop a treatment if the p-value for testing any pairwise hypothesis is less than $\frac{.05}{M}\binom{K}{2}^{-1}$ which should have a type I error rate below .05 due to Bonferroni correction. If all pairwise p-values for testing whether the PO effect is 1.05 is above .50 for all $\mathcal{A}_n$, we stop the trial and declare all treatments equally optimal. This setup creates a similar version of the decision structure used for MABOUST.

We also compare MABOUST to a permutation based method using empirical utilities. We first calculate the pairwise difference in mean utilities between each treatment pair, and use permutation tests to establish pairwise p-values testing whether each treatment are $min(\boldsymbol{\Delta})$ equivalent. Formally, using the empirical probabilities of each treatment/outcome combination, calculate the mean utilities $\bar{U}_k$ and $\bar{U}_l$, and calculate a test statistic $\bar{U}_k - \bar{U}_l - min(\boldsymbol{\Delta})$. If this number is > 0, (indicating treatment k has superiority over treatment l), we permute data 1,000 times between treatments k and l, computing $\bar{U}_k - \bar{U}_l - min(\boldsymbol{\Delta})$ each time using the permuted empirical treatment/outcome probabilities. We calculate the permutation test p-value as the proportion of times where our observed test statistic $\bar{U}_k - \bar{U}_l - min(\boldsymbol{\Delta})$ is bigger than the permuted values. We stop enrolling patients in a treatment arm if this permutation p-value is less than the same threshold used for the proportional odds model, which are Bonferroni-adjusted to ensure type I error control across multiple pairwise comparisons and interim looks.

For operating characteristics (OCs) of each design, we explore: The familywise type I error rate, $FWER$, which is the probability of declaring a treatment l inferior, when $|U_l^{True} - max_k\, U_k^{True}| < min(\boldsymbol{\Delta})$. We also report the generalized power, GP. This is the probability that we correctly decide on the inferiority of every treatment based on the smallest value of Δ considered. For example, if $U_1^{true} = 45, U_2^{true} = 53$ and $U_3^{true} = 51$, the correct decision is to stop treatment arm 1 for inferiority and continue the entire trial until the end or treatments 2 and 3 have been declared equally optimal. We record the probability of dropping truly $min(\boldsymbol{\Delta})$ inferior treatments and display the average

probability of making a correct decision about a treatment in supplemental table 1. Lastly, we record the average sample size of the trial, $\bar{n}$, and show average sample size standard deviations along with .25, .75 quantiles in supplemental table 1.

For the remainder of this section we describe the simulation methods used to generate realistic ordinal outcome probabilities, based on clinical research in COVID-19 patients. We list the trial parameters $\boldsymbol{\gamma}$ in this subsection and discuss graphically how the decision boundaries behave for various n, Δ and $\mathcal{A}_n$. We then describe simulation results for a 3- and 5-armed MABOUST clinical trial under the 4 considered model settings and discuss sensitivity of the results to various design parameter changes.

4.2 Scenario Generation

Since we randomly generate a large set of scenarios under instead of examining 8-10 different choices, we avoid arbitrary choices that might show our method is particularly effective compared to the PO model. We simulate 1,000 replications of the trial under 1,000 randomly generated scenarios, where several treatments may have identical outcome probabilities.

Table 1 displays the outcome structure and elicited utility scores for each outcome in each scenario, and the trial being conducted at UMC. We also display two probability ranges, $\boldsymbol{R}_1 = (R_{11}, \ldots, R_{1J})$ which denotes the plausible range of probabilities for each outcome for a generic treatment, and $\boldsymbol{R}_2 = (R_{21}, \ldots, R_{2J})$ which denotes the plausible range of probabilities for each outcome for particularly frail or strong patients. We generate random scenarios under true proportional odds and truly non-proportional odds relationships. For each randomly generated scenario, we do the following:

1. Randomly generate the number of clusters of equivalent treatments from a discrete uniform distribution on $1, \ldots, K$.

2. Based on the randomly chosen number of clusters, randomly assign treatment labels $1, \ldots, K$ to each cluster.

3. For each treatment $k = 1, \ldots, K$, randomly generate the true probabilities of each outcome $\pi_j^k = P[Y_i = j | T_i = k]$, ensuring that $\pi_j^k \in R_{1j}$, and that $\pi_1^k + \ldots + \pi_J^k = 1$. Assign all marginal probabilities for a treatment l clustered with treatment k equal, i.e. $(\pi_1^l, \ldots, \pi_J^l) = (\pi_1^k, \ldots, \pi_J^k)$. This ensures that all marginal probabilities of each outcome for the K treatments fall in the range of possible outcomes. For example, it's unrealistic to expect a miracle cure, where the probability of event 6 (Discharge) is .95 or a terrible treatment with the probability of event 0 (Death or ECMO) of .9. These probabilities give us the true values of $\theta_{k,j} = logit(\pi_1^k + \ldots + \pi_j^k)$.

   For simulation scenarios generated under the proportional odds relationship, the values $\pi_1^1, \ldots, \pi_J^1$ are randomly generated according to the above procedure. Then we randomly generate proportional odds effects $\gamma_k \sim U[-1, 1]$. New marginal cumulative probabilities $P[Y_i \leq j | T_i = k]$ are calculated for treatment k using the

randomly generated values of $P[Y_i \leq j|T_i = 1]$ and γ_k via

$$logit[P[Y \leq j|T_i = k, \gamma_k] = logit[P[Y \leq j|T_i = 1] + \gamma_k,$$

and these probabilities are transformed to obtain $(\pi_1^k, \ldots, \pi_J^k)$ which are then checked to see if each $\pi_j^k \in R_{1j}$. If not, the value of γ_k is discarded due to leading to an implausible outcome structure in COVID-19. Afterwards, based on the randomly generated clustering configuration, some γ_k values are set to 0 and others are set to γ_l for some $l \neq k \neq 1$. This gives the frequentist proportional odds model a fair chance to compete with MABOUST, which assumes an NPO model, and also allows us to explore the robustness of each method to these assumptions.

4. Randomly generate, $\boldsymbol{\beta}$, the effect of $\boldsymbol{x}_i$ on outcome given treatment T_i. This is done by proposing a candidate vector $\boldsymbol{\beta}$ where each entry is randomly generated from a uniform distribution from -2 to 0, which implies that increased values of $\boldsymbol{x}_i$ are associated with poorer outcomes. This is done by using the true values of $\{\{\theta_{k,j}\}_{j=1}^J\}_{k=1}^K$ and generating probabilities for 10,000 randomly chosen $\boldsymbol{x}_i$ and T_i vectors, i.e.

$$logit[P[Y \leq j|T_i = k, \boldsymbol{X}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\beta}] = \theta_{k,j} + \boldsymbol{X}_i\boldsymbol{\beta}.$$

Finally, we check if all probabilities $P[Y = j|T_i, \boldsymbol{X}_i, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\beta}] \in R_{2j}$, which implies that all hypothetical patients generated in the study will have reasonable probabilities of each outcome.

5. Given the values $\{\{\theta_{k,j}\}_{j=1}^J\}_{k=1}^K$, $\boldsymbol{\theta}$, calculate the true values of $\bar{U}(1), \ldots, \bar{U}(K)$ and simulate 1,000 trial replications - including stopping treatments after $n_1, \ldots,$ n_M patients are enrolled. We will keep track of the OCs described above.

The ranges R_1 are used to specify true probabilities for each treatment, without considering covariate effects, for use in simulation. All generated true probabilities of outcome j and treatment k must fall in the range $R_{1,j}$ to be admissible. The ranges R_2 are used to calibrate the true effect of $\boldsymbol{X}_i$ on these probabilities. Given $\boldsymbol{X}_i$, all generated true probabilities of outcome j and treatment k must fall in the range $R_{2,j}$ to be admissable. This determines possible values used for $\boldsymbol{\beta}$. For the UMC trial, we have that $\boldsymbol{X}_i = \{Age_i, CCI_i, O_i\}$. Here Age_i is the numerical age group of patient i $\{0-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80+\}$ with these outcomes generated with probability $\{.11, .16, .18, .20, .18, .17\}$ which reflects empirical probabilities of each age group seen in Louisiana on April 12th (Whitfield and Swenson (2020)). We number these age groups from 1 to 7.

CCI_i is the modified Charleston Comorbidity Index (CCI) excluding age and including obesity and smoking status as an extra point added to the score. The individual CCI_i scores were generated from a Poisson distribution with mean 2, which gave probabilities of the scores $\{0, 1, 2, 3, 4, 5+\}$ of $\{.14, .27, .27, .18, .09, .05\}$ which reflects probabilities seen by the clinicians on the study. Finally $O_i = 1, 2, 3$ denotes the disease severity of each patient when they enroll in the study, as definite in order by events

(5) Hospitalization without supplemental oxygen, (3) Hospitalization with other supplemental oxygen, and (2) Hospitalized, on high flow oxygen therapy or non-invasive mechanical ventilation. The probabilities of each enrollment outcome were assumed to be .5, .4, .1, respectively to reflect the decreased likelihood that a patient will immediately need higher flow oxygen upon arrival. These probabilities were based on reporting outcomes in COVID-19 patients in an early study (Wang et al. (2020)) and determined from discussions of Dr. Clement, Dr. Bennani, and Julio Figueroa from Touro Medical Center. The elicited values shown in Table 1 reflect a clinicians opinion on relative patient clinical benefit for each outcome, which reflect patient disease courses that were studied early in the pandemic.

Outcome #	U_j	R_1	R_2	Description
1	0	.10-.35	0-.50	Death, on ECMO or invasive mechanical ventilation
2	35	.10-.30	0-.50	Hospitalized, on high flow oxygen therapy
3	65	.40-.70	0-.8	Hospitalization with other supplemental oxygen
4	75	0-.10	0-.45	Discharge with supplemental oxygen
5	85	.10-.30	0-.45	Hospitalization without supplemental oxygen
6	100	0-.10	0-.30	Discharge without supplemental oxygen

Table 1: Possible patient outcomes, utility scores for each outcome, and a range of plausible true probabilities.

The second column of Table 1 assigns a numerical utility score, denoted U_j for each outcome j, to each patient outcome, with 0 being the lowest and 100 being the highest. The goal will be to find the treatment with the highest mean utility score, as defined by (3). The distribution of these randomly generated marginal utilities based on the probability ranges R_1, R_2, and the utilities U_j are shown in Figure 2 (top left). The randomly generated $\boldsymbol{\beta}$ values associated with age and CCI are plotted in a scatterplot in the top right of Figure 2.

First we note that the majority of the simulated true utility values under this structure fall between 42 and 52, constituting a 10 point difference in utilities. The histogram sizes for randomly drawn utilities in the range 40-42 and 52-58 and are about half that as those between 42-52, with a very small amount of simulated utilities being ≥ 60 or < 40.

The top right of Figure 2 shows that as the age effect increases, the oxygen effect must decrease for the simulation scenario to be grounded in possible reality. This is because increasing either of these variables, where age groups are numbered from 1,...,8 and oxygen status is numbered 1, 2, 3, will drastically increase the true probability of death of mechanical ventilation for a hypothetical patient. Given that $R_{2j} = 0 - .50$, it is not permitted that the probability of death/ECMO is above .50 for any patient, regardless of treatment status or covariates. While we do not plot the randomly drawn coefficients here for the CCI effect on the ordinal outcome, we note that there are similar shapes when comparing CCI to age group and O2 status.

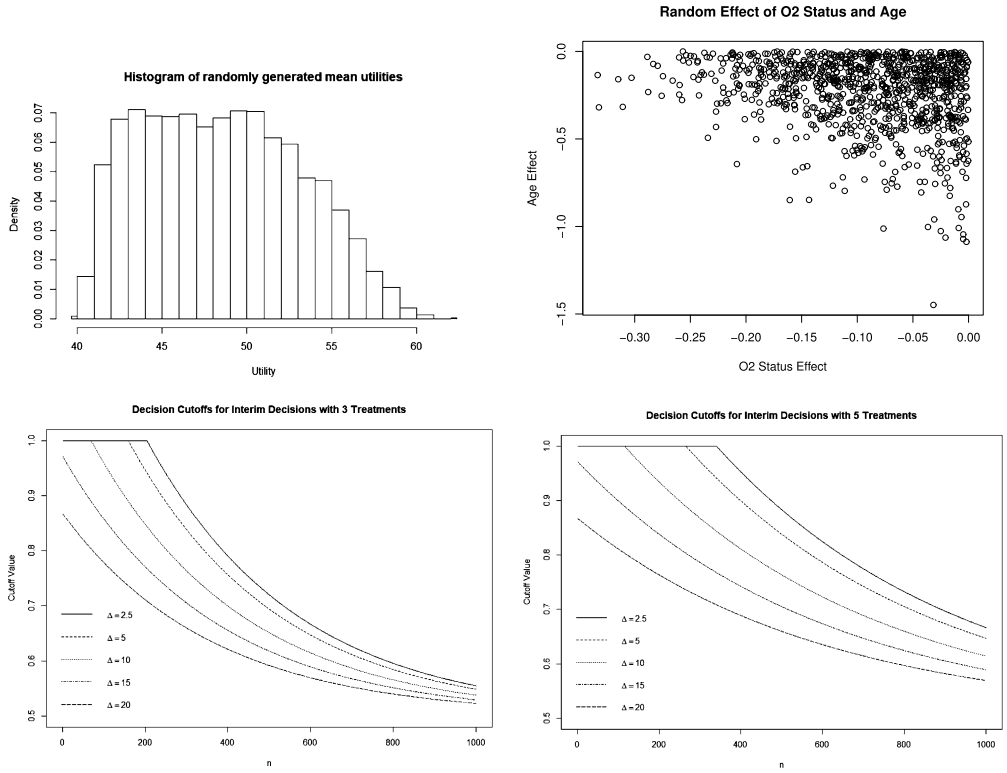

Figure 2: Simulation Scenario Settings. Top Left: Marginal distribution of randomly generated utilities according to Table 1. Top Right: Effect sizes on age and O2 status for randomly generated scenarios according to Table 1. Bottom Left: Cutoffs for interim decisions for $K = 3$. Bottom Right: Cutoffs for interim decisions for $K = 5$.

4.3 Results for 3 Treatment Arm

For three subgroups, we mimicked the trial design proposed for conduct at UMC medical center. Figure 3 displays the average operating characteristics of the 4 MAOBUST designs considered for proportional odds (top) and non-proportional odds (bottom) data generation structures. We also show the average operating characteristics for the frequentist proportional odds model and the permutation test based on empirical probabilities. We examine overall trends, as well as average operating characteristics when 1, 2, or 3 null clusters are present. Note that even when none of the treatment-outcome relationships are identical (3 null clusters), we could have true utilities that do not differ by $min(\boldsymbol{\Delta}) = 2.5$. Within each null cluster grouping, we display in order:

- The frequentist proportional odds results.
- The permutation test results that uses empirical utilities.

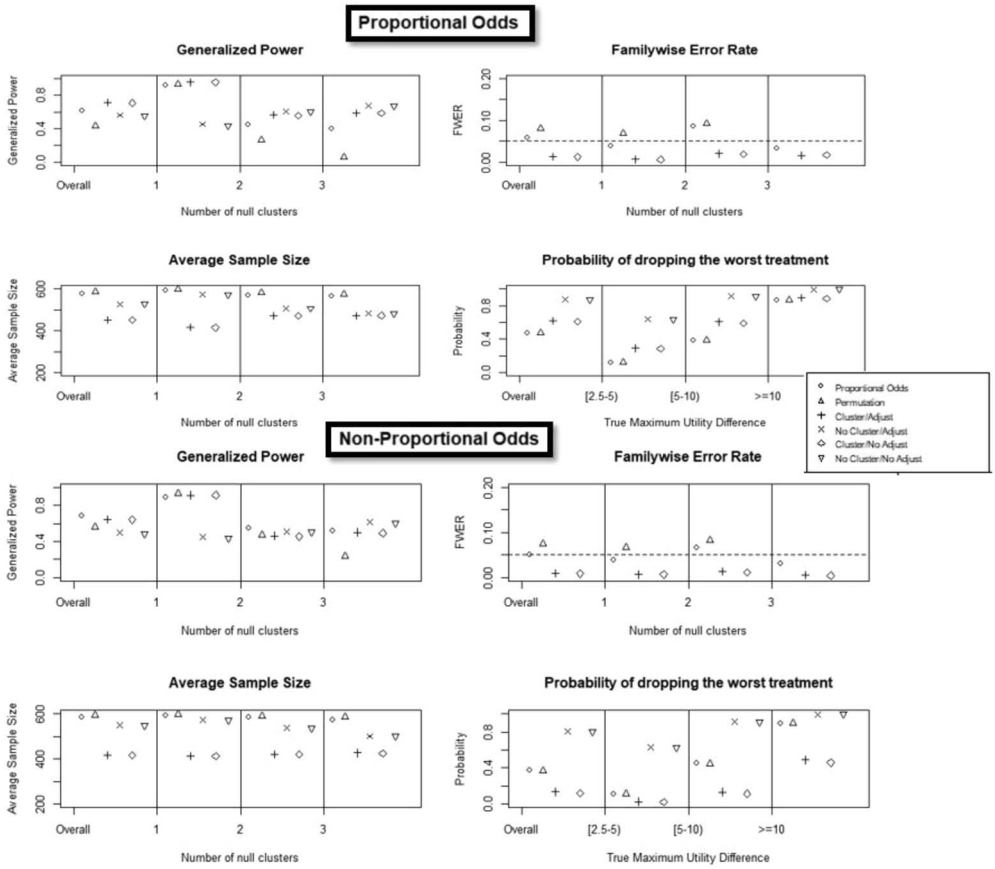

Figure 3: 3-armed design operating characteristics for the 4 MABOUST settings considered and the PO simulation truth (top) and NPO simulation truth (bottom). For each operating characteristic shown, the average across 1,000 simulations is reported for each bin. Here the number of null clusters refers to the number of treatment groups that have a true utility distances above $min(\boldsymbol{\Delta}) = 2.5$ of each other. The probability of dropping the worst treatment displays the average probability of dropping the treatment with the smallest utility, given there exists a treatment with an improvement of at least $min(\boldsymbol{\Delta}) = 2.5$ exists.

- MABOUST with clustering ($p_\zeta = .9$) and covariate adjustment.
- MABOUST without clustering and covariate adjustment.
- MABOUST with clustering ($p_\zeta = .9$) and no covariate adjustment.
- MABOUST without clustering and no covariate adjustment.

We first turn our attention to the proportional odds simulation truths in Figure 3, which are represented by the top 4 graphs. These overall numerical results are also shown in Table 2. Overall, for a proportional odds truth, the average generalized power (top left) is .69 for the PO model and .56 for the permutation model. This is compared to .65 for the MABOUST design with clustering and adjustment and .64 for clustering without adjustment. MABOUST designs without clustering had average generalized power of .49 and .48, which was much less desirable - most likely due to high family wise error rates (FWER) of .43 and .44 for adjustment and non-adjustment models, respectively. Average generalized power was .02 higher when only one null-cluster was present for MABOUST with clustering compared to the PO model, but was lower than the permutation method by .02. FWER was well controlled by the PO method (average of .05) and the two MABOUST designs with clustering (.01 each), it was slightly higher in the permutation model (.07).

The permutation and PO methods also had higher probabilities of dropping the worst treatment when there is a clear improvement (.37) compared to .14 for MABOUST with clustering and .81 for MABOUST without clustering. This latter value should not be interpreted as favoring MABOUST without clustering, since the FWER for both of these are above .40, which is unacceptable. When the data truly has a proportional odds treatment-outcome structure, the PO and permutation methods provide slight improvements in terms of generalized power and a doubling in the probability of dropping inferior treatments. However, the average sample size is drastically lower for MABOUST with clustering (417.9) vs it's frequentist competitors (587 and 593, respectively). Supplemental table 1 displays the average standard deviations of the trial sample size along with .25 and .75 quantiles. The .25 and .75 quantiles on the trial sample size were much lower for MABOUST with clustering (400 and 408) than for the frequentist comparators ($\geq$ 593 and 600) which is also reflected in a larger average standard deviation for MABOUST with clustering (56 and 85) compared to the frequentist methods (45 and 30). The correct decision percentages (i.e. average % of correct treatment conclusions across trials) are also shown in this table, which show a gain of about .04 for the PO model compared to the two MABOUST clustering approaches.

For non-proportional odds simulation truths, MABOUST with clustering has higher average generalized power (.71 for both adjustment schemes) compared to the proportional odds model (GP = .62) and the permutation test (GP = .43). The standard deviations of these generalized powers are also lower for MABOUST than it's competitors (.32 vs $\geq$ .39). The MABOUST designs that do not allow clustering performed more poorly than the PO model, with an average $GP = .55, .56$. This further illustrates the strength of the pairwise clustering in MABOUST, which is the biggest innovation presented in this manuscript. Consistently, the MABOUST designs with clustering outperformed the PO and Permutation based alternatives, with an improvement in GP of at least .04 for 1 null-cluster, .11 for 2 null-clusters, and .19 for 3 null-clusters (i.e. all different treatment-outcome relationships). The standard deviations on the generalized power across these null-cluster subsets are also smaller for MABOUST than for the PO model, indicating a more consistent performance in terms of generalized power across these random scenarios.

Average FWER rates for the PO and Permutation models were .06 and .08, respectively, compared to .01 for the MABOUST clustering designs. The MABOUST designs that did not allow clustering had terrible type I error control with FWER rates of .42 and .43, respectively, and are not shown in Figure 3. FWER for MABOUST was controlled below .05 for simulation truths with varying numbers of null-clusters. The probability of dropping the worst treatment (when one exists, i.e. has a utility at least 2.5 less than the optimal treatment) was also consistently higher for MABOUST with clustering (.61, .62) than for the PO and Permutation models (.47, .47). This probability was much higher for MABOUST without pairwise-null clustering (.88, .88), but this is coupled with a FWER that is too high to recommend using this design. MABOUST with clustering had higher probabilities of dropping the worst treatment for true maximum utility differences of $(2.5, 5]$, $(5, 10]$, and 10+.

Lastly, the average trial sample size across these randomly generated scenarios was much lower for MABOUST designs with clustering (451.6, 451.8) compared to the PO model (579.1) and permutation model (585.6). Average standard deviations and .25, .75 quantiles are shown in supplemental table 1 across the random scenarios. The average .25 quantile, and .75 quantiles were all lower for MABOUST with clustering compared to the PO and permutation models. For the standard deviations, MABOUST had higher values (64,65) compared to the PO and permutation tests (53 and 37). This is likely due to MABOUST's average sample size of 451, which is much lower than the average sample sizes for the other methods that are at least 580, and the fact that the maximum trial sample size allowed is 600. The correct decision percentage is about .04 higher than the PO model under the NPO simulation truth (supplemental table 1).

Collectively, these results suggest that MABOUST without clustering is an unreliable method in $K = 3$ treatments due to unreasonably high FWER. However, MABOUST with clustering provides an improvement in generalized power under the non-proportional odds truth (.09) that is higher than this loss in the proportional odds truth (.04). Across both the proportional and non-proportional odds simulation truths, MABOUST with clustering produces smaller average sample sizes and .25, .75 quantiles of average sample sizes. The type I error rate is also well controlled. Adjusting for covariates did not seem to make a difference in the two MABOUST models in terms of generalized power or average sample size, but produced a smaller sample size standard deviation.

4.4 Results for 5 Treatment Arms

Figure 4 displays the average operating characteristics across 1,000 randomly generated scenarios with $K = 5$ treatments. Results for truly proportional odds treatment-outcome relationships are shown in the top 4 plots, while results for non-proportional odds relationships are shown in the bottom 4 plots.

For $K = 5$ and a true proportional odds relationship, the average generalized power for both the PO model and MABOUST with clustering and covariate adjustment are both .47. $GP = .46$ for MABOUST without covariate adjustment. Similar to $K = 3$, MABOUST without clustering performs poorly with $GP = .24, .23$ and FWER of

.62, .63 for the methods with and without covariate adjustment, respectively. FWER was higher, on average for the PO and permutation models (.08 and .11, respectively), which were also reflected by higher standard deviations of FWER (.11 for both). FWER was much better controlled for the MABOUST designs with clustering (.03 each, with a standard deviation of .04). Overall 81.6% of scenarios for clustering MABOUST with covariate adjustment (83% without adjustment) had FWER $< .05$ compared to 66% for the PO model and 31.2% for the permutation method. The probability of dropping the worst treatment (when one exists) was still higher for PO models (.34) than for MABOUST with clustering (.17). Here with $K = 5$, the high FWER rate for PO and permutation models despite Bonferroni indicate superiority of MABOUST, since generalized powers are similar.

MABOUST with clustering had an overall improvement in average sample sizes (596.8 and 592.3) compared to the PO and permutation models (997.7 and 997.3). The average .25 and .75 quantiles for trial sample size (shown in supplemental table 1) were also lower for MABOUST with clustering and adjustment (479 and 661) compared to the PO model (999 and 1000) and permutation model (998 and 1000). Interestingly, when covariate adjustment was not done, MABOUST with clustering had higher average .25 and .75 quantiles of trial sample size (759 and 970). This is the first suggestion presented that covariate adjustment can be helpful in terms of operating characteristics when MABOUST with clustering is used. Average standard deviations of the trial sample size reflect these quantiles for MABOUST (> 160 for all 4 approaches). The PO and permutation models have lower average sample size deviations (15 and 12), which reflects the .25 and .75 quantiles being close to the maximum sample size of 1,000. Despite similar generalized power, average correct decision rates were higher for the PO model (.76) than the two MABOUST models with clustering (.73). When weighing the generalized power, correct decision rates, FWER, and trial sample sizes, MABOUST with clustering provides an improvement for $K = 5$ when compared to the PO and permutation models when the data generation structure truly has a proportional odds treatment-outcome relationship.

Looking across simulations based on the number of null-clusters, we see that the generalized power is higher with 1, 2, and 3 null clusters compared to the PO model, but is slightly lower for 4 and 5 null-clusters. Average FWER is controlled for 1,2,3, and 4 null-clusters for MABOUST with clustering. There is a higher average FWER for the PO model with 4 null-clusters. Clustering MABOUST's average sample size was consistently the lowest across any number of null-clusters. MABOUST with clustering did suffer when the largest utility difference was greater than 10 in terms of dropping the worst treatment (.49 and .45) versus the PO model and permutation model (.83 each).

The bottom of Figure 4 displays the results for non-proportional odds treatment-outcome structures. Here MABOUST with clustering provides a drastic improvement in average generalized power (.62 and .61) compared to the PO model (.39) and permutation model (.25). MABOUST with clustering also had better control of average FWER (.06 and .06) compared to the PO model (.12) and the permutation model (.14). MABOUST without clustering again produced high average FWER (.56 and .57) making these approaches impractical in modern research.

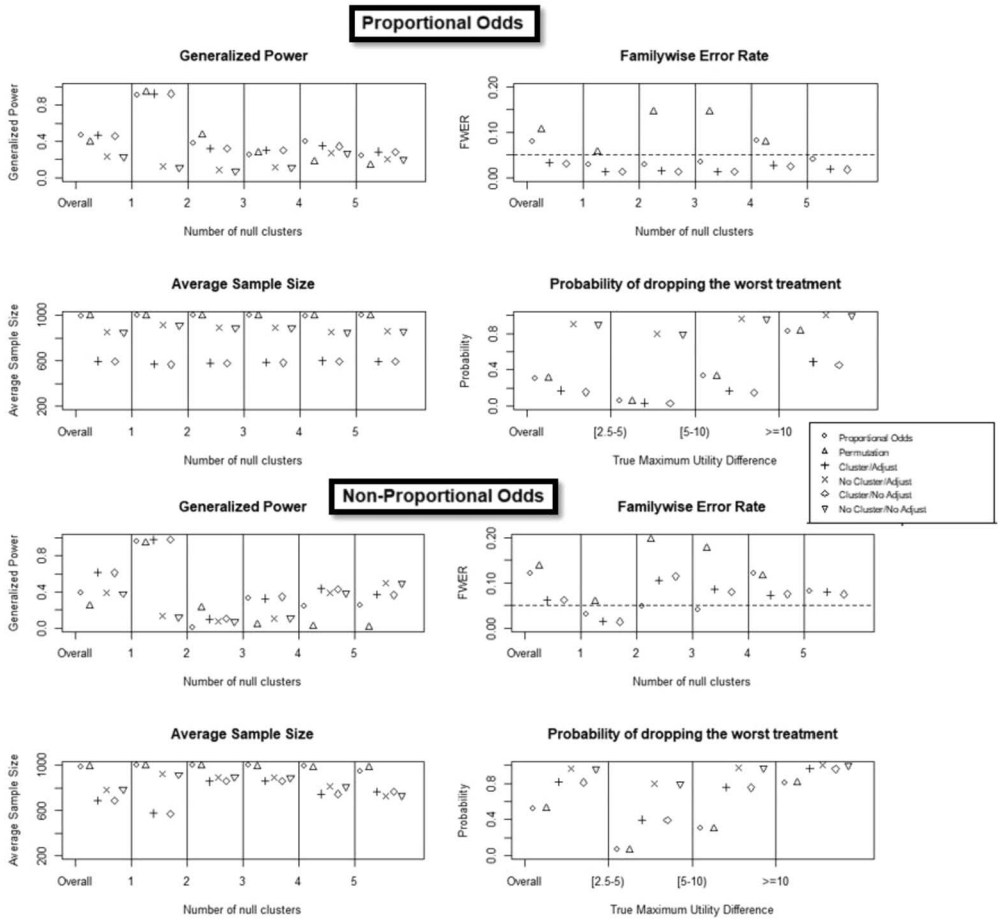

Figure 4: 5-armed design operating characteristics for the 4 MABOUST settings considered and the PO simulation truth (top) and NPO simulation truth (bottom). For each operating characteristic shown, the average across 1,000 simulations is reported for each bin. Here the number of null clusters refers to the number of treatment groups that have a true utility distances above $min(\boldsymbol{\Delta}) = 2.5$ of each other. The probability of dropping the worst treatment displays the average probability of dropping the treatment with the smallest utility, given there exists a treatment with an improvement of at least $min(\boldsymbol{\Delta}) = 2.5$ exists.

The average probability of dropping the worst treatment was also higher for MABOUST with clustering (.82 and .81) compared to the PO and permutation models (both .52). Average mean trial sample size was lower for MABOUST (690 for both) compared to the PO and permutation models (990 each). This is again reflected in supplemental table 1 by lower average .25 and .75 quantiles of trial sample size, and a higher average sample size standard deviation. Similar to results for proportional

odds treatment-outcome relationships, the average .25 and .75 sample size quantiles were lower for the models that used covariate adjustment (564 and 810) compared to MABOUST without covariate adjustment (685 and 906).

Across varying numbers of null-clusters, MABOUST with clustering had higher generalized power than the PO model. It also had lower average FWER, higher average probabilities of dropping an inferior treatment and lower average sample size. The permutation method had higher generalized power than MABOUST (.22 compared to .11) with 2 null clusters, but also had higher FWER (.2 vs .11). The results for $K = 5$ paint a stronger picture for the relative OCs improvement of MABOUST with clustering compared to the PO and permutation models, whether or not the true treatment-outcome relationship was proportional odds.

4.5 Sensitivity to $\boldsymbol{\gamma}$, Utility Choice, and p_ζ

In this section, we investigate sensitivity of MABOUST results to several design parameter choices, with average operating characteristics reported in Table 2 and supplemental table 1. These include $\boldsymbol{\gamma}$, which governs the adaptive decision boundary for MABOUST, p_ζ which is the pairwise null clustering probability, and a different utility function. For a differing utility function, we set utilities for events $1, \ldots, J$ to be $100 * (j-1)/(J-1)$ for $j = 1, \ldots, J$ which mimics a utility function that only uses the outcome indices but is scaled to $0, 20, 40, 60, 80, 100$ so that the previous choice of $\boldsymbol{\Delta}$ is applicable, instead of needing to shift this to the 1-6 range. The permutation test that uses empirical utility values also has sensitivity explored to this choice, listed as U_2 PERM in Table 2.

We investigate choices of $p_\zeta = .5$ and $p_\zeta = .75$ for the MABOUST design to determine how this hyperparameter, which governs conservativeness of the design in terms of pairwise comparisons, affects overall operating characteristics. Here a smaller value of p_γ indicate a decreased probability *a priori* of a pairwise-null relationship for any two treatments. Noting that the design vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ may have an effect on operating characteristics, we explored how various adjustments to γ_1 and γ_2 changed results for MABOUST. Since $\boldsymbol{\gamma}$ is used in the decision boundary function:

$$c(n, \Delta, J, \mathcal{A}_n) = \gamma_0 + exp\left(-\gamma_1 \Delta - \gamma_2 \frac{n}{J|\mathcal{A}_n|}\right),$$

increased values of γ_1 will allow for easier stopping for larger tested values of Δ. Increased values of γ_2 will allow for easier stopping as the trial enrolls more patients (i.e. n increases) or reduces the number of active treatments $|\mathcal{A}_n|$. The boundary function above also controls the likelihood of stopping a trial early for futility, and declaring treatments equally optimal - so increased values of γ_1, γ_2 may increase the likelihood of a type I error, while also increasing the likelihood of stopping for equivalency.

The results of the sensitivity analysis for testing $\gamma_1 = .05, .10$ and $\gamma_2 = .05, .10$ are shown in Table 2. We did not vary γ_0 since even though $c(n, \Delta, J, \mathcal{A}_n)$ is used solely for a boundary function, stopping one treatment for inferiority if $\max_{l \neq k} P[\bar{U}(l) > \bar{U}(k) + min(\boldsymbol{\Delta})|\mathcal{D}_n] < .50$ seemed unethical. If we can't demonstrate a $min(\boldsymbol{\Delta})$ utility

Setting	Generalized Power	FWER	P[Dropping Worst]	Nbar
		PO: 3 Treatments		
$\boldsymbol{\gamma} = (.05, .05)$	(.65,.49,.64,.48)	(.01,.01,.43,.44)	(.14,.12,.81,.81)	(418,417,550,549)
$\boldsymbol{\gamma} = (.05, .10)$	(.64,.30,.64,.29)	(.02,.02,.67,.68)	(.14,.13,.89,.89)	(403,403,434,436)
$\boldsymbol{\gamma} = (.10, .05)$	(.65,.45,.64,.44)	(.01,.01,.48,.49)	(.14,.13,.83,.83)	(414,413,531,530)
$\boldsymbol{\gamma} = (.10, .10)$	(.64,.29,.64,.28)	(.02,.02,.68,.69)	(.14,.13,.89,.89)	(402,402,420,422)
$p_\zeta = .75$	(.67,–,.67,–)	(.03,–,.03,–)	(.26,–,.26,–)	(448,–,448,–)
$p_\zeta = .50$	(.69,–,.68,–)	(.07,–,.07,–)	(.43,–,.43,–)	(506,–,506,–)
PO	.69	.05	.37	587
Perm	.56	.07	.01	593
U_2	(.67,.52,.66,.51)	(.01,.01,.4,.41)	(.14,.12,.81,.81)	(416,415,552,552)
U_2 Perm	.58	.07	0	594
		PO: 5 Treatments		
$\boldsymbol{\gamma} = (.05, .05)$	(.47,.24,.46,.23)	(.03,.03,.62,.63)	(.17,.15,.91,.90)	(597,592,851,852)
$\boldsymbol{\gamma} = (.05, .10)$	(.44,.15,.44,.15)	(.04,.04,.71,.71)	(.13,.12,.95,.95)	(439,437,612,614)
$\boldsymbol{\gamma} = (.10, .05)$	(.46,.21,.45,.2)	(.03,.03,.65,.66)	(.16,.14,.92,.92)	(521,517,796,797)
$\boldsymbol{\gamma} = (.10, .10)$	(.44,.15,.44,.14)	(.04,.04,.71,.71)	(.13,.12,.95,.95)	(432,431,572,573)
$p_\zeta = .75$	(.5,–,.49,–)	(.09,–,.09,–)	(.38,–,.38,–)	(770,–,770,–)
$p_\zeta = .50$	(.48,–,.47,–)	(.23,–,.23,–)	(.61,–,.61,–)	(890,–,890,–)
PO	.47	.08	.31	998
PERM	.39	.11		997
U_2	(.5,.25,.49,.23)	(.03,.03,.6,.61)	(.17,.15,.91,.91)	(563,559,850,850)
U_2 Perm	.42	.11	0	998
		NPO: 3 Treatments		
$\boldsymbol{\gamma} = (.05, .05)$	(.71,.56,.71,.55)	(.01,.01,.42,.43)	(.62,.61,.88,.88)	(452,452,529,528)
$\boldsymbol{\gamma} = (.05, .10)$	(.69,.41,.69,.40)	(.02,.02,.64,.65)	(.62,.61,.93,.93)	(410,411,413,414)
$\boldsymbol{\gamma} = (.10, .05)$	(.71,.53,.71,.52)	(.01,.01,.46,.47)	(.63,.62,.90,.89)	(442,442,507,507)
$\boldsymbol{\gamma} = (.10, .10)$	(.69,.4,.69,.39)	(.02,.02,.65,.66)	(.62,.61,.93,.93)	(407,408,399,400)
$p_\zeta = .75$	(.75,–,.74,–)	(.03,–,.03,–)	(.72,–,.72,–)	(473,–,473,–)
$p_\zeta = .50$	(.75,–,.74,–)	(.08,–,.08,–)	(.79,–,.79,–)	(505,–,505,–)
PO	.62	.06	.47	579
PERM	.43	.08	.01	586
U_2	(.72,.58,.71,.56)	(.01,.01,.38,.39)	(.6,.6,.87,.86)	(452,452,535,536)
U_2 Perm	.45	.08	0	589
		NPO: 5 Treatments		
$\boldsymbol{\gamma} = (.05, .05)$	(.62,.39,.61,.38)	(.06,.06,.56,.57)	(.82,.81,.96,.96)	(690,691,785,786)
$\boldsymbol{\gamma} = (.05, .10)$	(.54,.31,.54,.30)	(.08,.08,.65,.66)	(.74,.73,.98,.98)	(510,511,563,565)
$\boldsymbol{\gamma} = (.10, .05)$	(.6,.37,.6,.36)	(.07,.07,.59,.59)	(.8,.79,.97,.97)	(632,632,728,729)
$\boldsymbol{\gamma} = (.10, .10)$	(.54,.30,.53,.30)	(.09,.09,.66,.66)	(.73,.72,.98,.98)	(495,496,526,527)
$p_\zeta = .75$	(.65,–,.64,–)	(.12,–,.12,–)	(.89,–,.89,–)	(759,–,759,–)
$p_\zeta = .50$	(.60,–,.59,–)	(.24,–,.24,–)	(.93,–,.93,–)	(801,–,801,–)
PO	.39	.12	.52	989
PERM	.25	.14		990
U_2	(.61,.38,.6,.37)	(.06,.06,.56,.56)	(.79,.78,.96,.96)	(682,684,792,793)
U_2 Perm	.26	.14	0	995

Table 2: Average operating characteristics across 1,000 randomly generated scenarios. Generalized Power = average probability of making all comparative treatment decisions correctly, $FWER$ = average family wise type I error rate, P[Dropping Worst] = average probability of dropping the a truly suboptimal treatment, Nbar = average mean trial sample size. We report this for the PO (PO) and Permutation models (PERM and U_2 PERM), and for different MABOUST designs based on $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, p_ζ, and a different utility choice U_2. For MABOUST designs, we report the OCs, in order, for designs that cluster and adjust, don't cluster but adjust, cluster but don't adjust, and neither cluster nor adjust. Blanks $--$ exist for $p_\zeta = .75$ and $p_\zeta = .5$ since this parameter only explores MABOUST with clustering.

improvement is more probable than not based on the data, we should not make decisions regarding dropping treatments.

Table 2 displays the average operating characteristics for each of the adjustments to design considerations for MABOUST, permutation tests, and additionally displays the numerical results for the PO method. Within all rows related to design changes for MABOUST, we list the operating characteristics in parentheses (in order) for MABOUST (1) with clustering and adjustment, (2) without clustering and adjustment, (3) with clustering but no adjustment, and (4) without clustering and no adjustment. Since we have already demonstrated that MABOUST without clustering produces unreasonably high FWER and low generalized power, we focus discussion on the 1st and 3rd entries of each operating characteristic for each design parameter shift. We do not list values of (2) and (4) for different choices of p_ζ since this choice only effects MABOUST designs where clustering is allowed.

Generalized power for different choices of $\boldsymbol{\gamma}$ and MABOUST clustering models did not change by more than .02 for all simulation truths except for $K = 5$ treatments and an NPO truth. Here the generalized power was .62 for $(\gamma_1 = .05, \gamma_2 = .05)$ but dropped to .54 when $\gamma_2 = .10$ was used. This is partially reflected in the change in average FWER, which went from .06 for the displayed results in Figures 3 and 4 to .09 for an aggressive choice of $(\gamma_1 = .10, \gamma_2 = .10)$. Increasing γ_1 to .10 resulted in a smaller average sample sizes, which was more apparent for $K = 5$ than $K = 3$.

Decreasing p_ζ increased the average FWER in each K and PO/NPO relationship group. This was drastic for $K = 5$ treatments and $p_\zeta = .5$ which increased FWER to .23 and .24, respectively, for PO and NPO simulation truths. This prior setting equally favors pairwise null and alternative hypotheses. For $p_\zeta = .75$, these rates were .07 and .12 for $K = 5$ treatments, but were controlled below the .05 level for $K = 3$ treatments. Decreasing p_ζ was associated with an increase in the probability of dropping a truly inferior treatment, but results for generalized power were mixed - which relates to both null and alternative conclusions about treatment comparisons. Decreasing p_ζ also resulted in larger average sample sizes for each K and simulation truth, likely due to a decreased chance of stopping the trial early and declaring a set of treatments equally optimal.

Lastly, adjustments in the utility structure (U_2 and U_2 PERM) to reflect the ordinal outcome labels did not have a large effect on generalized power, FWER, or probability of dropping the worst treatment (maximum difference of .03, and in both directions). The effect of changing the utilities on average sample size was also not drastic (maximum difference of 30). Similar results were seen for the permutation test and a different utility function. Taken together these results indicate that the utility function has less of an impact on operating characteristics shown in Table 2 than the choice of $\boldsymbol{\gamma}$ and p_ζ.

Supplemental table 1 displays how the average trial standard deviations and .25, .75 quantiles vary for different design choices. We see that in general, the choice of $\boldsymbol{\gamma}$ does not have a large impact on the quantiles for 3 treatments, but has a sizable impact for $K = 5$ treatments, with $(\gamma_1, \gamma_2) = (.10, .10)$ producing the smallest values of these quantiles, and $(\gamma_1, \gamma_2) = (.05, .10)$ producing the second smallest. Recall that

γ_2 is related to current trial sample size and the set of active treatments, whereas γ_1 is related to values of Δ tested. Smaller p_ζ led to increased sample size quantiles for $K = 3, 5$ and both PO and NPO truths, likely due to a deceased prior probability of clustering and hence stopping to declare a set of treatments equally optimal. Quantiles of sample sizes did not change much for the different choice of utilities for MABOUST or the permutation test. Standard deviation values in this table reflect general shifts in these quantiles.

We recommend $p_\zeta = .9$ to keep FWER rates and average sample size low, and ($\gamma_2 = .05$) to give a better generalized power. The choice of γ_1 had less of an impact than these two parameters. However, we suggest that the trial statistician simulate operating characteristics under various plausible simulation truths before conducting a trial, which can better calibrate the choices of p_ζ and $\boldsymbol{\gamma}$. The insensitivity of the results to the choice of the utility function is positive, as this suggests that clinicians can build this function to reflect clinical benefit for patients.

5 Discussion

We proposed a Multi-Armed Bayesian Ordinal Outcome Utility-based Sequential Trial (MABOUST) that repeatedly makes interim decisions to remove inferior treatment arms if another arm provides a clinically relevant improvement in patient outcomes, and also declare a set of treatments equally optimal. An ordinal outcome approach was used similar to Murray (2018) that uses utilities to take an ordinal outcome and reduce it to a single optimality score, via utilities. This approach differs in that marginally for each treatment a flexible non-proportional odds approach is used to estimating the outcome probabilities for each treatment, which also allows better estimation of the utilities. By using the Bayesian pairwise null *house-party-prior*, we avoided the need for a baseline treatment group, and controlled family wise error rates across 1,000 randomly generated scenarios for 3 and 5 treatments.

Under non-proportional odds simulation truths, this design outperforms a frequentist proportional odds model, a permutation based utility method, and a Bayesian utility based design that does not allow clustering. The design also performs relative to the proportional odds model in accuracy when the data generation scheme truly is proportional odds, with a reduced average sample size. We should note that an adaptive decision boundary was not used for the proportional odds or permutation tests, and could be explored in future studies. One benefit of the utility based approach is the translatability of a proportional (or non-proportional) odds treatment effect into a numerical score which might be better understood by clinicians - and improve the outreach of MABOUST. We explored covariate adjustment in this context, which did lead to a decrease in average trial sample size and .25, .75 quantiles, suggesting that covariate adjustment might give a better chance of stopping the trial early, without decreasing overall accuracy.

Sensitivity of the operating characteristics were explored for MABOUST based on the choice of $\boldsymbol{\gamma}$ which characterises the adaptive decision boundary and p_ζ which is the

pairwise-null prior clustering probability. Choices of $p_\zeta < .9$ resulted in increased family-wise error rates, which shows that a conservative choice is needed for the clustering prior to be effective. Increasing γ_2, which is related to trial sample size and number of active treatments, also increases FWER and the probability of dropping inferior treatments. Changes in the choice of the utility function did not have a drastic change on MABOUST's operating characteristics, which means that clinicians can choose a utility function that reflects their perceived patient impact of each outcome.

The primary novelty of this paper is proposing a Bayesian clustering prior on the treatment specific ordinal outcome parameters - which encourages the pairwise hypothesis of no difference between each treatment and reduces towards the global null of no treatment difference. This Bayesian *house-party-prior* has been used for subgroup specific parameters in the past, but has yet to be used to actively encourage a conservative group sequential design in terms of pairwise comparisons. We also proposed stopping rules that weed out ineffective treatments repeatedly if any other treatment provides a clinically relevant improvement in comparison, rather than stopping the loser or graduating the winner. Sets of optimal treatments might advance to public usage through a futility rule, which stops sets of treatments if they are equally optimal in terms of clinical significance. This rule is actively encouraged by the Bayesian clustering prior. Finally, this design proposes a dynamic decision boundary which is a function of the current sample size, set of active treatments in the trial, and a cascade of increasing clinically relevant improvements in true mean utility.

This design gives researchers a way to search through several treatments with ordinal outcomes and eventually determine a set of optimal treatments. While the motivation of this trial was COVID-19 related, this structure could be applied to many areas in medicine. For example, in pregnant mothers, this ordinal scale could be incorporated via extremely premature, premature, term, and late-term deliveries. Similarly, after a total knee replacement with a novel knee implant - one might ask patients to rate the difference they are experiencing in terms of pain after 5 months - which is ordinal. We provide user friendly R code in the package *MABOUST* found on CRAN. In the supplemental material, we provide detailed instructions to use functions needed to make interim decision, simulate proposed trials, and generate sets of random scenarios that are grounded in clinical reality - much like what was done for the COVID-19 trial at UMC. Unfortunately that trial was stopped due to lack of enrollment, but MABOUST could be used for other sets of drugs even in a multi-institutional setting. This might allow a faster resolution of this pandemic by considering a more sophisticated outcome structure, statistical model, and group sequential multi-armed decision-making.

Supplementary Material

Supplemental Material for 'A Multi-Armed Bayesian Ordinal Outcome Utility-Based Sequential Trial with a Pairwise Null Clustering Prior'
(DOI: 10.1214/22-BA1316SUPP; .pdf). A detailed description of how to use the package *MABOUST* and supplemental operating characteristics table is available with this paper.

References

Amsbaugh, M. and et al (2019). "A phase 1/2 trial of reirradiation for diffuse intrinsic pontine gliomas." *International Journal of Radiation Oncology, Biology, Physics*, 104 (1): 144–148. 520

Brock, K., Billingham, L., Copland, M., and et al. (2017). "Implementing the EffTox dose-finding design in the Matchpoint trial." *BMC Med Res Methodol*, 17(112). 520

Chapple, A. (2021). "Bayesian subgroup clustering in phase I clinical trials." *SAGE Research Methods Cases*. 523, 524

Chapple, A. G., Bennani, J. and Clement, M. (2022). "Supplementary Material for "A Multi-Armed Bayesian Ordinal Outcome Utility-Based Sequential Trial with a Pairwise Null Clustering Prior"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1316SUPP. 522

Chapple, A. and Thall, P. (2018). "Subgroup-specific dose finding in phase I clinical trials based on time to toxicity allowing adaptive subgroup combination." *Journal of Pharmaceutical Statistics*, 17(6): 734–749. 523, 524

Charlson, M., Pompei, P., Ales, K., and et al. (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation." *Journal of Chronic Disease*, 405: 373–383. 522

Chu, C., Cheng, V., Hung, I., and et al. (2020). "Role of lopinavir/ritonavir in the treatment of SARS: initial virological and clinical findings." *Thorax*, 59(3): 252–256. 519

ClinicalTrials.gov (2020). "Treatment in patients with suspected or confirmed COVID-19 with early moderate or severe disease." *https://clinicaltrials.gov/ct2/show/study/NCT04344444?term=NCT04344444&draw=2&rank=1*. 519

Dunnett, C. (1955). "Multiple Comparison Procedure for Comparing Several Treatments with a Control." *Journal of the American Statistical Association*, 50(272): 1096–1121. 521

Harrell, F. and Lindsell, C. (2020). "Statistical design and analysis plan for randomized trial of Hydroxychloroquine for treatment of COVID-19: ORCHID." *https://hbiostat.org/proj/covid19/bayesplan.html*. 522, 526

Hoftsetter, W. (2020). "Nutritional Supplementation in Reducing Complications in Patients With Locally Advanced Esophageal Cancer Undergoing Chemotherapy, Radiation Therapy, and/or Surgery." *https://www.clinicaltrials.gov/ct2/show/NCT04029857?term=Hofstetter&draw=2&rank=2*. 520

Liu, J., Cao, R., Xu, M., and et al (2020). "Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro." *Cell Discoveries*, 6:16. 519

Marshall, J. and et al (2020). "A minimal common outcome measure set for COVID-19 clinical research." *The Lancet Infectious Diseases*, 8: e192–e197. 520

Murray, T., Thall, P., and Yuan, Y. (2016). "Utility-based designs for randomized comparative trials with categorical outcomes." *Statistics in Medicine*, 35 (4): 4285–4305. MR3554963. doi: https://doi.org/10.1002/sim.6989. 520

Murray, T., Yuan, Y., Thall, P., Elizondo, J., and Hoffstetter, W. (2018). "A utility-based design for randomized comparative trials with ordinal outcomes and prognostic subgroups." *Biometrics*, 74: 1095–1103. MR3860730. doi: https://doi.org/10.1111/biom.12842. 520

Pordes, R. e. a. (2007). "The Open Science Grid." *J. Phys. Conf. Ser. 78*. 519

Sfiligoi, I., Bradley, D., Holzman, B., Mhashilkar, P., Padhi, S., and Wurthwein, F. (2009). "The Pilot Way to Grid Resources Using glideinWMS." *2009 WRI World Congress on Computer Science and Information Engineering*, 2: 428–432. 519

Shah, N., Thall, P., Fox, P., Bashir, Q., Qazilbash, M., and et al (2015). "Phase I/II trial of lenalidomide and high dose melphalan with autologous stem cell transplantation for relapsed myeloma." *Leukemia*, 29: 1945–1948. 520

Stallard, N. and Todd, S. (2003). "Sequential Designs for Phase III Clinical Trials Incorporating Treatment Selection." *Statistics in Medicine*, 22(5): 689–703. 521

Thall, P. and Cook, J. (2004). "Dose-finding based on efficacy-toxicity trade-offs." *Biometrics*, 60(3): 684–693. MR2089444. doi: https://doi.org/10.1111/j.0006-341X.2004.00218.x. 520

Thall, P., Simon, R., and Ellenberg, S. (1989). "A two-stage design for choosing among several experimental treatments and a control in clinical trials." *Biometrics*, 45(2): 537–547. MR1010517. doi: https://doi.org/10.2307/2531495. 521

Wang, D., Hu, B., Hu, C., and et al (2020). "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China." *JAMA*, 323(11): 1061–1069. 533

Whitfield, K. and Swenson, D. (2020). "Coronavirus In Louisiana: 21,518 Cases — 1,013 Dead — 1,977 In Hospital. [online] NOLA.com." *https://www.nola.com/news/coronavirus/*. 523, 532

Yao, X., Ye, F., Zhang, M., and et al. (2020). "In Vitro Antiviral Activity and Projection of Optimized Dosing Design of Hydroxychloroquine for the Treatment of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)." *Clin Infect Dis*, 71(15): 732–739. 519