A Utility-Based Design for Randomized Comparative Trials with Ordinal Outcomes and Prognostic Subgroups

Thomas A. Murray,[1] Ying Yuan[2], Peter F. Thall[2],

Joan H. Elizondo[3] and Wayne L. Hofstetter[4]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, U.S.A.

[2]Department of Biostatistics, [3]Department of Clinical Nutrition, and

[4]Department of Thoracic and Cardiovascular Surgery,

The University of Texas M.D. Anderson Cancer Center, Houston, Texas, U.S.A.

Abstract

A design is proposed for randomized comparative trials with ordinal outcomes and prognostic subgroups. The design accounts for patient heterogeneity by allowing possibly different comparative conclusions within subgroups. The comparative testing criterion is based on utilities for the levels of the ordinal outcome and a Bayesian probability model. Designs based on two alternative models that include treatment-subgroup interactions are considered, the proportional odds model and a non-proportional odds model with a hierarchical prior that shrinks toward the proportional odds model. A third design that assumes homogeneity and ignores possible treatment-subgroup interactions also is considered. The three approaches are applied to construct group sequential designs for a trial of nutritional prehabilitation versus standard of care for esophageal cancer patients undergoing chemoradiation and surgery, including both untreated patients and salvage patients whose disease has recurred following previous therapy. A simulation study is presented that compares the three designs, including evaluation of within-subgroup type I and II error probabilities under a variety of scenarios including different combinations of treatment-subgroup interactions.

Keywords: Group sequential; Hierarchical model; Non-proportional odds; Ordinal response; Precision medicine.

1 Background

This paper describes a design for a small single-center randomized controlled clinical trial to evaluate the effectiveness of nutritional prehabilitation (Nuprehab) for esophageal cancer patients who undergo esophageal resection preceded and followed by chemoradiation therapy. Common postoperative morbidities for patients who undergo esophageal resection include anastomotic leak and stricture, chylothorax, delayed emptying or dumping syndrome, pulmonary complications such as pneumonia, and cardiac complications such as atrial fibrillation (Parekh and Iannettoni, 2007; Chen, 2014). The Nuprehab is given prior to surgery as well as seven days after surgery, with the aim to achieve oral immunomodulation with an L-arginine based enteral formula. The motivation for the trial is the hypothesis that providing patients with Nuprehab may reduce the incidence of postoperative morbidity and mortality via nutritional supplementation (see, e.g., Braga et al., 2002; Waitzberg et al., 2006).

Patients will be randomized to receive either Nuprehab or control, which is the standard of care. All patients will be monitored for morbidity and mortality for thirty days following their surgery. The trial's primary outcome is Clavien-Dindo postoperative morbidity (POM) score (Clavien et al., 1992; Dindo et al., 2004; Clavien et al., 2009), which is ordinal with six levels: 0=normal recovery, 1=minor complication, 2=complication requiring pharmaceutical intervention, 3=complication requiring surgical, endoscopic or radiological intervention, 4=life-threatening complication requiring intensive care, and 5=death. The worst POM score during 30 days post surgery will be recorded. The trial will enroll approximately 60% primary and 40% salvage patients. Primary patients are treatment naive, whereas salvage patients have been treated previously with chemoradiation therapy, but not surgery, and their disease has recurred. Salvage patients are expected to have fewer preoperative nutritional deficiencies, but more preoperative comorbidities and worse prognosis. Consequently, it is plausible that the efficacy of Nuprehab may differ substantially for primary and salvage patients.

In this paper, we describe a design that accounts for the possibility that Nuprehab may

be clinically beneficial for one of the subgroups but not the other. This is in sharp contrast with a more traditional "one-size-fits-all" approach that ignores prognostic information and makes one recommendation for all patients about whether Nuprehab is clinically beneficial. We evaluate the design based on its probabilities of recommending Nuprehab to each subgroup in four key scenarios: (i) the "complete null" scenario where Nuprehab does not improve POM scores for patients in either subgroup; (ii) the "partial null" scenario where Nuprehab does not improve POM scores for primary patients, but achieves a targeted benefit for salvage patients; (iii) the "partial null" scenario where Nuprehab achieves a targeted benefit for primary patients, but does not improve POM scores for salvage patients; and (iv) the "complete alternative" scenario where Nuprehab achieves targeted benefits in POM score reduction for both primary and salvage patients. The proposed design addresses the concern that, in the partial null scenarios (ii) and (iii), a one-size-fits-all design will have an unacceptably high (low) probability for recommending Nuprehab to the non-benefiting (benefiting) subgroup.

The proposed design is frequentist in that it is specified to provide specific probabilities of recommending Nuprehab to a non-benefiting subgroup (i.e. type I error) and to a benefiting subgroup for a particular targeted benefit (i.e. power). However, the decision to recommend Nuprehab for a particular subgroup is based on a posterior probability from a Bayesian model. Designs have been proposed that are similar to our design in that they facilitate subgroup specific recommendations, and stopping subsequent enrollment for particular subgroups, see, e.g., Brannath et al. (2009); Wang et al. (2009); Rosenblum et al. (2016). These designs do not involve an ordinal outcome, however.

We consider two alternative Bayesian probability models that facilitate subgroup specific recommendations. The first is the proportional odds (PO) cumulative logistic regression model of McCullagh (1980) with a treatment-subgroup interaction parameter. The second is a non-proportional odds (NPO) model with a hierarchical prior that shrinks toward the PO model. Although NPO models have been proposed, see, e.g., Peterson and Harrell (1990);

3

Bender and Grouven (1998); Ishwaran (2000); Agresti (2010), as far as we are aware our formulation of the NPO model is novel, and moreover this is the first proposal to use an NPO model as the basis for comparing treatments in a randomized clinical trial. Guo and Yuan (2017) use the dispersed cumulative probit model of McCullagh (1980), which is a type of NPO model, for personalized dose finding in a phase I/II study of molecularly targeted agents. As we describe below, compared with the PO model, treatment comparison based on the proposed NPO model is more robust but more complex. To obtain a practical design that deals with this complexity, we propose a comparative testing criterion based on elicited numerical utilities of the six POM scores. Our approach may be considered a generalization of the utility-based design proposed by Murray et al. (2016), which does not accommodate prognostic subgroups and uses a Dirichlet-multinomial model.

The remainder of the paper is organized as follows. In Section 2, we discuss treatment comparison with ordinal outcomes in general and our utility-based comparative criterion in particular. In Section 3, we describe the PO and NPO models. In Section 4, we discuss practical design considerations, including specifying targeted alternatives, analysis and monitoring plan, controlling the probability of committing a type I error, and sample size. In Section 5, we present the results of a simulation study comparing the proposed design based on either the PO or NPO model, and also a more traditional design based on a PO model without a treatment-subgroup interaction parameter. We conclude with a brief discussion in Section 6.

2   Treatment Comparison

Each design compares the efficacy of Nuprehab relative to standard of care using the six-level ordinal POM score. Comparing treatments based on an ordinal outcome is complicated by the fact that, even when the probability of each outcome level is known, it is not always clear whether one treatment is superior to the other. A simple example is a three-level

outcome (Good, Intermediate, Poor) where treatment A gives probabilities $(0.30, 0.50, 0.20)$ and treatment B gives probabilities $(0.40, 0.30, 0.30)$. Since B has larger probabilities of both Good and Poor compared to A, it is not clear whether one treatment is superior to the other. Comparing the two treatments requires additional information, such as a quantification of the relative desirabilities of the three possible events.

Accounting for prognostic subgroups further complicates matters. To see this, denote Y = POM score, P = primary, S = salvage, N = Nuprehab, C = control, and

$$\pi_y(\text{Sgp}, \text{Trt}) = \text{Prob}(Y = y \,|\, \text{Sgp}, \text{Trt}) \text{ and } \pi_y^+(\text{Sgp}, \text{Trt}) = \text{Prob}(Y \le y \,|\, \text{Sgp}, \text{Trt}),$$

for $y = 0, \ldots, 5$, Sgp $\in \{P, S\}$ and Trt $\in \{N, C\}$. Indexing Sgp by x and Trt by a or a', if

$$\pi_y^+(x, a) \ge \pi_y^+(x, a'), \text{ for } y = 0, 1, \ldots, 4, \text{ and}$$

$$\pi_y^+(x, a) > \pi_y^+(x, a'), \text{ for some } y = 0, 1, \ldots, 4,$$

then clearly treatment a is superior to a' for patients in subgroup x. By contrast, if

$$\pi_y^+(x, a) < \pi_y^+(x, a'), \text{ for some } y = 0, 1, \ldots, 4, \text{ and}$$

$$\pi_y^+(x, a) > \pi_y^+(x, a'), \text{ for some } y = 0, 1, \ldots, 4,$$

then it is not clear whether a is superior to a' for patients in subgroup x. Stated formally, for a particular patient subgroup, if the POM score distributions corresponding to each treatment arm are *stochastically ordered* with strict inequality $\pi_y^+(x, a) > \pi_y^+(x, a')$ for at least one level y, then it is clear which treatment is superior for that subgroup. By contrast, if the POM score distributions corresponding to each treatment arm are not stochastically ordered, then it is not clear whether one treatment is superior to the other for that subgroup.

To provide a criterion for determining whether one treatment is superior, we elicit numerical utilities $U(Y = y)$ for all levels of Y, and compare treatments using mean utilities,

$$\overline{U}(\boldsymbol{\pi}(\text{Sgp}, \text{Trt})) = \sum_{y=0}^{5} U(Y = y) \times \pi_y(\text{Sgp}, \text{Trt}).$$

These depend on the subgroup-treatment specific outcome probabilities $\boldsymbol{\pi}(\text{Sgp}, \text{Trt})$ and a

utility function $U(Y = y)$ that quantifies the desirability of each outcome level. Following, Houede et al. (2010), Thall and Nguyen (2012) and Murray et al. (2016), we elicited $\{U(Y = y), y = 0, \ldots, 5\}$ from the trial's Principal Investigator, WH, so that the numerical utilities reflect his familiarity with postoperative complications following esophageal resection. To do this, we first set $U(Y = 0) = 100$ and $U(Y = 5) = 0$, and then asked WH to specify numerical values for the intermediate levels, $y = 1, \ldots, 4$, that reflect their desirability relative to the best and worst levels. The numerical values that WH chose are:

$$U(Y = 0) = 100, \quad U(Y = 1) = 80, \quad U(Y = 2) = 65,$$
$$U(Y = 3) = 25, \quad U(Y = 4) = 10, \quad U(Y = 5) = 0.$$

These reflect that POM scores ≤ 2 are substantially more desirable than POM scores ≥ 3. Because a larger mean utility corresponds to better patient outcomes on average, if $\overline{U}(\boldsymbol{\pi}(x, a)) > \overline{U}(\boldsymbol{\pi}(x, a'))$, then treatment a is superior to a' for patients in subgroup x. Therefore, regardless of whether $\boldsymbol{\pi}(x, a)$ and $\boldsymbol{\pi}(x, a')$ are stochastically ordered, the mean utilities provide an unambiguous criterion for comparing treatments.

One important property of the mean utilities is a consequence of the following theorem.

Theorem 1. *If $\boldsymbol{\pi}(x, a)$ stochastically dominates $\boldsymbol{\pi}(x, a')$, then $\overline{U}(\boldsymbol{\pi}(x, a)) > \overline{U}(\boldsymbol{\pi}(x, a'))$ for all admissible $U(Y)$ such that $U(Y = 0) > U(Y = 1) > \cdots > U(Y = 4) > U(Y = 5)$.*

Theorem 1 follows from first-order stochastic dominance (Quirk and Saposnik, 1962); nonetheless, we provide a proof in the supplementary materials. Consequently, when the POM score distributions are stochastically ordered—and thus, it is clear which treatment is superior without appealing to the mean utilities—the proposed utility-based comparison is not sensitive to the elicited numerical values in that a different set of admissible values will result in the same conclusion. By contrast, when the POM score distributions are not stochastically ordered, eliciting numerical values is necessary to determine whether one treatment is superior. The proposed utility-based comparison necessarily is sensitive to the elicited values in that a different set of admissible values may result in a different conclusion.

Since the POM score probabilities are unknown, we learn about these using a Bayesian model with unknown parameter $\boldsymbol{\theta}$ and model-based mean utilities $\overline{U}(\boldsymbol{\pi}(\mathrm{Sgp}, \mathrm{Trt}; \boldsymbol{\theta}))$. Given interim or final data $\boldsymbol{D}$, our comparative testing criterion is as follows. If

$$\mathrm{Prob}\big\{\overline{U}(\boldsymbol{\pi}(x, a; \boldsymbol{\theta})) > \overline{U}(\boldsymbol{\pi}(x, a'; \boldsymbol{\theta})) \,|\, \boldsymbol{D}\big\} > p_{cut},$$

then we declare treatment a is superior to a' for patients in subgroup x. We specify p_{cut} to control subgroup specific type I error probabilities. We describe how to do this in Section 4.

3 Probability Models

During the process of designing the trial, we considered two Bayesian cumulative logistic regression models that both include treatment-subgroup interaction parameters. The first is a PO model, which is a popular regression model for ordinal response variables, see, e.g., McCullagh (1980); Walters et al. (2001); Abreu et al. (2008). The restrictive parametric assumption underlying the PO model often is unrealistic, however. The second is a NPO model that relaxes this assumption at the cost of greater model complexity.

3.1 Proportional Odds Model

Denote $\mathrm{logit}(q) = \log\{q/(1-q)\}$. The PO model that we considered assumes

$$\mathrm{logit}\left\{\pi_y^+(\mathrm{X}, \mathrm{A}; \alpha_y, \boldsymbol{\beta})\right\} = \alpha_y + \beta_1\, X + \beta_2\, A + \beta_3\, X\, A, \quad \text{for} \quad y = 0, \ldots, 4, \tag{1}$$

where $\alpha_0 \leq \cdots \leq \alpha_4$, $X = -0.5$ for primary, $X = 0.5$ for salvage, $A = -0.5$ for control and $A = 0.5$ for Nuprehab. This model is parsimonious in that it accounts for all treatment and subgroup effects using three parameters, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$.

Let $n_y(X, A)$ denote the number of patients with a POM score equal to y in subgroup X and treatment arm A. We assume observations are mutually independent, so that the

likelihood function for the unknown model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{D}) = \prod_{y \in \{0,\ldots,5\}} \prod_{X \in \{-.5,.5\}} \prod_{A \in \{-.5,.5\}} \pi_y(X, A;\ \boldsymbol{\alpha}, \boldsymbol{\beta})^{n_y(X,A)}, \tag{2}$$

where

$$\pi_0(X, A;\ \boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi_0^+(X, A;\ \alpha_0, \boldsymbol{\beta}),$$

$$\pi_y(X, A;\ \boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi_y^+(X, A;\ \alpha_y, \boldsymbol{\beta}) - \pi_{y-1}^+(X, A;\ \alpha_{y-1}, \boldsymbol{\beta}), \text{ for } y = 1, \ldots, 4,$$

$$\pi_5(X, A;\ \boldsymbol{\alpha}, \boldsymbol{\beta}) = 1 - \pi_4^+(X, A;\ \alpha_4, \boldsymbol{\beta}),$$

and $\pi_y^+(X, A;\ \alpha_y, \boldsymbol{\beta})$, $y = 0, \ldots, 4$, is defined in (1).

We specify the prior distribution for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that $p_0(\boldsymbol{\alpha}, \boldsymbol{\beta}) = p_0(\boldsymbol{\alpha}) \times p_0(\boldsymbol{\beta})$, where

$$p_0(\boldsymbol{\alpha}) = p_0(\alpha_0) \times \prod_{s=1}^{4} p_0(\alpha_y \mid \alpha_{y-1}) \ \text{ and } \ p_0(\boldsymbol{\beta}) = p_0(\beta_1) \times p_0(\beta_2) \times p_0(\beta_3).$$

The exact prior distributional forms that we assume are

$$\alpha_0 \sim t_5(\alpha_0^*, 2.5), \quad \alpha_y \mid \alpha_{y-1}, \sim t_5(\alpha_y^*, 2.5)[\alpha_{y-1}, \infty], \ \text{ for } \ y = 1, 2, 3, 4,$$
$$\beta_1 \sim t_5(\beta_1^*, 2.5), \quad \beta_2 \sim t_5(0, 2.5), \quad \beta_3 \sim t_5(0, 2.5), \tag{3}$$

where we write $p_0(\theta)[L, U]$ to denote that $p_0(\theta)$ has support on the interval $[L, U]$. The above prior restricts $\alpha_0 \leq \cdots \leq \alpha_4$ so that $0 \leq \pi_1^+(X, A;\ \alpha_1, \boldsymbol{\beta}) \leq \cdots \leq \pi_4^+(X, A;\ \alpha_4, \boldsymbol{\beta}) \leq 1$ for all $X \in \{-.5, .5\}$ and $A \in \{-.5, .5\}$, which is necessary to ensure that the probability model is admissible. Following the recommendations of Ghosh et al. (2017), we specify t-distributions with a scale of 2.5 and five degrees of freedom. This specification places about 90% of the prior probability mass on the range of values within 5 of the prior mean, while the heavy tails do not preclude more extreme values should the data demand this. Because an effect size of 5 corresponds to a shift from 0.01 to 0.99 on the probability domain between subgroups or treatment arms, the proposed prior specification allows the observed data to dominate posterior inference.

We specify non-zero prior means, $\{\alpha_y^*\}_{y=0}^{4}$ and β_1^*, to reflect the prior information that WH provided about the POM score probabilities of each subgroup in the control arm, which

Table 1: Elicited prior mean POM score probabilities in the control arm for primary and salvage patients.

		POM Score				
Subgroup	0	1	2	3	4	5
Primary	0.50	0.20	0.10	0.10	0.05	0.05
Salvage	0.30	0.25	0.10	0.10	0.10	0.15

we report in Table 1. Using $\pi_y^*(x)$ to denote the prior probabilities that WH provided for the subgroup corresponding to $X = x$, and $\pi_y^{+,*}(x) = \sum_{\ell=1}^{y} \pi_\ell^*(x)$, we set

$$\alpha_y^* = \left[\log \left(\frac{\pi_y^{+,*}(P)}{1 - \pi_y^{+,*}(P)} \right) + \log \left(\frac{\pi_y^{+,*}(S)}{1 - \pi_y^{+,*}(S)} \right) \right] \Big/ 2, \text{ for } y = 0, \dots, 4, \text{ and}$$

$$\beta_1^* = \left[\sum_{y=0}^{4} \log \left(\frac{\pi_y^{+,*}(P)}{1 - \pi_y^{+,*}(P)} \right) - \log \left(\frac{\pi_y^{+,*}(S)}{1 - \pi_y^{+,*}(S)} \right) \right] \Big/ 5.$$

Because we set the prior means for β_2 and β_3 equal to zero, *a priori* $\overline{U}(\boldsymbol{\pi}(P, N)) = \overline{U}(\boldsymbol{\pi}(P, C))$ and $\overline{U}(\boldsymbol{\pi}(S, N)) = \overline{U}(\boldsymbol{\pi}(S, C))$. Therefore, *a posteriori* a non-zero mean utility difference between treatment arms in either subgroup will reflect the observed data, and not the prior.

The PO model assumes that the regression coefficients, and thus the log-odds ratios, do not differ with the level of the response variable. This is a strong parametric assumption that often is unrealistic in practice, including the present context. The PO model likely is popular since it facilitates treatment comparison in that a utility function need not be elicited from the clinician(s). To see this, note that for $y = 0, \dots, 4$ and $X = x$,

$$\pi_y^+(x, N; \alpha_y, \boldsymbol{\beta}) = \frac{\pi_y^+(x, C; \alpha_y, \boldsymbol{\beta}) \exp\{\beta_2 + \beta_3 \, x\}}{1 - \pi_y^+(x, C; \alpha_y, \boldsymbol{\beta}) + \pi_y^+(x, C; \alpha_y, \boldsymbol{\beta}) \exp\{\beta_2 + \beta_3 \, x\}},$$

$\pi_y^+(x, N; \alpha_y, \boldsymbol{\beta})$ is monotonically increasing in $(\beta_1 + \beta_3 \, x)$ such that $\pi_y^+(x, N; \alpha_y, \boldsymbol{\beta}) = \pi_y^+(x, C; \boldsymbol{\alpha}, \boldsymbol{\beta})$ when $(\beta_1 + \beta_3 \, x) = 0$. Therefore, for any $U(Y = 0) < \cdots < U(Y = 5)$,

$$\text{if } (\beta_1 + \beta_3 \, x) > 0, \text{ then } \overline{U}(\boldsymbol{\pi}(x, N; \boldsymbol{\beta}, \boldsymbol{\alpha})) > \overline{U}(\boldsymbol{\pi}(x, C; \boldsymbol{\beta}, \boldsymbol{\alpha})),$$

and conversely,

$$\text{if } (\beta_1 + \beta_3\, x) < 0, \text{ then } \overline{U}(\boldsymbol{\pi}(x, N; \boldsymbol{\beta}, \boldsymbol{\alpha})) < \overline{U}(\boldsymbol{\pi}(x, C; \boldsymbol{\beta}, \boldsymbol{\alpha})).$$

Consequently, when posterior inference is based on the PO model defined in (1), the utility function is superfluous. However, when the actual response distributions do not satisfy the PO assumption, e.g., they are not stochastically ordered, the PO model may be misleading.

3.2 Non-Proportional Odds Model

To relax the assumption required by the PO model, that the log-odds ratios do not differ with the response level, we propose a hierarchical cumulative logistic regression model that assumes

$$\text{logit}\left\{\pi_y^+(X, A;\, \alpha_y, \boldsymbol{\gamma}_y)\right\} = \alpha_y + \gamma_{1,y}\, X + \gamma_{2,y}\, A + \gamma_{3,y}\, X\, A, \text{ for } y = 0, \ldots, 4, \qquad (4)$$

where X and A are defined similarly as for the PO model. In contrast with the PO model, the NPO model in (4) allows different log-odds ratios at each response level y. We assume that observations are mutually independent such that the likelihood function for the unknown parameters $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ has the same general form (2) as for the PO model.

We specify a hierarchical prior for $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ as follows,

$$\alpha_0 \sim t_5(\alpha_0^*, 2.5), \quad \alpha_y \mid \alpha_{y-1}, \boldsymbol{\gamma}_{y-1}, \boldsymbol{\gamma}_y \sim t_5(\alpha_y^*, 2.5)[\alpha_{y-1} + 0.5|\gamma_{1,y-1} - \gamma_{1,y}| +$$

$$0.5|\gamma_{2,y-1} - \gamma_{2,y}| + 0.25|\gamma_{3,y-1} - \gamma_{3,y}|, \infty], \text{ for } y = 1, 2, 3, 4,$$

$$\gamma_{1,y} \mid \beta_1, \sigma_1 \sim N(\beta_1, \sigma_1^2), \quad \beta_1 \sim t_5(\beta_1^*, 2.5), \quad \sigma_1 \sim N(0, 1)[0, \infty], \qquad (5)$$

$$\gamma_{2,y} \mid \beta_2, \sigma_2 \sim N(\beta_2, \sigma_2^2), \quad \beta_2 \sim t_5(0, 2.5), \quad \sigma_2 \sim N(0, 1)[0, \infty],$$

$$\gamma_{3,y} \mid \beta_3, \sigma_3 \sim N(\beta_3, \sigma_3^2), \quad \beta_3 \sim t_5(0, 2.5), \quad \sigma_3 \sim N(0, 1)[0, \infty].$$

The prior constraints on $\{\alpha_y\}_{y=1}^4$ ensure that $0 \leq \pi_1^+(X, A;\, \alpha_1, \boldsymbol{\gamma}_1) \leq \cdots \leq \pi_4^+(X, A;\, \alpha_4, \boldsymbol{\gamma}_4) \leq$

1 for all $X \in \{-0.5, 0.5\}$ and $A \in \{-0.5, 0.5\}$. These inequalities hold when

$$\alpha_y \geq \alpha_{y-1} + 0.5|\gamma_{1,y-1} - \gamma_{1,y}| + 0.5|\gamma_{2,y-1} - \gamma_{2,y}| + 0.25|\gamma_{3,y-1} - \gamma_{3,y}|, \ y = 1, 2, 3, 4,$$

which is reflected in our specification of the prior. The hierarchical structure that we propose in (5) shrinks toward the PO model defined in (1). As $\sigma_1^2 \to 0$, $\sigma_2^2 \to 0$ and $\sigma_3^2 \to 0$, then $\gamma_{1,y} \to \beta_1$, $\gamma_{2,y} \to \beta_2$ and $\gamma_{3,y} \to \beta_3$, for $y = 0, \ldots, 4$, and thus the log-odds ratios become invariant to the outcome level. Essentially, the NPO model in (4) adds a layer of additional structure to the PO model in (1) that allows each effect to deviate from the PO assumption. By using half-normal distributions with a unit standard deviation for σ_1, σ_2 and σ_3, *a priori* the proposed NPO model prefers small deviations from the PO model. This type of hierarchical NPO model was alluded to as an alternative for PO models in the discussion of McKinley et al. (2015), but they neither implemented nor fully specified such a model. Because the proposed NPO model allows the log-odds ratios to differ with the response level, the model-based estimates of the response distributions need not be stochastically ordered. Consequently, when posterior inference is based on the NPO model, our proposed utility-based comparative testing criterion facilitates treatment comparison.

3.3   Posterior Estimation

We carry out posterior estimation for the PO and NPO models using `JAGS` via the `R` package `R2jags` (Plummer, 2003). Posterior convergence tends to be immediate and autocorrelation tends to be low, likely due, in part, to our balanced specification of the design matrix. We use the posterior samples from `JAGS` to calculate the four posterior probabilities required for the utility-based comparative testing criterion given in Section 2. Because the mean utilities are tractable functions of the unknown model parameters for both the PO and NPO models, obtaining posterior samples of the mean utilities is straightforward. We provide freely-available, user-friendly `R` software for implementation, see Supplementary Materials.

4 Design Considerations

Although we use a Bayesian probability model for statistical inference, we design the trial to ensure certain desirable frequentist operating characteristics (OCs), e.g., 0.80 power under the targeted alternative with 0.05 probability of making a type I error. We are concerned with the subgroup specific power and type I error probability. That is, when Nuprehab reduces the number and severity of postoperative complications for a particular subgroup by the targeted amount, we want our design to have 0.80 probability of correctly declaring N superior to C for patients in that subgroup. By contrast, when Nuprehab does not reduce the number and severity of postoperative complications for a particular subgroup, we want our design only to have 0.05 probability of incorrectly declaring N superior or inferior to C for patients in that subgroup.

We met with WH to determine a practical targeted difference between the treatment arms for primary and salvage patients. During our discussion, WH expressed his desire for the trial to be powered to detect a 75% reduction in POM scores ≥ 3 in each subgroup. We then derived a targeted mean utility difference corresponding to this 75% reduction in POM scores ≥ 3 in each subgroup as follows. For the anticipated POM score probabilities in the control arm for each subgroup, a 75% reduction corresponds to shifts in the probability of POM scores ≥ 3 from 0.20 to 0.05 for primary patients, and from 0.35 to 0.09 for salvage patients. Under the proportional odds assumption, a 75% reduction in POM scores ≥ 3 corresponds to the cumulative POM score probabilities and mean utilities in Table 2. We designed the trial to target mean utility differences of $17.3 = 92.8 - 75.5$ in primary patients and $27.6 = 87.6 - 60.0$ in salvage patients.

Whitehead (1993) derived a sample size formula for a traditional fixed-sample design with ordinal outcomes based on a PO model, which is

$$ n = 12 \left[\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta) \right]^2 \Big/ \left\{ \delta^2 \left[1 - \sum\nolimits_{y=0}^{5} (\pi_y^*)^3 \right] \right\}, $$

where $\Phi(x)$ is the standard normal distribution evaluated at x, α and β are the desired type

Table 2: Targeted alternative for each subgroup in terms of the cumulative POM score probabilities and the corresponding mean utility.

| | | POM Score | | | | | Mean |
Subgroup	Treatment Arm	≤ 0	≤ 1	≤ 2	≤ 3	≤ 4	Utility
Primary	Nutritional Prehabilitation	0.83	0.92	0.95	0.98	0.99	92.8
Primary	Control	0.50	0.70	0.80	0.90	0.95	75.5
Salvage	Nutritional Prehabilitation	0.71	0.87	0.91	0.94	0.97	87.6
Salvage	Control	0.30	0.55	0.65	0.75	0.85	60.0

I and II error probabilities, δ is the targeted log-odds ratio, and $\pi_y^* = \Pr(Y = y)$ under the targeted alternative. Although we use our utility-based comparative testing criterion proposed in Section 2, we demonstrated earlier that for the PO model this criterion is equivalent to a particular contrast of the regression coefficients that is also the basis for Whitehead's sample size formula. Viewing each subgroup as a separate trial, under their respective targeted alternatives, we need to enroll 56 primary and 38 salvage patients for $\alpha = 0.05$ and $\beta = 0.20$. Assuming 60% of the enrollees will be primary patients, we need to enroll 94 patients to achieve these subgroup sample sizes. Because the PO model borrows strength across subgroups for estimating the intercept parameters, $\{\alpha_y\}_{y=0}^4$, we expect the above sample size calculation to be conservative. When the PO assumption holds we expect our NPO model to have less power than our PO model, though only slightly less as we specify an informative half-standard normal distribution as the prior for σ_1, σ_2 and σ_3 in (5).

In the trial, patients are assigned to the two treatment arms using stratified block random-ization with blocks of size four. Thus, for each block of four patients within each subgroup, two patients will receive Nuprehab and two will receive control. This will ensure that the treatment arms will have similar numbers of patients from each subgroup throughout the trial. Given the modest sample size requirements, one interim analysis will be done half-way through the trial. At this point, using our utility-based comparative testing criterion, the design will decide whether to continue enrolling patients from each subgroup. If one treat-ment is declared superior to the other for a certain subgroup at the interim analysis, then no

additional patients from that subgroup will be enrolled. Otherwise, enrollment of patients from that subgroup will continue until the final analysis.

To control the probability of committing a type I error, we use a maximum duration alpha-spending approach such that $f(t) = \alpha \times (t/T_{max})^3$, where T_{max} denotes the maximum trial duration, see Jennison and Turnbull (1999, Section 7.2.3). To do this, we set $p_{cut} = \Phi(z)$ where z corresponds to the relevant threshold for the test statistic in a frequentist group sequential analysis, which we calculate using the R package `gsDesign`. With one interim analysis at the mid-point of the trial, this gives probability thresholds at the interim and final analyses of 0.997 and 0.976, respectively. Due to the asymptotic normality of the posterior distribution in general, see, e.g., Gelman et al. (2014, Section 4), these thresholds control the type I error asymptotically. However, we use computer simulation to verify that our design controls type I error for the planned sample size. To be conservative, we aim to enroll up to 100 patients. Given the anticipated accrual rate of 2 patients per month, the interim and final analyses are expected to be performed at 26 and 51 months, respectively.

5 Simulation Study

In this section, we describe a simulation study that we carried out to evaluate and compare the frequentist OCs of the proposed design, under each of the PO and NPO models defined in Section 3. For further comparison, we considered a one-size-fits-all design based on a PO model that assumes

$$\text{logit}\left\{\pi_y^+(X, A; \alpha_y, \boldsymbol{\beta})\right\} = \alpha_y + \beta_1 X + \beta_2 A, \ \ \text{for} \ \ y = 0, \ldots, 4.$$

We specify the same prior distributions for $\boldsymbol{\alpha}$, β_1 and β_2 as for the PO model defined in (3). For this design, using the same two-stage group sequential structure that controls the probability of committing a type I error at 0.05, if $\text{Prob}(\beta_2 > 0 \mid \boldsymbol{D}) > p_{cut}$, then we declare N superior to C for patients in both subgroups. Conversely, if $\text{Prob}(\beta_2 < 0 \mid \boldsymbol{D}) > p_{cut}$, then

Table 3: Simulation scenarios defined by the POM score distributions in each subgroup and treatment arm, and the corresponding mean utility.

True POM Score Distribution	≤ 0	≤ 1	POM Score ≤ 2	≤ 3	≤ 4	Mean Utility
$\boldsymbol{P}_0$	0.50	0.70	0.80	0.90	0.95	75.5
$\boldsymbol{P}_1$	0.83	0.92	0.95	0.98	0.99	92.8
$\boldsymbol{P}_2$	0.67	0.85	0.95	0.98	0.99	88.8
$\boldsymbol{S}_0$	0.30	0.55	0.65	0.75	0.85	60.0
$\boldsymbol{S}_1$	0.71	0.87	0.91	0.94	0.97	87.6
$\boldsymbol{S}_2$	0.53	0.80	0.91	0.94	0.97	82.8

Scenario	P,C	POM Score Distribution P,N	S,C	S,N	Mean Utility Difference P	S
1	$\boldsymbol{P}_0$	$\boldsymbol{P}_0$	$\boldsymbol{S}_0$	$\boldsymbol{S}_0$	0.0	0.0
2	$\boldsymbol{P}_0$	$\boldsymbol{P}_1$	$\boldsymbol{S}_0$	$\boldsymbol{S}_0$	17.3	0.0
3	$\boldsymbol{P}_0$	$\boldsymbol{P}_0$	$\boldsymbol{S}_0$	$\boldsymbol{S}_1$	0.0	27.6
4	$\boldsymbol{P}_0$	$\boldsymbol{P}_1$	$\boldsymbol{S}_0$	$\boldsymbol{S}_1$	17.3	27.6
5	$\boldsymbol{P}_0$	$\boldsymbol{P}_2$	$\boldsymbol{S}_0$	$\boldsymbol{S}_0$	13.3	0.0
6	$\boldsymbol{P}_0$	$\boldsymbol{P}_0$	$\boldsymbol{S}_0$	$\boldsymbol{S}_2$	0.0	22.8
7	$\boldsymbol{P}_0$	$\boldsymbol{P}_2$	$\boldsymbol{S}_0$	$\boldsymbol{S}_2$	13.3	22.8

we declare N inferior to C for patients in both subgroups. We compare the designs based on their probabilities of declaring N superior (or inferior) to C across a range of scenarios. To assess the decision criteria, we used 10,000 posterior samples following 500 warm-up samples. For all three models, posterior sampling took about 4 seconds for an interim analysis with 50 observations, and about 7 seconds for a final analysis with up to 100 observations.

Table 3 reports the true POM score distributions for each scenario. We used the same POM score distribution to generate observations for the control arm in each subgroup for every scenario, i.e. $\boldsymbol{P}_0$ for primary patients and $\boldsymbol{S}_0$ for salvage patients. Each patient had a 60% chance of belonging to the primary subgroup throughout. Scenario 1 is the complete null case where N provides no benefit to patients in either subgroup. Scenarios 2 and 3 are treatment-subgroup interaction cases where N provides the targeted benefit to patients in one subgroup, and no benefit to patients in the other subgroup. Scenario 4 is the complete alternative case where N provides the targeted benefit to patients in both subgroups. The

PO assumption holds for Scenarios 1-4. Scenarios 5 and 6 are treatment-subgroup interaction cases, and Scenario 7 is a complete alternative case, but the PO assumption does not hold. In particular, the log-odds ratio comparing treatment arms corresponding to POM scores ≤ 0 and ≤ 1 are smaller than those corresponding to POM scores ≤ 2, ≤ 3 and ≤ 4, but the targeted 75% reduction in POM scores ≥ 3 is still achieved in each subgroup. This reflects a benefit that greatly reduces severe postoperative complications, but affects the rate of minor postoperative complications to a lesser degree.

Table 4 reports the proportion of trials in which N was declared superior (inferior) to C, and the average sample size. In Scenario 1, i.e. the complete null case, all three designs control the within subgroup probability of making a type I error near 0.05. In Scenario 4, i.e. the complete alternative case, the two stratified designs have greater than 0.80 power of declaring N superior to C in each subgroup. Excepting Scenario 4, the three competing designs are unlikely to stop early, which is reflected by the average sample sizes near 100. Compared to the stratified design based on the PO model, when the PO assumption holds, e.g., Scenarios 2, 3 and 4, the stratified design based on the more flexible NPO model has similar power for declaring N superior to C in the benefiting subgroup(s), and when the PO assumption does not hold, e.g., Scenarios 5, 6 and 7, the NPO model has larger power of declaring N superior to C in the benefiting subgroup(s). The power for each design is lower when the PO assumption does not hold, which is not surprising as the mean utility differences are smaller in these cases. Compared to the traditional design, when both subgroups benefit, e.g., Scenarios 4 and 7, the stratified designs have less power for declaring N superior to C in both subgroups, but when only one subgroup benefits, e.g., Scenarios 2, 3, 5 and 6, the stratified designs have greater power of declaring N superior to C in the benefiting subgroup, and are far less likely to incorrectly declare N superior to C in the non-benefiting subgroup.

16

Table 4: Simulation results. We report the proportion of simulated trials in which N was declared superior (inferior) to C, and the average sample size. The correct decision is indicated in boldface. Reported figures are based on 5,000 simulated trials.

Design	Primary Sup.	Inf.	Salvage Sup.	Inf.	Average Sample Size
Scenario 1 (Complete Null)					
Traditional	0.022	0.025	0.022	0.025	99.7
Stratified PO	0.022	0.025	0.020	0.025	99.8
Stratified NPO	0.029	0.028	0.029	0.032	99.5
Scenario 2 (Partial Null - PO)					
Traditional	**0.509**	0.000	0.509	0.000	95.5
Stratified PO	**0.774**	0.000	0.038	0.017	94.9
Stratified NPO	**0.763**	0.000	0.048	0.022	93.7
Scenario 3 (Partial Null - PO)					
Traditional	0.417	0.000	**0.417**	0.000	96.6
Stratified PO	0.040	0.015	**0.784**	0.000	96.5
Stratified NPO	0.047	0.016	**0.782**	0.000	95.4
Scenario 4 (Complete Alternative - PO)					
Traditional	**0.978**	0.000	**0.978**	0.000	74.7
Stratified PO	**0.841**	0.000	**0.850**	0.000	87.4
Stratified NPO	**0.842**	0.000	**0.853**	0.000	84.6
Scenario 5 (Partial Null - NPO)					
Traditional	**0.215**	0.001	0.215	0.001	98.5
Stratified PO	**0.314**	0.001	0.032	0.019	98.9
Stratified NPO	**0.347**	0.001	0.043	0.025	98.0
Scenario 6 (Partial Null - NPO)					
Traditional	0.239	0.001	**0.239**	0.001	98.3
Stratified PO	0.036	0.019	**0.462**	0.000	98.7
Stratified NPO	0.042	0.021	**0.503**	0.000	97.7
Scenario 7 (Complete Alternative - NPO)					
Traditional	**0.710**	0.000	**0.710**	0.000	92.1
Stratified PO	**0.387**	0.000	**0.528**	0.000	96.6
Stratified NPO	**0.434**	0.000	**0.584**	0.000	94.8

6 Discussion

In this paper, we have proposed a design for comparing treatments in two prognostic subgroups based on ordinal outcomes. The design was motivated by a trial comparing the effectiveness of nutritional prehabilitation (NuPrehab) against the standard of care for im-

proving postoperative outcomes in primary and salvage patients who undergo esophageal resection. We considered two Bayesian cumulative logistic regression models for statistical inference, a proportional odds (PO) model and a hierarchical non-proportional odds (NPO) model that shrinks toward the PO model. Based on the results of our simulation study, we determined that the design based on the NPO model has preferable frequentist operating characteristics. In particular, when the PO assumption is satisfied the NPO model has similar probability of recommending NuPrehab in the benefiting subgroup(s), whereas when the PO assumption is not satisfied the NPO model can have substantially higher probability of recommending NuPrehab in the benefiting subgroup(s).

We also compared our stratified medicine design with a more traditional design that does not facilitate subgroup specific recommendations. If both subgroups benefit from NuPrehab, then the traditional design is more likely to recommend its use in each subgroup. However, when only one of the two subgroups benefits from NuPrehab, the traditional design is less likely to recommend its use in the benefiting subgroup and more likely to recommend its use in the non-benefiting subgroup. Because substantial treatment effect heterogeneity between primary and salvage patients is plausible in the motivating trial, we find the stratified medicine design based on the NPO model appealing. In another context, if this design is not appealing, then including a model selection between a model with and a model without an interaction term may provide an avenue for achieving a design with more appealing operating characteristics. Another alternative is to consider a prior for the interaction parameter(s) that facilitates borrowing substantial strength for the treatment effect across subgroups, perhaps in a data-dependent manner. These are avenues for future research.

7 Supplementary Materials

Web Appendices with a proof of Theorem 1 referenced in Section 2, and R software referenced in Section 3 for implementing the probability models and reproducing the simulation study

is available with this paper at the Biometrics website on Wiley Online Library.

References

Abreu, M. N. S., Siqueira, A. L., Cardoso, C. S., and Caiaffa, W. T. (2008). Ordinal logistic regression models: Application in quality of life studies. *Cadernos de Saúde Pública* **24,** s581–s591.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data.* Wiley.

Bender, R. and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology* **51,** 809–816.

Braga, M., Gianotti, L., Nespoli, L., Radaelli, G., and Di Carlo, V. (2002). Nutritional approach in malnourished surgical patients: A prospective randomized study. *Archives of Surgery* **137,** 174–80.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* **28,** 1445–1463.

Chen, K.-N. (2014). Managing complications I: Leaks, strictures, emptying, reflux, chylothorax. *Journal of Thoracic Disease* **6,** S355–63.

Clavien, P. A., Barkun, J., de Oliveira, M. L., Vauthey, J. N., Dindo, D., Schulick, R. D., de Santibañes, E., Pekolj, J., Slankamenac, K., Bassi, C., Graf, R., Vonlanthen, R., Padbury, R., Cameron, J. L., and Makuuchi, M. (2009). The Clavien-Dindo classification of surgical complications. *Annals of Surgery* **250,** 187–196.

Clavien, P. A., Sanabria, J. R., and Strasberg, S. M. (1992). Proposed classification of complications of surgery with examples of utility in cholecystectomy. *Surgery* **111,** 518–26.

Dindo, D., Demartines, N., and Clavien, P.-A. (2004). Classification of surgical complications: A new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of Surgery* **240,** 205–13.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis, 3rd Edition.* Chapman & Hall/CRC Press.

Ghosh, J., Li, Y., and Mitra, R. (2017). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis* .

Guo, B. and Yuan, Y. (2017). Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *Journal of the American Statistical Association* **0,** 0–0.

Houede, N., Thall, P. F., Nguyen, H., Paoletti, X., and Kramar, A. (2010). Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Biometrics* **66,** 532–40.

Ishwaran, H. (2000). Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *Canadian Journal of Statistics* **28,** 715–730.

Jennison, C. and Turnbull, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall/CRC.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* **42,** 109–142.

McKinley, T. J., Morters, M., and Wood, J. L. N. (2015). Bayesian model choice in cumulative link ordinal regression models. *Bayesian Analysis* **10,** 1–30.

Murray, T. A., Thall, P. F., and Yuan, Y. (2016). Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in Medicine* **35,** 4285–4305.

Parekh, K. and Iannettoni, M. D. (2007). Complications of esophageal resection and reconstruction. *Seminars in Thoracic and Cardiovascular Surgery* **19,** 79–88.

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C* **39,** 205–17.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Quirk, J. P. and Saposnik, R. (1962). Admissibility and measurable utility functions. *The Review of Economic Studies* **29,** 140–146.

Rosenblum, M., Luber, B., Thompson, R. E., and Hanley, D. (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine* **35,** 3776–3791. sim.6957.

Thall, P. F. and Nguyen, H. Q. (2012). Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of Biopharmaceutical Statistics* **22,** 785–801.

Waitzberg, D. L., Saito, H., Plank, L. D., Jamieson, G. G., Jagannath, P., Hwang, T.-L., Mijares, J. M., and Bihari, D. (2006). Postsurgical infections are reduced with specialized nutrition support. *World Journal of Surgery* **30,** 1592–1604.

Walters, S., Campbell, M., and Lall, R. (2001). Design and analysis of trials with quality of life as an outcome: A practical guide. *Journal of Biopharmaceutical Statistics* **11,** 155–176.

Wang, S.-J., James Hung, H. M., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* **51,** 358–374.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine* **12,** 2257–2271.

Web-based Supplementary Materials for "A Utility-Based Design for Randomized Comparative Trials with Ordinal Outcomes and Prognostic Subgroups" by Thomas A. Murray, Ying Yuan, Peter F. Thall, Joan H. Elizondo and Wayne L. Hofstetter

Re-statement and Proof of Theorem 1

Theorem 2. *If $\boldsymbol{\pi}(x, a)$ stochastically dominates $\boldsymbol{\pi}(x, a')$, then $\overline{U}(\boldsymbol{\pi}(x, a)) > \overline{U}(\boldsymbol{\pi}(x, a'))$ for all admissible $U(Y)$ such that $U(Y = 0) > U(Y = 1) > \cdots > U(Y = 4) > U(Y = 5)$.*

Proof. Note that

$$\overline{U}(\boldsymbol{\pi}(x, a)) - \overline{U}(\boldsymbol{\pi}(x, a')) = \sum_{y=0}^{5} U(Y = y)\pi_y(x, a) - \sum_{y=0}^{5} U(Y = y)\pi_y(x, a')$$

$$= U(Y = 0)[\pi_0^+(x, a)] + \sum_{y=1}^{4} U(Y = y)[\pi_y^+(x, a) - \pi_{y-1}^+(x, a)] + U(Y = 5)[1 - \pi_4^+(x, a)]$$

$$- U(Y = 0)[\pi_0^+(x, a')] - \sum_{y=1}^{4} U(Y = y)[\pi_y^+(x, a') - \pi_{y-1}^+(x, a')] - U(Y = 5)[1 - \pi_4^+(x, a')]$$

$$= \sum_{y=0}^{4} [U(Y = y) - U(Y = y + 1)] \left[\pi_y^+(x, a) - \pi_y^+(x, a') \right].$$

Therefore, if for $y = 0, \ldots, 4$, $U(Y = y) > U(Y = y + 1)$, $\pi_y^+(x, a) \geq \pi_y^+(x, a')$, and for some $y = 0, \ldots, 4$, $\pi_y^+(x, a) > \pi_y^+(x, a')$, then

$$\sum_{y=0}^{4} [U(Y = y) - U(Y = y + 1)] \left[\pi_y^+(x, a) - \pi_y^+(x, a') \right] > 0,$$

and thus, $\overline{U}(\boldsymbol{\pi}(x, a)) > \overline{U}(\boldsymbol{\pi}(x, a'))$. □

R Software

7.1 Readme.txt

```
### Author: Thomas A Murray (8tmurray@gmail.com)
### Date: 14 June 2017


This folder contains two files for replicating the simulation study reported in Murray et
    al. (2017).


(i) functions.R, which contains the relevant R functions.
(iv) simulation.R, which implements the simulation study reported in Murray et al. (2017)


Because simulation.R 'sources' functions.R, these files need to be kept in the same
    working directory.


This software requires R 3.2.0 (or newer), which is freely available for download here:
    https://www.r-project.org/
JAGS 4.2.0 (or newer), which is freely available for download here: https://sourceforge.
    net/projects/mcmc-jags/files/


This software also requires the R packages 'R2jags' (version 0.7 or newer) and 'gsDesign'
    (version 3.1 or newer).


This software was written and executed on a 64-bit machine running Windows 7 Enterprise.
functions.R, which is also indicated in the file.
```

7.2   functions.R

```
#   This file contains R functions for implementing the Nuprehab Trial
#
#   Requirements, R version 3.3.2 or newer (available at https://www.r-project.org/)
#              JAGS version 4.2.0 or newer (available at https://sourceforge.net/
    projects/mcmc-jags/files/)
#              R packages, "R2jags" version 0.7 or newer
#                          "gsDesign" version 3.1 or newer
#
```

```
#   (Originally written for a machine with 64-Bit Windows 7 Enterprise)




### Function to generate datasets using stratified randomization (this function is
    required for simulations and testing, but not for trial conduct)
# Inputs:
# n - sample size
# pis - 2x2x6 array defining POM score probabilities in each patient subgroup and
    treatment arm
#         i.e., pis[sgp, trt, response level]
# prob.primary - probability that a particular subject is a primary patient, default is
    0.6
# block.size - block size for stratified randomization, default is 4
#Outputs:
# data.frame with the three variables
#    y - POM Scores
# sgp - patient subgroup: 1-Primary, 2-Salvage
# trt - treatment arm: 1-Control, 2-Nutritional Prehabilitation
get.dataset <- function(n,pis,prob.primary=0.6,block.size=4){
sgp = rbinom(n,1,1-prob.primary)+1
trt.prim = as.vector(sapply(1:ceiling(n/block.size),function(x) sample(rep(c(1,2),each=
    block.size/2))))
trt.salv = as.vector(sapply(1:ceiling(n/block.size),function(x) sample(rep(c(1,2),each=
    block.size/2))))
patid.prim = sapply(1:n,function(i) sum(sgp[1:i]==1))
patid.salv = sapply(1:n,function(i) sum(sgp[1:i]==2))
trt = sapply(1:n,function(i) ifelse(sgp[i]==1,trt.prim[patid.prim[i]],trt.salv[patid.salv
    [i]]))
y = sapply(1:n,function(i) sample(0:5,1,prob=pis[sgp[i],trt[i],]))
return(data.frame(y=y,trt=trt,sgp=sgp))
}
```

```
### Function to fit non-proportional odds cumulative logistic regression model in JAGS

library('R2jags')

npo.model <- function(){

### Likelihood ###

# Cumulative Clavien-Dindo Probabilities, i.e., Prob(y <= s | Sgp, Trt), s=0,...,4

for(s in 1:5){

logit(Q[1,1,s]) <- alpha[s] - 0.5*gamma[1,s] - 0.5*gamma[2,s] + 0.25*gamma[3,s] #Primary,
    Control

logit(Q[2,1,s]) <- alpha[s] + 0.5*gamma[1,s] - 0.5*gamma[2,s] - 0.25*gamma[3,s] #Salvage,
    Control

logit(Q[1,2,s]) <- alpha[s] - 0.5*gamma[1,s] + 0.5*gamma[2,s] - 0.25*gamma[3,s] #Primary,
    NuPrehab

logit(Q[2,2,s]) <- alpha[s] + 0.5*gamma[1,s] + 0.5*gamma[2,s] + 0.25*gamma[3,s] #Salvage,
    NuPrehab

}

# Clavien-Dindo Probabilities, i.e., Prob(y = s | Sgp, Trt), s=0,...,5

for(j in 1:2){

for(k in 1:2){

pi[j,k,1] <- Q[j,k,1]

for(s in 2:5){

pi[j,k,s] <- Q[j,k,s]-Q[j,k,s-1]

}

pi[j,k,6] <- 1-Q[j,k,5]

}

}

# Likelihood Contributions

for(i in 1:n){

y.plus.one[i] ~ dcat(pi[sgp[i],trt[i],])

}


### Prior ###

alpha[1] ~ dt(alpha.star[1],0.4,5)
```

```
for(s in 2:5){
alpha[s] ~ dt(alpha.star[s],0.4,5)%_%T(alpha[s-1]+0.5*(abs(gamma[1,s-1]-gamma[1,s])+abs(
    gamma[2,s-1]-gamma[2,s])+0.5*abs(gamma[3,s-1]-gamma[3,s])),)
}
for(m in 1:3){
for(s in 1:5){
gamma[m,s] ~ dnorm(beta[m],tau[m])
}
beta[m] ~ dt(beta.star[m],0.4,5)
tau[m] <- 1/sigma[m]^2
sigma[m] ~ dnorm(0,4)%_%T(0,)
}


### Mean Utilities ###
for(j in 1:2){
for(k in 1:2){
u[j,k] <- inprod(U[1:6],pi[j,k,1:6])
}
}
}




### Function to fit NPO model in JAGS and return the posterior probabilities for the
    decision rules
# Inputs:
#   dataset - data.frame with the three variables, (y, sgp, trt)
#   U - numerical utilities, default corresponds to the values for the IMPACT trial
#   pi.star - prior POM score probabilities, default corresponds to the values for the
      IMPACT trial
#   nsamps - number of posterior samples to generate, default is 10000
```

```
#    warmup - number of warmup samples, default is 500
# Outputs:
#    vector of the four posterior probabilities,
#       1 - posterior probability that nutritional prehabilitation is superior to the
       control for primary patients
#       2 - posterior probability that nutritional prehabilitation is superior to the
       control for salvage patients
#       3 - posterior probability that nutritional prehabilitation is inferior to the
       control for primary patients
#       4 - posterior probability that nutritional prehabilitation is inferior to the
       control for salvage patients
npo.analysis <- function(dataset,U=c(100,80,65,25,10,0),pi.star=rbind(c
    (0.50,0.20,0.10,0.10,0.05,0.05),c(0.30,0.25,0.10,0.10,0.10,0.15)),nsamps=10000,warmup
    =500){
# Data preparation
n = nrow(dataset); y = dataset$y; sgp = dataset$sgp; trt = dataset$trt; y.plus.one = y+1
dat = list("n","y.plus.one","sgp","trt","alpha.star","beta.star","U"); params = c("u")


# Prior Specification
alpha.star <- colMeans(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))))
beta.star <- c(mean(apply(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))),2,function(x) diff(x))),0,0)


# Fit model with interaction
fit <- jags(dat,inits=NULL, parameters.to.save=params, model.file=npo.model, n.iter=
    nsamps+warmup, n.burnin=warmup, n.chains=1, n.thin=1)


# Calculate Posterior Probabilities
pp.prim.sup = mean(fit$BUGSoutput$sims.list$u[,1,2] > fit$BUGSoutput$sims.list$u[,1,1])
pp.salv.sup = mean(fit$BUGSoutput$sims.list$u[,2,2] > fit$BUGSoutput$sims.list$u[,2,1])
```

```
pp.prim.inf = mean(fit$BUGSoutput$sims.list$u[,1,2] < fit$BUGSoutput$sims.list$u[,1,1])

pp.salv.inf = mean(fit$BUGSoutput$sims.list$u[,2,2] < fit$BUGSoutput$sims.list$u[,2,1])


return(round(c(pp.prim.sup,pp.salv.sup,pp.prim.inf,pp.salv.inf),4))

}




### Function to get the decision for the NPO model-based stratified medicine design at
    the current analysis
# Inputs:
#   dataset - data.frame with the three variables, (y, sgp, trt)
#   analysis - "Interim" or "Final"
#   active.subgroups - active subgroups, i.e,. the subgroups for which no decision has
    been made at a previous analysis
#                     at the interim analysis this will be c("P","S"), i.e., both Primary
     (P) and Salvage (S) subgroups are active
#                     at the final analysis this may be either c("P","S"), c("P"), or c("
    S")
#   U - numerical utilities, default corresponds to the values for the IMPACT trial
#   pi.star - prior POM score probabilities, default corresponds to the values for the
      IMPACT trial
#   nsamps - number of posterior samples to generate, default is 10000
#   warmup - number of warmup samples, default is 500
#   alpha - type I error, default is 0.05
# Outputs: decision
library('gsDesign')
get.decision <- function(dataset,analysis=c("Interim","Final"),active.subgroups=c("P","S
    "),U=c(100,80,65,25,10,0),pi.star=rbind(c(0.50,0.20,0.10,0.10,0.05,0.05),c
    (0.30,0.25,0.10,0.10,0.10,0.15)),nsamps=10000,warmup=500,alpha=0.05){


#Calculate Relevant Posterior Probabilities
post.probs <- npo.analysis(dataset=dataset,U=U,pi.star=pi.star,nsamps=nsamps,warmup=
```

```
    warmup)


#Get Cutoffs for Declaring Superiority/Inferiority at Each Analysis
cutoffs <- pnorm(gsDesign(alpha=alpha/2,k=2,test.type=2,sfu=sfPower,sfupar=3)$upper$bound
    )


#Output what subgroups are still active and whether this is an iterim or final analysis
if(analysis == "Interim"){ cat("This is an Interim Analysis. \n\n"); cut = cutoffs[1]}
if(analysis == "Final"){ cat("This is the Final Analysis. \n\n"); cut = cutoffs[2]}
if("P" %in% active.subgroups) cat("No Decision has previously been made for Primary
    Patients. \n")
if(!("P" %in% active.subgroups)) cat("A Decision was made at an earlier analysis for
    Primary Patients. \n")
if("S" %in% active.subgroups) cat("No Decision has previously been made for Salvage
    Patients. \n\n")
if(!("S" %in% active.subgroups)) cat("A Decision was made at an earlier analysis for
    Salvage Patients. \n\n")


#Output Posterior Probabilities Corresponding to Active Subgroup and the Cut-off
cat(paste("The Posterior Probability Cutoff for Declaring Superiority/Inferiority is",
    round(cut,4),"\n\n",sep=" "))
if("P" %in% active.subgroups){
cat(paste("The Posterior Probability that NuPrehab is Superior to Placebo for Primary
    Patients is",post.probs[1],"\n",sep=" "))
cat(paste("The Posterior Probability that NuPrehab is Inferior to Placebo for Primary
    Patients is",post.probs[3],"\n\n",sep=" "))
}
if("S" %in% active.subgroups){
cat(paste("The Posterior Probability that NuPrehab is Superior to Placebo for Salvage
    Patients is",post.probs[2],"\n",sep=" "))
cat(paste("The Posterior Probability that NuPrehab is Inferior to Placebo for Salvage
    Patients is",post.probs[4],"\n\n",sep=" "))
}
```

```
#Output Decision
cat("Therefore, \n")
#Interim Analysis Decisions
if(analysis == "Interim" & "P" %in% active.subgroups & post.probs[1] > cut) cat("Stop
    Enrolling Primary Patients: NuPrehab is Superior to Placebo for Primary Patients \n")
if(analysis == "Interim" & "P" %in% active.subgroups & post.probs[3] > cut) cat("Stop
    Enrolling Primary Patients: NuPrehab is Inferior to Placebo for Primary Patients \n")
if(analysis == "Interim" & "P" %in% active.subgroups & max(post.probs[c(1,3)]) <= cut)
    cat("Continue Enrolling Primary Patients \n")
if(analysis == "Interim" & "S" %in% active.subgroups & post.probs[2] > cut) cat("Stop
    Enrolling Salvage Patients: NuPrehab is Superior to Placebo for Salvage Patients \n")
if(analysis == "Interim" & "S" %in% active.subgroups & post.probs[4] > cut) cat("Stop
    Enrolling Salvage Patients: NuPrehab is Inferior to Placebo for Salvage Patients \n")
if(analysis == "Interim" & "S" %in% active.subgroups & max(post.probs[c(2,4)]) <= cut)
    cat("Continue Enrolling Salvage Patients \n")
#Final Analysis Decisions
if(analysis == "Final" & "P" %in% active.subgroups & post.probs[1] > cut) cat("NuPrehab
    is Superior to Placebo for Primary Patients \n")
if(analysis == "Final" & "P" %in% active.subgroups & post.probs[3] > cut) cat("NuPrehab
    is Inferior to Placebo for Primary Patients \n")
if(analysis == "Final" & "P" %in% active.subgroups & max(post.probs[c(1,3)]) <= cut) cat
    ("The Trial is Inconclusive for Primary Patients \n")
if(analysis == "Final" & "S" %in% active.subgroups & post.probs[2] > cut) cat("NuPrehab
    is Superior to Placebo for Salvage Patients \n")
if(analysis == "Final" & "S" %in% active.subgroups & post.probs[4] > cut) cat("NuPrehab
    is Inferior to Placebo for Salvage Patients \n")
if(analysis == "Final" & "S" %in% active.subgroups & max(post.probs[c(2,4)]) <= cut) cat
    ("The Trial is Inconclusive for Salvage Patients \n")
}
```

```
### Function to fit proportional odds cumulative logistic regression model in JAGS
library('R2jags')
po.model <- function(){
### Likelihood ###
# Cumulative Clavien-Dindo Probabilities, i.e., Prob(y <= s | Sgp, Trt), s=0,...,4
for(s in 1:5){
logit(Q[1,1,s]) <- alpha[s] - 0.5*beta[1] - 0.5*beta[2] + 0.25*beta[3] #Primary, Control
logit(Q[1,2,s]) <- alpha[s] - 0.5*beta[1] + 0.5*beta[2] - 0.25*beta[3] #Primary, NuPrehab
logit(Q[2,1,s]) <- alpha[s] + 0.5*beta[1] - 0.5*beta[2] - 0.25*beta[3] #Salvage, Control
logit(Q[2,2,s]) <- alpha[s] + 0.5*beta[1] + 0.5*beta[2] + 0.25*beta[3] #Salvage, NuPrehab
}
# Clavien-Dindo Probabilities, i.e., Prob(y = s | Sgp, Trt), s=0,...,5
for(j in 1:2){
for(k in 1:2){
pi[j,k,1] <- Q[j,k,1]
for(s in 2:5){
pi[j,k,s] <- Q[j,k,s]-Q[j,k,s-1]
}
pi[j,k,6] <- 1-Q[j,k,5]
}
}
# Likelihood Contributions
for(i in 1:n){
y.plus.one[i] ~ dcat(pi[sgp[i],trt[i],])
}


### Prior ###
alpha[1] ~ dt(alpha.star[1],0.4,5)
for(s in 2:5){
alpha[s] ~ dt(alpha.star[s],0.4,5)%_%T(alpha[s-1],)
}
for(m in 1:3){
beta[m] ~ dt(beta.star[m],0.4,5)
```

```
}


### Mean Utilities ###

for(j in 1:2){

for(k in 1:2){

u[j,k] <- inprod(U[1:6],pi[j,k,1:6])

}

}

}
```

```
### Function to fit PO model in JAGS and return the posterior probabilities for the
    decision rules
# Inputs:
#   dataset - data.frame with the three variables, (y, sgp, trt)
#   U - numerical utilities, default corresponds to the values for the IMPACT trial
#   pi.star - prior POM score probabilities, default corresponds to the values for the
      IMPACT trial
#   nsamps - number of posterior samples to generate, default is 10000
#   warmup - number of warmup samples, default is 500
# Outputs:
#   vector of the four posterior probabilities,
#     1 - posterior probability that nutritional prehabilitation is superior to the
      control for primary patients
#     2 - posterior probability that nutritional prehabilitation is superior to the
      control for salvage patients
#     3 - posterior probability that nutritional prehabilitation is inferior to the
      control for primary patients
#     4 - posterior probability that nutritional prehabilitation is inferior to the
      control for salvage patients
po.analysis <- function(dataset,U=c(100,80,65,25,10,0),pi.star=rbind(c
```

```r
    (0.50,0.20,0.10,0.10,0.05,0.05),c(0.30,0.25,0.10,0.10,0.10,0.15)),nsamps=10000,warmup
    =500){
# Data preparation
n = nrow(dataset); y = dataset$y; sgp = dataset$sgp; trt = dataset$trt; y.plus.one = y+1
dat = list("n","y.plus.one","sgp","trt","alpha.star","beta.star","U"); params = c("u")


# Prior Specification
alpha.star <- colMeans(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))))
beta.star <- c(mean(apply(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))),2,function(x) diff(x))),0,0)


# Fit model with interaction
fit <- jags(dat,inits=NULL, parameters.to.save=params, model.file=po.model, n.iter=nsamps
    +warmup, n.burnin=warmup, n.chains=1, n.thin=1)


# Calculate Posterior Probabilities
pp.prim.sup = mean(fit$BUGSoutput$sims.list$u[,1,2] > fit$BUGSoutput$sims.list$u[,1,1])
pp.salv.sup = mean(fit$BUGSoutput$sims.list$u[,2,2] > fit$BUGSoutput$sims.list$u[,2,1])
pp.prim.inf = mean(fit$BUGSoutput$sims.list$u[,1,2] < fit$BUGSoutput$sims.list$u[,1,1])
pp.salv.inf = mean(fit$BUGSoutput$sims.list$u[,2,2] < fit$BUGSoutput$sims.list$u[,2,1])


return(round(c(pp.prim.sup,pp.salv.sup,pp.prim.inf,pp.salv.inf),4))
}




### Function to faciliate fitting the propotional odds logistic regression model in JAGS
    (no input or output)
library('R2jags')
```

```
trad.model <- function(){
### Likelihood ###
# Cumulative Clavien-Dindo Probabilities, i.e., Prob(y <= s | Sgp, Trt), s=0,...,4
for(s in 1:5){
logit(Q[1,1,s]) <- alpha[s] - 0.5*beta[1] - 0.5*beta[2] #Primary, Control
logit(Q[1,2,s]) <- alpha[s] - 0.5*beta[1] + 0.5*beta[2] #Primary, NuPrehab
logit(Q[2,1,s]) <- alpha[s] + 0.5*beta[1] - 0.5*beta[2] #Salvage, Control
logit(Q[2,2,s]) <- alpha[s] + 0.5*beta[1] + 0.5*beta[2] #Salvage, NuPrehab
}
# Clavien-Dindo Probabilities, i.e., Prob(y = s | Sgp, Trt), s=0,...,5
for(j in 1:2){
for(k in 1:2){
pi[j,k,1] <- Q[j,k,1]
for(s in 2:5){
pi[j,k,s] <- Q[j,k,s]-Q[j,k,s-1]
}
pi[j,k,6] <- 1-Q[j,k,5]
}
}
# Likelihood Contributions
for(i in 1:n){
y.plus.one[i] ~ dcat(pi[sgp[i],trt[i],])
}


### Prior ###
alpha[1] ~ dt(alpha.star[1],0.4,5)
for(s in 2:5){
alpha[s] ~ dt(alpha.star[s],0.4,5)%_%T(alpha[s-1],)
}
for(m in 1:2){
beta[m] ~ dt(beta.star[m],0.4,5)
}
}
```

```
### Function to fit traditional PO model in JAGS and return the posterior probabilities
    for the decision rules
# Inputs:
#   dataset - data.frame with the three variables, (y, sgp, trt)
#   pi.star - prior POM score probabilities, default corresponds to the values for the
    IMPACT trial
#   nsamps - number of posterior samples to generate, default is 10000
#   warmup - number of warmup samples, default is 500
# Outputs:
#   vector of the two posterior probabilities,
#      1 - posterior probability that nutritional prehabilitation is superior to the
    control
#      2 - posterior probability that nutritional prehabilitation is inferior to the
    control
trad.analysis <- function(dataset,pi.star=rbind(c(0.50,0.20,0.10,0.10,0.05,0.05),c
    (0.30,0.25,0.10,0.10,0.10,0.15)),nsamps=10000,warmup=500){
# Data preparation
n = nrow(dataset); y = dataset$y; sgp = dataset$sgp; trt = dataset$trt; y.plus.one = y+1
dat = list("n","y.plus.one","sgp","trt","alpha.star","beta.star"); params = c("beta")


# Prior Specification
alpha.star <- colMeans(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))))
beta.star <- c(mean(apply(rbind(sapply(1:5,function(s) log(sum(pi.star[1,1:s])/(1-sum(pi.
    star[1,1:s])))),sapply(1:5,function(s) log(sum(pi.star[2,1:s])/(1-sum(pi.star[2,1:s])
    )))),2,function(x) diff(x))),0)


# Fit model with interaction
```

```
fit <- jags(dat,inits=NULL, parameters.to.save=params, model.file=trad.model, n.iter=
    nsamps+warmup, n.burnin=warmup, n.chains=1, n.thin=1)


# Calculate Posterior Probabilities

pp.sup = mean(fit$BUGSoutput$sims.list$beta[,2] > 0)

pp.inf = mean(fit$BUGSoutput$sims.list$beta[,2] < 0)


return(round(c(pp.sup,pp.inf),4))
}
```

7.3   simulation.R

```
### This file is for replicating the simulation study reported in Murray et al. (2017), "
    A Bayesian Utility-Based Stratified Medicine Design for the Effectiveness of
    Nutritional Prehabilitation in Thoracic Surgery"
### Written by Thomas A Murray on 14 June 2017
### E-mail: 8tmurray@gmail.com


# Set working directory
setwd("C:/Users/")


# Load functions
source('functions.R')


# Analysis Sample Sizes, probability of each subject being a primary patient
ns = c(50,100); prob.primary = 0.6


# Elicited Utilities
U = c(100,80,65,25,10,0)


# Prior mean probabilities in the control arm of the two patient subgroups
pi.star = rbind(c(0.50,0.20,0.10,0.10,0.05,0.05), #Primary Patients
c(0.30,0.25,0.10,0.10,0.10,0.15)) #Salvage Patients
```

```
# MCMC Specifications
nsamps=10000; warmup=500


# Probability Thresholds
# for parallel you need to supply the cutoffs by hand since the server doesn't have
    gsDesign package...or you can request to have this added. To do this, just calculate
    the cutoffs on your own machine. right now these are 0.997 and 0.976 at the interim
    and final analyses
alpha = 0.05; cutoffs <- round(pnorm(gsDesign(alpha=alpha/2,k=2,test.type=2,sfu=sfPower,
    sfupar=3)$upper$bound),3)


# Number of trial simulations per scenario
niters = 2 #We use 5000 iterations in the paper. We recommend parallelizing the code in
    the sequel.


# Number of scenarios
scens = 1:7


# Run Simulation
for(scen in scens){
# True POM Score Probabilities in each scenario
pis = array(NA,c(2,2,6)); #Cumulative POM Score Probabilities <0 (= 0.00), <=0, <=1, ...,
    <=4, <=5 (= 1.00)
pis[1,1,] = diff(c(0.00,0.50,0.70,0.80,0.90,0.95,1.00)) #Primary, Control Arm
pis[2,1,] = diff(c(0.00,0.30,0.55,0.65,0.75,0.85,1.00)) #Salvage, Control Arm


if(scen == 1){
pis[1,2,] = diff(c(0.00,0.50,0.70,0.80,0.90,0.95,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.30,0.55,0.65,0.75,0.85,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 2){
pis[1,2,] = diff(c(0.00,0.50,0.70,0.80,0.90,0.95,1.00)) #Primary, Nuprehab Arm
```

```
pis[2,2,] = diff(c(0.00,0.71,0.87,0.91,0.94,0.97,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 3){
pis[1,2,] = diff(c(0.00,0.83,0.92,0.95,0.98,0.99,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.30,0.55,0.65,0.75,0.85,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 4){
pis[1,2,] = diff(c(0.00,0.83,0.92,0.95,0.98,0.99,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.71,0.87,0.91,0.94,0.97,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 5){
pis[1,2,] = diff(c(0.00,0.50,0.70,0.80,0.90,0.95,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.53,0.80,0.91,0.94,0.97,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 6){
pis[1,2,] = diff(c(0.00,0.67,0.85,0.95,0.98,0.99,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.30,0.55,0.65,0.75,0.85,1.00)) #Salvage, Nuprehab Arm
}
if(scen == 7){
pis[1,2,] = diff(c(0.00,0.67,0.85,0.95,0.98,0.99,1.00)) #Primary, Nuprehab Arm
pis[2,2,] = diff(c(0.00,0.53,0.80,0.91,0.94,0.97,1.00)) #Salvage, Nuprehab Arm
}


# True log-odds ratios compartion Nuprehab to Control at each POM score level in Primary
    then Salvage
#log(cumsum(pis[1,2,1:5])/(1-cumsum(pis[1,2,1:5])))-log(cumsum(pis[1,1,1:5])/(1-cumsum(
    pis[1,1,1:5])))
#log(cumsum(pis[2,2,1:5])/(1-cumsum(pis[2,2,1:5])))-log(cumsum(pis[2,1,1:5])/(1-cumsum(
    pis[2,1,1:5])))


set.seed(1985)
for(iter in 1:niters){
### Create Storage Objects
```

```
pps.npo = pps.po = matrix(NA,nrow=length(ns),ncol=4)
pps.trad = matrix(NA,nrow=length(ns),ncol=2)


### Generate a Dataset
dataset = get.dataset(n=max(ns),pis=pis,prob.primary=prob.primary)


### Run each analysis function at the interim analysis
pps.npo[1,] = npo.analysis(U=U,dataset=dataset[1:ns[1],],pi.star=pi.star,nsamps=nsamps,
    warmup=warmup)
pps.po[1,]  = po.analysis(U=U,dataset=dataset[1:ns[1],],pi.star=pi.star,nsamps=nsamps,
    warmup=warmup)
pps.trad[1,] = trad.analysis(dataset=dataset[1:ns[1],],pi.star=pi.star,nsamps=nsamps,
    warmup=warmup)


### Run each analysis function at the final analysis based on the interim decision
# NPO Model-based Stratified Medicine Design
if((pps.npo[1,1] > cutoffs[1] | pps.npo[1,3] > cutoffs[1]) & (pps.npo[1,2] > cutoffs[1] |
    pps.npo[1,4] > cutoffs[1])) {
pps.npo[2,] = NA
}
if((pps.npo[1,1] > cutoffs[1] | pps.npo[1,3] > cutoffs[1]) & (pps.npo[1,2] <= cutoffs[1]
    & pps.npo[1,4] <= cutoffs[1])) {
pps.npo[2,] = npo.analysis(U=U,dataset=rbind(dataset[1:ns[1],],subset(dataset[(ns[1]+1):
    ns[2],],sgp==2)),pi.star=pi.star,nsamps=nsamps,warmup=warmup)
pps.npo[2,c(1,3)] = NA
}
if((pps.npo[1,1] <= cutoffs[1] & pps.npo[1,3] <= cutoffs[1]) & (pps.npo[1,2] > cutoffs[1]
    | pps.npo[1,4] > cutoffs[1])) {
pps.npo[2,] = npo.analysis(U=U,dataset=rbind(dataset[1:ns[1],],subset(dataset[(ns[1]+1):
    ns[2],],sgp==1)),pi.star=pi.star,nsamps=nsamps,warmup=warmup)
pps.npo[2,c(2,4)] = NA
}
if((pps.npo[1,1] <= cutoffs[1] & pps.npo[1,3] <= cutoffs[1]) & (pps.npo[1,2] <= cutoffs
```

```
     [1] & pps.npo[1,4] <= cutoffs[1])) {
pps.npo[2,] = npo.analysis(U=U,dataset=dataset,pi.star=pi.star,nsamps=nsamps,warmup=
    warmup)
}


# PO Model-based Stratified Medicine Design
if((pps.po[1,1] > cutoffs[1] | pps.po[1,3] > cutoffs[1]) & (pps.po[1,2] > cutoffs[1] |
    pps.po[1,4] > cutoffs[1])) {
pps.po[2,] = NA
}
if((pps.po[1,1] > cutoffs[1] | pps.po[1,3] > cutoffs[1]) & (pps.po[1,2] <= cutoffs[1] &
    pps.po[1,4] <= cutoffs[1])) {
pps.po[2,] = po.analysis(U=U,dataset=rbind(dataset[1:ns[1],],subset(dataset[(ns[1]+1):ns
    [2],],sgp==2)),pi.star=pi.star,nsamps=nsamps,warmup=warmup)
pps.po[2,c(1,3)] = NA
}
if((pps.po[1,1] <= cutoffs[1] & pps.po[1,3] <= cutoffs[1]) & (pps.po[1,2] > cutoffs[1] |
    pps.po[1,4] > cutoffs[1])) {
pps.po[2,] = po.analysis(U=U,dataset=rbind(dataset[1:ns[1],],subset(dataset[(ns[1]+1):ns
    [2],],sgp==1)),pi.star=pi.star,nsamps=nsamps,warmup=warmup)
pps.po[2,c(2,4)] = NA
}
if((pps.po[1,1] <= cutoffs[1] & pps.po[1,3] <= cutoffs[1]) & (pps.po[1,2] <= cutoffs[1] &
     pps.po[1,4] <= cutoffs[1])) {
pps.po[2,] = po.analysis(U=U,dataset=dataset,pi.star=pi.star,nsamps=nsamps,warmup=warmup)
}


# PO Model-based Traditional Design
if(pps.trad[1,1] > cutoffs[1] | pps.trad[1,2] > cutoffs[1]){
pps.trad[2,] = NA
}
if(pps.trad[1,1] <= cutoffs[1] & pps.trad[1,2] <= cutoffs[1]){
pps.trad[2,] = trad.analysis(dataset=dataset,pi.star=pi.star,nsamps=nsamps,warmup=warmup)
```

```
}


### Write posterior probabilities to .txt files

write.table(rbind(as.vector(pps.npo)),file=paste("NPO-Res-Scen",scen,".txt",sep=""),sep
    =",",append=TRUE,row.names=FALSE,col.names=FALSE)
write.table(rbind(as.vector(pps.po)),file=paste("PO-Res-Scen",scen,".txt",sep=""),sep
    =",",append=TRUE,row.names=FALSE,col.names=FALSE)
write.table(rbind(as.vector(pps.trad)),file=paste("Trad-Res-Scen",scen,".txt",sep=""),sep
    =",",append=TRUE,row.names=FALSE,col.names=FALSE)
}
}




### Results
library('gsDesign')
#For alpha = 0.05 type I error, the monitoring Boundaries at interim and final analyses
    are
alpha = 0.05; cutoffs <- round(pnorm(gsDesign(alpha=alpha/2,k=2,test.type=2,sfu=sfPower,
    sfupar=3)$upper$bound),3)


### Load NPO Results
res = read.csv("NPO-Res-Scen1.txt",header=FALSE)


# Calculate Power Figures
mean(res[,1] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1] & res[,2] >
    cutoffs[2])) #N superior for P patients
mean(res[,5] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1] & res[,6] >
    cutoffs[2])) #N inferior for P patients
mean(res[,3] > cutoffs[1] | (res[,3] <= cutoffs[1] & res[,7] <= cutoffs[1] & res[,4] >
    cutoffs[2])) #N superior for S patients
mean(res[,7] > cutoffs[1] | (res[,3] <= cutoffs[1] & res[,7] <= cutoffs[1] & res[,8] >
```

```
    cutoffs[2])) #N inferior for S patients


# Calculate Expected Sample Size
50*mean((res[,1] > cutoffs[1] | res[,5] > cutoffs[1]) & (res[,3] > cutoffs[1] | res[,7] >
     cutoffs[1])) +
70*mean((res[,1] > cutoffs[1] | res[,5] > cutoffs[1]) & (res[,3] <= cutoffs[1] & res[,7]
    <= cutoffs[1])) +
80*mean((res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1]) & (res[,3] > cutoffs[1] | res[,7]
     > cutoffs[1])) +
100*mean((res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1]) & (res[,3] <= cutoffs[1] & res
    [,7] <= cutoffs[1]))



### Load PO Results
res = read.csv("PO-Res-Scen1.txt",header=FALSE)


# Calculate Power Figures
mean(res[,1] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1] & res[,2] >
    cutoffs[2])) #N superior for P patients
mean(res[,5] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1] & res[,6] >
    cutoffs[2])) #N inferior for P patients
mean(res[,3] > cutoffs[1] | (res[,3] <= cutoffs[1] & res[,7] <= cutoffs[1] & res[,4] >
    cutoffs[2])) #N superior for S patients
mean(res[,7] > cutoffs[1] | (res[,3] <= cutoffs[1] & res[,7] <= cutoffs[1] & res[,8] >
    cutoffs[2])) #N inferior for S patients


# Calculate Expected Sample Size
50*mean((res[,1] > cutoffs[1] | res[,5] > cutoffs[1]) & (res[,3] > cutoffs[1] | res[,7] >
     cutoffs[1])) +
70*mean((res[,1] > cutoffs[1] | res[,5] > cutoffs[1]) & (res[,3] <= cutoffs[1] & res[,7]
    <= cutoffs[1])) +
80*mean((res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1]) & (res[,3] > cutoffs[1] | res[,7]
     > cutoffs[1])) +
```

```
100*mean((res[,1] <= cutoffs[1] & res[,5] <= cutoffs[1]) & (res[,3] <= cutoffs[1] & res
    [,7] <= cutoffs[1]))



### Load Trad Results
res = read.csv("Trad-Res-Scen1.txt",header=FALSE)


# Calculate Power Figures
mean(res[,1] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,3] <= cutoffs[1] & res[,2] >
    cutoffs[2])) #N superior
mean(res[,3] > cutoffs[1] | (res[,1] <= cutoffs[1] & res[,3] <= cutoffs[1] & res[,4] >
    cutoffs[2])) #N inferior


# Calculate Expected Sample Size
50*mean((res[,1] > cutoffs[1] | res[,3] > cutoffs[1])) + 100*mean(res[,1] <= cutoffs[1] &
     res[,3] <= cutoffs[1])
```