Utility-Based Designs for Randomized Comparative Trials with Categorical Outcomes

Thomas A. Murray[*], Peter F. Thall[†] and Ying Yuan[‡]

Department of Biostatistics, The University of Texas MD Anderson Cancer Center

[*]TAMurray@MDAnderson.org, [†]Rex@MDAnderson.org, [‡]YYuan@MDAnderson.org

This work was partially funded by NIH/NCI grant 5-R01-CA083932.

April 26, 2016

Abstract

A general utility-based testing methodology for design and conduct of randomized comparative clinical trials with categorical outcomes is presented. Numerical utilities of all elementary events are elicited to quantify their desirabilities. These numerical values are used to map the categorical outcome probability vector of each treatment to a mean utility, which is used as a one-dimensional criterion for constructing comparative tests. Bayesian tests are presented, including fixed sample and group sequential procedures, assuming Dirichlet-multinomial models for the priors and likelihoods. Guidelines are provided for establishing priors, eliciting utilities, and specifying hypotheses. Efficient posterior computation is discussed, and algorithms are provided for jointly calibrating test cutoffs and sample size to control overall type I error and achieve specified power. Asymptotic approximations for the power curve are used to initialize the algorithms. The methodology is applied to re-design a completed trial that compared two chemotherapy regimens for chronic lymphocytic leukemia, in which an ordinal efficacy outcome was dichotomized and toxicity was ignored to construct the trial's design. The Bayesian tests also are illustrated by several types of categorical outcomes arising in common clinical settings. Freely available computer software for implementation is provided.

Keywords: Bayesian Methods; Dirichlet-multinomial; Multiple Outcomes; Oncology; Randomized Comparative Trials; Utility Elicitation.

1 Introduction

Medical outcomes often are complex and multivariate. Physicians routinely select each patient's treatment based on consideration of risk-benefit tradeoffs between desirable and undesirable clinical outcomes. Conventional designs for randomized comparative trials (RCTs) seldom reflect this aspect of medical practice. Rather, most designs in clinical trial protocols are based on one outcome, identified as "primary," with all other outcomes given the nominal status of "secondary." This dichotomy often is codified in institutionally required protocol formats. For example, in cancer studies of chemotherapies for solid tumors, the primary outcome may be *objective response*, defined as 30% or greater tumor shrinkage compared to baseline evaluation, while regimen-related adverse events, called "toxicities," are listed as secondary outcomes Eisenhauer et al. (2009). This approach is convenient because it facilitates sample size and power computations in terms of the probabilities of a one-dimensional outcome in the treatment arms. It does not reflect the way that practicing physicians actually think and behave, however. Alternative design approaches include defining a composite outcome that treats efficacy and safety events equally Sankoh et al. (2003); Pocock (1997), using a test statistic that is a weighted average Freedman et al. (1996), or basing a test on a quadratic form, such as Hotelling's T-squared statistic, with weights estimated to reflect variability Hotelling (1931). These approaches ignore the relative clinical importance of beneficial and adverse outcomes, however.

Safety is never a secondary concern in a clinical trial. In actual trial conduct, if interim data from a randomized clinical trial (RCT) show that one treatment has a much higher adverse event rate than the other, or that both arms are unacceptably toxic in a trial comparing two experimental agents, the physicians conducting the trial will terminate accrual whether the protocol's design includes a formal safety stopping rule or not. Such a decision shows that, due to their unwillingness to continue the trial, the physicians have decided that one treatment is inferior to the other in terms of safety. While stopping a trial due to an unacceptably high adverse event rate is an ethical decision, it also is part of the general consideration of how much risk of an adverse outcome is acceptable as a tradeoff for a given level of therapeutic benefit.

This paper is motivated by the consideration that, because clinical trial conduct must accommodate medical practice, a trial design should account formally for risk-benefit tradeoffs between all clinically relevant outcomes. That is, in actual trial design and conduct, scientific and ethical considerations should not be separated. We provide a practical framework for including such tradeoffs explicitly in the treatment comparison underlying the design of two-arm RCTs. We focus on settings where the clinically relevant events are categorical, and thus the outcome Y is a realization from a finite set of elementary patient outcomes. The clinically relevant events, and the resulting set of elementary outcomes, are determined in collaboration with the physician(s) planning the trial. The proposed framework accommodates most discrete outcome structures that occur in practice, including univariate ordinal, bivariate binary indicators of efficacy and safety, bivariate ordinal variables, and such bivariate variables with death as a separate event.

1.1 A Trial in Chronic Lymphocytic Leukemia

We illustrate the proposed methodology by applying it to re-design a RCT reported by Flinn et al. (2007) that compared two chemotherapy regimens for untreated chronic lymphocytic leukemia (CLL), FC = fludarabine plus cyclophosphamide versus F = fludarabine alone. Patients in this study were treated for up to six 28-day cycles. Following the recommended guidelines at the time of the trial Cheson et al. (1996), patients were monitored for *clinical response*, with categories CR = Complete response, PR = Partial response, SD = Stable disease, and PD = Progressive disease.

Patients also were monitored for several adverse events (AEs), including infections, with severity grades {None, Minor, Major, Fatal}, hematological toxicities with severity grades 0-5, and non-hematological toxicities graded 0-5, according to the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE). Detailed definitions of the levels of clinical response and the AEs are given in Cheson et al. (1996).

In the CLL trial design, CR was designated as the primary outcome, with all other outcomes designated as secondary. Thus, the comparison of FC to F was based on the probabilities of CR in the two arms. For this comparison, since clinical response was not evaluable for patients that died during the observation period, these patients were counted as non-responders. This approach is sensible since it counts death during response evaluation as a treatment failure. In contrast, the non-fatal AEs were not included in the study design, despite the fact that the safety of FC was an important concern. Because the above approach to constructing the design for this trial is quite typical, it serves as a useful illustration of our proposed methodology.

To apply our methodology to design this trial would have required working with the physicians planning the trial to determine the clinically relevant outcomes and elicit their utilities. Thus, for the sake of illustration, we first assume that the physicians decided that the relevant outcomes were clinical response, specifically the ordinal variable with possible values {CR, PR, SD, PD}, and also the worst AE with levels {Minimal, Moderate, Severe, Fatal}. Here, "minimal" is defined as no AE requiring medical intervention, "moderate" as a non-life-threatening AE requiring medical intervention without hospitalization, "severe" as an imminently life-threatening AE requiring hospitalization, and "fatal" as an AE resulting in death. Using these definitions, a moderate AE includes grade 3 hematologic and non-hematologic toxicities and minor infections, and a severe AE includes grade 4 hematologic and non-hematologic toxicities and major infections. To define the values of Y, we denote the 12 = 4×3 non-fatal elementary patient outcomes by the pairs (r, s), for $r = \{CR, PR, SD, PD\}$ and $s = \{Min, Mod, Sev\}$, with the 13^{th} elementary event D = a fatal AE. Thus, for example, (PR, Mod) is the elementary outcome that the patient had a partial response and a moderate worst AE level. Our design requires numerical utilities for the 13 elementary outcomes, which in practice would be elicited from the physicians. Since we cannot do this retrospectively, we specify numerical utilities (Table 1) for the CLL trial's 13 outcomes that may be considered a reasonable representation of what would be obtained in practice. In Section 6, once our methodology has been established, we will compare our proposed design to a design that compares the two regimens based on the probabilities of CR. Because the numerical utilities are a key component our methodology, we also include an analysis of the sensitivity of the final inferences to alternative utilities (Table 6).

1.2 Mean Utilities

For the general development, we index the elementary outcomes by $k = 1, \ldots, K$, and denote their numerical utilities by $U_k = U(Y = k)$, with $\boldsymbol{U} = (U_1, U_2, \cdots, U_K)'$. These are elicited from the physician(s) planning the trial. For some specific examples, we will replace these integer indices with more descriptive indexing schemes. For convenience, we assign the most desirable outcome utility 100, the least desirable outcome utility 0, with all other outcomes assigned utilities between these two extremes. The domain [0, 100] is chosen to facilitate communication with the physician(s), although in general any compact domain will work. In Section 2, we provide practical strategies for utility elicitation, and illustrate them for the case of bivariate-ordinal outcomes that include the possibility of death.

For treatments $j = A$ and B, we denote the patient response probabilities $\theta_{j,k} = \Pr(Y =$

$k \mid trt = j)$, with $\boldsymbol{\theta}_j = (\theta_{j,1}, \theta_{j,2}, \cdots, \theta_{j,K})'$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$. The mean utility of treatment j is

$$\overline{U}\,(\boldsymbol{\theta}_j) = \boldsymbol{U}'\boldsymbol{\theta}_j = \sum_{k=1}^{K} U_k\,\theta_{j,k}. \tag{1}$$

Our testing methodology relies on the mean utilities $\overline{U}\,(\boldsymbol{\theta}_A)$ and $\overline{U}\,(\boldsymbol{\theta}_B)$ as one-dimensional criteria to compare overall treatment effects, since $\overline{U}\,(\boldsymbol{\theta}_A) > \overline{U}\,(\boldsymbol{\theta}_B)$ corresponds to the mean clinical desirability of patient outcome being higher for A than B, and conversely. The Bayesian comparative test relies on the posterior of $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = \overline{U}(\boldsymbol{\theta}_A) - \overline{U}(\boldsymbol{\theta}_B)$.

As a first illustration, suppose that the clinically relevant outcome is trinary where a treatment may result in response (R), failure (F), or neither response nor failure (N) so, temporarily suppressing j, for a single treatment $\boldsymbol{\theta} = (\theta_R, \theta_N, \theta_F)$. In particular, R and F are not complementary events. Since $U_R = 100$ and $U_F = 0$ in this case, only $U_N \in (0, 100)$ need be elicited, and the mean utility is $\overline{U}(\boldsymbol{\theta}) = \theta_R \times 100 + \theta_N \times U_N$, which increases with U_N for any $\boldsymbol{\theta}$. If, for example, $U_N = 60$ and the true outcome probabilities are $(\theta_R,\ \theta_N,\ \theta_F)' = (0.30,\ 0.60,\ 0.10)'$ then the mean utility is $\overline{U}(\boldsymbol{\theta}) = \boldsymbol{U}'\boldsymbol{\theta} = 0.30{\times}100 + 0.60{\times}60 + 0.10{\times}0 = 66$. Next, consider a trial to compare two clot dissolving agents, A and B, for rapid treatment of stroke, with the outcome evaluated within 24 hours from the start of treatment. Response, R, is defined as the clot that caused the stroke being dissolved without a brain hemorrhage or death, failure, F, is defined as a brain hemorrhage or death, and N is the third event that no brain hemorrhage occurred, the patient did not die, but the clot was not dissolved. Suppose that the true outcome probabilities are $\boldsymbol{\theta}_A = (\theta_{A,R},\ \theta_{A,N},\ \theta_{A,F})'$ $= (0.50,\ 0.30,\ 0.20)'$ and $\boldsymbol{\theta}_B = (\theta_{B,R},\ \theta_{B,N},\ \theta_{B,F})' = (0.60,\ 0.30,\ 0.10)'$. Since B has both a larger response probability and a smaller failure probability compared to A, it is clear that B is clinically superior to A. The mean utilities reflect this, since $\overline{U}(\boldsymbol{\theta}_B) = 60 + \theta_{B,N}U_N$ and $\overline{U}(\boldsymbol{\theta}_A) =$ $50 + \theta_{B,N}U_N$, so $\overline{U}(\boldsymbol{\theta}_B) - \overline{U}(\boldsymbol{\theta}_A) = 10$ for all $U_N \in (0, 100)$.

If a third agent, C, has $\boldsymbol{\theta}_C = (0.60,\ 0.10,\ 0.30)'$, then C has a larger response probability than A but also a larger failure probability, so it is not obvious which of the treatments A or C is superior. If $U_N = 50$, then $\overline{U}(\boldsymbol{\theta}_A) = 65$ compared to $\overline{U}(\boldsymbol{\theta}_B) = 75$, so B is superior to A for this utility. The large difference $\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta}) = \overline{U}(\boldsymbol{\theta}_B)$ - $\overline{U}(\boldsymbol{\theta}_A)$ = 75 - 65 = 10 is due to the fact that B increases $\theta_{A,R}$ by 0.10 and also decreases $\theta_{A,F}$ by 0.10. This might be described as a "win-win" scenario for B versus A. Comparing C to A, since $\overline{U}(\boldsymbol{\theta}_C) = \overline{U}(\boldsymbol{\theta}_A) = 65$, that is, A and C have identical mean utilities with $\delta_{\boldsymbol{U},C-B}(\boldsymbol{\theta}) = 0$, they are equally desirable despite the fact that $\boldsymbol{\theta}_A \neq \boldsymbol{\theta}_C$. This is because the increases in both the response and failure probabilities with C compared to A, specifically $\theta_{C,R} - \theta_{A,R}$ = 0.60 - 0.50 = 0.10 and $\theta_{C,F} - \theta_{A,F}$ = 0.30 - 0.20 = 0.10, cancel each other out if $U_N = 50$. If $U_N = 20$ rather than 50, however, then $\delta_{\boldsymbol{U},C-A}(\boldsymbol{\theta}) = 62 - 56 = 6$, so for this utility C is slightly superior to A since the increase in failure probability with C versus A is considered a favorable tradeoff for the increase in response probability.

1.3 Utility-Based Design Framework

Given this general categorical outcome and utility structure, since $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ are not known they must be estimated, and data for doing this must be obtained. The statistical problem thus is how to design and conduct a clinical trial to obtain the necessary data. This requires specification of decision rules, a trial design, and a practical method for establishing a consensus among the investigators for the numerical values in $\boldsymbol{U}$, since the methodology requires one utility and one utility only. This provides a transparent, formal structure that reflects what physicians actually do in practice, rather than constructing a trial design that focuses on a single primary outcome and then, formally or informally, also monitors secondary outcomes. For the Bayesian version, we call the methodology categorical outcome Bayesian utility-based (CAT-BUB) tests. To implement the proposed design framework, in cooperation with the physician(s) planning the trial, one should take the following steps:

(a) Specify the clinically relevant outcomes and resulting set of elementary patient responses.

(b) Elicit numerical utilities.

(c) Specify design parameters, including targeted alternative treatment differences that will be identified with a specified power, type I error, timing of interim analyses for a group sequential test, and test cut-offs.

(d) Implement the design algorithm, developed below, to determine maximum and interim sample sizes and operating characteristics.

(e) Repeat steps (a-d) until a design with satisfactory operating characteristics is identified.

1.4 Outline

In Section 2 we provide practical guidelines for utility elicitation, illustrated for bivariate categorical outcomes. The Dirichlet-multinomial model is reviewed in Section 3. In Section 4 we present the Bayesian utility-based comparative testing procedure. For the Bayesian test, we provide a scaled-beta approximation for the posterior distribution of the mean utility to facilitate calculation of the test statistic and derive frequentist properties, including an approximate sample size calculation that we use to initialize our computational algorithms. In Section 5, we discuss designs for a single test or a group sequential procedure, and provide guidelines for eliciting targeted alternatives, and computational algorithms to derive a CAT-BUB design having given overall type I error and power. In Section 6, we illustrate how to implement the CAT-BUB procedure in several settings and report simulation results, including comparison of the CAT-BUB design for the CLL trial to the design based on a binary indicator of CR. We conclude with a brief discussion in Section 7. The Web Supplement provides additional illustrations for several categorical outcome structures often encountered in practice. To facilitate application, freely available user-friendly software is provided (see Supplementary Materials).

2 Utility Elicitation

Since a utility function is required for implementing the proposed methods, we provide practical utility elicitation guidelines. In our experience, specifying $\boldsymbol{U}$ is an intuitive process for the physician(s) that they find to be quite natural. An extension of the previously discussed trinary outcome case with elementary events $\{R, N, F\}$ is an ordinal Y with four or more categories. For example, in oncology trials it is very common to characterize solid tumor response from the start of chemotherapy as an ordinal variable. Following the RECIST tumor evaluation guidelines Eisenhauer et al. (2009), the outcome may be defined using tumor size relative to baseline, with a 100% decrease a complete response (CR), a 30% to 99% decrease a partial response (PR), a 19% increase to 19% decrease stable disease (SD), and a 20% or greater increase progressive disease (PD). In this and similar contexts, the statistician can simply provide each physician a spreadsheet with the outcomes ordered by desirability and, given $U(CR) = 100$ and $U(PD) = 0$, the physicians can specify numerical utilities for the intermediate outcomes. When there are multiple physicians planning the trial, one approach to establish a consensus utility is the "Delphi" method Dalkey (1969); Brook et al. (1986), wherein one asks each physician independently to specify their numerical utilities, then shows the mean of all elicited utilities to all physicians and allows them to adjust their utilities if desired on that basis, and if needed iterates the process until a consensus is reached.

Another common categorical structure is a bivariate binary (efficacy, toxicity) outcome. An example from chemotherapy for acute myelogenous leukemia (AML) defines efficacy as complete remission, C, in terms of recovery of circulating white cells, platelets, and blastic (undifferentiated) cells to normal levels, and toxicity, T, as severe (NCI grade 3 or 4) non-hematologic toxicity, both scored within 42 days. Denoting the respective complementary events by $\overline{C}$ and $\overline{T}$, the statistician can again simply provide each physician with a spreadsheet that contains a 2×2 utility table with $U(C,\overline{T}) = 100$, $U(\overline{C},T) = 0$, so only the two intermediate utilities $U(\overline{C},\overline{T})$ and $U(C,T)$ must be specified.

A refinement of the bivariate binary (efficacy, toxicity) outcomes is to define these events for patients who are alive, and include death as a fifth event. This is appropriate for treatment of rapidly fatal diseases, such as AML, where death during therapy has a non-trivial probability. In the AML example, the four elementary events determined by C and T are defined only for patients alive at day 42, and the fifth event is D = [death within 42 days]. This structure may motivate the question of whether assigning a finite utility to death is ethically appropriate, since the utilities will be the basis for medical decision making. If the value $U_D = -\infty$ were assigned, however, the mean utility is $-\infty$ whenever the probability of D is non-zero, so in practice a single death would terminate the trial. Thus, when death has a non-trivial probability, if one wishes to actually do utility-based decision making then death must be assigned a finite numerical utility having magnitude comparable to the numerical utilities of the other possible patient outcomes. We recommend that the physician(s) first specify $U(\overline{C},T)$, i.e., the worst outcome for a patient who is alive, relative to $U(C,\overline{T}) = 100$, i.e., the best outcome, and $U(D) = 0$, and then specify $U(\overline{T},\overline{C})$ and $U(C,T)$ relative to the $U(C,\overline{T}) = 100$ and the selected $U(\overline{C},T)$. To implement this, the statistician may ask the physician(s) to fill in the following two tables sequentially,

$(C,\overline{T})$	$(\overline{C},T)$	D
100		0

	C	$\overline{C}$
$\overline{T}$	100	
T		$U(\overline{C},T)$

where $U(\overline{C},T)$ in the right-hand table takes the specified value from the left-hand table. This sequence decomposes utility elicitation into two intuitive steps. It also provides a partial motivation for establishing utilities for our re-design of the CLL trial.

To establish or elicit utilities for the CLL trial outcomes, and in general for bivariate ordinal outcomes with death as a separate event, we propose the following two alternative strategies, one *direct* and the other *indirect*. The *direct* elicitation strategy simply requires the statistician to provide the physician(s) with a utility table and suggest a specification order. For the CLL outcomes, using the direct strategy one would provide the physician(s) the table below and tell them to fill the empty cells in alphabetical order.

	CR	PR	SD	PD	
Min	100	C	C	B	Death
Mod	C	D	D	C	0
Sev	B	C	C	A	

The basic idea is to first specify the utility of the worst non-fatal outcome, then the two most extreme (efficacy, toxicity) trade-off outcomes, then the intermediate outcomes where either the best efficacy or worst toxicity event occurs, and finally the remaining outcomes in the interior portion of the table.

In contrast, the *indirect* strategy decomposes elicitation into a series of intuitive, mutually independent steps that induce numerical utilities. For the CLL outcome, we would implement the indirect strategy by having the physician(s) specify the following sub-tables,

(CR,Min)	(PD,Sev)	D
100	$100\times\nu$	0

	CR	PD
Min	100	$100 \times \zeta_1$
Sev	$100 \times \zeta_2$	0

(CR,Min)	(PR,Min)	(SD,Min)	(PD,Min)
100	$100 \times \phi_{1,1}$	$100 \times \phi_{1,2}$	0

(CR,Sev)	(PR,Sev)	(SD,Sev)	(PD,Sev)
100	$100 \times \phi_{2,1}$	$100 \times \phi_{2,2}$	0

(CR,Min)	(CR,Mod)	(CR,Sev)
100	$100 \times \xi_1$	0

(PD,Min)	(PD,Mod)	(PD,Sev)
100	$100 \times \xi_2$	0

In the above sub-tables, we denote the proportions that will be specified by the physician with Greek symbols, e.g. ν and ζ_1, which we use to determine the induced numerical utilities later. When the statistician provides the sub-tables to the physician(s), these entries will be left blank for the physician(s) to fill in, with the instruction that, for example, ν is the proportion quantifying the desirability of (PD,Sev) relative to (CR,Min), and so on. The sub-tables are mutually independent, i.e., the values in a particular sub-table are not restricted by, or dependent on the values from any other sub-table. Therefore, the sub-tables can be specified in whatever order the physicians prefer, and each can be revisited and adjusted during the specification process until the physicians are satisfied.

Based on the previous sub-tables, the induced numerical utilities can be determined sequentially as follows,

$$\begin{aligned}
U(CR, Min) &= 100, \quad U(D) = 0, \quad U(PD, Sev) = 100\nu, \\
U(PD, Min) &= \zeta_1[U(CR, Min) - U(PD, Sev)] + U(PD, Sev), \\
U(CR, Sev) &= \zeta_2[U(CR, Min) - U(PD, Sev)] + U(PD, Sev), \\
U(PR, Min) &= \phi_{1,1}[U(CR, Min) - U(PD, Min)] + U(PD, Min), \\
U(SD, Min) &= \phi_{1,2}[U(CR, Min) - U(PD, Min)] + U(PD, Min), \\
U(PR, Sev) &= \phi_{2,1}[U(CR, Sev) - U(PD, Sev)] + U(PD, Sev), \\
U(SD, Sev) &= \phi_{2,2}[U(CR, Sev) - U(PD, Sev)] + U(PD, Sev), \\
U(CR, Mod) &= \xi_1[U(CR, Min) - U(CR, Sev)] + U(CR, Sev), \\
U(PD, Mod) &= \xi_2[U(PR, Min) - U(PR, Sev)] + U(PR, Sev), \\
U(PR, Mod) &= \left[\frac{\xi_2(\phi_{1,1} - \phi_{2,1}) + \phi_{2,1}}{1 - (\xi_1 - \xi_2)(\phi_{1,1} - \phi_{2,1})}\right][U(CR, Mod) - U(PD, Mod)] + U(PD, Mod), \text{ and} \\
U(SD, Mod) &= \left[\frac{\xi_2(\phi_{1,2} - \phi_{2,2}) + \phi_{2,2}}{1 - (\xi_1 - \xi_2)(\phi_{1,2} - \phi_{2,2})}\right][U(CR, Mod) - U(PD, Mod)] + U(PD, Mod).
\end{aligned}$$

To aid elicitation, we recommend that the statistician provide the physician(s) with a spreadsheet that contains the relevant sub-tables and a numerical utility table that automatically populates based on the physician's specified values. As an example, we provide such a spreadsheet for the CLL outcome (see Supplementary Materials). In the Web Appendix A, we provide a generalization and detailed derivation of the induced numerical utilities for the indirect elicitation strategy with a $K \times L$ bivariate ordinal outcome plus death.

Table 1: Numerical utilities for the CLL trial's 13 elementary outcomes.

Level of Worst	**Clinical Response**				
Adverse Event	*CR*	*PR*	*SD*	*PD*	
Minimal	100	84	35	19	**Death**
Moderate	93	77	29	14	0
Severe	28	24	14	10	

The proposed indirect strategy facilitates utility elicitation in several important ways. First, when an individual physician is selecting numerical utilities, they can adjust the values in any sub-table and the resulting numerical utilities will repopulate automatically while preserving the partial ordering constraints. In contrast, for the direct strategy, adjusting a single numerical utility may require changing several other values, perhaps even the entire table, which may become impractical if the elementary patient outcome set is large. Second, when the physicians convene to obtain consensus utilities, each sub-table can be addressed independently in turn. Therefore, should a disagreement arise, the physicians can focus on a specific low-dimensional sub-table rather than the entire numerical utility table. Third, the indirect strategy requires the physician(s) to specify fewer values than the direct strategy, which can be a great practical advantage when K and L are both moderately large, say ≥ 4. An advantage of the indirect approach for the statistician is that the sub-tables provide low-dimensional bases for conducting a utility sensitivity assessment, which we discuss below in Section 6.

For our re-design of the CLL trial, suppose that the physician(s) specified sub-table entries corresponding to the following parameters: $\nu = 0.10$, $\zeta_1 = 0.10$, $\zeta_2 = 0.20$, $\phi_{1,1} = \phi_{2,1} = 0.80$, $\phi_{1,2} = \phi_{2,2} = 0.20$, $\xi_1 = 0.90$, and $\xi_2 = 0.40$. The numerical utilities induced by these values are given in Table 1. Our choice to specify $\nu = 0.10$ in this illustration reflects the belief that (PD, Sev) is very undesirable relative to (CR, Min). Specifying $\zeta_1 = 0.10$ and $\zeta_2 = 0.20$ reflects that (CR, Sev) is more desirable than (PD, Min), yet both responses have desirabilities more similar to (PD, Sev) than (CR, Min), i.e., both are undesirable outcomes with utilities < 30. Specifying $\phi_{1,1} = \phi_{2,1} = 0.80$ and $\phi_{1,2} = \phi_{2,2} = 0.20$ reflects the belief that PR and SD have desirabilities similar to CR and PD, respectively, and moreover their desirabilities relative to CR and PD are invariant across the AE severity levels. In contrast, specifying $\xi_1 = 0.90$ and $\xi_2 = 0.40$ reflects the belief that a moderate AE is more tolerable given an efficacious clinical response. Conditional on CR, a moderate AE has similar desirability compared to a minimal AE, whereas, conditional on PD, it has desirability more similar to a severe AE than to a minimal AE. In summary, these choices reflect the general belief that (CR, Min), (CR, Mod), (PR, Min), and (PR, Mod) are all desirable patient outcomes with numerical utilities > 75, whereas all other patient outcomes are relatively undesirable with numerical utilities ≤ 35.

3 Dirichlet-Multinomial Model

Let $\boldsymbol{X}_j = (X_{j,1} \; X_{j,2} \; \cdots \; X_{j,K})'$ denote the count vector of patient outcomes, with probabilities $\boldsymbol{\theta}_j = (\theta_{j,1} \; \cdots \; \theta_{j,K})'$ and $n_j = \sum_{k=1}^{K} X_{j,k}$ the number of observations for treatment $j = A, B$. For the Bayesian tests presented in Section 4, we will assume the Dirichlet-multinomial model

$$\begin{aligned} \boldsymbol{X}_j \,|\, \boldsymbol{\theta}_j \;&\sim\; Mult(n_j,\; \boldsymbol{\theta}_j), \qquad &\text{(Likelihood)} \\ \boldsymbol{\theta}_j \;&\sim\; Dir(n_j^* \boldsymbol{\theta}_j^*), &\text{(Prior)} \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_j^* = (\theta_{j,1}^* \ \cdots \ \theta_{j,K}^*)'$ is the prior mean of $\boldsymbol{\theta}_j$ and n_j^* is the effective sample size (ESS) of the prior (cf. Morita, et al., 2008, 2010). This is well known for the important special case $K = 2$, which is the beta distribution, where the ESS of $f(\boldsymbol{\theta}_j | n_j^*, \boldsymbol{\theta}_j^*)$ is $n_j^* = n_j^*(\theta_{j,1}^* + \theta_{j,2}^*)$. The model (2) has a simple conjugate structure, with each $\boldsymbol{\theta}_j \,|\, \boldsymbol{X}_j \sim Dir(\boldsymbol{X}_j + n_j^* \boldsymbol{\theta}_j^*)$, *a posteriori*, which greatly facilitates posterior computation. The Dirichlet-multinomial model is quite general, and accommodates any categorical outcome structure. The multinomial pdf is

$$f(\boldsymbol{X}_j | \boldsymbol{\theta}_j) = \Gamma\left(n_j + 1\right) \prod_{k=1}^{K} \frac{\theta_{j,k}^{X_{j,k}}}{\Gamma\left(X_{j,k} + 1\right)}, \tag{3}$$

and the Dirichlet pdf is

$$f(\boldsymbol{\theta}_j | n_j^* \boldsymbol{\theta}_j^*) = \Gamma\left(n_j^*\right) \prod_{k=1}^{K} \frac{\theta_{j,k}^{n_j^* \theta_{j,k}^* - 1}}{\Gamma\left(n_j^* \theta_{j,k}^*\right)}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function. The Dirichlet has $E(\boldsymbol{\theta}_j | n_j^*, \boldsymbol{\theta}_j^*) = \boldsymbol{\theta}_j^*$, $Var(\theta_{j,k} | n_j^*, \boldsymbol{\theta}_j^*) = \theta_{j,k}^*(1 - \theta_{j,k}^*)/(n_j^* + 1)$, and $Cov(\theta_{j,k}, \theta_{j,\ell} | n_j^*, \boldsymbol{\theta}_j^*) = \theta_{j,k}^* \theta_{j,\ell}^* / (n_j^* + 1)$, for $k \neq \ell = 1, \ldots, K$. The posterior has a conjugate form with pdf

$$p(\boldsymbol{\theta}_j | \boldsymbol{X}_j, n_j^*, \boldsymbol{\theta}_j^*) = \Gamma\left(n_j + n_j^*\right) \prod_{k=1}^{K} \frac{\theta_{j,k}^{X_{j,k} + n_j^* \theta_{j,k}^* - 1}}{\Gamma\left(X_{j,k} + n_j^* \theta_{j,k}^*\right)}, \tag{5}$$

and posterior mean $E(\boldsymbol{\theta}_j | \boldsymbol{X}_j, n_j^*, \boldsymbol{\theta}_j^*) = (\boldsymbol{X}_j + n_j^* \boldsymbol{\theta}_j^*)/(n_j + n_j^*)$.

For prior specification, the two priors for $\boldsymbol{\theta}_j$ should match, i.e. $n_A^* = n_B^*$ and $\boldsymbol{\theta}_A^* = \boldsymbol{\theta}_B^*$, so that any statistical comparisons are unbiased, and the priors should not include an inappropriate amount of information, which is quantified by ESS. For this reason, we drop the treatment subscript on these hyperparameters in the sequel. As a default prior, i.e., in the absence of prior information, we will assume $n^* = 1$ and $\theta_k^* = K^{-1}$ so that each ESS $= 1$ and all elementary events are equally likely *a priori*. This default choice allows the accruing data to quickly overwhelm the prior while shrinking response probabilities away from 0 and 1 in small samples. When historical information is available, n^* and $\boldsymbol{\theta}^*$ can be specified to reflect that experience and its relevance to the investigation. Alternatively, for a more robust use of historical information, power priors Ibrahim and Chen (2000) or commensurate prior methods Murray et al. (2015) could be applied. Because the use of historical data to construct informative priors for Bayesian models underlying RCTs is a complex and controversial issue, however, we will not use such priors here, and assume $n^* = 1$ in the sequel.

4 Comparative Tests

Treatment differences are characterized by the mean utility difference, and we test the hypotheses

$$H_0 : \ \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0 \qquad versus \qquad H_1 : \ \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) \neq 0. \tag{6}$$

If desired, a one-sided version of (6) may be appropriate. For example, to test whether A is superior to B, the hypotheses would be $H_0 : \ \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) \leq 0$ *versus* $H_1 : \ \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) > 0$. In what follows, we will focus on two-sided hypotheses, since the one-sided case is a straightforward modification.

Let $\boldsymbol{X} = (\boldsymbol{X}_A, \ \boldsymbol{X}_B)$ denote the observed elementary outcome count data. We conduct a CAT-BUB comparative test using the following symmetric decision criteria. If

$$T_{A>B}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*) = Pr\{\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) > 0 | \boldsymbol{X}, n^*, \boldsymbol{\theta}^*\} > p_{cut} \tag{7}$$

then conclude superiority of A over B, denoted by $A > B$. If

$$T_{B>A}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*) = Pr\{\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) < 0|\boldsymbol{X}, n^*, \boldsymbol{\theta}^*\} > p_{cut}, \tag{8}$$

then conclude superiority of B over A, denoted by $B > A$. We select the probability cutoff, p_{cut}, to ensure an approximate level α test for all $\boldsymbol{\theta} = (\boldsymbol{\theta}_A,\ \boldsymbol{\theta}_B)$ with $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0$. We discuss technical details for doing this below.

4.1 Efficient Posterior Computation

While the posterior distributions of $\overline{U}(\boldsymbol{\theta}_A)$ and $\overline{U}(\boldsymbol{\theta}_B)$ are not analytically tractable, because mean utilities are linear combinations of Dirichlet random vectors there are several feasible numerical approximations. With a Monte Carlo (MC) approach, one would generate M samples from the posterior mean utility (PMU) distribution for treatment j, i.e. $p\left\{\overline{U}(\boldsymbol{\theta}_j)|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*\right\}$, by drawing $\boldsymbol{\theta}_j^{(m)} \sim Dir(\boldsymbol{X}_j + n^*\boldsymbol{\theta}^*)$, since $p(\boldsymbol{\theta}_j|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*) \equiv Dir(\boldsymbol{X}_j + n^*\boldsymbol{\theta}^*)$ (see (5)), and defining $\overline{U}(\boldsymbol{\theta}_j^{(m)}) = \boldsymbol{U}'\boldsymbol{\theta}_j^{(m)}$, for $m = 1, \ldots, M$ and $j = A, B$ (see Carlin and Louis (2009), Chapter 3.3). These samples provide estimates of $T_{A>B}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*)$ and $T_{B>A}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*)$ in (7) and (8). For data analysis, any desired level of accuracy can be obtained by increasing M, since it only needs to be conducted once. In contrast, the MC approach is computationally expensive for constructing a clinical trial design, since it requires iterative simulations to assess frequentist operating characteristics in a variety of scenarios, and thus a very large number of MC calculations.

For the CAT-BUB design, a more computationally efficient method for estimating the posterior quantities in (7) and (8) is a parametric approximation to the PMU distribution based on a scaled-beta distribution. To implement this approach, we exploit the following well known forms of the posterior moments of a Dirichlet:

$$\begin{aligned} E[\boldsymbol{\theta}_j|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*] \ &= \ \tilde{\boldsymbol{\theta}}_j \ = \ \frac{\boldsymbol{X}_j + n^*\boldsymbol{\theta}^*}{(n_j + n^*)}, \\ Var[\theta_{j,k}|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*] \ &= \ \frac{\tilde{\theta}_{j,k}\ (1 - \tilde{\theta}_{j,k})}{(n_j + n^* + 1)} \ = \ \frac{(X_{j,k} + n^*\theta_k^*)[(n_j + n^*) - (X_{j,k} + n^*\theta_k^*)]}{(n_j + n^*)^2(n_j + n^* + 1)}, \text{ and} \\ Cov[\theta_{j,k}, \theta_{j,\ell}|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*] \ &= \ \frac{-\ \tilde{\theta}_{j,k}\ \tilde{\theta}_{j,\ell}}{(n_j + n^* + 1)} \ = \ \frac{-(X_{j,k} + n^*\theta_k^*)(X_{j,\ell} + n^*\theta_\ell^*)}{(n_j + n^*)^2(n_j + n^* + 1)}, \text{ for } k \neq \ell. \end{aligned} \tag{9}$$

It follows that

$$\tilde{\mu}_j = E\left[\overline{U}(\boldsymbol{\theta}_j)|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*\right] = \boldsymbol{U}'\tilde{\boldsymbol{\theta}}_j, \text{ and } \tilde{\sigma}_j^2 = Var\left[\overline{U}(\boldsymbol{\theta}_j)|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*\right] = \boldsymbol{U}'\tilde{\boldsymbol{\Sigma}}_j\boldsymbol{U}, \tag{10}$$

where $\tilde{\boldsymbol{\Sigma}}_j = Var[\boldsymbol{\theta}_j|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*]$ with entries defined in (9). Using (10), we match the support, mean, and variance of each PMU distribution with those of a scaled-beta distribution. Let $Beta(\lambda, \gamma)$ denote a beta distribution with mean $\mu = \lambda/(\lambda + \gamma)$ and variance $\sigma^2 = \mu(1 - \mu)/(\lambda + \gamma + 1)$. We approximate $p\left\{\overline{U}(\boldsymbol{\theta}_j)|\boldsymbol{X}_j, n^*, \boldsymbol{\theta}^*\right\}$ with $100 \times Beta(\tilde{\lambda}_j, \tilde{\gamma}_j)$, where

$$\tilde{\lambda}_j = \tilde{\mu}_j \left[\frac{\tilde{\mu}_j(1 - \tilde{\mu}_j)}{\tilde{\sigma}_j^2} - 1\right], \quad \tilde{\gamma}_j = (1 - \tilde{\mu}_j)\left[\frac{\tilde{\mu}_j(1 - \tilde{\mu}_j)}{\tilde{\sigma}_j^2} - 1\right], \tag{11}$$

and the mean $\tilde{\mu}_j$ and variance $\tilde{\sigma}_j^2$ are defined in (10). When $K = 2$ the PMU distribution is precisely this scaled-beta distribution. We provide the derivation for (11) in Web Appendix B.

Using this approximation, the posterior decision criterion is

$$T_{A>B}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*) \;\approx\; \int_0^1 [1-B(x|\tilde{\lambda}_A,\tilde{\gamma}_A)]b(x|\tilde{\lambda}_B,\tilde{\gamma}_B)dx, \tag{12}$$

where $B(x|\lambda,\gamma)$ and $b(x|\lambda,\gamma)$ denote the cdf and pdf of a $Beta(\lambda,\gamma)$ distribution. The approximation for $T_{B>A}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*)$ follows by symmetry. We use adaptive quadrature via the `integrate()` function in `R` to evaluate (12) efficiently Piessens et al. (1983). In Web Appendix B, we confirm the validity of (12) by comparing it to the usual MC approach using simulation under a variety of settings. The scaled-beta approximation is 1,000 times faster than the usual MC approximation with M = 100,000, and it works well even with very small sample sizes, such as $n_A = n_B = 10$. We will use the scaled-beta approximation for the remainder of the paper, and recommend its use in practice.

4.2 Type I Error, Power, and Sample Size

We derive an expression for the approximate power function of the CAT-BUB procedure based on (7) and (8), and use this result to show control of type I error and to obtain a sample size formula. We first apply the Bayesian central limit theorem, and use the resulting posterior asymptotic normality to obtain tractable expressions for $T_{A>B}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*)$ and $T_{B>A}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*)$. We will show that the resulting approximate test statistics are tractable functions of the data, $\boldsymbol{X}$. We then take the frequentist perspective, treating $\boldsymbol{\theta} = (\boldsymbol{\theta}_A,\ \boldsymbol{\theta}_B)$ as a fixed quantity, and apply the classical central limit theorem to derive the asymptotic sampling distributions of $T_{A>B}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*)$ and $T_{B>A}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*)$, and an approximate power function.

Since $\boldsymbol{X}_j$ is multinomial with parameter $\boldsymbol{\theta}_j$, the MLE is $\hat{\boldsymbol{\theta}}_j = \boldsymbol{X}_j/n_j$ and the estimated Fischer information is $n_j\hat{\boldsymbol{\Sigma}}_j^{-1}$, where $\hat{\boldsymbol{\Sigma}}_j$ has k-th diagonal entry $\hat{\theta}_{j,k}(1-\hat{\theta}_{j,k})$ and (k,ℓ)-th off-diagonal entry $-\hat{\theta}_{j,k}\hat{\theta}_{j,\ell}$, $k,\ell = 1,\ldots,K$, $j = A,B$. Applying the Bayesian central limit theorem (see Gelman et al. (2014), Chapter 4)

$$\boldsymbol{\theta}_j|\boldsymbol{X}_j,n^*,\boldsymbol{\theta}^* \;\dot{\sim}\; \mathcal{N}_K(\hat{\boldsymbol{\theta}}_j,\ n_j^{-1}\hat{\boldsymbol{\Sigma}}_j), \quad j = A,B.$$

Since $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are independent, $\boldsymbol{U}'(\boldsymbol{\theta}_A-\boldsymbol{\theta}_B)|\boldsymbol{X},n^*,\boldsymbol{\theta}^* \;\dot{\sim}\; \mathcal{N}(\hat{\delta}_{\boldsymbol{U},A-B},\ \hat{\sigma}^2_{+,n})$, where $\hat{\delta}_{\boldsymbol{U},A-B} = \boldsymbol{U}'\left(\hat{\boldsymbol{\theta}}_A - \hat{\boldsymbol{\theta}}_B\right)$ and $\hat{\sigma}^2_{+,n} = \boldsymbol{U}'\left(\hat{\boldsymbol{\Sigma}}_A/n_A + \hat{\boldsymbol{\Sigma}}_B/n_B\right)\boldsymbol{U}$. It follows that

$$T_{A>B}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*) \approx \Phi\left(\frac{\hat{\delta}_{\boldsymbol{U},A-B}}{\hat{\sigma}_{+,n}}\right) \;\text{ and }\; T_{B>A}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*) \approx \Phi\left(-\frac{\hat{\delta}_{\boldsymbol{U},A-B}}{\hat{\sigma}_{+,n}}\right), \tag{13}$$

where $\Phi(\cdot)$ denotes the standard normal cdf. We use the notation "$\approx$" to mean that an approximation can be made arbitrarily accurate for sufficiently large sample size.

To derive an approximate power function, we treat the posterior quantities in (13) as functions of the data $\boldsymbol{X}$ given a fixed $\boldsymbol{\theta}$, and derive asymptotic approximations for their sampling distributions. First, the exact power function is the probability of rejecting the null for a fixed $\boldsymbol{\theta}$, i.e.

$$\psi(\boldsymbol{\theta}) = Pr\left\{T_{A>B}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*) > p_{cut}\,|\,\boldsymbol{\theta}\right\} + Pr\left\{T_{B>A}(\boldsymbol{X};n^*,\boldsymbol{\theta}^*) > p_{cut}\,|\,\boldsymbol{\theta}\right\}. \tag{14}$$

Applying the classical central limit theorem, $(\hat{\delta}_{\boldsymbol{U},A-B} - \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}))/\hat{\sigma}_{+,n} \;\dot{\sim}\; \mathcal{N}(0,1)$, so plugging (13) into (14) gives the approximate power function

$$\psi(\boldsymbol{\theta})^{approx} \;=\; \Phi\left[\left(\frac{\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta})}{\sigma_{+,n}(\boldsymbol{\theta})}\right) - \Phi^{-1}(p_{cut})\right] + \Phi\left[-\left(\frac{\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta})}{\sigma_{+,n}(\boldsymbol{\theta})}\right) - \Phi^{-1}(p_{cut})\right], \tag{15}$$

where $\sigma_{+,n}(\boldsymbol{\theta})^2 = \boldsymbol{U}'\left(\boldsymbol{\Sigma}_A(\boldsymbol{\theta})/n_A + \boldsymbol{\Sigma}_B(\boldsymbol{\theta})/n_B\right)\boldsymbol{U}$. is a function of $\boldsymbol{\theta}$.

The type I error is $sup\{\psi(\boldsymbol{\theta}) : \delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0\}$, and since $\psi(\boldsymbol{\theta})^{approx} = 2(1-p_{cut})$ for all $\boldsymbol{\theta}$ with $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0$, using $p_{cut} = 1-\alpha/2$ provides an asymptotic level α test. To derive an approximate sample size formula, we set $p_{cut} = 1-\alpha/2$ and define $n = n_A = n_B$. If desired, one could instead define $n = n_A$ and $n_B = \eta \times n_A$, where η controls the randomization ratio. For a given fixed target alternative $\boldsymbol{\theta}^{(Alt)}$, e.g., the hypothesized outcome probabilities, we equate $\psi(\boldsymbol{\theta}^{(Alt)})^{approx} = 1-\beta$ and solve for n, which gives approximate sample size

$$n_f\left(\boldsymbol{\theta}^{(Alt)}, \alpha, \beta\right) = \frac{\left[\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\right]^2 \sigma_+^2(\boldsymbol{\theta}^{(Alt)})}{\delta^2_{\boldsymbol{U},A-B}(\boldsymbol{\theta}^{(Alt)})}. \tag{16}$$

We discuss elicitation of $\boldsymbol{\theta}^{(Alt)}$ in Section 5.

5 Designing a CAT-BUB Trial

In this section, we derive design parameters that control overall type I error at level α and provide 1-β power for targeted alternatives, i.e., the set of treatment effects that we want to identify with the specified power. For this computation, we distinguish between fixed sample designs with one comparative test at the end of the trial, and group sequential designs with up to S comparative analyses over the course of the trial, allowing early termination with rejection of the null at each interim analysis. We first present guidelines for eliciting targeted alternatives, then discuss fixed sample CAT-BUB designs, followed by group sequential CAT-BUB designs. For each design setting, we provide a computational algorithm for deriving the probability cut-offs and sample size, given α, β and the targeted alternatives.

5.1 Eliciting Targeted Alternatives

Consider a fixed sample CAT-BUB test with type I error α for all $\boldsymbol{\theta}$ for which $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0$, and power $1-\beta$ for a set of fixed targeted alternatives with $|\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta})| > 0$. The approximate power function in (15) shows that selecting p_{cut} to control type I error for one fixed null response probability vector, say $\boldsymbol{\theta}^{(Null)} = \left(\boldsymbol{\theta}_A^{(Null)}, \boldsymbol{\theta}_B^{(Null)}\right)$ with $\delta^{(Null)}_{\boldsymbol{U},A-B} = 0$, will control type I error for all fixed $\boldsymbol{\theta}$ with $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta}) = 0$. In contrast, the power varies with both the targeted utility difference and the fixed $\boldsymbol{\theta}$ from which this difference arises, via $\sigma_{+,n}(\boldsymbol{\theta})$. Therefore, targeted alternatives must be elicited in the $\boldsymbol{\theta}$ domain. We denote a fixed target by $\boldsymbol{\theta}^{(Alt)} = \left(\boldsymbol{\theta}_A^{(Alt)}, \boldsymbol{\theta}_B^{(Alt)}\right)$ and its utility difference by $\left|\delta^{(Alt)}_{\boldsymbol{U},A-B}\right| > 0$.

Since it may not be intuitively obvious how to specify $\boldsymbol{\theta}^{(Alt)}$, we provide the following guidelines, which require a discussion between the statistician and the physicians. For simplicity, we will treat A as the null or standard treatment, although the algorithm works if A and B are both experimental and considered to be symmetric. The statistician begins by eliciting an expected probability vector corresponding to historical experience with standard therapy, say $\boldsymbol{\theta}_A^{(Alt)}$, which may be based on the physician(s)' experience or analysis of historical data. Given $\boldsymbol{\theta}_A^{(Alt)}$, the statistician asks the physician(s) to specify one or more alternative probability vectors, $\boldsymbol{\theta}_B^{(Alt,1)}, \cdots, \boldsymbol{\theta}_B^{(Alt,m)}$, that are considered equally desirable improvements over $\boldsymbol{\theta}_A^{(Alt)}$. In practice, m should be reasonably small, in the range $1 \le m \le K$. Each elicited alternative $\boldsymbol{\theta}_B^{(Alt,r)}$ gives standardized utility difference $s^{(Alt,r)} = \delta^{(Alt,r)}_{\boldsymbol{U},B-A}/\sigma_+^{(Alt,r)}$, where $\delta^{(Alt,r)}_{\boldsymbol{U},B-A}$ and $\sigma_+^{(Alt,r)}$ are evaluated at $\boldsymbol{\theta}^{(Alt,r)} = \left(\boldsymbol{\theta}_A^{(Alt)}, \boldsymbol{\theta}_B^{(Alt,r)}\right)$

for $r = 1, \ldots, m$. For the sample size calculation in (16), one then selects the targeted alternative $\boldsymbol{\theta}_B^{(Alt)}$ giving smallest $s^{(Alt,r)}$, formally

$$\boldsymbol{\theta}^{(Alt)} = \left\{ \left(\boldsymbol{\theta}_A^{(Alt)},\ \boldsymbol{\theta}_B^{(Alt,r^*)} \right) : s^{(Alt,r^*)} = min\{s^{(Alt,1)}, \ldots, s^{(Alt,m)}\} \right\}. \tag{17}$$

This choice is conservative since it ensures the test will achieve the desired power for all elicited $\boldsymbol{\theta}^{(Alt,r)}$.

In practice, if this computation gives a sample size that is not feasible, then the physician(s) should be asked to re-consider their set of specified alternatives. This is not unlikely, since it may not be intuitively obvious, when specifying one or more fixed target probability vectors, how they translate into a required sample size. To help guide the physician(s) in this process, one should show them the numerical values of $\delta_{\boldsymbol{U},B-A}^{(Alt,r)}$, $s^{(Alt,r)}$, and $\boldsymbol{\theta}_B^{(Alt,r)}$, for $r = 1, \cdots, m$, possibly as a table with m rows and three columns to facilitate comparison and interpretation. Since smaller values of $s^{(Alt,r)}$ and $\delta_{\boldsymbol{U},B-A}^{(Alt,r)}$ require a larger sample size to detect the corresponding $\boldsymbol{\theta}_B^{(Alt,r)}$, this provides a quantitative index of the relative difficulty of detecting each target, and it also identifies the targeted alternative $\boldsymbol{\theta}_B^{(Alt,r^*)}$ having the smallest $s^{(Alt,r)}$ that produced the sample size. If a modified set of targets is specified, the sample size may be recomputed, with this process iterated if desired. This may be considered a multidimensional analog of a conventional power and sample size computation in terms of a one-dimensional parameter. If desired, the CAT-BUB test's power function computed over a set of $(\boldsymbol{\theta}_A,\ \boldsymbol{\theta}_B)$ values also may be examined.

Recall the trinary outcome example where $\boldsymbol{U} = (U_R,\ U_N,\ U_F)' = (100,\ 60,\ 0)'$. Given standard vector $\boldsymbol{\theta}_A^{(Alt)} = (0.30,\ 0.50,\ 0.20)'$, suppose that the three equally desirable targets $\boldsymbol{\theta}_B^{(Alt,1)} = (0.40,\ 0.50,\ 0.10)'$, $\boldsymbol{\theta}_B^{(Alt,2)} = (0.50,\ 0.35,\ 0.15)'$, and $\boldsymbol{\theta}_B^{(Alt,3)} = (0.35,\ 0.60,\ 0.05)'$ are elicited. Then $s^{(Alt,1)} = 10/45.8 = 0.218$, $s^{(Alt,2)} = 11/49.2 = 0.224$, and $s^{(Alt,3)} = 11/42.6 = 0.258$, so we would take $\boldsymbol{\theta}^{(Alt)} = \left(\boldsymbol{\theta}_A^{(Alt)},\ \boldsymbol{\theta}_B^{(Alt,1)}\right)$. If the standardized utility differences, $s^{(Alt,r)}$, differ substantially, then the physician(s) may instead select the *a priori* most likely alternative, perhaps sacrificing power for some alternatives as a trade-off for a smaller sample size. If the utility differences, $\delta_{\boldsymbol{U},B-A}^{(Alt,r)}$, differ substantially, then the physician(s) may wish to reconsider the choices of equally desirable targets, or possibly may decide to modify some entries of the numerical utility vector $\boldsymbol{U}$.

5.2 Computational Algorithm for Fixed Sample CAT-BUB Design

Given α, β and the targeted alternative $\boldsymbol{\theta}^{(Alt)}$ defined in (17), we jointly select a sample size and cutoff p_{cut} for a fixed sample CAT-BUB design using the following algorithm:

Step 0. Set $\hat{n} = n_f(\boldsymbol{\theta}^{(Alt)}, \alpha, \beta)$, where $n_f(\cdot)$ is defined by (16).

Step 1. Generate G_0 null datasets as follows. For $g_0 = 1, \ldots, G_0$,

(i) generate $\boldsymbol{X}_j^{(g_0)} \sim Mult\left(\hat{n},\ \boldsymbol{\theta}_A^{(Alt)}\right)$ for $j = A, B$.

(ii) store $\boldsymbol{X}^{(Null,g_0)} = \left(\boldsymbol{X}_A^{(g_0)},\ \boldsymbol{X}_B^{(g_0)}\right)$.

(iii) calculate and store

$$T^{(Null,g_0)} = \max\left\{T_{A>B}\left(\boldsymbol{X}^{(Null,g_0)}; n^*, \boldsymbol{\theta}^*\right),\ T_{B>A}\left(\boldsymbol{X}^{(Null,g_0)}; n^*, \boldsymbol{\theta}^*\right)\right\}.$$

Step 2. Set $\hat{p}_{cut}$ to the empirical $(1-\alpha)$%-tile of $\left\{T^{(Null,1)},\ \cdots,\ T^{(Null,G_0)}\right\}$.

Step 3. Generate G_1 alternative datasets as follows. For $g_1 = 1, \ldots, G_1$,

(i) generate $\boldsymbol{X}_j^{(g_1)} \sim Mult\left(\hat{n},\ \boldsymbol{\theta}_j^{(Alt)}\right)$ for $j = A, B$.

(ii) store $\boldsymbol{X}^{(Alt,g_1)} = \left(\boldsymbol{X}_A^{(g_1)},\ \boldsymbol{X}_B^{(g_1)}\right)$.

(iii) If $\delta_{\boldsymbol{U},A-B}^{(Alt)} > 0$, calculate and store $T^{(Alt,g_1)} = T_{A>B}\left(\boldsymbol{X}^{(Alt,g_1)}; n^*, \boldsymbol{\theta}^*\right)$.
Otherwise, calculate and store $T^{(Alt,g_1)} = T_{B>A}\left(\boldsymbol{X}^{(Alt,g_1)}; n^*, \boldsymbol{\theta}^*\right)$.

Step 4. Set $\hat{\beta} = G_1^{-1} \sum_{g_1=1}^{G_1} \left[T^{(Alt,g_1)} \leq \hat{p}_{cut}\right]$, where $[E] = 1$ if E is true, and 0 otherwise.

Step 5. If $\hat{\beta} \in [\beta - \epsilon,\ \beta + \epsilon]$, stop and select $n = \hat{n}$ and $p_{cut} = \hat{p}_{cut}$.
Otherwise, update $\hat{n} = \hat{n}\left(\frac{\Phi^{-1}(1-\beta)+\Phi^{-1}(\hat{p}_{cut})}{\Phi^{-1}(1-\hat{\beta})+\Phi^{-1}(\hat{p}_{cut})}\right)^2$ and return to Step 1.

In practice, $\hat{n}$ is rounded to its nearest integer value and $\hat{p}_{cut}$ is rounded to its nearest larger thousandth. We use default values $\epsilon = 0.005$, $G_0 = 50,000$ and $G_1 = 25,000$. Since choosing G_0 and G_1 is non-intuitive, detailed guidelines are given in Web Appendix C. Briefly, these default values allow us to estimate p_{cut} accurately to three digits, and be certain that the power for $\boldsymbol{\theta}^{(Alt)}$ at the selected n is within $2 \times \epsilon = 0.01$ of $1 - \beta$. The sample size adjustment in step 5 is motivated by (16), and allows $\hat{n}$ to be increased or decreased by a magnitude proportional to the current discrepancy between the estimated and desired power.

5.3 Computational Algorithm for Group Sequential CAT-BUB Design

In typical practice, RCTs require group sequential tests Jennison and Turnbull (2000). Here, we discuss implementation of the CAT-BUB test in this context, denoting the sample sizes where an analysis occurs by n_s, $s = 1, \ldots, S$. We take use the α-spending approach proposed by Slud and Wei (1982) and extended by Lan and DeMets (1983), with an α-spending function $f(n_s; \alpha, \rho, n_S)$ $= \alpha(n_s/n_S)^\rho$ suggested by Kim and DeMets (1987). The design parameter $\rho \geq 0$ controls the α-spending rate, with larger values spending less α at early looks. This approach is appealing in practice because the actual analysis schedule need not follow the planned schedule. At the first interim look with n_1 of the planned n_S samples, we calibrate the probability threshold, $p_{cut,1}$, to spend $f(n_1; \alpha, \rho, n_S)$ of the overall type I error. Similarly, at s-th interim look, we calibrate $p_{cut,s}$ to spend $f(n_s; \alpha, \rho, n_S)$ - $f(n_{s-1}; \alpha, \rho, n_S)$ of the overall type I error. So if the trial reaches a final analysis at n_S samples, the overall type I error is exactly α.

To determine a maximum sample size, n_S, for a group sequential CAT-BUB design with up to S tests, power $1 - \beta$ for the elicited alternative $\boldsymbol{\theta}^{(Alt)}$, we specify a complete analysis schedule using the proportions of n_S, denoted by t_s, $s = 1, \ldots, S$. We determine n_S using the following algorithm:

Step 0. Set $\hat{n}_S = n_f(\boldsymbol{\theta}^{(Alt)}, \alpha, \beta)$, where $n_f(\cdot)$ is defined by (16), and $\hat{n}_s = t_s \times \hat{n}_S$, for $s = 1, \ldots, S - 1$.

Step 1. Generate G_0 null sequential datasets as follows. For $g_0 = 1, \ldots, G_0$,

(i) generate $\boldsymbol{X}_{j,s}^{(g_0)} \sim Mult\left(\hat{n}_s,\ \boldsymbol{\theta}_A^{(Alt)}\right)$ for $j = A, B$ and $s = 1, \ldots, S$.

(ii) store $\boldsymbol{X}_{s,+}^{(Null,g_0)} = \left(\boldsymbol{X}_{A,s,+}^{(g_0)},\ \boldsymbol{X}_{B,s,+}^{(g_0)}\right)$, where $\boldsymbol{X}_{j,s,+}^{(g_0)} = \sum_{m=1}^{s} \boldsymbol{X}_{j,m}^{(g_0)}$ for $j = A, B$ and $s = 1, \ldots, S$.

(iii) calculate and store, for $s = 1, \ldots, S$,

$$T_s^{(Null,g_0)} = \max\left\{T_{A>B}\left(\boldsymbol{X}_{s,+}^{(Null,g_0)}; n^*, \boldsymbol{\theta}^*\right),\ T_{B>A}\left(\boldsymbol{X}_{s,+}^{(Null,g_0)}; n^*, \boldsymbol{\theta}^*\right)\right\}.$$

Step 2. Calculate $\hat{p}_{cut,1}, \ldots, \hat{p}_{cut,S}$ as follows.

(i) Set $\hat{p}_{cut,1}$ to the empirical $\{1 - f(n_1; \alpha, \rho, n_S)\}$%-tile of $\left\{T_1^{(Null,1)}, \ldots, T_1^{(Null,G_0)}\right\}$.

(ii) Set $\hat{p}_{cut,s}$ to the empirical $[\{1 - f(n_s; \alpha, \rho, n_S)\}/\{1 - f(n_{s-1}; \alpha, \rho, n_S)\}]$%-tile of $\left\{T_s^{(Null,g_0)} : T_1^{(Null,g_0)} \leq \hat{p}_{cut,1}, \cdots, T_{s-1}^{(Null,g_0)} \leq \hat{p}_{cut,s-1},\ g_0 = 1, \ldots, G_0\right\}$ for $s = 2, \ldots, S$.

Step 3. Generate G_1 alternative sequential datasets as follows. For $g_1 = 1, \ldots, G_1$,

(i) generate $\boldsymbol{X}_{j,s}^{(g_1)} \sim Mult\left(\hat{n}_s,\ \boldsymbol{\theta}_j^{(Alt)}\right)$ for $j = A, B$ and $s = 1, \ldots, S$.

(ii) store $\boldsymbol{X}_{s,+}^{(Alt,g_1)} = \left(\boldsymbol{X}_{A,s,+}^{(g_1)},\ \boldsymbol{X}_{B,s,+}^{(g_1)}\right)$, where $\boldsymbol{X}_{j,s,+}^{(g_1)} = \sum_{m=1}^{s} \boldsymbol{X}_{j,m}^{(g_1)}$ for $j = A, B$ and $s = 1, \ldots, S$.

(iii) If $\delta_{\boldsymbol{U},A-B}^{(Alt)} > 0$, calculate and store $T_s^{(Alt,g_1)} = T_{A>B}\left(\boldsymbol{X}_{s,+}^{(Alt,g_1)}; n^*, \boldsymbol{\theta}^*\right)$, for $s = 1, \ldots, S$.

Otherwise, calculate and store $T_s^{(Alt,g_1)} = T_{B>A}\left(\boldsymbol{X}_{s,+}^{(Alt,g_1)}; n^*, \boldsymbol{\theta}^*\right)$, for $s = 1, \ldots, S$.

Step 4. Set $\hat{\beta} = G_1^{-1} \sum_{g_1=1}^{G_1} \left[T_1^{(Alt,g_1)} \leq \hat{p}_{cut,1}, \cdots, T_S^{(Alt,g_1)} \leq \hat{p}_{cut,S}\right]$, where $[E] = 1$, if E is true, and 0, otherwise.

Step 5. If $\hat{\beta} \in [\beta - \epsilon,\ \beta + \epsilon]$, stop and select $n_S = \hat{n}_S$.

Otherwise, update $\hat{n}_S = \hat{n}_S \left(\frac{\Phi^{-1}(1-\beta)+\Phi^{-1}(\hat{p}_{cut,S})}{\Phi^{-1}(1-\hat{\beta})+\Phi^{-1}(\hat{p}_{cut,S})}\right)^2$, $\hat{n}_s = t_s \times \hat{n}_S$ for $s = 1, \ldots, S-1$, and return to Step 1.

We use the same default values as the fixed sample algorithm, that is $\epsilon = 0.005$, $G_0 = 50,000$ and $G_1 = 25,000$. Using the planned analysis schedule, we can assess the operating characteristics at a variety of alternatives. The actual analysis schedule may differ from the planned schedule, so the realized power may differ from $1-\beta$; however, Jennison and Turnbull (2000) show that the realized power is quite robust to deviations from the planned analysis schedule. During an actual trial, we can follow steps 1–2 to re-estimate $p_{cut,s}$ for the actual n_s being used, given the previous interim analysis sample sizes $n_1, \ldots, n_{s-1}$ and their corresponding $p_{cut,1}, \ldots, p_{cut,s-1}$ values.

6 Illustrations

In this section, we illustrate CAT-BUB tests and report results of various simulation studies comparing both fixed sample and group sequential CAT-BUB designs with beta-binomial designs. We investigate the proposed procedure in the contexts of a trinary outcome, a bivariate-binary outcome, and the CLL trial, which actually had a bivariate ordinal outcome including death. We also report the results of utility sensitivity analyses.

6.1 Trinary Outcomes

6.1.1 Fixed Sample Tests

Returning to the example involving clot dissolving agents for rapid treatment of stroke with a trinary outcome and utility $\boldsymbol{U} = (100,\ 50,\ 0)'$, we investigate the frequentist operating characteristics of the proposed CAT-BUB approach for a variety of fixed response probability vectors. We consider a CAT-BUB test with type I error $\alpha = 0.05$, and power $1-\beta = 0.80$ for targeted alternative $\boldsymbol{\theta}^{(Alt)} = \left(\boldsymbol{\theta}_A^{(Alt)},\ \boldsymbol{\theta}_B^{(Alt)}\right) = ((0.50,\ 0.30,\ 0.20)',\ (0.60,\ 0.30,\ 0.10)')$ with $\delta_{\boldsymbol{U},B-A}^{(Alt)} = 10$. In this context, the fixed sample CAT-BUB design algorithm, given in Section 5.2, gives $p_{cut} = 0.976$ and $n = 208$.

Table 2: Power of a fixed sample CAT-BUB design for trinary outcome $\{R, N, F\}$ versus a beta-binomial design based on "success" probability $\pi_j = \theta_{j,R}$, for $j = A, B$. In all scenarios, $\boldsymbol{\theta}_A = (0.50,\ 0.30,\ 0.20)'$ and $n_A = n_B = 208$. Results in the first row are based on 50,000 simulations (std.err. ≈0.001), whereas all other results are based on 25,000 simulations (std.err. <0.0032).

	Scenario		**CAT-BUB Design**		**Beta-Bin Design**	
	$\boldsymbol{\theta}_B$	$\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta})$	$B > A$	$A > B$	$B > A$	$A > B$
1.0:	(0.50, 0.30, 0.20)	0	0.025	0.025	0.025	0.025
2.1:	(0.60, 0.00, 0.40)	-5	0.001	0.206	0.552	0.000
2.2:	(0.60, 0.10, 0.30)	0	0.024	0.025		
2.3:	(0.60, 0.20, 0.20)	5	0.246	0.001		
2.4:	(0.60, 0.30, 0.10)	10	0.798	0.000		
2.5:	(0.60, 0.40, 0.00)	15	0.997	0.000		
3.1:	(0.65, 0.05, 0.30)	2.5	0.088	0.006	0.877	0.000
3.2:	(0.65, 0.15, 0.20)	7.5	0.485	0.000		
3.3:	(0.65, 0.25, 0.10)	12.5	0.936	0.000		
3.4:	(0.65, 0.35, 0.00)	17.5	1.000	0.000		
4.1:	(0.70, 0.00, 0.30)	5	0.217	0.001	0.989	0.000
4.2:	(0.70, 0.10, 0.20)	10	0.720	0.000		
4.3:	(0.70, 0.20, 0.10)	15	0.987	0.000		
4.4:	(0.70, 0.30, 0.00)	20	1.000	0.000		

We compare the CAT-BUB approach for trinary outcomes $\{R, N, F\}$ with a Bayesian design that follows the more common approach of combining the events N and F so that outcome may be considered binary, specifically $R =$ "success," versus $N \cup F =$ "failure," and compares therapies in terms of the probabilities $\pi_j = \Pr(Y = R \,|\, j)$ for j=A,B. For this design, we assume a Bayesian beta-binomial model with common beta priors $\pi_j \,|\, \boldsymbol{q}_j \sim Beta(q_{j,1} = 0.50,\ q_{j,2} = 0.50)$ for $j = A, B$, which has ESS = 1. The posterior is $Beta(S_j + 0.50,\ n_j - S_j + 0.50)$, where S_j is the number of successes out of n_j in arm j. Denoting $\boldsymbol{W} = (S_A,\ n_A - S_A,\ S_B,\ n_B - S_B)$ and $\boldsymbol{q} = (q_{A,1},\ q_{A,2},\ q_{B,1},\ q_{B,2})$, we use the test statistic $S_{A>B}(\boldsymbol{W}; \boldsymbol{q}) = Pr(\pi_A > \pi_B | \boldsymbol{W}, \boldsymbol{q})$, which we calculate similarly to (12). This is the special case of the Dirichlet-multinomial model and CAT-BUB test with $K = 2$, since the mean utility for treatment j is $100 \times \pi_j$, so the utility is superfluous. To ensure comparability, for the binary test we also use $n = 208$, and set $p_{cut} = 0.975$ to obtain a 0.05-level test when $\pi_A = \pi_B$.

Operating characteristics of the fixed sample CAT-BUB and beta-binomial tests are given in Table 2. Scenario 1.0 is the null case used to calibrate p_{cut} for each design, so the type I error for both designs is 0.05, with equal probabilities for concluding $A > B$ or $B > A$. Scenario 2.4 is the alternative used to select a sample size that provides power 0.80, so the estimated power is in the interval $[0.80 - \epsilon, 0.80 + \epsilon]$. For Scenarios 2.1-2.5, π_B is fixed at 0.60 versus $\pi_A = 0.50$, so the beta-binomial design always has power 0.55, despite obvious differences between these four scenarios. For example, in Scenarios 2.1 and 2.2, the beta-binomial design fails by concluding $B > A$ 55% of the time even though A is clinically superior or equal to B in terms of $\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta})$. In contrast, the CAT-BUB test distinguishes between these scenarios, correctly concluding $A > B$ 21% of the time in Scenario 2.1 and controlling type I error at 0.05 in Scenario 2.2. Scenarios 2.3-2.5 exhibit various tradeoffs that favor B over A in an increasing manner in terms of $\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta}) = 5$, 10, 15 and the CAT-BUB test reflects this with increasing power figures 0.246, 0.798, 0.997. In particular, the CAT-BUB test has substantially more power than the beta-binomial test for the "win-win" Scenarios 2.4 and 2.5. Scenarios 3.1-3.4 and 4.1-4.4 respectively fix the probability of response at 0.65 or 0.70, for which the beta-binomial test has 0.88 and 0.99 power figures. In contrast, the power of the CAT-BUB test increases as the true utility difference $\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta})$ increases, and equals or exceeds that of the beta-binomial design in "win-win" scenarios where the probability of failure is also reduced (Scenarios 3.3-3.4 and 4.3-4.4). The failure of the beta-binomial design is due to B providing an unfavorable trade-off between the probability of response and failure versus A. Such tradeoffs cannot be identified by the naive binary outcome design, which is used very commonly. The price of the CAT-BUB approach is potentially less power for "tradeoff" scenarios when the treatment redistributes probability away from N to R and/or F (Scenarios 2.3, 3.2, 4.1 and 4.2). However, we feel that this price is well worth being able to distinguish between, for example, Scenarios 2.1-2.5 in practice. Lastly, the CAT-BUB test has varying power over the set of $\boldsymbol{\theta}$ with the same utility difference. For example, Scenarios 2.3 and 4.1 have utility difference 5, yet power figures 0.25 and 0.22, respectively. The CAT-BUB design's power varies more substantially with $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta})$ than over the set of $\boldsymbol{\theta}$ with the same utility difference.

6.1.2 Sensitivity to Elicited Utilities

The power of the CAT-BUB design at the targeted alternatives, and thus the required sample size, depend on the particular elicited utilities. The sensitivity of the CAT-BUB test's power to the elicited utilities can be assessed by fixing the sample size and calculating the power for targeted alternatives using other numerical utilities. Continuing with the example involving clot dissolving agents, we fix $n = 208$, $p_{cut} = 0.976$ and $\boldsymbol{\theta}_A = (0.50,\ 0.30,\ 0.20)'$, and calculate the CAT-BUB test's power for alternative Scenarios 2.1, 2.2, 2.4 and 2.5 in Table 2 over the entire domain $U_N \in [0,\ 100]$. We calculate power using (15), which is quite accurate when $n = 208$.

Figure 1 plots the *overall* power as a function of U_N at each alternative; that is, we do not explicitly distinguish between the decisions $A > B$ and $B > A$. Scenarios 2.1 and 2.2 are trade-off scenarios, wherein B relative to A has a higher probability of R and F, and lower probability of N, so $\delta_{\boldsymbol{U},A-B}(\boldsymbol{\theta})$ varies substantially with U_N and the power is thus quite sensitive. This sensitivity is a desirable property, because the numerical value of U_N determines whether a particular trade-off favors $B > A$ or $A < B$. For Scenarios 2.1 and 2.2, although not explicitly depicted, the CAT-BUB test has power primarily for $A > B$ ($B > A$) to the left (right) of the numerical utility with minimum power. In Scenario 2.4, the probability of N is equal for both A and B, so $\delta_{\boldsymbol{U},B-A} = 10$ for all U_N, and the sensitivity merely reflects the relationship between U_N and σ_+. In win-win Scenario 2.5, power increases with U_N because $\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta})$ increases with U_N. Lastly, each scenario in Figure 1 fixes $\theta_{R,B} = 0.60$ and $\theta_{R,A} = 0.50$, and for $U_N = 0$ the CAT-BUB test is based exclusively on $100 \times \theta_R$, so it is identical to the usual beta-binomial test, providing 54% power. Therefore, even in settings where selecting a particular U_N may be challenging, the proposed CAT-BUB test obviously is more sensible than the usual beta-binomial test, which implicitly sets $U_N = 0$.

It is useful to consider sensitivity of inferences to the elicited utilities. We illustrate how this may be done for bivariate binary outcomes. Figure 2 depicts the posterior probability for $B > A$, defined by (8), for the AML bivariate binary example from Section 1, given three different realizations of $\boldsymbol{X}_B$ while fixing $\boldsymbol{X}_A = (\boldsymbol{X}_{A,[C,\overline{T}]},\ \boldsymbol{X}_{A,[C,T]},\ \boldsymbol{X}_{A,[\overline{C},\overline{T}]},\ \boldsymbol{X}_{A,[\overline{C},T]})' = (15, 20, 25, 40)'$ and enumerating over $(U_{C,T},\ U_{\overline{C},\overline{T}}) \in [0, 100]^2$, i.e. all possible intermediate utilities. In Scenario 1, $\boldsymbol{X}_B = (10, 40, 20, 30)'$ and $\boldsymbol{X}_A = (15, 20, 25, 40)'$. For these data, inference is more sensitive to $U_{C,T}$ than $U_{\overline{C},\overline{T}}$, because the two treatments appear to differ greatly for the probability of $[C, T]$ (40 versus 20 observations) and little for the probability of $[\overline{C}, \overline{T}]$ (20 versus 25 observations). The data in Scenario 2 reflect a similar yet smaller treatment difference, and inference is less sensitive to the utilities. In Scenario 3, the data suggest that B is a win-win relative to A in that B has both higher marginal probability of C and lower marginal probability of T, and the CAT-BUB design's

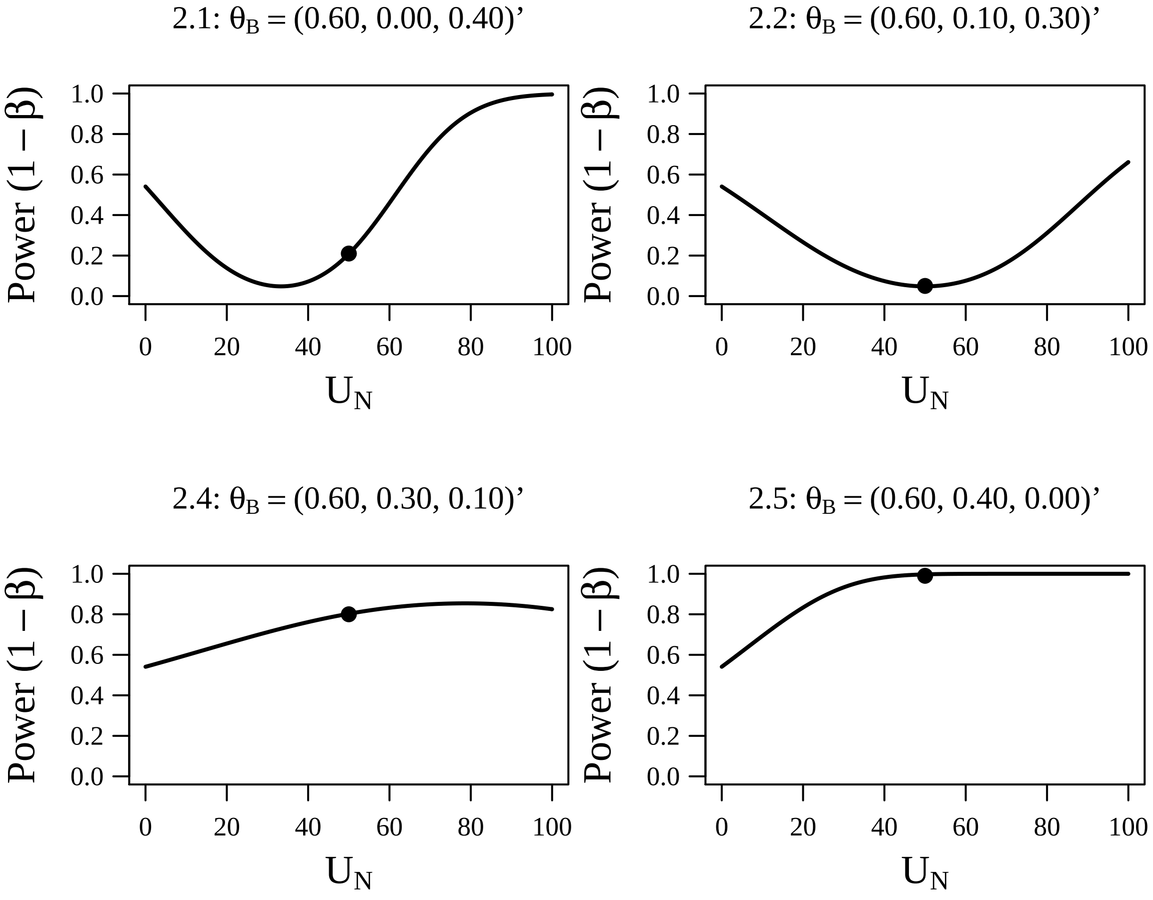

Figure 1: Sensitivity to U_N of the fixed sample CAT-BUB design's power for various alternative $\boldsymbol{\theta}_B$s, fixing $n = 208$, $p_{cut} = 0.976$, and $\boldsymbol{\theta}_A = (0.50,\ 0.30,\ 0.20)'$. The thick dot denotes power for the elicited utilities, i.e. $U_N = 50$.

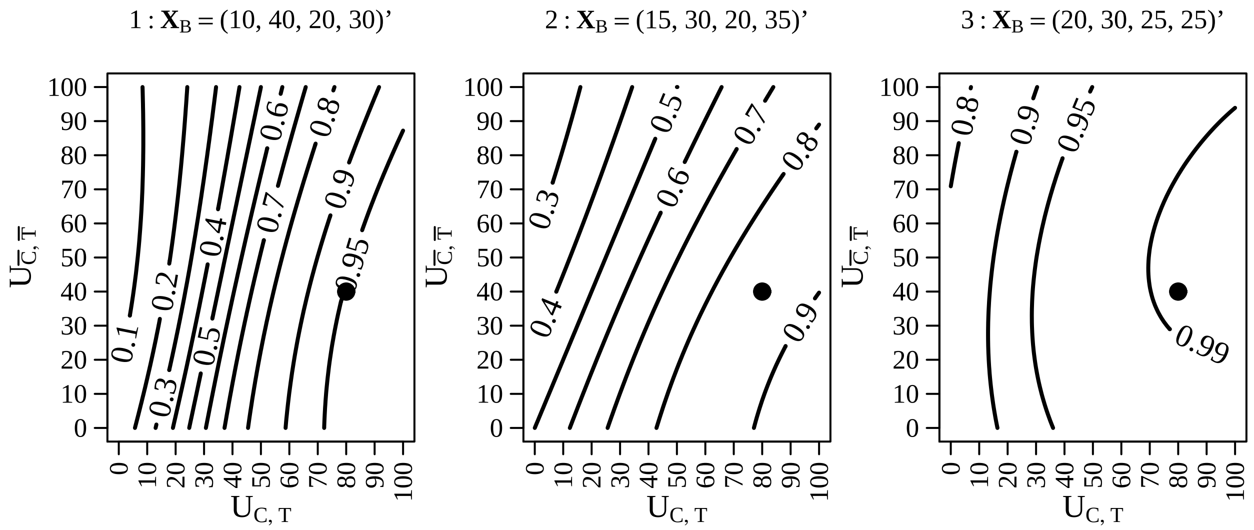

Figure 2: Posterior probability of $B > A$ while varying ($U_{C,T}$, $U_{\overline{C},\overline{T}}$) for three different realizations of $\boldsymbol{X}_B$ and $\boldsymbol{X}_A = (\boldsymbol{X}_{A,[C,\overline{T}]}, \boldsymbol{X}_{A,[C,T]}, \boldsymbol{X}_{A,[\overline{C},\overline{T}]}, \boldsymbol{X}_{A,[\overline{C},T]})' = (15, 20, 25, 40)'$. The thick dot denotes our inferential result at the elicited utilities, i.e. ($U_{C,T} = 80$, $U_{\overline{C},\overline{T}} = 40$).

inference always supports the conclusion $B > A$. For these data, posterior evidence supporting $B > A$ becomes stronger as $U_{C,T}$ is increased.

6.1.3 Group Sequential Tests

To assess the operating characteristics of the group sequential tests, we continue with the trinary versus binary example. We assume the analysis schedule has $S = 3$ equally spaced looks at $t_1 = 0.33$, $t_2 = 0.66$ and $t_3 = 1$. We use the same targeted alternative as the fixed sample design for calibration, and compare the operating characteristics of the group sequential versions of the CAT-BUB design and beta-binomial design for Scenarios 1.0 and 2.1-2.5 used for the fixed sample simulation. We applied the group sequential CAT-BUB design algorithm, given in Section 5.3, to maintain $\alpha \leq 0.05$ with $\rho = 3$. This gave $n_S = 213$, $p_{cut,1} = 0.999$, $p_{cut,2} = 0.993$ and $p_{cut,3} = 0.978$. Scenarios 1.0 and 2.4 were used to jointly calibrate the planned sample size and probability thresholds to provide type I error of 0.05 and overall power of 0.80, respectively. For the beta-binomial design, to maintain size 0.05 we used $p_{cut,1} = 0.999$, $p_{cut,2} = 0.992$ and $p_{cut,3} = 0.979$. We used $n_S = 213$ for both designs to ensure comparability.

The results of the group sequential simulations are reported in Table 3. For the null Scenario 1.0, the operating characteristics of the CAT-BUB and beta-binomial designs are practically identical. Both designs have an average sample size of 212 and overall type I error of 0.05. In contrast, for Scenarios 2.1-2.5, the operating characteristics of the two designs differ dramatically. The beta-binomial design does not distinguish between these scenarios because $\pi_{B,R} = 0.60$ and $\pi_{A,R} = 0.50$ for all 5 scenarios, whereas the CAT-BUB test distinguishes between them quite well. In Scenario 2.1, A is preferred over B due to an unfavorable tradeoff between R and F. The beta-binomial design incorrectly selects B over A 54% of the time with an average sample size of 193, whereas the CAT-BUB design correctly selects A over B 21% of the time with an average sample size of 208. In Scenario 2.2, B and A are equivalent due to the increase in response probability being canceled

Table 3: Power figures of a group sequential CAT-BUB design for a trinary outcome $\{R.N, F\}$ versus a beta-binomial design based on "success" probabilities $\pi_j = \theta_{j,R}$, for $j = A, B$. In each scenario, $\boldsymbol{\theta}_A = (0.50,\ 0.30,\ 0.20)'$, $n_1 = 71$, $n_2 = 142$, $n_3 = 213$, and $\rho = 3$.

Scenario Specification			**CAT-BUB Design**			**Beta-Binomial Design**		
	$\boldsymbol{\theta}_B$	$\delta_{\boldsymbol{U},B-A}(\boldsymbol{\theta})$	Ave SS	$B > A$	$A > B$	Ave SS	$B > A$	$A > B$
1.0:	(0.50, 0.30, 0.20)	0	211.9	0.025	0.025	211.8	0.026	0.024
2.1:	(0.60, 0.00, 0.40)	-5	207.7	0.001	0.214	192.8	0.541	0.000
2.2:	(0.60, 0.10, 0.30)	0	211.8	0.026	0.025			
2.3:	(0.60, 0.20, 0.20)	5	206.6	0.250	0.001			
2.4:	(0.60, 0.30, 0.10)	10	177.8	0.800	0.000			
2.5:	(0.60, 0.40, 0.00)	15	123.8	0.998	0.000			

out by the increase in failure probability. Here, the CAT-BUB design controls type I error at level 0.05. Scenarios 2.3-5 have increasing magnitudes of the benefit for B over A, and the CAT-BUB design has increasing power for concluding $B > A$. As the true benefit of B over A increases, the average sample size of the CAT-BUB design decreases because the probability of early termination increases. In the most favorable Scenario 2.5, the CAT-BUB design has power 0.998 and terminates early nearly 95% of the time, with average sample size 124 that is 42% smaller than the planned maximum sample size. In contrast, the beta-binomial design has 54% power and average sample size 193 in this case, as in all Scenarios 2.1 - 2.5, essentially because it ignores the distinction between N and F.

6.2 Redesigning the CLL Trial

Returning to the CLL trial, we illustrate how to implement the CAT-BUB design in this context. We assume that the elicited numerical utilities are those in Table 1. Recall that, since we cannot elicit utilities for this trial retrospectively, as explained in Section 1 the utilities in Table 1 are specified to be a reasonable representation of what one actually would elicit in practice. We compare the CAT-BUB design with a beta-binomial design based on an efficacy test, which we denote by BB-EO. Like the actual trial, the BB-EO design defines efficacy using a binary indicator for CR, where the comparative test was based on targeted alternative CR probability $\pi_{CR,FC} = 0.45$ versus null $\pi_{CR,F} = 0.25$ Flinn et al. (2007). We also compare the CAT-BUB design to an alternative approach that is based on a hierarchical testing procedure. This alternative design first compares the probabilities of efficacy (here, CR) as the primary endpoint and, if this test fails to reject the null, then the procedure compares the probabilities of toxicity (here, severe or fatal AE) in a second test. This design, which we denote by BB-ET, assumes independent beta-binomial models for the two outcomes. Based on this hierarchical testing procedure, the BB-ET design recommends a treatment if it is found to have either better efficacy or toxicity compared to the other treatment.

Because the actual CLL trial outcome is bivariate ordinal plus death, to implement the CAT-BUB design, a practical approach for eliciting the targeted alternative(s) is as follows. First, ask the physicians to hypothesize the marginal probabilities of the AE levels, {Min, Mod, Sev, Fatal}, in each treatment group. Denote these probabilities by

$$\boldsymbol{\theta}_{j,T} = (\theta_{j,Min}, \theta_{j,Mod}, \theta_{j,Sev}, \theta_{j,Fatal}), \text{ where } \theta_{j,Min} + \theta_{j,Mod} + \theta_{j,Sev} + \theta_{j,Fatal} = 1,\ j = F, FC.$$

Next, ask the physicians to hypothesize probabilities of the clinical response events, {CR, PR, SD,

Table 4: Response probabilities for the scenarios considered in our CLL trial simulation study. Toxicity probabilities correspond to {Min, Mod, Sev, Fatal}, and efficacy probabilities correspond to {CR, PR, SD, PD}, given that the patient is alive.

Scenarios	**Abbreviation**	**Response Probabilities**
All	NA	$\boldsymbol{\theta}_{F,T} = (0.67, 0.25, 0.05, 0.03)$
1.0, 2.0, 3.0, 4.0	=	$\boldsymbol{\theta}_{FC,T} = (0.67, 0.25, 0.05, 0.03)$
1.1, 2.1, 3.1, 4.1	>	$\boldsymbol{\theta}_{FC,T} = (0.44, 0.40, 0.10, 0.06)$
1.2, 2.2, 3.2, 4.2	>>	$\boldsymbol{\theta}_{FC,T} = (0.26, 0.45, 0.20, 0.09)$
All	NA	$\boldsymbol{\theta}_{F,E} = (0.25, 0.35, 0.20, 0.20)$
1.0, 1.1, 1.2	=	$\boldsymbol{\theta}_{FC,E} = (0.25, 0.35, 0.20, 0.20)$
2.0, 2.1, 2.2	>	$\boldsymbol{\theta}_{FC,E} = (0.35, 0.35, 0.15, 0.15)$
3.0, 3.1, 3.2	>>	$\boldsymbol{\theta}_{FC,E} = (0.45, 0.35, 0.10, 0.10)$
4.0, 4.1, 4.2	>>>	$\boldsymbol{\theta}_{FC,E} = (0.60, 0.30, 0.05, 0.05)$

PD}, conditional on being alive. Denote these conditional probabilities by

$$\boldsymbol{\theta}_{j,E} = (\theta_{j,CR}, \theta_{j,PR}, \theta_{j,SD}, \theta_{j,PD}), \text{ where } \theta_{j,CR} + \theta_{j,PR} + \theta_{j,SD} + \theta_{j,PD} = 1, \ j = F, FC.$$

Assuming independence for simplicity, set $\theta^{(Alt)}_{j,Fatal} = \theta_{j,Fatal}$ and $\theta^{(Alt)}_{j,k,\ell} = \theta_{j,k}\theta_{j,\ell}$ for $j = F, FC$, $k = \{Min, Mod, Sev\}$, and $\ell = \{CR, PR, SD, PD\}$. We assume that the targeted alternative arises from $\boldsymbol{\theta}_{F,T} = \boldsymbol{\theta}_{FC,T} = (0.67, 0.25, 0.05, 0.03)$, i.e., FC and F have equivalent toxicity, and $\boldsymbol{\theta}_{F,E} = (0.25, 0.35, 0.20, 0.20)$ versus $\boldsymbol{\theta}_{FC,E} = (0.45, 0.35, 0.10, 0.10)$, i.e., FC compared to F has higher efficacy. This alternative maintains similar marginal CR probabilities $\pi_{CR,FC} = 0.4365$ versus $\pi_{CR,F} = 0.2425$ specified for the actual trial design, and it results in a large mean utility difference $\delta^{(Alt)}_{\boldsymbol{U},FC-F} = 13.5$ for the utilities given in Table 1. Specifying $n^* = 1$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_F$ for the Dirichlet priors, a fixed sample CAT-BUB test requires slightly more patients than a beta-binomial test to achieve 90% power, $n_F = n_{FC} = 127$ versus 120. To ensure comparability, we determine the power figures for all three designs using the larger sample size 128. We compare the designs for 12 scenarios covering a wide range of different possibilities. The response probabilities for F and FC in each scenario are reported in Table 4. These probabilities for F are fixed at the targeted alternative values throughout, whereas the probabilities for FC vary with the combinations of the ordinal efficacy and toxicity outcomes. For each outcome, we characterize these numerical probability vector pairs nominally as being "equivalent" (=), or having "moderate" (>), "large" (>>), or "very large" (>>>) differences.

The results of our simulation reported in Table 5 show that, in general, the CAT-BUB design is sensitive to efficacy-toxicity tradeoffs characterized by the utilities, while the beta-binomial design with an efficacy test (BB-EO) is not. In contrast, if there is low power for detecting an efficacy difference between the two treatments, then the hierarchical beta-binomial design (BB-ET) is sensitive to toxicity, otherwise it is not. Because the BB-ET design is based on two tests, rather than one test like the BB-EO design, it requires a more stringent cut-off to control the type I error, and thus has lower power compared to the BB-EO design for selecting the treatment with superior efficacy, which is FC in all the scenarios we considered. Scenario 1.0 is the null, i.e., $\boldsymbol{\theta}_{FC} = (\boldsymbol{\theta}_{FC,E}, \boldsymbol{\theta}_{FC,T}) \equiv (\boldsymbol{\theta}_{F,E}, \boldsymbol{\theta}_{F,T}) = \boldsymbol{\theta}_F$, and the cut-off for all three designs was calibrated to control type I error at the $\alpha = 0.05$-level, where F and FC are selected with the same 0.025 probabilities. In Scenarios 1.1 and 1.2, FC has equivalent efficacy with moderate and high toxicity, respectively, and both the CAT-BUB design and the BB-ET design are increasingly likely to select F, whereas the BB-EO design is unable to distinguish between these clinically very different scenarios since it

Table 5: Power figures for the CLL trial based on the CAT-BUB design, the beta-binomial design with an efficacy test only (BB-EO), and the hierarchical beta-binomial design with an efficacy test followed by a toxicity test (BB-ET). Comparisons of $\boldsymbol{\theta}_{FC,E}$ vs $\boldsymbol{\theta}_{F,E}$ and $\boldsymbol{\theta}_{FC,T}$ vs $\boldsymbol{\theta}_{F,T}$ are characterized as being "equivalent" (=), or having "moderate" (>), "large" (>>), or "very large" (>>>) differences.

	Scenario			**Probability of Final Conclusion**					
						Beta-Binomial Designs			
	Efficacy	**Toxicity**		**CAT-BUB Design**		**Efficacy Only**		**Efficacy then Toxicity**	
	FC vs F	*FC vs F*	$\delta_{U,FC-F}$	$FC > F$	$F > FC$	$FC > F$	$F > FC$	$FC > F$	$F > FC$
1.0:	=	=	0.0	0.025	0.025	0.026	0.024	0.024	0.025
1.1:	=	>	-5.2	0.001	0.222	0.019	0.035	0.007	0.388
1.2:	=	>>	-11.9	0.000	0.782	0.012	0.047	0.005	0.982
2.0:	>	=	6.8	0.352	0.000	0.402	0.000	0.289	0.010
2.1:	>	>	1.1	0.041	0.012	0.331	0.000	0.226	0.308
2.2:	>	>>	-6.5	0.000	0.314	0.278	0.000	0.173	0.818
3.0:	>>	=	13.5	0.903	0.000	0.910	0.000	0.846	0.002
3.1:	>>	>	7.3	0.397	0.000	0.873	0.000	0.778	0.097
3.2:	>>	>>	-1.1	0.041	0.015	0.816	0.000	0.716	0.281
4.0:	>>>	=	21.0	1.000	0.000	1.000	0.000	1.000	0.000
4.1:	>>>	>	14.2	0.917	0.000	1.000	0.000	0.999	0.001
4.2:	>>>	>>	5.0	0.201	0.001	0.999	0.000	0.997	0.003

ignores toxicity. The BB-ET design is more likely to correctly select F than the CAT-BUB design, 0.39 versus 0.22 and 0.98 versus 0.78, respectively. In Scenario 2.0, FC has a moderate efficacy advantage with equivalent toxicity, and the BB-EO, CAT-BUB, and BB-ET designs select FC with probabilities 0.40, 0.35, and 0.29, respectively. In Scenario 2.1, FC has a moderate efficacy advantage and toxicity disadvantage, where this tradeoff that slightly favors FC for the assumed utilities. The CAT-BUB design is unlikely to select either treatment, whereas the BB-EO design selects FC with probability 0.33, and the BB-ET design selects FC and F with probabilities 0.23 and 0.31, respectively. In Scenario 2.2, because the toxicity disadvantage for FC increases, the tradeoff moderately favors F for the assumed utilities. The CAT-BUB design selects F with higher probability 0.31 and does not select FC, whereas the BB-EO design selects FC with probability 0.28 and does not select F, and the BB-ET design selects F with probabilities 0.82 and FC with probability 0.17. Scenario 3.0 is the targeted alternative, which is a case where FC has higher efficacy and equivalent toxicity compared to F. In this ideal case, the CAT-BUB design has 90% power compared to 91% power for the BB-EO design and 85% for the BB-ET design. Scenarios 3.1, 3.2, 4.1, and 4.2 are tradeoff settings where FC has an a large or very large efficacy advantage, and either a moderate or large toxicity disadvantage. In these cases, the CAT-BUB design selects treatments with probabilities that are sensitive to the assumed utilities, whereas the beta-binomial designs consistently select FC with high probability, regardless of the toxicity burden of FC.

Scenarios 1.2, 2.2, and 3.2 show very undesirable potential consequences of using the BB-EO design, which completely ignores toxicity. The BB-EO design based on the probability of CR treats Scenario 1.2 like a null case since $\pi_{CR,FC} = 0.2275$ versus $\pi_{CR,F} = 0.2425$, while in fact the two pairs of toxicity probability vectors $\boldsymbol{\theta}_{FC}$ and $\boldsymbol{\theta}_{F}$ are very different, with $\boldsymbol{\theta}_{FC,T} = (0.26, 0.45, 0.20, 0.09)$ versus $\boldsymbol{\theta}_{F,T} = (0.67, 0.25, 0.05, 0.03)$, so that FC has a much lower minor toxicity probability but much higher moderate, severe, and fatal AE probabilities compared to F. The CAT-BUB design recognizes this, concluding that $F > FC$ with power 0.78 compared to 0.05 for the BB-EO design. Scenario 2.2 is an intermediate case, since FC has moderate efficacy with $\boldsymbol{\theta}_{FC,E} = (0.35, 0.35, 0.15, 0.15)$ versus $\boldsymbol{\theta}_{F,E} = (0.25, 0.35, 0.20, 0.20)$ but also high toxicity, with

$\boldsymbol{\theta}_{FC,T} = (0.26, 0.45, 0.20, 0.09)$ versus $\boldsymbol{\theta}_{F,T} = (0.67, 0.25, 0.05, 0.03)$. The BB-EO design detects the $0.3185 - 0.2425 = 0.076$ difference in CR probabilities in favor of FC with power 0.28, but since the probability of severe toxicity or death is 0.29 with FC versus 0.08 with F, the CAT-BUB design concludes F is superior to FC with power 0.314 and never concludes that FC is superior to F. In Scenario 3.2, FC has a large efficacy advantage but also high toxicity burden, with $\boldsymbol{\theta}_{FC,E} = (0.45, 0.35, 0.10, 0.10)$ versus $\boldsymbol{\theta}_{F,E} = (0.25, 0.35, 0.20, 0.20)$ but, as in Scenario 2.2, $\boldsymbol{\theta}_{FC,T} = (0.26, 0.45, 0.20, 0.09)$ versus $\boldsymbol{\theta}_{F,T} = (0.67, 0.25, 0.05, 0.03)$. The BB-EO design has power 0.82 of concluding that FC is superior to F, whereas the CAT-BUB design recognizes both the much better efficacy and much worse toxicity with FC compared to F, and based on the assumed utilities does not recommend either treatment over the other with probability $1 - (0.041 + 0.015) = 0.94$. Scenario 4.2 shows that the BB-TE design can have a similar undesirable behavior as the BB-EO design. If the efficacy advantage of FC is very large, because the efficacy test will detect a difference with high probability, the BB-TE design effectively ignores toxicity, since the toxicity test is unlikely to be applied. In Scenarios 2.1 and 3.2, which have less extreme tradeoffs than Scenario 4.2, the BB-TE design is likely to recommend a particular treatment, despite that neither treatment may be strongly preferred under the assumed utilities. For example, in Scenario 3.2, the BB-TE design recommends FC and F with probabilities 0.72 and 0.28, respectively, and thus recommends either treatment with probability 0.99. Lastly, the BB-TE design has lower power than the CAT-BUB design for the targeted alternative in the CLL trial, i.e., Scenario 3.0.

There are several key points in these comparisons. First, basing a test on the probability of CR is equivalent to using a degenerate utility-based test that assigns utilities 100 to CR and 0 to its complement, while completely ignoring toxicity. An elaboration of this that accounts for the ordinal categories of efficacy is the two-sample test of Whitehead (1993), although this test still suffers from the fact that it ignores toxicity. Considering Scenarios 3.0, 3.1, and 3.2 together shows how the CAT-BUB design adjusts its conclusions depending on the varying $\boldsymbol{\theta}_{FC,T}$ vectors, essentially agreeing with the BB-EO design when toxicity is equivalent but very likely reaching the opposite conclusion when FC has much higher toxicity than F. The same pattern can be seen when considering Scenarios 4.0, 4.1, and 4.2 together. This also illustrates the benefits from considering the ordinal level of each outcome rather than dichotomizing it, since the probabilities of concluding that FC is superior to F vary from 1 to 0.20 as the probabilities of the levels of each outcome change across scenarios. In practice, if a conventional design based on efficacy alone is used one might hope that, in such cases, at some point during actual trial conduct someone would notice an excessively higher toxicity rate in one arm compared to the other, and ask the Principal Investigator or Institutional Review Board to halt accrual to the trial. Hope is not a strategy, however. Moreover, if in fact a trial designed based on efficacy alone will be stopped due to such a toxicity difference, then the nominal size and power of the design are incorrect, and in fact they are conditional on the assumption that there will be no difference in toxicity sufficiently large that it would cause the trial to be stopped early. All of these concerns are taken care of automatically by the group sequential CAT-BUB test's structure. For the group sequential design of the CLL trial (see Web Supplement), its interim decision rules will stop the trial early with high probability when there is a large difference in terms of either efficacy or toxicity, as quantified by the joint utilities of the elementary (efficacy, toxicity) outcomes.

The utilities in Table 1 used for the CAT-BUB re-design of the CLL trial emphasize both avoiding severe toxicity or death and achieving good clinical response by specifying the relative utility parameters (Section 2) to be $\zeta_1 = 0.1$ and $\zeta_2 = 0.2$ near zero, respectively. To assess the CAT-BUB design's sensitivity to the utilities, we also considered the two alternative utilities given in Table 6. One alternative places greater importance on achieving better efficacy, and the other places greater importance on achieving lower toxicity. We obtained the first (second) set of

Table 6: Three alternative utilities for the CLL trial's outcomes.

Level of Worst Adverse Event	**Clinical Response**				**Death**
Original Utilities from Table 1					
	CR	*PR*	*SD*	*PD*	
Minimal	100	84	35	19	
Moderate	93	77	29	14	0
Severe	28	24	14	10	
Utilities Giving Better Efficacy Higher Value					
Minimal	100	84	35	19	
Moderate	98	81	31	14	0
Severe	82	68	24	10	
Utilities Giving Lower Toxicity Higher Value					
Minimal	100	93	71	64	
Moderate	93	81	44	32	0
Severe	28	24	14	10	

Table 7: Power figures for the CLL trial based on a fixed sample CAT-BUB design using two alternative sets of numerical utilities that place greater importance on either improving efficacy or reducing toxicity, compared to the original utilities.

	Scenario		**Alternative Utility Giving Efficacy Higher Value**			**Alternative Utility Giving Lower Toxicity Higher Value**		
	Efficacy	Toxicity	$\delta_{\boldsymbol{U},FC-F}$	$FC > F$	$F > FC$	$\delta_{\boldsymbol{U},FC-F}$	$FC > F$	$F > FC$
1.0:	=	=	0.0	0.025	0.025	0.0	0.024	0.026
1.1:	=	>	-3.2	0.003	0.100	-8.4	0.000	0.931
1.2:	=	>>	-6.7	0.000	0.287	-18.3	0.000	1.000
2.0:	>	=	7.1	0.349	0.000	3.6	0.365	0.000
2.1:	>	>	3.7	0.122	0.003	-4.6	0.000	0.470
2.2:	>	>>	-0.1	0.026	0.025	-14.6	0.000	1.000
3.0:	>>	=	14.2	0.897	0.000	7.3	0.902	0.000
3.1:	>>	>	10.6	0.647	0.000	-0.8	0.011	0.049
3.2:	>>	>>	6.5	0.288	0.007	-11.0	0.000	0.986
4.0:	>>>	=	22.1	0.999	0.000	11.3	0.999	0.000
4.1:	>>>	>	18.2	0.985	0.000	3.4	0.290	0.000
4.2:	>>>	>>	13.8	0.859	0.000	-7.0	0.000	0.748

alternative utilities by changing ζ_2 from 0.20 to 0.80 (ζ_1 from 0.10 to 0.60), while retaining all other indirect elicitation parameter values as detailed in Section 2. For the first alternative, the utilities of moderate or severe toxicity for CR or PR levels of efficacy are substantially increased from the original utility, for example, $U(CR, Sev)$ is increased from 28 to 82. For the second alternative, the utilities of minimal or moderate toxicity for SD or PD levels of efficacy are substantially increased from the original utility, for example, $U(SD, Min)$ is increased from 35 to 71. The required sample sizes for the two alternative utilities are $n_F = n_{FC} = 110$ and 271, respectively, which is in contrast to $n_F = n_{FC} = 127$ for the elicited numerical utilities. The CAT-BUB design based on the utilities that place greater importance on higher efficacy thus requires a smaller sample size to achieve 90% power for detecting the targeted alternative compared with the beta-binomial design that was used for the actual trial. The numerical utilities that place greater importance on higher efficacy (lower toxicity) provide more (less) power for detecting the targeted alternative, which has a large treatment difference for the marginal probabilities of clinical response and zero difference for AE probabilities.

Table 7 reports the power figures for the same scenarios as in Table 5 for the CAT-BUB design based on the two alternative utilities with the above sample sizes. These power figures illustrate that the numerical utilities affect the power of detecting specific treatment differences and determine which therapy is preferred for tradeoff scenarios. Scenario 4.2 is an extreme tradeoff example, wherein FC has very high efficacy and high toxicity compared to F. As shown by the mean utility difference and power figures, this tradeoff slightly favors FC under the original utility function, whereas FC is strongly favored under the alternative utilities that place greater importance on efficacy, and conversely, under the alternative utilities that place greater importance on lower toxicity. The dependence of the proposed CAT-BUB design on the numerical utilities underscores the importance of eliciting values that actually reflect the clinical desirabilities of each patient response.

6.3 Additional Illustrations

In Section 1, we introduced several categorical outcome structures, including trinary, bivariate binary, bivariate binary plus death, ordinal, and bivariate ordinal. The computational algorithms given previously readily accommodate all of these cases. In Web Appendix D, we provide detailed illustrations of both fixed sample and group sequential CAT-BUB designs for a bivariate binary outcome, and for a bivariate ordinal outcome having $16 = 4 \times 4$ elementary events. We also include a group sequential CAT-BUB re-design of the CLL trial.

7 Discussion

Because clinical trial conduct must accommodate medical practice, a trial design should account formally for risk-benefit tradeoffs between all clinically relevant outcomes. The utility-based tests that we have proposed provide a practical approach for comparing treatments based on categorical outcomes in a RCT. We have provided both fixed sample and group sequential procedures, computational algorithms to derive design parameters, and freely-available user-friendly software. The CAT-BUB test directly addresses the problem of comparing treatments for all *clinically relevant* differences. The method deals with the problem of deciding whether one therapy is clinically superior to another based on its outcome probability vector by exploiting the elicited utilities of the elementary outcomes to reduce the multidimensional outcome to a one dimensional mean utility. This is used to construct comparative tests. The elicited utilities provide a rigorous framework for treatment comparison that makes explicit any subjective tradeoffs between outcomes.

We have demonstrated that designs based on a single binary outcome for efficacy often are unsafe, and do not reflect medical practice. Because safety is never a secondary concern in any clinical trial, conventional designs only based on an efficacy test presumably rely on informal stopping criteria for safety, which makes it difficult to determine operating characteristics. We also have demonstrated that designs based on a hierarchical testing procedure may be unsafe in scenarios where one treatment is more efficacious but also is more toxic than the other drug. These designs also do not reflect medical practice. If an efficacy difference is detected, it is naive to believe that the toxicity profiles of the available treatment options will not be considered by physicians when deciding which treatment is actually clinically preferable. In the proposed method, we have provided a practical tool to explicitly account for the tradeoffs between disparate outcomes that physicians routinely assess, and thus for designing RCTs that better reflect medical practice.

Supplementary Materials

The Web Appendices referenced in Sections 2, 4, 5, and 6 are available with this paper at the journal's website. An example spreadsheet mentioned in Section 2 for utility elicitation in the context of the CLL example, and a suite of R functions for implementing the computational algorithms in Sections 5.2 and 5.3, along with annotated example R programs for replicating each illustration are available at:

https://biostatistics.mdanderson.org/SoftwareDownload/

References

Brook, R. H., M. R. Chassin, A. Fink, D. H. Solomon, J. Kosecoff, and R. E. Park (1986, 1). A method for the detailed assessment of the appropriateness of medical technologies. *International Journal of Technology Assessment in Health Care 2*, 53–63.

Carlin, B. P. and T. A. Louis (2009). *Bayesian Methods for Data Analysis, 3rd edition.* Boca-Raton, FL: Chapman & Hall/CRC Press.

Cheson, B., J. Bennett, M. Grever, N. Kay, M. Keating, S. O'Brien, and K. Rai (1996). National Cancer Institute-Sponsored Working Group guidelines for chronic lymphocytic leukemia: revised guidelines for diagnosis and treatment. *Blood 87*(12), 4990–4997.

Dalkey, N. C. (1969). *The Delphi Method: An Experimental Study of Group Opinion.* Santa Monica, CA: RAND Corporation.

Eisenhauer, E., P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij (2009). New response evaluation criteria in solid tumours: Revised {RECIST} guideline (version 1.1). *European Journal of Cancer 45*(2), 228 – 247. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

Flinn, I. W., D. S. Neuberg, M. R. Grever, G. W. Dewald, J. M. Bennett, E. M. Paietta, M. A. Hussein, F. R. Appelbaum, R. A. Larson, D. F. Moore, and M. S. Tallman (2007). Phase III trial of fludarabine plus cyclophosphamide compared with fludarabine for patients with previously untreated chronic lymphocytic leukemia: US Intergroup Trial E2997. *Journal of Clinical Oncology 25*(7), 793–798.

Freedman, L., G. Anderson, V. Kipnis, R. Prentice, C. Wang, J. Rossouw, J. Wittes, and D. DeMets (1996). Approaches to monitoring the results of long-term disease prevention trials: Examples from the Women's Health Initiative. *Controlled Clinical Trials 17*(6), 509 – 525.

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

Hotelling, H. (1931). The economics of exhaustible resources. *Journal of Political Economy 39*(2), pp. 137–175.

Ibrahim, J. G. and M.-H. Chen (2000, 02). Power prior distributions for regression models. *Statist. Sci. 15*(1), 46–60.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods Applications to Clinical Trials.* Boca-Raton, FL: Chapman & Hall/CRC Press.

Kim, K. and D. L. DeMets (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika 74*(1), 149–154.

Lan, K. K. G. and D. L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*(3), pp. 659–663.

Murray, T., B. Hobbs, and B. Carlin (2015). Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *In Press: Annals of Applied Statistics*.

Piessens, R., E. D. Doncker-Kapenga, and C. W. Überhuber (1983). *QUADPACK: a subroutine package for automatic integration.* Springer.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials 18*(6), 530 – 545. Eighth International Symposium on Long-Term Clinical Trials.

Sankoh, A. J., R. B. D'Agostino, and M. F. Huque (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine 22*(20), 3133–3150.

Slud, E. and L. J. Wei (1982). Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association 77*(380), pp. 862–868.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine 12*(24), 2257–2271.

Web-based Supplement for “Utility-Based Designs for Randomized Comparative Trials with Categorical Outcomes”

Thomas A. Murray[1,*], Peter F. Thall[1,†] and Ying Yuan[1,‡]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center
[*]TAMurray@MDAnderson.org, [†]Rex@MDAnderson.org, [‡]YYuan@MDAnderson.org

April 26, 2016

Outline

In Web Appendix A, we provide the generalization of the indirect utility elicitation strategy described in Section 2 for a $K \times L$ (plus death) bivariate ordinal outcome. In Web Appendix B, we provide the derivation for the moment matching technique used in Section 4 to implement our scaled-beta approximation of the PMU distribution, and we confirm the validity of this approximation using a numerical comparison with the usual MC approach and a normal approximation in a variety of settings. In Web Appendix C, we provide guidelines for selecting the Monte Carlo sample sizes used in the computational algorithms defined in Section 5. In Web Appendix D, we present the detailed illustrations referenced in Section 6 for implementing the proposed CAT-BUB procedures with bivariate binary and bivariate ordinal outcomes.

Web Appendix A: Generalization of the Indirect Utility Elicitation Strategy

Suppose the clinically relevant outcome is bivariate ordinal (plus death), where there are $K \times L$ (+ 1) possible elementary responses. Without loss of generality, we partially order the elementary events such that $U(1,1) = 100$, $U(1,\ell) \geq U(2,\ell) \geq \cdots \geq U(K,\ell) \geq U(D) = 0$, and $U(k,1) \geq \cdots \geq U(k,L) \geq U(D) = 0$, for $k = 1,\ldots,K$ and $\ell = 1,\ldots,L$. To implement the indirect utility elicitation strategy described in Section 2, the statistician would create a spreadsheet with the following sub-tables for the physician(s) to fill in,

(A)

$(1,1)$	(K,L)	D
100	ν	0

(B)

	1	L
1	100	$100 \times \zeta_1$
K	$100 \times \zeta_2$	0

(C1)

$(1,1)$	$(1,2)$	$\cdots$	$(1,\ell)$	$\cdots$	$(1,L-1)$	$(1,L)$
100	$100 \times \phi_{1,1}$	$\cdots$	$100 \times \phi_{1,\ell-1}$	$\cdots$	$100 \times \phi_{1,L-2}$	0

(C2)

$(K,1)$	$(K,2)$	$\cdots$	(K,ℓ)	$\cdots$	$(K,L-1)$	(K,L)
100	$100 \times \phi_{2,1}$	$\cdots$	$100 \times \phi_{2,\ell-1}$	$\cdots$	$100 \times \phi_{2,L-2}$	0

(D1)

$(1,1)$	$(2,1)$	$\cdots$	$(k,1)$	$\cdots$	$(K-1,1)$	$(K,1)$
100	$100 \times \xi_{1,1}$	$\cdots$	$100 \times \xi_{k-1,1}$	$\cdots$	$100 \times \xi_{K-2,1}$	0

(D2)

$(1,L)$	$(2,L)$	$\cdots$	(k,L)	$\cdots$	$(K-1,L)$	(K,L)
100	$100 \times \xi_{1,2}$	$\cdots$	$100 \times \xi_{k-1,2}$	$\cdots$	$100 \times \xi_{K-2,2}$	0

If death is not included as a potential patient response, sub-table (A) will be omitted. If $L = 2$, sub-tables (C1) and (C2) will be omitted. Similarly, if $K = 2$, sub-tables (D1) and (D2) will be omitted. The numerical utilities that are induced by the entries in the above sub-tables can be determined sequentially as follows,

$$\begin{aligned}
U(1,1) &= 100, \quad \text{If } D, \text{ then } U(D) = 0 \text{ and } U(K,L) = \nu, \text{ otherwise, } U(K,L) = 0,\\
U(1,L) &= \zeta_1[U(1,1) - U(K,L)] + U(K,L),\\
U(K,1) &= \zeta_2[U(1,1) - U(K,L)] + U(K,L),\\
U(1,\ell+1) &= \phi_{1,\ell}[U(1,1) - U(1,L)] + U(1,L) \text{ and}\\
U(K,\ell+1) &= \phi_{2,\ell}[U(K,1) - U(K,L)] + U(K,L) \text{ for } \ell = 1,\ldots,L-2,\\
U(k+1,1) &= \xi_{k,1}[U(1,1) - U(K,1)] + U(K,1) \text{ and}\\
U(k+1,L) &= \xi_{k,2}[U(1,L) - U(K,L)] + U(K,L) \text{ for } k = 1,\ldots,K-2, \text{ and}\\
U(k+1,\ell+1) &= \left[\frac{\xi_{k,2}(\phi_{1,\ell} - \phi_{2,\ell}) + \phi_{2,\ell}}{1 - (\xi_{k,1} - \xi_{k,2})(\phi_{1,\ell} - \phi_{2,\ell})}\right][U(k+1,1) - U(k+1,L)] + U(k+1,L),\\
&\qquad \text{for } k = 1,\ldots,K-2, \text{ and } \ell = 1,\ldots,L-2.
\end{aligned}$$

The above induced numerical utilities, aside from those in the final line, i.e., $U(k+1,\ell+1)$, arise straightforwardly from the below identities,

$$\begin{aligned}
\nu &= [U(1,1) - U(K,L)]/[U(K,L) - U(D)],\\
\zeta_1 &= [U(1,1) - U(1,L)]/[U(1,L) - U(K,L)],\\
\zeta_2 &= [U(1,1) - U(K,1)]/[U(K,1) - U(K,L)],\\
\phi_{1,\ell} &= [U(1,\ell+1) - U(1,L)]/[U(1,1) - U(1,L)] \text{ and}\\
\phi_{2,\ell} &= [U(K,\ell+1) - U(K,L)]/[U(K,1) - U(K,L)] \text{ for } \ell = 1,\ldots,L-2,\\
\xi_{k,1} &= [U(k+1,1) - U(K,1)]/[U(1,1) - U(K,1)] \text{ and}\\
\xi_{k,2} &= [U(k+1,L) - U(K,L)]/[U(1,L) - U(K,L)] \text{ for } k = 1,\ldots,K-2.
\end{aligned}$$

Using linear interpolation, the remaining numerical utilities, i.e., $U(k+1,\ell+1)$, arise from the identities,

$$\begin{aligned}
U(k+1,\ell+1) &= \eta_{k,\ell}[U(k+1,1) - U(k+1,L)] + U(k+1,L),\\
\eta_{k,\ell} &= \eta'_{k,\ell}[\phi_{1,\ell} - \phi_{2,\ell}] + \phi_{2,\ell}, \text{ and}\\
\eta'_{k,\ell} &= \eta_{k,\ell}[\xi_{k,1} - \xi_{k,2}] + \xi_{k,2}.
\end{aligned}$$

Therefore, as desired,

$$\eta_{k,\ell} = [\xi_{k,2}(\phi_{1,\ell} - \phi_{2,\ell}) + \phi_{2,\ell}]/[1 - (\xi_{k,1} - \xi_{k,2})(\phi_{1,\ell} - \phi_{2,\ell})],$$

for $k = 1, \ldots, K-2$, and $\ell = 1, \ldots, L-2$.

Web Appendix B: Moment Matching and Validation for the Scaled-Beta Posterior Approximation

We derive the moment matching technique used for the scaled-beta approximation of the posterior distribution for mean utility. If $X \sim Beta(\lambda, \gamma)$, where $\mu = E[X] = \frac{\lambda}{\lambda+\gamma}$ and $\sigma^2 = Var[X] = \frac{\lambda\gamma}{(\lambda+\gamma)^2(\lambda+\gamma+1)}$, then $\gamma = \frac{\lambda(1-\mu)}{\mu}$, $\lambda + \gamma = \frac{\lambda}{\mu}$, and therefore $\sigma^2 = \frac{\mu^2(1-\mu)}{\lambda+\mu}$. Solving the last identity for λ, we have the first identity defined by equation (11) in the paper, i.e.

$$\lambda = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right).$$

Plugging this result into the identity $\gamma = \frac{\lambda(1-\mu)}{\mu}$, we have the second identity defined by equation (11) in the paper, i.e.

$$\gamma = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right).$$

We address the validation of the scaled-beta posterior approximation with two inquiries. In the first, we investigate how well the scaled-beta distribution approximates the actual PMU distribution, using the MC approach with M=1,000,000 as the gold standard, given various realizations of $\boldsymbol{X}$. In the second, we investigate the frequentist properties of the proposed CAT-BUB comparative testing procedure using the posterior quantities defined by equations (7) and (8) in the paper based on the scaled-beta approximation versus that based on the usual MC approach and the normal approximation defined by equation (13).

For our first assessment, we consider a single treatment with a trinary outcome and utilities $\boldsymbol{U} = (100,\ 50,\ 0)'$. For Scenario A we assume the observed data are $\boldsymbol{X} = n \times (0.50,\ 0.30\ ,0.20)$; whereas, for Scenario B we assume $\boldsymbol{X} = n \times (0.10,\ 0.10,\ 0.80)$. For both scenarios, we fix n equal to each of the values in the set $\{10,\ 25,\ 50,\ 100\}$ and use the usual MC approach with M=1,000,000 to estimate the actual PMU distribution. We then compare this precise estimate of the PMU distribution with the proposed scaled-beta approximation. The results of this assessment are displayed in Figure 1. As evidenced by Figure 1, the scaled-beta approximation is nearly indistinguishable from MC estimate even in small sample sizes (e.g. n=10 and 25) and non-negligible posterior density near the boundaries of the PMU distribution's domain (i.e. Scenario B).

For our second assessment, we continue with the example involving clot dissolving agents for rapid treatment of stroke with a trinary outcome and utility $\boldsymbol{U} = (100,\ 50,\ 0)'$. We investigate the power of the CAT-BUB approach for a variety of fixed response probability vectors. We consider one CAT-BUB test based on the usual MC approach with M=100,000, one based on the proposed scaled-beta approximation, and one based on a normal approximation matching the PMU distribution's mean and variance using the results from equation (10) in the paper. Using simulation, we tuned the threshold p_{cut} for each Bayesian test to control type I error at 0.05 under point null hypothesis $\boldsymbol{\theta}_A = \boldsymbol{\theta}_B = (0.50,\ 0.30,\ 0.20)'$ with 100 observations in each therapy arm.

The results of this simulation are reported in Table 1, and they show that scaled-beta approximation is nearly indistinguishable from the MC approach, while the normal approximation differs

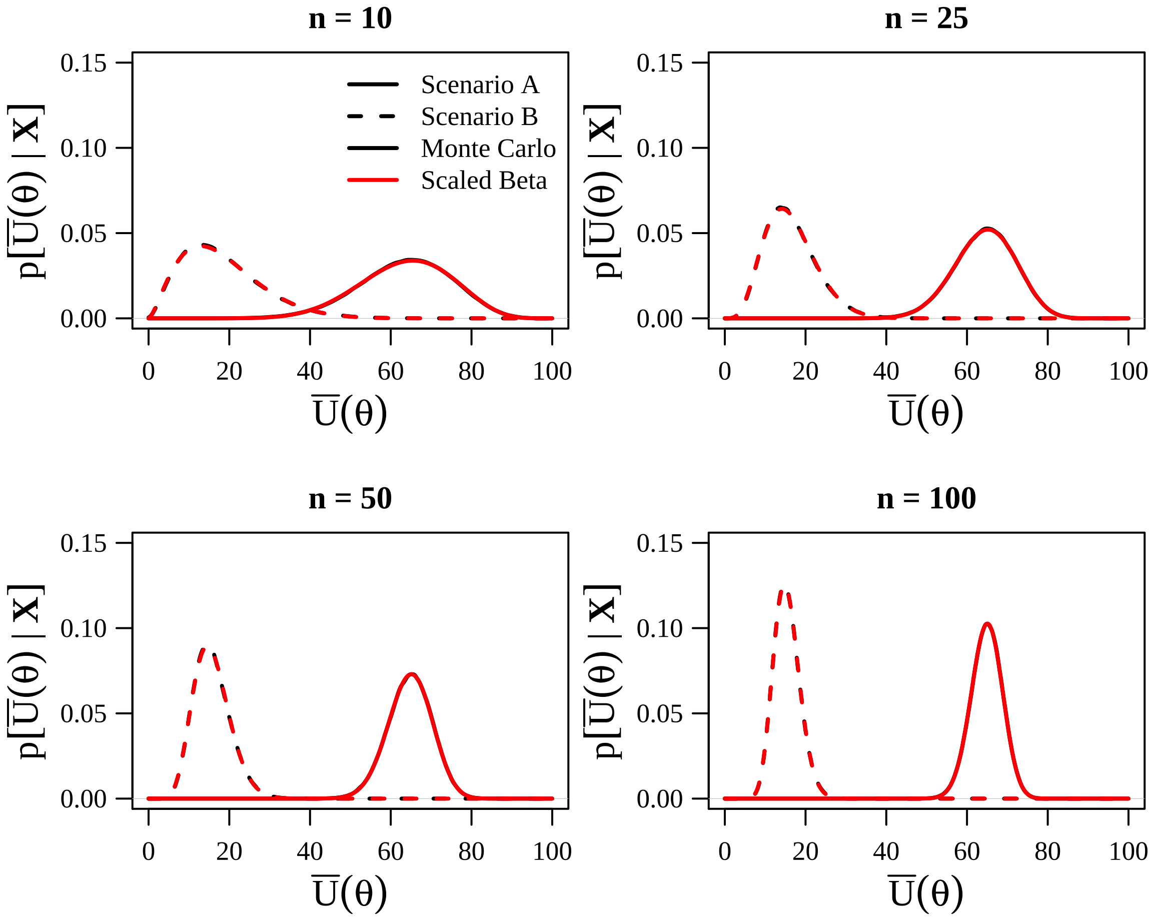

Figure 1: Numerical assessment of the proposed beta approximation for the posterior distribution of mean utility.

Table 1: Simulation-based power assessment under various fixed response probability vectors of the proposed utility-based tests based on the usual MC approach, the proposed scaled-beta approximation, and a normal approximation. All scenarios use $\boldsymbol{\theta}_A = (0.50,\ 0.30,\ 0.20)'$, and $n_A = n_B = 100$. All results are based on 2,500 simulation runs.

Scenario	$\boldsymbol{\theta}_B$	$\delta_{\boldsymbol{U},B-A}$	*MC*	*Beta*	*Normal*
	p_{cut}		0.976	0.976	0.976
1.0:	(0.50, 0.30, 0.20)	0	0.05	0.05	0.05
2.1:	(0.60, 0.00, 0.40)	-5	0.12	0.12	0.12
2.2:	(0.60, 0.10, 0.30)	0	0.05	0.05	0.05
2.3:	(0.60, 0.20, 0.20)	5	0.15	0.15	0.15
2.4:	(0.60, 0.30, 0.10)	10	0.49	0.48	0.48
2.5:	(0.60, 0.40, 0.00)	15	0.92	0.91	0.90
3.1:	(0.65, 0.05, 0.30)	3	0.07	0.07	0.07
3.2:	(0.65, 0.15, 0.20)	8	0.26	0.26	0.27
3.3:	(0.65, 0.25, 0.10)	13	0.68	0.68	0.68
3.4:	(0.65, 0.35, 0.00)	18	0.98	0.98	0.97
4.1:	(0.70, 0.00, 0.30)	5	0.13	0.13	0.13
4.2:	(0.70, 0.10, 0.20)	10	0.42	0.42	0.43
4.3:	(0.70, 0.20, 0.10)	15	0.82	0.82	0.82
4.4:	(0.70, 0.30, 0.00)	20	1.00	1.00	1.00

slightly for $n_A = n_B = 100$. The normal approximation works surprisingly well, providing a more powerful procedure in some scenarios, and a less powerful procedure in others. Similar simulations with smaller sample sizes also show good agreement between the MC and scaled-beta approach, whereas the normal approximation can suffer a loss of power relative to these procedures.

Web Appendix C: Computational Algorithm Guidelines

In step 3 of the fixed sample and group sequential CAT-BUB design computational algorithms defined in Section 5, we use G_0 MC replications to estimate p_{cut}, i.e. the $(1-\alpha)$%-tile of the sampling distribution for $T(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*) = \max\{T_{A>B}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*),\ T_{B>A}(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*)\}$ under the null. $\hat{p}_{cut}$ thus has standard error $\sqrt{\frac{\alpha(1-\alpha)}{G_0 f(F^{-1}(1-\alpha))^2}}$, where $f(\cdot)$ and $F(\cdot)$ denote the pdf and cdf for the sampling distribution of $T(\boldsymbol{X}; n^*, \boldsymbol{\theta}^*)$ under the null. Using equation (15) in the paper, this sampling distribution is approximately uniform between 0.5 and 1 when $\delta = 0$ (i.e. under the null), and thus $f(F^{-1}(1-\alpha)) \approx 2$. Therefore, the standard error for $\hat{p}_{cut}$ when $G_0 = 50{,}000$ is approximately 0.0005, which we feel is sufficiently accurate for estimating p_{cut} to three digits.

In contrast, G_1 is used to calculate $\hat{\beta}$, the power estimate for the alternative $\boldsymbol{\theta}^{(Alt)}$, which has standard error $\sqrt{\frac{\beta(1-\beta)}{G_1}}$, where $1-\beta$ is the true power. We use a default $G_1 = 25{,}000$, which ensures the standard error of $\hat{\beta}$ for $\beta = 0.20$ is about $0.0025 \approx \frac{\epsilon}{\Phi^{-1}(0.975)}$ for $\epsilon = 0.005$. Therefore, $Pr[\hat{\beta} \in (\beta-\epsilon, \beta+\epsilon) | \boldsymbol{\theta}^{(Alt)}] \approx 0.95$ when the true power for $\boldsymbol{\theta}^{(Alt)}$ is indeed $1-\beta$, and $Pr[\hat{\beta} \in (\beta-\epsilon, \beta+\epsilon) | \boldsymbol{\theta}^{(Alt)}] \approx 0.05$ or less when the true power for $\boldsymbol{\theta}^{(Alt)}$ is less than $(1-\beta) - 2\epsilon$ or greater than $(1-\beta) + 2\epsilon$. For general β and ϵ, G_1 can be adjusted accordingly.

Table 2: Power figures of a fixed sample CAT-BUB design for a bivariate binary outcome versus a beta-binomial design based on the probability of "success" defined as $[C, \overline{T}]$. In each scenario, $\boldsymbol{\theta}_A = (0.15,\ 0.20,\ 0.25,\ 0.40)'$ and $n_A = n_B = 284$.

			CAT-BUB Test		**Beta-Bin Test**	
Scenario	$\boldsymbol{\theta}_B$	$\delta_{\boldsymbol{U},B-A}$	$B > A$	$A > B$	$B > A$	$A > B$
1:	(0.15, 0.20, 0.25, 0.40)	0	0.025	0.025	0.025	0.025
2:	(0.10, 0.40, 0.20, 0.30)	9	0.805	0.000	0.000	0.438
3:	(0.15, 0.30, 0.20, 0.35)	6	0.445	0.000	0.025	0.025
4:	(0.20, 0.30, 0.25, 0.25)	13	0.980	0.000	0.351	0.000

Table 3: Power figures of a group sequential CAT-BUB design for a bivariate binary outcome. In each scenario, $\boldsymbol{\theta}_A = (0.15,\ 0.20,\ 0.25,\ 0.40)'$, $n_1 = 117$, $n_2 = 176$, $n_3 = 234$, $n_4 = 292$ and $\rho = 3$.

Scenario	$\boldsymbol{\theta}_B$	$\delta_{\boldsymbol{U},B-A}$	Ave SS	$B > A$	$A > B$
1:	(0.15, 0.20, 0.25, 0.40)	0	289.7	0.024	0.026
2:	(0.10, 0.40, 0.20, 0.30)	9	228.8	0.802	0.000
3:	(0.15, 0.30, 0.20, 0.35)	6	266.6	0.439	0.000
4:	(0.20, 0.30, 0.25, 0.25)	13	175.9	0.982	0.000

Web Appendix D: Additional Illustrations

In this section we provide detailed illustrations for how to design a CAT-BUB comparative trial with the general categorical outcome structures of most cases that are likely to be encountered in practice, including a bivariate binary outcome. For each outcome structure, we discuss utility elicitation prior to designing a fixed sample trial as well as a group sequential trial, and we report the resulting design's frequentist operating characteristics for a variety of alternative scenarios.

Bivariate Binary Outcome

Recall the acute myelogenous leukemia (AML) example that defines efficacy as complete remission, C, achieved within 42 days, and toxicity, T, as severe (National Cancer Institute grade 3 or 4) non-hematologic toxicity within 42 days. Denoting the complements by $\overline{C}$ and $\overline{T}$, suppose $U(C, \overline{T}) = 100$, $U(C, T) = 80$, $U(\overline{C}, \overline{T}) = 40$ and $U(\overline{C}, T) = 0$, where targeted alternative has probabilities of $([C, \overline{T}],\ [C, T]\ [\overline{C}, \overline{T}], [\overline{C}, T])$ for treatment A of $\boldsymbol{\theta}_A^{(Alt)} = (0.15,\ 0.20,\ 0.25,\ 0.40)'$, and for treatment B of $\boldsymbol{\theta}_B^{(Alt)} = (0.10,\ 0.40,\ 0.20,\ 0.30)'$. Therefore, the targeted alternative we will use to design our CAT-BUB trial is $\boldsymbol{\theta}^{(Alt)} = \left(\boldsymbol{\theta}_A^{(Alt)},\ \boldsymbol{\theta}_B^{(Alt)}\right)$ with a utility difference of $\delta_{\boldsymbol{U},B-A}^{(Alt)} = 9$.

Fixed Sample CAT-BUB Design

Using the fixed sample CAT-BUB design computational algorithm, we require n=284 subjects assigned to each treatment to achieve 80% power at the targeted alternative, while using a probability threshold of $p_{cut} = 0.975$ to control type I error at the 0.05 level. Table 2 reports the operating characteristics of this fixed sample CAT-BUB design for a few scenarios compared to the beta-binomial design for probability of "success" defined as $[C, \overline{T}]$.

Table 4: Elicited Utilities for Bivariate Ordinal Outcome.

	Disease Status			
Toxicity	CR/PR	SD2	SD1	PD
Mild	100	80	55	25
Moderate	90	70	35	20
High	70	50	25	10
Severe	40	25	10	0

Group Sequential CAT-BUB Design

Suppose we instead want to design a group sequential test with three interim analyses planned at sampling fractions $t_1 = 0.40$, $t_2 = 0.60$ and $t_3 = 0.80$ of the maximum sample size. Using the group sequential CAT-BUB design computational algorithm, a planned maximum sample size of $n_S =$ 292 subjects in each treatment group will provide 80% power at $\boldsymbol{\theta}^{(Alt)}$ while controlling type I error at the 0.05 level. Table 3 reports the operating characteristics of this group sequential CAT-BUB design for a few scenarios, assuming the planned analysis schedule is followed.

Bivariate Ordinal Outcome

Our motivating bivariate ordinal example involves a targeted agent plus chemotherapy for solid tumors, measuring efficacy and toxicity each with four levels. Efficacy events are partial or complete response (PR/CR) defined $>$ 30% reduction in tumor size from baseline, stable disease levels 2 ($SD2$) and 1 ($SD1$) defined respectively by 0–30% reduction in tumor size and 0–20% increase in tumor size, and progressive disease (PD) defined by >20% increase in tumor size. Toxicity events are mild, moderate, high and severe. The elicited utilities are provided in Table 4. We illustrate how to design both a fixed sample CAT-BUB test and a group sequential CAT-BUB test in this context.

For each type of CAT-BUB test, we use the same single targeted alternative. To ease elicitation, we consider marginal probabilities of efficacy and toxicity outcomes for each treatment, and assume independence to obtain the probabilities of the 16 possible elementary outcomes. For treatment A, we assume marginal efficacy probabilities of

$$\boldsymbol{\theta}^{(Alt)}_{A,Eff} = \left(\theta^{(Alt)}_{A,CR/PR} = 0.10,\ \theta^{(Alt)}_{A,SD2} = 0.10,\ \theta^{(Alt)}_{A,SD1} = 0.10,\ \theta^{(Alt)}_{A,PD} = 0.70\right)',$$

and marginal toxicity probabilities of

$$\boldsymbol{\theta}^{(Alt)}_{A,Tox} = \left(\theta^{(Alt)}_{A,Mild} = 0.70,\ \theta^{(Alt)}_{A,Mod} = 0.20,\ \theta^{(Alt)}_{A,High} = 0.05,\ \theta^{(Alt)}_{A,Severe} = 0.05\right)'.$$

Therefore, $\theta_{A,[CR/PR,Mild]} = 0.10 \times 0.70 = 0.07$, and the probabilities for the remaining 15 elementary outcomes follow similarly, or more concisely as $\boldsymbol{\theta}^{(Alt)}_{A} = \boldsymbol{\theta}^{(Alt)}_{A,Eff}\left(\boldsymbol{\theta}^{(Alt)}_{A,Tox}\right)'$. For treatment B, we use targeted marginal probabilities of $\boldsymbol{\theta}^{(Alt)}_{B,Eff} = (0.30,\ 0.20,\ 0.20,\ 0.30)'$ and $\boldsymbol{\theta}^{(Alt)}_{B,Tox} = (0.50,\ 0.20,\ 0.15,\ 0.15)'$ that correspond to a targeted utility difference of $\delta^{(Alt)}_{\boldsymbol{U},B-A} = 14.9$.

Fixed Sample CAT-BUB Design

The fixed sample CAT-BUB test providing 80% power at the targeted alternative, while controlling type I error at the 0.05 level requires 64 subjects in each treatment group, and has operating characteristics reported in Table 5.

Table 5: Power figures of a fixed sample CAT-BUB design with $n_A = n_B = 64$ for a bivariate ordinal outcome.

Scenario	$\boldsymbol{\theta}_B = \boldsymbol{\theta}_{B,Eff}\boldsymbol{\theta}'_{B,Tox}$	$\delta_{\boldsymbol{U},B-A}$	$B > A$	$A > B$
1:	$\boldsymbol{\theta}_{B,Eff} = (0.10,\ 0.10,\ 0.10,\ 0.70)'$ $\boldsymbol{\theta}_{B,Tox} = (0.70,\ 0.20,\ 0.05,\ 0.05)'$	0.0	0.026	0.024
2:	$\boldsymbol{\theta}_{B,Eff} = (0.30,\ 0.20,\ 0.20,\ 0.30)'$ $\boldsymbol{\theta}_{B,Tox} = (0.50,\ 0.20,\ 0.15,\ 0.15)'$	14.9	0.803	0.000
3:	$\boldsymbol{\theta}_{B,Eff} = (0.50,\ 0.20,\ 0.20,\ 0.10)'$ $\boldsymbol{\theta}_{B,Tox} = (0.60,\ 0.20,\ 0.10,\ 0.10)'$	32.3	1.000	0.000
4:	$\boldsymbol{\theta}_{B,Eff} = (0.40,\ 0.20,\ 0.20,\ 0.20)'$ $\boldsymbol{\theta}_{B,Tox} = (0.30,\ 0.20,\ 0.30,\ 0.20)'$	15.1	0.826	0.000
5:	$\boldsymbol{\theta}_{B,Eff} = (0.50,\ 0.10,\ 0.20,\ 0.20)'$ $\boldsymbol{\theta}_{B,Tox} = (0.20,\ 0.20,\ 0.30,\ 0.30)'$	12.0	0.633	0.000

Table 6: Power figures of a group sequential CAT-BUB design with $n_1 = 22$, $n_2 = 44$, $n_4 = 65$, and $\rho = 3$ for a bivariate ordinal outcome.

Scenario	$\boldsymbol{\theta}_B = \boldsymbol{\theta}_{B,Eff}\boldsymbol{\theta}'_{B,Tox}$	$\delta_{\boldsymbol{U},B-A}$	Ave SS	$B > A$	$A > B$
1:	$\boldsymbol{\theta}_{B,Eff} = (0.10,\ 0.10,\ 0.10,\ 0.70)'$ $\boldsymbol{\theta}_{B,Tox} = (0.70,\ 0.20,\ 0.05,\ 0.05)'$	0.0	64.6	0.025	0.25
2:	$\boldsymbol{\theta}_{B,Eff} = (0.30,\ 0.20,\ 0.20,\ 0.30)'$ $\boldsymbol{\theta}_{B,Tox} = (0.50,\ 0.20,\ 0.15,\ 0.15)'$	14.9	53.9	0.800	0.000
3:	$\boldsymbol{\theta}_{B,Eff} = (0.50,\ 0.20,\ 0.20,\ 0.10)'$ $\boldsymbol{\theta}_{B,Tox} = (0.60,\ 0.20,\ 0.10,\ 0.10)'$	32.3	29.5	1.000	0.000
4:	$\boldsymbol{\theta}_{B,Eff} = (0.40,\ 0.20,\ 0.20,\ 0.20)'$ $\boldsymbol{\theta}_{B,Tox} = (0.30,\ 0.20,\ 0.30,\ 0.20)'$	15.1	52.7	0.825	0.000
5:	$\boldsymbol{\theta}_{B,Eff} = (0.50,\ 0.10,\ 0.20,\ 0.20)'$ $\boldsymbol{\theta}_{B,Tox} = (0.20,\ 0.20,\ 0.30,\ 0.30)'$	12.0	57.6	0.631	0.000

Group Sequential CAT-BUB Design

Suppose we instead want to design a group sequential test with two interim analyses planned at equally spaced sampling fractions $t_1 = 1/3$ and $t_2 = 2/3$ of the maximum sample size. In this case, a planned maximum sample size of $n_S = 65$ subjects in each treatment group will provide 80% power at the targeted alternative while controlling type I error at the 0.05-level. Table 6 reports the operating characteristics of this group sequential CAT-BUB design for a few of other scenarios, assuming the planned analysis schedule is followed.

CLL Trial: Group Sequential CAT-BUB Re-design

We compare the operating characteristics a group sequential CAT-BUB design versus a group sequential beta-binomial design with an efficacy test for the CLL trial discussed in the main paper, each with two interim analyses planned at 40% and 70% accrual. We compare these designs across the same scenarios considered in the main paper and using the utilities reported in Table 1. Using our computational algorithm, the CAT-BUB design with interim looks when 53 and 92 patients are enrolled in each treatment regimen and a planned maximum sample size of 131 patients assigned to each regimen achieves 90% power for the targeted alternative. The operating characteristics of the two designs using the above sample sizes are reported in Table 7. In general, the results show that

Table 7: Power figures of the group sequential CAT-BUB design versus a beta-binomial design based on CR probability for the CLL trial, with tests when 53, 92, and 131 patients are enrolled in each regimen.

	Scenario			**Probability of Final Conclusion**					
	Efficacy	**Toxicity**		**CAT-BUB Design**			**Beta-Binomial Design**		
	FC vs F	*FC vs F*	$\delta_{\boldsymbol{U},FC-F}$	Ave SS	$FC > F$	$F > FC$	Ave SS	$FC > F$	$F > FC$
1.0:	=	=	0.0	130.2	0.024	0.025	130.2	0.024	0.025
1.1:	=	>	-5.2	129.2	0.001	0.219	130.2	0.018	0.036
1.2:	=	>>	-11.9	100.9	0.000	0.788	130.2	0.012	0.045
2.0:	>	=	6.8	125.1	0.359	0.000	124.0	0.398	0.000
2.1:	>	>	1.1	129.1	0.043	0.013	125.6	0.334	0.000
2.2:	>	>>	-6.5	121.0	0.000	0.316	126.7	0.267	0.000
3.0:	>>	=	13.5	99.2	0.902	0.000	101.0	0.910	0.000
3.1:	>>	>	7.3	123.4	0.400	0.000	104.8	0.872	0.000
3.2:	>>	>>	-1.1	130.2	0.013	0.042	108.7	0.816	0.000
4.0:	>>>	=	21.0	67.9	0.999	0.000	68.5	1.000	0.000
4.1:	>>>	>	14.2	97.1	0.919	0.000	71.5	1.000	0.000
4.2:	>>>	>>	5.0	127.5	0.205	0.001	74.3	0.999	0.000

the CAT-BUB design is sensitive to toxicity differences between the two regimens whereas the beta-binomial design is not. For example, the beta-binomial design has similar operating characteristics across Scenarios 3.0-3.2, i.e., between 0.82 and 0.91 probability of concluding $FC > F$ and between 101 and 109 average sample size. In contrast, the CAT-BUB design has similar power and average sample size in Scenario 3.0, wherein FC and F have equivalent toxicity burden, but is unlikely to conclude either regimen is superior in Scenario 3.2, wherein FC versus F has high toxicity that cancels out its high efficacy, based on these utilities.