Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE

Bayesian Safety and Futility Monitoring in Phase II Trials Using One Utility-Based Rule

Juhee Lee[1] | Peter F. Thall[2]

[1]Department of Statistics, University of California, Santa Cruz, Santa Cruz, California, USA | [2]Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston, Texas, USA

Correspondence: Juhee Lee (juheelee@soe.ucsc.edu)

Received: 21 April 2024 | **Revised:** 26 August 2024 | **Accepted:** 2 October 2024

Funding: Juhee Lee's research was supported by NSF grant DMS-1662427. Peter F. Thall's research was supported by NIH/NCI grants 1R01CA261978 and 5 P30 CA016672 45.

ABSTRACT

For phase II clinical trials that determine the acceptability of an experimental treatment based on ordinal toxicity and ordinal response, most monitoring methods require each ordinal outcome to be dichotomized using a selected cut-point. This allows two early stopping rules to be constructed that compare marginal probabilities of toxicity and response to respective upper and lower limits. Important problems with this approach are loss of information due to dichotomization, dependence of treatment acceptability decisions on precisely how each ordinal variable is dichotomized, and ignoring association between the two outcomes. To address these problems, we propose a new Bayesian method, which we call U-Bayes, that exploits elicited numerical utilities of the joint ordinal outcomes to construct one early stopping rule that compares the mean utility to a lower limit. U-Bayes avoids the problems noted above by using the entire joint distribution of the ordinal outcomes, and not dichotomizing the outcomes. A step-by-step algorithm is provided for constructing a U-Bayes rule based on elicited utilities and elicited limits on marginal outcome probabilities. A simulation study shows that U-Bayes greatly improves the probability of determining treatment acceptability compared to conventional designs that use two monitoring rules based on marginal probabilities.

1 | Introduction and Motivation

In a Phase II clinical trial of an experimental treatment, E, a futility monitoring rule stops accrual early if interim data show that E is not sufficiently promising. Many Phase II designs have been proposed, most based on the probability of a binary efficacy outcome, "response." Frequentist test-based Phase II designs include a group sequential procedure proposed by Chang et al. [1], and the two-stage optimal and minimax designs of Simon [2]. Thall and Simon [3] proposed a Bayesian Phase II design for trials with one binary response outcome that stops accrual if, based on interim data, it is unlikely a posteriori that the probability of response with E provides a specified level of improvement over a standard treatment. This rule is applied after successive cohorts of a specified size. Wathen et al. [4] refined this design to accommodate patient heterogeneity by defining subgroup-specific stopping rules, assuming a Bayesian analysis of covariance model. For biologically similar diseases in a basket trial [5], disease-specific futility monitoring may be done by assuming a Bayesian hierarchical model to induce correlation among diseases [6–8], or by performing frequentist tests [9].

Following introduction of Bayesian designs based on a Dirichlet-multinomial model for Phase II trials with multiple outcomes [10, 11], it has become common practice to include two stopping rules, a futility rule for binary response and a

safety rule for binary toxicity. Because the problem of monitoring multiple outcomes in clinical trials is quite common, many other Bayesian methods have been proposed, for a wide variety of clinical settings. The BOP2 method of Zhou, Lee, and Yuan [12] accommodates complex combinations of discrete and continuous endpoints, assuming a Dirichlet-multinomial model, while explicitly controlling overall Type I error rate. Sambucini [13] proposed a method based on predictive probabilities. Jiang et al. [14] studied the use of different types of stopping boundaries. Similarly, Bayesian Phase I–II dose finding designs include pairs of rules that stop accrual to a dose with an unacceptably high toxicity rate or low response rate, while continuing accrual to acceptable doses [14–18]. Extensions account for patient heterogeneity by using stopping rules that are specific to both dose and subgroup [19, 20].

In contrast to designs that use conventional rules based on marginal outcome probabilities, utility-based designs incorporate explicit trade-offs between outcomes through a utility function. As a simple illustration of how a utility function works in a phase II trial, consider a trial where the outcome is a bivariate binary variable consisting of indicators of toxicity and response, $(Y_T, Y_R) \in \{(0,0),(0,1),(1,0),(1,1)\}$, where $Y_T = 1$ if toxicity occurs and $Y_R = 1$ if response occurs. Numerical utilities can be established for the four possible outcome pairs and used to build a unified monitoring rule, as follows. For convenience, we let the utility $U(y_T, y_R)$ of an outcome (y_T, y_R) lie in the domain $[0, 100]$, and set $U(1,0) = 0$ for the worst possible outcome of toxicity and no response, and $U(0,1) = 100$ for the best possible outcome of response and no toxicity. We next elicit values for the two intermediate outcomes, (0,0) and (1,1). Suppose that the physician specifies $U(0,0) = 60$ and $U(1,1) = 70$. This implies that, if a response is achieved, toxicity reduces the utility from 100 to 70, but if no response is achieved, the absence of toxicity increases the utility from 0 to 60. This quantifies a significant penalty for toxicity without response. For each possible outcome pair (y_T, y_R), denote the joint probability by $\pi_{y_T,y_R} = P(Y_T = y_T, Y_R = y_R)$, and the vector $\boldsymbol{\pi} = (\pi_{0,0}, \pi_{0,1}, \pi_{1,0}, \pi_{1,1})$ with $\sum_{y_T=0}^{1}\sum_{y_E=0}^{1}\pi_{y_T,y_R} = 1$, so the mean utility is

$$\overline{U}(\boldsymbol{\pi}) = \sum_{y_T=0}^{1}\sum_{y_R=0}^{1}\pi_{y_T,y_R}U(y_T, y_R).$$

For example, if $\boldsymbol{\pi}_1 = (0.6, 0.1, 0.0, 0.3)$, then $\overline{U}(\boldsymbol{\pi}_1) = 67$, while if $\boldsymbol{\pi}_2 = (0.3, 0.4, 0.3, 0.0)$ then $\overline{U}(\boldsymbol{\pi}_2) = 58$. While $\boldsymbol{\pi}_1$ is more desirable than $\boldsymbol{\pi}_2$ in terms of mean utility, the two $\boldsymbol{\pi}$'s have identical marginal probabilities for toxicity and response, $P(Y_T = 1) = 0.3$ and $P(Y_R = 0) = 0.4$. Consequently, on average, a monitoring rule based on $\overline{U}$ is likely to lead to different decisions under $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ but, in contrast, a conventional rule that evaluates each outcome separately using its marginal probability gives the same decisions for $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$.

Bayesian utility-based designs have been proposed for various clinical trial settings. Thall and Nguyen [21] presented a dose-finding design based on the numerical utilities of all possible values of a bivariate ordinal outcome. Thall et al. [22] proposed a design that jointly optimizes dose and schedule based on two time-to-event variables, with the utility characterized as a surface obtained by smoothing utilities of pairs of event times elicited on a grid of rectangles. Murray, Thall, and Yuan [23] presented a randomized comparative trial design based on numerical utilities of post-operative morbidity severity levels. Lee, Thall, and Rezvani [24] presented a design to optimize the dose of natural killer cells to treat advanced hematologic malignancies, based on elicited joint utilities of five co-primary outcomes. Lee, Thall, and Msaouel [25] proposed a randomized design for treatment screening and selection based on subgroup-specific utilities of ordinal categorical response and toxicity.

In this article, we propose a utility-based Bayesian monitoring procedure, which we call U-Bayes, that uses one early stopping criterion to accommodate ordinal toxicity and ordinal efficacy. U-Bayes provides an alternative to first defining binary outcomes and then using two rules based on the resulting marginal probabilities of toxicity and response. Since, as described above, there are many different settings, in order to focus on comparing U-Bayes to the use of two marginal probability-based rules, we consider single-arm Phase II trials where toxicity and response both are ordinal, and we assume that patients are homogeneous. To construct a U-Bayes rule, a utility function that numerically quantifies the desirability of all possible combinations of toxicity and response must be elicited from the physicians planning the trial. Important advantages of this utility-based approach are that (1) it accounts for association between the multiple outcomes by using the joint distribution of the two ordinal outcomes, rather than only using the marginal distributions of binary variables obtained by dichotomizing each outcome, and (2) it obviates the problem of deciding where to dichotomize each ordinal variable. An additional advantage is that (3) the utility provides an explicit representation of toxicity-efficacy trade-offs, quantified by the subjective numerical desirability of each possible (toxicity, efficacy) outcome pair. While all methods for constructing monitoring rules require subjective decisions in their constructions, U-Bayes makes the subjective trade-off explicit through the elicited utility.

The remainder of the article is organized as follows. In Section 2, we define the conventional two-rule stopping criteria and provide motivating examples for the U-Bayes design. In Sections 3.1 and 3.2, we describe a utility function that varies with toxicity and response levels and define U-Bayes based on the utility function and the bivariate ordinal outcome distribution. In Section 3.3, we provide a step-by-step algorithm for calibrating the lower limit on mean utility required by U-Bayes. Section 3.4 presents a Bayesian bivariate probit model for the outcomes, and Section 3.5 presents the U-Bayes design. In Section 4, we evaluate operating characteristics of U-Bayes and two conventional comparators by simulation. We conclude with a brief discussion in Section 5.

2 | Constructing Monitoring Rules

We consider single-arm phase II trials where the goal is to decide whether an experimental treatment E has is sufficiently promising. If E is deemed unacceptable based on interim monitoring, then accrual to the trial is stopped early. Conventional Bayesian monitoring methods require dichotomizing each outcome, and computing two posterior probabilities, one that the probability of toxicity is too high, and the other that the probability of response

is too low, each compared to a selected threshold. To formalize this, denote $\boldsymbol{Y} = (Y_T, Y_R)$, where Y_T is ordinal toxicity and Y_R is ordinal response. Let $\{0, \ldots, K_j - 1\}$ denote the domain of Y_j for $j = T$ and R, with $K_j \geq 2$. We will use lowercase $\boldsymbol{y} = (y_T, y_R)$ to denote values taken on by $\boldsymbol{Y}$. To construct conventional Bayesian safety and futility stopping rules, for each $j = T$ and R, if $K_j > 2$ then Y_j must be dichotomized to define a binary variable. This is done by asking the physicians conducting the trial to specify the lowest level $h_T \in \{1, \ldots, K_T - 1\}$ of Y_T that is considered unacceptable, and the lowest level $h_R \in \{1, \ldots, K_R - 1\}$ of Y_R that is considered acceptable. Denoting the indicators $Z_{T,h_T} = I(Y_T \geq h_T)$ and $Z_{R,h_R} = I(Y_R \geq h_R)$ for specified cutoffs (h_T, h_R), conventional Bayesian rules determine the acceptability of E by using two posterior criteria, defined in terms of the marginal probabilities $\xi_{T,h_T} = P(Z_{T,h_T} = 1)$ and $\xi_{R,h_R} = P(Z_{R,h_R} = 1)$. The selected cut-offs are subjective and, as shown below, different (h_T, h_R) pairs may lead to designs that behave very differently.

To make things concrete, we will illustrate the conventional rules and proposed U-Bayes rule using the following prototype of a phase II trial. Toxicity is defined as Low (Grade 0 or 1), Moderate (Grade 2), High (Grade 3) or Severe (Grade 4, or 5). Response is defined by the disease status levels CR = complete response, PR = partial response, SD = stable disease, and PD = progressive disease. Thus, $K_T = K_R = 4$ and the levels of each Y_j are represented by the integers $\{0, 1, 2, 3\}$. In this example, each Y_j may be dichotomized in three ways, depending on the chosen cut-point h_j. For example, if $h_T = 2$ and $h_R = 2$, this defines $Z_{T,2} = 1$ for High or severe (Grade 3, 4, or 5) toxicity, and $Z_{T,2} = 0$ for Low or Moderate (Grade 0, 1, or 2), with $Z_{R,2} = 1$ if CR or PR is achieved, and $Z_{R,2} = 0$ for PD or SD. Alternatively, one may set $h_T = 3$ and $h_R = 3$ to define $Z_{T,3} = 1$ for Severe toxicity and $Z_{R,3} = 1$ for CR. There are a total of $3 \times 3 = 9$ possible ways to define (Z_{T,h_T}, Z_{R,h_R}), with each combination giving different meanings for "toxicity" and "response." This underscores the inherent subjectivity of the conventional dichotomization-based approach, and the fact that it greatly reduces the available information by replacing (Y_T, Y_R) with (Z_{T,h_T}, Z_{R,h_R}).

Let the vector $\boldsymbol{\pi} = (\pi_{0,0}, \ldots, \pi_{K_T-1,K_R-1})$ denote the joint distribution of (Y_T, Y_R) with $\pi_{y_T,y_R} = P(Y_T = y_T, Y_R = y_R)$. To specify conventional safety and futility stopping rules, given selected cutpoints h_T and h_R, the marginal probabilities of the binary outcomes are computed as follows:

$$\text{Toxicity/safety: } \xi_{T,h_T} = P(Z_{T,h_T} = 1) = P(Y_T \geq h_T) = \sum_{y_T=h_T}^{K_T-1} \sum_{y_R=0}^{K_R-1} \pi_{y_T,y_R},$$

$$\text{Response/futility: } \xi_{R,h_R} = P(Z_{R,h_R} = 1) = P(Y_R \geq h_R) = \sum_{y_T=0}^{K_T-1} \sum_{y_R=h_R}^{K_R-1} \pi_{y_T,y_R}.$$

These may be used to define two conventional monitoring rules,

$$\begin{aligned} &\text{Stop for safety if} && P(\xi_{T,h_T} > \overline{\xi}_{T,h_T} | data) > c_T(t), \\ &\text{Stop for futility if} && P(\xi_{R,h_R} < \underline{\xi}_{R,h_R} | data) > c_R(t). \end{aligned} \tag{1}$$

The probabilities $\overline{\xi}_{T,h_T}$ and $\underline{\xi}_{R,h_R}$ are fixed limits corresponding to h_T and h_R, which must be specified by the clinical investigators, while $c_T(t)$ and $c_R(t)$ denote decision cut-offs. Following Jiang et al. [14], to improve design performance, denoting the maximum sample size by N_{max}, and $n(t)$ the sample size at t, we define $c_T(t) = 1 - \{n(t)/N_{max}\}(1 - c_T^\star)$, where $c_T^\star$ is fixed. Thus, as $n(t)$ approaches N_{max}, $c_T(t)$ approaches $c_T^\star$ from above, and equals $c_T^\star$ when $n(t) = N_{max}$. We specify $c_R(t)$ similarly and let $c_R(t) = 1 - \{n(t)/N_{max}\}(1 - c_R^\star)$ with fixed $c_R^\star$. The fixed values $c_T^\star$ and $c_R^\star$ usually may be chosen in the range $0.80 - 0.95$, and are calibrated by computer simulation to obtain a design with desirable operating characteristics (OCs). These two rules are applied after successive cohorts of patients have been treated and their (Z_{T,h_T}, Z_{R,h_R}) values have been evaluated. The rules given by (1) are slightly simplified versions of the rules originally introduced by Thall, Simon, and Estey [10], where $\overline{\xi}_{T,h_T}$ and $\underline{\xi}_{R,h_R}$ each included random components chosen based on historical data on standard therapy as a comparator.

Using the rules in (1), desirable OCs of a design should include (i) a small early stopping probability and large expected sample size if the true marginal probabilities $(\xi_{T,h_T}^{TR}, \xi_{R,h_R}^{TR})$ satisfy the inequalities $\xi_{T,h_T}^{TR} \leq \overline{\xi}_{T,h_T}$ and $\xi_{R,h_R}^{TR} \geq \underline{\xi}_{R,h_R}$, and (ii) a large early stopping probability and small expected sample size if $\xi_{T,h_T}^{TR} > \overline{\xi}_{T,h_T} + \delta_T$ for nontrivial $\delta_T > 0$, such as 0.10 or larger, or $\xi_{R,h_R}^{TR} < \underline{\xi}_{R,h_R} - \delta_R$ for nontrivial $\delta_R > 0$. That is, the rules should give a small early stopping probability if E is both safe and efficacious, and a large stopping probability if E is too toxic or inefficacious. To calibrate a design's parameters, the trial is simulated for several candidate pairs of cutoffs $(c_T^\star, c_R^\star)$ and each of a set of different combinations of assumed true marginal probabilities $(\xi_{T,h_T}^{TR}, \xi_{R,h_R}^{TR})$. For convenience, to obtain OCs, it often is assumed that Z_{T,h_T} and Z_{R,h_R} are independent to facilitate computing joint probabilities from each pair of marginals.

While this commonly used paradigm for constructing two early stopping rules based on marginal probabilities may appear effective in practice, it has potentially severe limitations. These limitations can result in a design with undesirable properties in cases where multiple outcomes Y_T and Y_R should be fully and jointly accounted for in decision-making. First, the stopping rules given in (1) only use the marginal probabilities of dichotomized outcomes, $\xi_{T,h_T} = P(Z_{T,h_T} = 1)$ and $\xi_{R,h_R} = P(Z_{R,h_R} = 1)$. Because replacing (Y_T, Y_R) with (Z_{T,h_T}, Z_{R,h_R}) and considering only their marginals reduces information, it may lead to undesirable decisions when determining the acceptability of E. To illustrate this, we consider two of the simulation scenarios that will be described in Section 4. In Scenario 1 of Table 1a, for $h_T = h_R = 2$, the true binary marginal outcome probabilities are $\xi_{T,2}^{TR} = 0.25$ and $\xi_{R,2}^{TR} = 0.45$, obtained from the joint distribution $\boldsymbol{\pi}$. Scenario 2, given in Table 1b, has the exactly same marginals $\xi_{T,2}^{TR} = 0.25$ and $\xi_{R,2}^{TR} = 0.45$ as Scenario 1, consequently the conventional marginal rules will behave identically under Scenarios 1 and 2. For fixed limits $\overline{\xi}_{T,2} = 0.30$ and $\underline{\xi}_{R,2} = 0.40$, the rules in (1) conclude that E is unacceptable if $\xi_{T,2} > 0.30$ is likely, or if $\xi_{R,2} < 0.40$ is likely. Comparing the marginal probabilities to the limits, E is likely to be found acceptable using the marginal rules in (1) in both scenarios. The choice of cutoffs matters a great deal, however. While the true marginal severe toxicity probability with $h_T = 3$, $\xi_{T,3}^{TR} = P^{TR}(Z_{T,3} = 1)$, equals 0.05 in Scenario 1 and 0.20 in Scenario 2, the marginal probability of (high or severe toxicity) equals $\xi_{T,2}^{TR} = 0.25$ in both scenarios. This shows that the

TABLE 1 | Numerical illustration.

	(a) Scenario 1					(b) Scenario 2				
	Response (y_R)					Response (y_R)				
Toxicity	**PD**	**SD**	**PR**	**CR**		**PD**	**SD**	**PR**	**CR**	
Severity (y_T)	**(0)**	**(1)**	**(2)**	**(3)**		**(0)**	**(1)**	**(2)**	**(3)**	
Low (0)	0.04	0.35	0.03	0.28	0.70	0.04	0.00	0.01	0.00	0.05
Moderate (1)	0.00	0.02	0.01	0.02	0.05	0.40	0.04	0.24	0.02	0.70
High (2)	0.01	0.10	0.01	0.08	0.20	0.02	0.00	0.03	0.00	0.05
Severe (3)	0.00	0.03	0.00	0.02	0.05	0.04	0.01	0.12	0.03	0.20
	0.05	0.50	0.05	0.40	1.00	0.50	0.05	0.40	0.05	1.00
Binary	**Binary response**					**Binary response**				
Toxicity	**No**	**Yes**				**No**	**Yes**			
No	0.41	0.34	0.75			0.48	0.27	0.75		
Yes	0.14	0.11	0.25			0.07	0.18	0.25		
	0.55	0.45	1.00			0.55	0.45	1.00		

Note: Two joint distributions of the bivariate ordinal outcomes $\boldsymbol{Y} = (Y_T, Y_R)$ and binary variables $Z_T = \mathrm{I}[Y_T \geq 2]$ and $Z_R = \mathrm{I}[Y_R \geq 2]$. In each of the two scenarios, $\boldsymbol{Z} = (Z_T, Z_R)$ have the marginals $\xi_{T,2} = 0.25$ and $\xi_{R,2} = 0.45$.

choice of whether one uses the cutoff $h_T = 3$ or 2 substantively changes the meaning of "toxicity". Moreover, the joint probabilities $\boldsymbol{\pi}$ differ between the two scenarios. In Scenario 2, there is a higher probability of better response outcomes occurring with worse toxicity, compared to Scenario 1. Specifically, for the highly desirable combination (low toxicity, CR) $= (Y_T = 0, Y_R = 3)$, in Scenario 1 this has joint probability $\pi_{0,3} = 0.28$, compared to 0.00 in Scenario 2. In contrast, $\pi_{1,2} = \Pr$(Moderate Toxicity, PR) equals 0.01 in Scenario 1 and 0.24 in Scenario 2. These large differences strongly suggest that quantifying the desirability of each possible outcome combination (y_T, y_R) may be very useful. In the simulations presented in Section 4, under Scenario 1, the utility based rule of U-Bayes declares E promising with probability 1.00, compared to 0.96 using either of the marginal probability based rules. In contrast, under Scenario 2, U-Bayes declares E promising with probability 0.00, compared to 0.92 using the marginal probability based rules. These large differences between the decision probabilities of the two monitoring approaches are due to the facts that (1) U-Bayes retains information on $\boldsymbol{Y}$ that is lost by dichotomizing and considering only marginal probabilities, and (2) U-Bayes accounts for the desirability of each outcome pair (y_T, y_R).

Another important problem is that dichotomizing each ordinal Y_j to define binary Z_{j,h_j} and formulate a monitoring rule may be done in more than one way. Different (h_T, h_R) pairs may lead to very different decisions about the acceptability of E for the same data. For example, suppose that Severe toxicity is used to define $Z_{T,3} = I[Y_T = 3]$ and CR is used to define $Z_{R,3} = I[Y_R = 3]$. That is, $h_T = h_R = 3$, so $\xi_{T,3} = \mathrm{P}(Z_{T,3} = 1) = \mathrm{P}(Y_T = 3)$ and $\xi_{R,3} = \mathrm{P}(Z_{R,3} = 1) = \mathrm{P}(Y_R = 3)$. In this case, values of $\overline{\xi}_{T,3}$ and $\underline{\xi}_{R,3}$ must be elicited to correspond to the marginal probabilities determined by these cut-points. Under $\boldsymbol{\pi}^{TR}$ in Scenario 2 of Table 1b, $\xi^{TR}_{T,3} = 0.20$ and $\xi^{TR}_{R,3} = 0.05$. If the upper limit $\overline{\xi}_{T,3} = 0.10$ and lower limit $\underline{\xi}_{R,3} = 0.30$ are elicited then, in this scenario, E is considered too toxic and inefficacious, so the two marginal probability based rules are likely to conclude that E is not acceptable. This is the opposite of the conclusion that is likely to be reached if, instead, the definitions $Z_{j,2} = \mathrm{I}(Y_j \geq 2)$ for $h_j = 2$, $j = T$ and R are used. In contrast, under Scenario 1 in Table 1a, the two different ways of dichotomizing $\boldsymbol{Y}$, with either $\boldsymbol{h} = (2, 2)$ or $\boldsymbol{h} = (3, 3)$, are likely to lead to the same conclusion that E is acceptable.

A key point is that, because the U-Bayes design is based on the utility of the joint outcomes, while the conventional design relies on two marginal dichotomized outcomes, they use qualitatively different early stopping criteria. Consequently, as illustrated by Scenario 2, the two approaches may disagree greatly with regard to what is considered a desirable or undesirable scenario in terms of $\boldsymbol{\pi}$. Recall the example in the bivariate binary case, given earlier, where $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ differed but they had identical marginal probabilities. For a given $\boldsymbol{\pi}$, what are considered desirable OCs under the U-Bayes design may differ from what are considered desirable OCs under the conventional design based on two marginal probabilities. For example, consider $\boldsymbol{\pi}$ having marginals for which $\xi_{T,h_T} = \overline{\xi}_{T,h_T} + 0.10$ and $\xi_{R,h_R} = \underline{\xi}_{R,h_R} + 0.20$. If the 0.10 increase in the probability of toxicity is considered a desirable trade-off for the 0.20 increase in the probability of response, then E is acceptable. In contrast, a design based on the two marginal probability criteria in (1) would consider E to be unacceptable due to its high toxicity rate, regardless of the response rate, and thus it would prefer a large early stopping probability in this case.

3 | Utility Based Rule and Trial Design

3.1 | Utility Function

The following unified monitoring rule avoids all of the problems described above. To construct the rule, one first must elicit the numerical utility $U(\boldsymbol{y})$ of each potential outcome pair $\boldsymbol{y}$,

TABLE 2 | Values of the numerical joint utilities $U(y_T, y_R)$ for $K_T = 4$ and $K_R = 4$ in the illustrative example.

		Response			
		PD	**SD**	**PR**	**CR**
Toxicity severity		**($y_R = 0$)**	**($y_R = 1$)**	**($y_R = 2$)**	**($y_R = 3$)**
Low: Grade 0 or 1	($y_T = 0$)	25	70	90	100
Moderate: Grade 2	($y_T = 1$)	10	50	70	90
High: Grade 3	($y_T = 2$)	5	30	40	60
Severe: Grade 4 or 5	($y_T = 3$)	0	10	20	30

which quantifies its desirability. Numerical utilities of the $K_T \times K_R$ pairs must be elicited from the clinical collaborators to reflect patient preferences. To elicit $U(\boldsymbol{y})$ in the context of our illustration, generalizing the approach described earlier for the bivariate binary outcome case, it is convenient to first fix $U(3, 0) = 0$ and $U(0, 3) = 100$, which are the respective utilities of the worst and best possible outcome pairs. One then may elicit utilities between 0 and 100 for the remaining outcomes, provided that they satisfy the consistency conditions $U(y_T, y_R) < U(y_T, y_R + 1)$ and $U(y_T, y_R) > U(y_T + 1, y_R)$, which formalize the requirement that either higher toxicity or worse response must have lower utility if the other outcome is fixed. Table 2 illustrates a utility function for the illustrative trial. For example, the values $U(0, 0) = U(\text{Low toxicity, PD}) = 25$ and $U(3, 3) = U(\text{Severe toxicity, CR}) = 30$ quantify a slightly larger desirability of an outcome with a high toxicity grade and CR compared to an outcome with low toxicity and PD. However, setting $U(1, 1) = 50$ says that (mild toxicity, SD) is much more desirable than either of these outcomes, so for this utility both severe toxicity and PD are very undesirable.

A utility function helps to quantify preferences for outcomes in a structured way, enabling decisions to be made by comparing the utilities associated with different options. As specific examples, consider a new treatment developed for an advanced stage of a life-threatening disease, such as acute leukemia or lymphoma. In such cases, a patient may be willing to endure higher grade toxicity events if it means achieving higher-level responses, since achieving a complete response (CR) is crucial for long-term survival. For these patients, the utility does not significantly decrease as the toxicity grade increases for CR ($y_R = 3$). On the other hand, when assessing a new antihypertensive drug designed to reduce high blood pressure, a patient might be far less willing to accept Grade 2 or Grade 3 toxicity as a desirable trade-off for response, defined as lowering systolic blood pressure by ≥ 10 mm Hg. In this case, the numerical value of U(Grade 2 toxicity, response) would be much lower than that of U(no toxicity, response). These trade-offs are quantified by utility functions, which facilitate decision making by offering an explicit objective function for decision makers to utilize in making the best choice. A rule based on $U(\mathbf{y})$, such as U-Bayes, provides a systematic procedure for understanding and making choices across various contexts. In simulation Scenarios 3 and 4, reflecting the trade-offs associated with advanced-stage life-threatening diseases, it will be demonstrated that U-Bayes tends to conclude that treatment E is acceptable (i) when its probabilities of high or severe toxicity slightly exceed specified limits but its probabilities of higher-level response are substantial, or (ii) when it exhibits low probabilities of high or severe toxicity alongside sufficiently large probabilities of higher-level response.

In practice, eliciting $U(\boldsymbol{y})$ is straightforward, since physicians understand what the utility means and readily provide their numerical values. During the elicitation process, discussing the implications of particular specified numerical values, as above, provides a simple way for investigators to adjust their numerical values, if desired. See, for example, Thall and Nguyen [21], Murray et al. [26], or Lee, Thall, and Rezvani [24].

3.2 | U-Bayes: A Utility-Based Bayesian Monitoring Rule

As an alternative to using two conventional monitoring rules of the forms in (1), U-Bayes uses a single monitoring rule, defined in terms of the utility function described in Section 3.1. This is constructed as follows. Recall that $\boldsymbol{\pi}$ denotes the vector of $K_T \times K_R$ joint probabilities for all possible outcome pairs (y_T, y_R). Given $\boldsymbol{\pi}$ and $U(\boldsymbol{y})$, the mean utility is

$$\overline{U}(\boldsymbol{\pi}) = \sum_{y_T=0}^{K_T-1} \sum_{y_R=0}^{K_R-1} U(y_T, y_R)\, \pi_{y_T, y_R}. \tag{2}$$

U-Bayes uses $\overline{U}(\boldsymbol{\pi})$ to define a single early stopping criterion. We employ a Bayesian model that utilizes outcome data from patients previously treated in the trial to generate a posterior distribution of $\boldsymbol{\pi}$. It is used to derive the posterior distribution of $\overline{U}(\boldsymbol{\pi})$ and define the stopping rule based on it. Details of the probability model will be provided in Section 3.4. Given a fixed lower limit $\underline{U}$, if E satisfies the following inequality it is considered unacceptable because, a posteriori, its mean utility is likely to be too low, and patient accrual is stopped:

$$\text{Utility-based stopping criterion:} \quad \Pr(\overline{U}(\boldsymbol{\pi}) < \underline{U} | data) > c_U(t). \tag{3}$$

Otherwise, E is considered acceptable and accrual is continued. Similarly to $c_T(t)$ and $c_R(t)$ in (1), the cutoff parameter $c_U(t)$ varies with trial time t. We define $c_U(t) = 1 - n(t)/N_{\max}(1 - c_U^\star)$, where $c_U^\star$ is fixed. A value between 0.80 and 0.95 can be used for $c_U^\star$, calibrated by simulation to control the rates of incorrect decisions. The lower limit $\underline{U}$ is a key design parameter that is determined based on the elicited upper and lower marginal probability limits, $\overline{\xi}_{T,h_T}$ and $\underline{\xi}_{R,h_R}$, and the elicited joint utilities, using the following calibration algorithm.

3.3 | Algorithm for Calibrating the Lower Utility Bound

In this subsection, we explain how to use elicited upper limits $\overline{\xi}_{\mathrm{T},h_T}$ for each $h_\mathrm{T} = 1, \ldots, K_\mathrm{T} - 1$ and lower limits $\underline{\xi}_{\mathrm{R},h_R}$ for $h_\mathrm{R} = 1, \ldots, K_\mathrm{R} - 1$, and the elicited utilities $U(\boldsymbol{y})$, to compute the lower limit $\underline{U}$ needed to define the U-Bayes stopping rule in (3). During the trial planning process, this calibration is carried out through computer simulation in the following steps.

Steps for Calibrating $\underline{U}$

Step 1: For each $h_\mathrm{T} = 1, \ldots, K_\mathrm{T} - 1$, elicit an upper limit $\overline{\xi}_{\mathrm{T},h_T}$ for $\mathrm{P}(Y_\mathrm{T} \geq h_\mathrm{T})$.

Step 2: For each $h_\mathrm{R} = 1, \ldots, K_\mathrm{R} - 1$, elicit a lower limit $\underline{\xi}_{\mathrm{R},h_R}$ for $\mathrm{P}(Y_\mathrm{R} \geq h_\mathrm{R})$.

Step 3: Simulate a large sample $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(B)}\}$ of joint probability vectors such that each $\boldsymbol{\pi}^{(b)}$ has marginal distributions for Y_T and Y_R that satisfy the following $K_\mathrm{T} + K_\mathrm{R} - 2$ constraints:

$$\Pr(Y_\mathrm{T} \geq h_\mathrm{T}) = \overline{\xi}_{\mathrm{T},h_T}, \quad h_\mathrm{T} = 1, \ldots, K_\mathrm{T} - 1, \quad \text{and}$$
$$\Pr(Y_\mathrm{R} \geq h_\mathrm{R}) = \underline{\xi}_{\mathrm{R},h_R}, \quad h_\mathrm{R} = 1, \ldots, K_\mathrm{R} - 1.$$

Step 4: Set the lower utility limit equal to the sample mean of the mean utilities evaluated at the B simulated joint probabilities,

$$\underline{U} = \frac{1}{B}\sum_{b=1}^{B}\overline{U}(\boldsymbol{\pi}^{(b)}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{y_T=0}^{K_T-1}\sum_{y_R=0}^{K_R-1} U(y_\mathrm{T}, y_\mathrm{R})\,\pi^{(b)}_{y_T,y_R}. \tag{4}$$

In Step 3, it is important that the simulated joint probability vectors $\boldsymbol{\pi}^{(1)}, \ldots \boldsymbol{\pi}^{(B)}$ represent varying degrees of association between Y_T and Y_R, since $\overline{U}(\boldsymbol{\pi})$ is based on the joint distribution rather than only the two marginals. This allows the simulated mean utilities $\overline{U}(\boldsymbol{\pi}^{(1)}), \ldots, \overline{U}(\boldsymbol{\pi}^{(B)})$ to reflect varying degrees of association, while all the $\boldsymbol{\pi}^{(b)}$'s have the same marginals. Step 3 thus requires a tractable method for generating $\boldsymbol{\pi}$'s that have varying degrees of negative and positive association between Y_T and Y_R, and we provide additional details for carrying out Step 3 below.

The rationale for deriving $\underline{U}$ in this way is that, due to the constraints on the marginal probabilities imposed in Step 3, each simulated joint probability $\boldsymbol{\pi}^{(b)}$ has marginal toxicity and response probabilities that are exactly at their respective elicited limits in Steps 1 and 2. Defining $\underline{U}$ by taking the mean over the B simulated probability vectors in (4) thus gives the corresponding one-dimensional lower limit on $\overline{U}(\boldsymbol{\pi})$ for computing the stopping rule (3), while also accounting for association between Y_T and Y_R.

While calibrating the lower utility limit $\underline{U}$ is not tied to a specific model for simulating $\boldsymbol{\pi}$ values, the assumed model must be both flexible and numerically tractable. To construct such a model, we follow Chib and Greenberg [27], and many others, by employing a multivariate probit model. This uses the computational device of defining $\boldsymbol{Y}$ in terms of latent real-valued variables $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_\mathrm{T}, \widetilde{Y}_\mathrm{R}) \in \mathbb{R}^2$ that follow a bivariate normal distribution, given by

$$\widetilde{\boldsymbol{Y}}|\boldsymbol{\mu}, \Sigma \sim \mathrm{N}_2(\boldsymbol{\mu}, \Sigma), \tag{5}$$

where

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_\mathrm{T}\\ \mu_\mathrm{R}\end{bmatrix} \text{ and } \Sigma = \begin{bmatrix}\sigma^2 & \rho\sigma^2\\ \rho\sigma^2 & \sigma^2\end{bmatrix}. \tag{6}$$

For each outcome $j = T$ or R, the value of observed Y_j is defined by using the cutoffs $e_{j,0} < e_{j,1} < \cdots < e_{j,K_j}$ with $e_{j,0} = -\infty$ and $e_{j,K_j} = \infty$, so that

$$Y_j = y_j \text{ if and only if } e_{j,y_j} < \widetilde{Y}_j \leq e_{j,y_j+1}, \text{ for } y_j = 0, \ldots, K_j - 1. \tag{7}$$

Under this construction, the bivariate normal distribution of the latent variables $\widetilde{\boldsymbol{Y}}$ and the cutoffs determine the distribution $\boldsymbol{\pi}$ of $\boldsymbol{Y}$, with the correlation ρ between $\widetilde{Y}_\mathrm{T}$ and $\widetilde{Y}_\mathrm{R}$ inducing association between Y_T and Y_R. For each possible observed outcome pair $(y_\mathrm{T}, y_\mathrm{R})$, we denote the rectangle

$$A_{y_T,y_R} = \{(\tilde{y}_\mathrm{T}, \tilde{y}_\mathrm{R}) : \ e_{\mathrm{T},y_T} \leq \tilde{y}_\mathrm{T} < e_{\mathrm{T},(y_T+1)} \text{ and } e_{\mathrm{R},y_R} \leq \tilde{y}_\mathrm{R} < e_{\mathrm{R},(y_R+1)}\} \subset \mathbb{R}^2. \tag{8}$$

Using (5), (7), and (8), we define the joint probabilities as bivariate normal probabilities of the rectangles,

$$\pi_{y_T,y_R} = \iint_{A_{y_T,y_R}} f_\mathrm{N}(\tilde{\boldsymbol{y}}|\boldsymbol{\mu}, \Sigma)\, d\tilde{y}_\mathrm{T}\, d\tilde{y}_\mathrm{R} \tag{9}$$

where $f_\mathrm{N}(\cdot|\boldsymbol{\mu}, \Sigma)$ denotes the bivariate normal pdf given above.

Below, we provide computational sub-steps that exploit this structure to carry out Step 3 of the $\underline{U}$ calibration given above. This is done by generating the $\boldsymbol{\pi}^{(b)}$'s from the assumed bivariate normal distribution of the latent variables using (9). To represent a range of associations between Y_T and Y_R, we vary the numerical value of ρ in (6) across the domain $(-1, 1)$ to generate the $\boldsymbol{\pi}^{(b)}$'s. This is done in the following substeps of calibration Step 3:

Steps for Simulating $\boldsymbol{\pi}^{(1)}, \cdots \boldsymbol{\pi}^{(B)}$ in $\underline{U}$ Calibration Step 3

Step 3.1: Fix values of $\boldsymbol{\mu} = [\mu_\mathrm{T}, \mu_\mathrm{R}]'$ and σ. Specify a grid of B equally spaced correlations $\rho^{(1)} < \rho^{(2)} < \cdots < \rho^{(B)}$ in the domain $(-1, 1)$. For each $\rho^{(b)}$ in the grid, denote

$$\Sigma^{(b)} = \begin{bmatrix}\sigma^2 & \rho^{(b)}\sigma^2\\ \rho^{(b)}\sigma^2 & \sigma^2\end{bmatrix}.$$

Step 3.2: Use the probability limit $\overline{\xi}_{\mathrm{T},h_T}$ to define the cutoff $e_{\mathrm{T},k} = \Phi^{-1}(1 - \overline{\xi}_{\mathrm{T},k}|\mu_\mathrm{T}, \sigma^2)$ for each $k = 1, \ldots, K_\mathrm{T} - 1$ in the definition of Y_T in terms of $\tilde{Y}_\mathrm{T}$, and the cutoffs in (7). Similarly, use $\underline{\xi}_{\mathrm{R},h_R}$ to define $e_{\mathrm{R},k} = \Phi^{-1}(1 - \underline{\xi}_{\mathrm{R},k}|\mu_\mathrm{R}, \sigma^2)$ for $k = 1, \ldots, K_\mathrm{R} - 1$. Set $u_{j,0} = -\infty$ and $u_{j,K_j} = \infty$ for $j = T, R$. That is, given $\boldsymbol{\mu}$ and σ^2, the cutoffs $\{e_{\mathrm{R},k}\}$ are determined such that the marginals $\pi_{\mathrm{T},h_T} = \overline{\xi}_{\mathrm{T},h_T}$ and $\pi_{\mathrm{R},h_R} = \underline{\xi}_{\mathrm{R},h_R}$.

Step 3.3: For each $\rho^{(b)}$ in the grid, $b = 1, \ldots, B$, compute

$$\pi^{(b)}_{y_T,y_R} = \mathrm{P}(\tilde{\boldsymbol{y}} \in A_{y_T,y_R}) = \iint_{A_{y_T,y_R}} f_\mathrm{N}(\tilde{\boldsymbol{y}}|\boldsymbol{\mu}, \Sigma^{(b)})\, d\tilde{y}_\mathrm{T}\, d\tilde{y}_\mathrm{R},$$

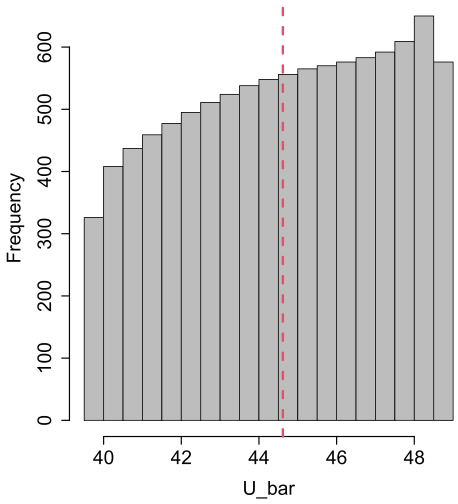

FIGURE 1 | Distribution of $\overline{U}^{(b)} = \overline{U}(\boldsymbol{\pi}^{(b)})$ based on the numerical utilities in Table 2, for $B = 10{,}000$. The red vertical dashed line represents the calibrated value $\underline{U} = 44.62$.

for all combinations of $y_T = 0, \ldots, K_T - 1$ and $y_R = 0, \ldots, K_R - 1$.

We use the numerical utilities given in Table 2, with the upper probability limits $\overline{\xi}_{T,1} = 0.50$, $\overline{\xi}_{T,2} = 0.30$, $\overline{\xi}_{T,3} = 0.10$ for the three toxicity severities and the lower probability limits $\underline{\xi}_{R,1} = 0.50$, $\underline{\xi}_{R,2} = 0.40$, $\underline{\xi}_{R,3} = 0.30$ for the three levels of response. Figure 1 gives a histogram of a sample of $B = 10,000$ values of $\overline{U}^{(b)} = \overline{U}(\boldsymbol{\pi}^{(b)})$, obtained from the equi-spaced grid of $\{\rho^{(b)}\}$ values on the interval [-0.999, 0.999]. In our illustration, we fix $\boldsymbol{\mu} = [\mu_T, \mu_R]' = [0, 0]'$ and $\sigma = 4$. The figure illustrates how a large sample of $\boldsymbol{\pi}^{(b)}$ vectors with the same marginals produces a sample of $\overline{U}^{(b)}$ values ranging from 39.54 to 48.96, due to the different degrees of association between the outcomes induced by the different $\rho^{(b)}$ values. The distribution of $\overline{U}^{(b)}$ is asymmetric because the toxicity and response outcomes are penalized differently by the numerical utility values in Table 2. The calibrated value $\underline{U} = 1/B\sum_{b=1}^{B}\overline{U}^{(b)} = 44.62$ is represented by the red vertical line in Figure 1. If desired, an alternative statistic, such as the median or some other quantile, can be used to determine $\underline{U}$. For example, if the upper quartile is used for $\underline{U}$, the design would make more conservative decisions by ensuring that the expected utility of E is smaller than the upper quartile with a probability of no more than $c_U(t)$. Moreover, rather than assuming equiprobable $\rho^{(b)}$ values, a probability distribution that reflects clinicians' beliefs about the outcomes' association can be employed. Finally, other probability models, such as a cumulative logit model, may be used to generate the sample $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(B)}$.

It is useful to consider how this construction works in the special case, discussed above, where both outcomes are binary. In this case, $Z_{j,1} = Y_j$ and $\xi_{j,1} = \Pr(Y_j = 1)$ for each $j = T, R$, giving $\boldsymbol{\pi} = (\pi_{0,0}, \pi_{0,1}, \pi_{1,0}, \pi_{1,1})$ and $U(\boldsymbol{y})$ for $y_T, y_R \in \{0, 1\}$. In this simple case, U-Bayes still provides a very useful alternative to using the two marginal rules (1) with $h_T = h_R = 1$. Recall the example in Section 1 with two joint distributions $\boldsymbol{\pi}_1 = (0.6, 0.1, 0.0, 0.3)$ and $\boldsymbol{\pi}_2 = (0.3, 0.4, 0.3, 0.0)$ that have the same marginal probabilities, $P(Y_T = 1) = 0.3$ and $P(Y_R = 1) = 0.4$ but very different association between the outcomes. The elicited utilities, $U(\boldsymbol{y}) = 60$, 100, 0, and 70 for $\boldsymbol{y} = (0, 0)$, (0, 1), (1, 0), and (1, 1), respectively, give mean utilities $\overline{U}(\boldsymbol{\pi}_1) = 67$ and $\overline{U}(\boldsymbol{\pi}_2) = 58$, which differ substantially despite the identical marginals. As a result, because the U-Bayes stopping criterion uses $\overline{U}(\boldsymbol{\pi})$ to incorporate between-outcome association through $\boldsymbol{\pi}$ and preferences of the outcomes through $U(\boldsymbol{y})$, it is more likely to conclude that E is acceptable for $\boldsymbol{\pi}_1$ than for $\boldsymbol{\pi}_2$. In contrast, the two marginal probability-based rules in (1), which ignore the association between Y_T and Y_R, yield the same conclusion for $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$.

3.4 | Inference Model

A probability model for $\boldsymbol{Y}$ and a prior for $\boldsymbol{\pi}$ must be specified in order to compute the posterior stopping criteria (1) or (3) used by the two methods during trial conduct. Any flexible models that yield a posterior distribution of $\boldsymbol{\pi}$ can be employed. We will assume an ordinal probit model to define a joint probability distribution of $\boldsymbol{Y}$, similar to the model used for calibrating $\underline{U}$ as part of the trial planning process given above. However, in contrast with its use to calibrate $\underline{U}$, this model is employed to make inferences about $\boldsymbol{\pi}$ using the observed data from a phase II trial.

Let $n(t)$ denote the number of patients accrued up to trial time t, and index patients by $i = 1, \ldots, n(t)$. For the i^{th} patient, denote the outcomes by $\boldsymbol{Y}_i = (Y_{i,T}, Y_{i,R})$. Let $\widetilde{\boldsymbol{Y}}_i = (\widetilde{Y}_{i,T}, \widetilde{Y}_{i,R}) \in \mathbb{R}^2$ denote a pair of latent real valued probit scores for the ith patient that follow a bivariate normal distribution,

$$\widetilde{\boldsymbol{Y}}_i | \boldsymbol{\mu}, \Sigma \overset{iid}{\sim} N_2(\boldsymbol{\mu}, \Sigma), \quad i = 1, \ldots, n(t), \tag{10}$$

where $\boldsymbol{\mu}$ and Σ are specified in (6). From (7)-(9), we assume

$$\begin{aligned} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\pi}) &= \pi_{y_{i,T}, y_{i,R}} \\ &= P(\widetilde{\boldsymbol{Y}}_i \in A_{y_{i,T}, y_{i,R}} | \boldsymbol{\mu}, \Sigma, e_{j,k}) \end{aligned}$$

where $A_{y_{i,T}, y_{i,R}}$ is defined as in (8) for $(y_{i,T}, y_{i,R})$. We assume that $\boldsymbol{\mu} \in \mathbb{R}^2$ and $\rho \in (-1, 1)$ are random. For the cutoffs $e_{j,0} < e_{j,1} < \cdots < e_{j,K_j}$, we set $e_{j,0} = -\infty$ and $e_{j,K_j} = \infty$ and allow $e_{j,2}, \ldots, e_{j,K_j-1}$ to be random for flexibility. For identifiability, we fixed $e_{j,1} = 0$ and $\sigma^2 = 25$ for our simulation studies.

To specify priors for the parameters $\boldsymbol{\mu}$ and ρ and random cut-offs $e_{j,2}, \ldots, e_{j,K_j-1}$, we first assume $\mu_j \overset{indep}{\sim} N(\overline{\mu}_j, \omega_j^2)$, with $\overline{\mu}_j$ and ω_j^2 fixed for each j. For $\rho \in (-1, 1)$, we assume $(\rho + 1)/2 \sim Be(a_\rho, b_\rho)$ with a_ρ and b_ρ fixed. For the cutoff distribution, denoting $\lambda_{j,k} = e_{j,k+1} - e_{j,k}$, $k = 1, \ldots, K_j - 2$, we assume that $\lambda_{j,k} \overset{indep}{\sim} Exp(1/\eta_j)$ with $E(\lambda_{j,k}) = \eta_j$, with η_j fixed. The inferential model includes random $\boldsymbol{\mu}$ and ρ, along with random cutoffs $e_{j,k}$, and can flexibly approximate any $\boldsymbol{\pi}$. The model facilitates incorporation of useful information into inferences through the prior distribution and generally is efficient, provided that the prior distribution is reasonable. Specifically, it obtains desirable frequentist properties such as posterior consistency, if the prior distribution assigns a nonzero probability to the true parameter values [28].

Collecting terms, denote the vector of all random model parameters by $\theta = (\{\mu_j\}, \rho, \{\lambda_{j,k}\})$, and the vector of fixed hyperparameter by $\tilde{\theta} = (\{\overline{\mu}_j\}, \{\omega_j^2\}, a_\rho, b_\rho, \{\eta_j\})$. The joint posterior distribution of the latent variables and parameters under the augmented model is

$$p(\tilde{\mathbf{y}}, \theta | \mathcal{D}_{n(t)}, \tilde{\theta}) \propto \prod_{i=1}^{n(t)} \left\{ f_{\mathrm{N}}(\tilde{\mathbf{y}}_i | \mu, \Sigma) \mathbf{1}(\tilde{\mathbf{y}}_i \in A_{y_{i,T}, y_{i,R}}) \right\} \times \prod_{j \in \{T,R\}} p(\mu_j | \overline{\mu}_j, \omega_j^2) \prod_{j \in \{T,R\}} \prod_{k=1}^{K_j - 2} p(\lambda_{j,k} | \eta_j)\ p(\rho | a_\rho, b_\rho),$$

denoting the data at trial time t by $\mathcal{D}_{n(t)} = \{\mathbf{y}_i, i = 1, \ldots, n(t)\}$. We use Markov chain Monte Carlo (MCMC) simulation to generate posterior samples of θ by iteratively drawing $(\mu_j, \rho, \lambda_{j,k})$, with each conditional on the values of the others at each iteration. Given the posterior distribution of θ, the posterior of $\boldsymbol{\pi}$ can be obtained easily using (9). Section 1 of the Supporting Information includes details of the posterior simulation, and provides an explanation on how to evaluate (3) using a posterior sample of θ. We specify the hyperparameters, $\tilde{\theta}$, so that the resulting priors are weakly informative. Specifically, for the simulation studies presented in Section 4 below, we set $(a_\rho, b_\rho) = (0.5, 0.5)$, $\overline{\mu}_j = -2.5$, $\omega_j^2 = 100$, and $\eta_j = 5$ for both $j = T$ and R. We performed prior sensitivity analyses by varying the hyperparameter values $\tilde{\theta}$, and observed only minimal changes in each design's performance for any change in $\tilde{\theta}$ within a reasonable range of values.

3.5 | Trial Conduct

Assume that the phase II design includes L interim analyses after successive cohorts of size $\lfloor \frac{1}{L+1} N_{\max} \rfloor$, with an accumulated sample size of $n_\ell = \lfloor \frac{\ell}{L+1} N_{\max} \rfloor$ at the ℓ^{th} analysis for $\ell = 1, \ldots, L$, and $n_{L+1} = N_{\max}$. The trial may be conducted as follows:

Trial Conduct

1. Treat the first cohort and observe their outcomes, $\boldsymbol{Y}_i$, $i = 1, \ldots, \lfloor \frac{1}{L+1} N_{\max} \rfloor$.
2. At each interim analysis $\ell = 1, \ldots, L$, compute the posterior distribution of $\boldsymbol{\pi}$ based on the current data $\mathcal{D}_{n(t)}$. Compute the posterior criterion (3) to decide whether E is acceptable. If E is determined to be acceptable, treat the next cohort and observe their outcomes. Otherwise, the trial is stopped early and E is declared unacceptable.
3. When $n(t) = N_{\max}$ is reached, the final decision on the acceptability of E is made by computing (3) using the final data $\mathcal{D}_{N_{\max}}$.

In practice, as with any phase II trial, $N_{\max}$ is limited primarily by resource availability. However, $N_{\max}$ must be large enough to ensure that, across a range of scenarios determined by values of $\boldsymbol{\pi}^{\mathrm{TR}}$, the early stopping probability (i) is large for $\overline{U}(\boldsymbol{\pi}^{\mathrm{TR}}) \leq \underline{U}$, and (ii) is small for $\overline{U}(\boldsymbol{\pi}^{\mathrm{TR}}) \geq \underline{U} + 10$. The maximum number of applications of the stopping rule, L, should be logistically feasible, while ensuring that the trial is monitored with sufficient frequency to ensure the above design properties. Given $N_{\max}$ and L, preliminary simulations should be done to calibrate the cutoff c_U in (3). In our simulation study, we used $N_{\max} = 60$ and $L = 3$. We also have conducted additional simulation studies in which both $N_{\max}$ and L are varied, studying $N_{\max} = 60, 90$, or 120 with various values of L.

A computer program "U-Bayes" for implementing the proposed U-Bayes design is available from https://sites.google.com/ucsc.edu/juheelee/software. The program also includes a function that implements the method for calibrating $\underline{U}$ when planning a trial.

4 | Simulation Study

4.1 | Simulation Design

To evaluate the U-Bayes design's performance and compare it to designs that use two marginal probability based rules, we simulated the trial under 10 scenarios. For each scenario, we assumed four-level ordinal outcomes for toxicity and efficacy, so $K_{\mathrm{T}} = K_{\mathrm{R}} = 4$. Each simulated trial had $N_{\max} = 60$ with cohort size 15, resulting in up to 3 interim looks at $n(t) = 15$, 30, and 45 (i.e., $L = 3$). For each scenario, we specified the true covariance matrix Σ^{TR} of the probit scores and true marginal probabilities $p_{j,k_j}^{\mathrm{TR}} = \mathrm{P}^{\mathrm{TR}}(Y_j = k_j)$, $j = T$ and R and $k_j = 0$, 1, 2, 3. We fixed $e_{j,0}^{\mathrm{TR}} = -\infty$, $e_{j,1}^{\mathrm{TR}} = 0$ and $e_{j,4}^{\mathrm{TR}} = \infty$. Using the marginal probability $p_{j,0}^{\mathrm{TR}}$, we determined μ_j^{TR} and $e_{j,k}^{\mathrm{TR}}$, for $k = 2, 3$ and $j = T, R$ by solving the equations

$$\mu_j^{\mathrm{TR}} = -\Phi^{-1}(p_{j,0}^{\mathrm{TR}} | 0, \sigma_j^{2,\mathrm{TR}}) \text{ and } e_{j,k}^{\mathrm{TR}} = \Phi^{-1}\left(\sum_{k'=0}^{k} p_{j,k'}^{\mathrm{TR}} | \mu_j^{\mathrm{TR}}, \sigma_j^{2,\mathrm{TR}} \right).$$

We set $\sigma_{\mathrm{T}}^{2,\mathrm{TR}} = \sigma_{\mathrm{R}}^{2,\mathrm{TR}} = 16$ for all scenarios. We then computed $\pi_{y_T, y_R}^{\mathrm{TR}}$ for all $(y_{\mathrm{T}}, y_{\mathrm{R}})$ pairs using (9). True values of the marginal probabilities $p_{j,k}^{\mathrm{TR}}$, joint probabilities $\boldsymbol{\pi}^{\mathrm{TR}}$, correlation ρ^{TR}, and mean utility $\overline{U}^{\mathrm{TR}}$ are given in Table 3. In the table, $p_{\mathrm{T},h_T}^{\mathrm{TR}}$ is given in red italics if $\xi_{\mathrm{T},h_T}^{\mathrm{TR}} > \overline{\xi}_{\mathrm{T},h_T}$, that is, if E is truly unacceptable due to excessive toxicity. Similarly, $p_{\mathrm{R},h_R}^{\mathrm{TR}} < \underline{\xi}_{\mathrm{R},h_R}$ is given in red italics, that is, if E is truly unacceptable due to a low response rate. Finally, true mean utility values $\overline{U}^{\mathrm{TR}} < \underline{U} = 44.62$ are marked in red italics since they are unacceptably low. Values of the true marginal probabilities $\mathbf{p}^{\mathrm{TR}}$ were specified arbitrarily to examine the robustness of the U-Bayes design.

As comparators, we used two conventional designs, each based on a pair of safety and efficacy monitoring rules given by (1), but defining binary toxicity and binary response using different cutoffs:

- **Prob-based I**: This design monitors occurrences of [Severe Toxicity] and CR, that is, $\boldsymbol{h} = (h_{\mathrm{T}}, h_{\mathrm{R}}) = (3, 3)$. The fixed limits $\overline{\xi}_{\mathrm{T},3} = 0.10$ and $\underline{\xi}_{\mathrm{R},3} = 0.30$ are used for the probability cutoffs with $\xi_{\mathrm{T},3}$ and $\xi_{\mathrm{R},3}$ compared to these limits in (1).
- **Prob-based II**: This design monitors occurrences of (High or Severe Toxicity) and (PR or CR), that is, $\boldsymbol{h} = (2, 2)$. The fixed limits $\overline{\xi}_{\mathrm{T},2} = 0.30$ and $\underline{\xi}_{\mathrm{R},2} = 0.40$ are used for the probability cutoffs with $\xi_{\mathrm{T},2}$ and $\xi_{\mathrm{R},2}$ compared to these limits in (1).

TABLE 3 | Simulation results.

	Scenario 1 ($\rho^{TR} = 0.0$)					Scenario 2 ($\rho^{TR} = 0.5$)				
Toxicity	**Response**					**Response**				
Severity	**PD**	**SD**	**PR**	**CR**		**PD**	**SD**	**PR**	**CR**	
Low	0.04	0.35	0.03	0.28	0.70	0.04	0.00	0.01	0.00	0.05
Moderate	0.00	0.02	0.01	0.02	0.05	0.40	0.04	0.24	0.02	0.70
High	0.01	0.10	0.01	0.08	0.20	0.02	0.00	0.03	0.00	0.05
Severe	0.00	0.03	0.00	0.02	0.05	0.04	0.01	0.12	0.03	*0.20*
	0.05	0.50	0.05	0.40	1.00	0.50	0.05	0.40	*0.05*	1.00

	$\overline{U}^{TR} = 68.93$		$\overline{U}^{TR} =$ *31.26*	
	p^{acc}	n^{trt}	p^{acc}	n^{trt}
U-Bayes	1.00	4.00	*0.00*	*2.10*
Prob-based I	0.98	3.94	*0.00*	*1.10*
Prob-based II	0.96	3.87	0.92	3.76

	Scenario 3 ($\rho^{TR} = -0.5$)					Scenario 4 ($\rho^{TR} = -0.5$)				
Toxicity	**Response**					**Response**				
Severity	**PD**	**SD**	**PR**	**CR**		**PD**	**SD**	**PR**	**CR**	
Low	0.08	0.09	0.08	0.25	0.50	0.11	0.21	0.11	0.17	0.60
Moderate	0.05	0.03	0.02	0.04	0.15	0.08	0.08	0.03	0.02	0.20
High	0.08	0.05	0.03	0.04	*0.20*	0.10	0.06	0.01	0.01	0.18
Severe	0.09	0.03	0.01	0.01	*0.15*	0.02	0.00	0.00	0.00	0.02
	0.30	0.20	0.15	0.35	1.00	0.30	0.35	*0.15*	*0.20*	1.00

	$\overline{U}^{TR} = 54.96$		$\overline{U}^{TR} = 56.08$	
	p^{acc}	n^{trt}	p^{acc}	n^{trt}
U-Bayes	1.00	3.99	1.00	3.99
Prob-based I	*0.31*	*3.36*	*0.07*	*2.50*
Prob-based II	*0.02*	*2.68*	*0.83*	*3.59*

	Scenario 5 ($\rho^{TR} = -0.3$)					Scenario 6 ($\rho^{TR} = -0.3$)				
Toxicity	**Response**					**Response**				
Severity	**PD**	**SD**	**PR**	**CR**		**PD**	**SD**	**PR**	**CR**	
Low	0.11	0.09	0.16	0.13	0.50	0.11	0.09	0.05	0.25	0.50
Moderate	0.08	0.05	0.07	0.04	0.25	0.03	0.02	0.01	0.04	0.10
High	0.08	0.04	0.05	0.02	0.20	0.15	0.08	0.04	0.11	*0.38*
Severe	0.03	0.01	0.01	0.00	0.05	0.01	0.00	0.00	0.00	0.02
	0.30	0.20	0.30	*0.20*	1.00	0.30	0.20	0.10	0.40	1.00

	$\overline{U}^{TR} = 55.28$		$\overline{U}^{TR} = 55.63$	
	p^{acc}	n^{trt}	p^{acc}	n^{trt}
U-Bayes	0.92	3.84	1.00	4.00
Prob-based I	*0.02*	*2.03*	0.97	3.94
Prob-based II	0.64	3.27	*0.03*	*3.52*

(Continues)

TABLE 3 | (Continued)

Toxicity / Severity	Scenario 7 ($\rho^{TR} = 0.8$) Response PD	SD	PR	CR		Scenario 8 ($\rho^{TR} = 0.8$) Response PD	SD	PR	CR	
Low	0.05	0.00	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.05
Moderate	0.43	0.04	0.14	0.08	0.70	0.44	0.03	0.03	0.09	0.60
High	0.01	0.00	0.03	0.05	0.10	0.06	0.02	0.02	0.24	*0.33*
Severe	0.00	0.00	0.02	0.12	*0.15*	0.00	0.00	0.00	0.02	0.02
	0.50	0.05	0.20	*0.25*	1.00	0.55	0.05	*0.05*	0.35	1.00

	$\overline{U}^{TR} = 33.66$		$\overline{U}^{TR} = 34.24$	
	p^{acc}	n^{trt}	p^{acc}	n^{trt}
U-Bayes	*0.00*	*2.04*	*0.00*	*2.32*
Prob-based I	*0.00*	*2.14*	0.95	3.86
Prob-based II	0.94	3.82	*0.71*	*3.23*

Toxicity / Severity	Scenario 9 ($\rho^{TR} = 0.0$) Response PD	SD	PR	CR		Scenario 10 ($\rho^{TR} = 0.0$) Response PD	SD	PR	CR	
Low	0.14	0.06	0.06	0.14	0.40	0.24	0.02	0.06	0.08	0.40
Moderate	0.14	0.06	0.06	0.14	0.40	0.15	0.01	0.04	0.05	0.25
High	0.05	0.02	0.02	0.05	0.15	0.09	0.01	0.02	0.03	*0.15*
Severe	0.02	0.01	0.01	0.02	0.05	0.12	0.01	0.03	0.04	*0.20*
	0.35	0.15	0.15	0.35	1.00	0.60	0.05	*0.15*	*0.20*	1.00

	$\overline{U}^{TR} = 54.04$		$\overline{U}^{TR} = 35.33$	
	p^{acc}	n^{trt}	p^{acc}	n^{trt}
U-Bayes	0.98	3.95	*0.00*	*2.29*
Prob-based I	0.95	3.87	*0.00*	*1.63*
Prob-based II	0.97	3.94	*0.00*	*2.36*

Note: p^{acc}=P(declare E acceptable), and n^{trt} = mean number of cohorts treated. U-Bayes uses the utility function in Table 2 with $\underline{U} = \mathbf{44.62}$. Prob-based I uses $\boldsymbol{h} = (3, 3)$ with $\overline{\pi}_{T,3} = 0.1$ and $\underline{\pi}_{R,3} = 0.3$. Prob-based II uses $\boldsymbol{h} = (2, 2)$ with $\overline{\pi}_{T,2} = 0.3$ and $\underline{\pi}_{R,2} = 0.4$. Values for *truly unacceptable* E are given in *red italics*.

The same inferential model given in Section 3.4 for ordinal outcomes is used for all three designs.

We evaluated and compared the three designs using the following two criteria:

p^{acc} = probability of declaring treatment E acceptable

n^{trt} = mean number of cohorts treated

In particular, $p^{acc} = 1-$ Pr(stop the trial early). Index the simulated trials under each design by $r = 1, \ldots, R$. For the rth trial, let $w^{(r)} = 1$ if a treatment is identified as acceptable and 0 if not, and let $N^{(r)}$ be the total number of cohorts treated. For each scenario and design, we summarized the simulation results using the following sample proportions:

$$p^{acc} = \frac{1}{R}\sum_{r=1}^{R} w^{(r)} \quad \text{and} \quad n^{trt} = \frac{1}{R}\sum_{r=1}^{R} N^{(r)}.$$

4.2 | Simulation Results

A total of $R = 1000$ trials with $N_{max} = 60$ and $L = 3$ interim decisions after cohorts of size 10 were simulated using each design under each scenario. The value 0.85 was used for all three cutoffs $c_T^\star$, $c_R^\star$, and $c_U^\star$. The simulation results are summarized in Table 3.

Compared to the two designs using a pair of marginal probability-based monitoring rules, across the 10 scenarios, on average U-Bayes is more likely to make much more reasonable decisions. When E is truly unacceptable, U-Bayes reliably determines this and stops the trial early with high probability, resulting

in fewer patients being treated. In general, U-Bayes often disagrees with the two-rule marginal probability based designs, and these often disagree with each other due to their use of different cutoffs to define binary outcomes.

Scenarios 1 and 2, which were discussed above in Section 2 as illustrative examples, have the same marginal distributions for the dichotomized outcomes $Z_{T,2}$ and $Z_{R,2}$, but the joint distributions $\boldsymbol{\pi}$ are very different and give the respective mean utilities, $\overline{U}^{TR} = 68.98$ and 31.26. For example, $\pi^{TR}_{0,3} = 0.28$ (low toxicity and CR) and $\pi^{TR}_{0,1} = 0.35$ (low toxicity and SD) under Scenario 1, but $\pi^{TR}_{1,0} = 0.40$ (moderate toxicity and PD) and $\pi^{TR}_{1,2} = 0.24$ (moderate toxicity and PR) under Scenario 2. While E is truly acceptable for both scenarios under Prob-based II, which is based on the marginal probabilities $\xi_{T,2}$ and $\xi_{R,2}$, under U-Bayes E is truly acceptable for Scenario 1 but not for Scenario 2. Prob-based II had $p^{acc} = 0.96$ for Scenario 1 and $p^{acc} = 0.92$ for Scenario 2, with averages of 3.81 and 3.76 cohorts treated, respectively. In sharp contrast, since U-Bayes accounts for the entire joint probability $\boldsymbol{\pi}$, it identified E as acceptable for all 1000 simulated trials in Scenario 1, giving $p^{acc} = 1.00$, but had $p^{acc} = 0.00$ in Scenario 2. In Scenario 1, U-Bayes always treated 4 cohorts, while in Scenario 2, an average of 2.10 cohorts were treated. Under Prob-based I, E is truly acceptable for Scenario 1, but not acceptable for Scenario 2. This design performs reasonably well for both scenarios. However, Prob-based I and Prob-based II have very different decision probabilities under Scenario 2, with $p^{acc} = 0.00$ for Prob-based I and $p^{acc} = 0.92$ for Prob-based II, due to their different dichotomization values $\boldsymbol{h} = (2,2)$ and (3,3). This sort of disagreement between Prob-based I and II is seen in many of the scenarios, which underscores the importance of how each ordinal variable is dichotomized.

In Scenarios 3 and 4, E is truly acceptable under U-Bayes, with nearly identical respective true mean utilities 54.96 and 54.08, but E is not acceptable under either of the marginal probability-based rules. Scenario 3 has $p^{TR}_{T,3} = 0.15$ and $p^{TR}_{T,2} = 0.20$ giving $\xi^{TR}_{T,3} = 0.15$ and $\xi^{TR}_{T,2} = 0.35$. Comparing these probabilities to the limits, $\overline{\xi}_{T,3} = 0.10$ and $\overline{\xi}_{T,2} = 0.30$, shows that E is truly unacceptable under both probability-based designs. Although the true toxicity probabilities slightly exceed their upper limits, E has a large mean utility due to a substantial increase in the response probability, and it is truly acceptable under U-Bayes. The desirable outcome (low or moderate toxicity, CR or PR) occurs with probability of 0.39, and the more desirable subevent (low toxicity, CR or PR) has probability 0.29. The U-Bayes design has $p^{acc} = 1.00$, and on average treats 3.99 cohorts. In contrast, Prob-based I has $p^{acc} = 0.31$ and Prob-based II has $p^{acc} = 0.02$, treating respective averages of 3.36 and 2.68 cohorts. Again, U-Bayes is much more likely to find E acceptable in scenarios where, because it accounts for the desirabilities of joint events, $\overline{U}(\boldsymbol{\pi})$ is large while, when considering only the marginals under Prob-based I or II, either $\pi_{T,h}$ is too large or $\pi_{R,h}$ is too small.

In Scenario 4, there is a significant reduction in toxicity probability at the cost of response probabilities being slightly smaller than the limits. Specifically, $\xi^{TR}_{R,3} = 0.20$ and $\xi^{TR}_{R,2} = 0.35$ with lower limits $\underline{\xi}_{R,3} = 0.30$, and $\underline{\xi}_{R,2} = 0.40$. The probability of Low toxicity occurring is very large, $p^{TR}_{T,0} = 0.60$, and Low toxicity occurs with a response above or equal to SD with probability of 0.49. Consequently, E is unacceptable under both marginal probability-based rules but is acceptable under U-Bayes since $\overline{U}^{TR} = 54.96$, which is much larger than $\underline{U} = 44.62$, so U-Bayes has $p^{acc} = 1.00$ and treats on average 3.99 cohorts. The two probability-based methods have $p^{acc} = 0.07$ and 0.83, respectively, again disagreeing with each other and with U-Bayes.

In Scenarios 5 and 6, E has nearly identical high true mean utilities 55.28 and 55.63, respectively, but E is unacceptable in terms of marginal probabilities. In Scenario 5, Prob-based I with $\boldsymbol{h} = (3,3)$ always identifies E as unacceptable, with $p^{acc} = 0.02$, because $\xi^{TR}_{R,3} = 0.20 < 0.30 = \underline{\xi}_{R,3}$. However, Scenario 5 also has $\xi^{TR}_{R,2} > \underline{\xi}_{R,2}$ and $\xi^{TR}_{T,2} < \overline{\xi}_{T,2}$, so Prob-based II has $p^{acc} = 0.64$. Since the true mean utility $\overline{U}^{TR} = 55.28 > 44.62 = \underline{U}$, U-Bayes has $p^{acc} = 0.92$. In Scenario 6, E is truly acceptable under Prob-based I but not acceptable under Prob-based II because $\xi^{TR}_{T,2} = 0.40 > 0.30 = \overline{\xi}_{T,2}$. For this scenario, U-Bayes has $p^{acc} = 1.00$, while Prob-based I has $p^{acc} = 0.97$ and Prob-based II has $p^{acc} = 0.03$. Thus, Prob-based I happens to agree with U-Bayes, but disagrees with Prob-based II.

In Scenarios 7 and 8, E has small mean utilities due to the large values of the probability of (moderate toxicity, PD), specifically $\pi^{TR}_{1,0} = 0.43$ and 0.44 for the two scenarios, respectively. In Scenario 7, the marginal probability of (CR or PR) is 0.45, while the probability of the joint event (high or severe toxicity, CR or PR) is 0.22. This results in an unacceptably small mean utility $\overline{U}^{TR} = 33.66$, and U-Bayes has $p^{acc} = 0.00$. In Scenario 7, under Prob-based I E is truly unacceptable and this design had $p^{acc} = 0.00$ after treating 2.14 cohorts on average. In contrast, Prob-based II has $p^{acc} = 0.94$, with an average of 3.82 cohorts treated.

Scenario 8 has marginal probability 0.40 for (CR or PR). Given (CR or PR), (High or Severe toxicity) occurs with conditional probability of 0.725 = 0.29/0.40. Consequently, the true mean utility $\overline{U}^{TR} = 34.24$ is unacceptably small and U-Bayes has $p^{acc} = 0.00$. However, E is truly acceptable under Prob-based I, which has $p^{acc} = 0.95$, with an average of 3.86 cohorts treated while E is truly unacceptable under Prob-based II, which has $p^{acc} = 0.71$.

In Scenario 9, the three designs all have high p^{acc} values, with U-Bayes having the highest value 0.98, and 3.95 out of 4 cohorts treated. In Scenario 10, E was not identified as acceptable in any of the simulated trials by any design, with $p^{acc} = 0.00$, and the trials were stopped early after 1.63 to 2.36 cohorts were treated.

Additionally, we varied the number L of interim looks and the maximum sample size N_{max} to examine how the performances of the three designs may change with different combinations of L and N_{max}. We let $L = 1$ or 2 with $N_{max} = 60$, $L = 1$, 2, or 5 with $N_{max} = 90$ and $L = 1$, 3 or 5 with $N_{max} = 120$. The results are summarized in Table 1 in the Supporting Information. Overall, the performance improves with smaller L and larger N_{max} for all three designs.

TABLE 4 | Simulation results of U-Bayes with different statistical models.

	Scenario 11 ($\rho^{TR} = -0.9$)					Scenario 15 ($\rho^{TR} = 0.9$)				
Toxicity	**Response**					**Response**				
Severity	**PD**	**SD**	**PR**	**CR**		**PD**	**SD**	**PR**	**CR**	
Low	0.01	0.04	0.02	0.33	0.40	0.33	0.05	0.01	0.01	0.40
Moderate	0.06	0.06	0.02	0.06	0.20	0.06	0.06	0.02	0.06	0.20
High	0.10	0.04	0.01	0.01	0.15	0.01	0.03	0.02	0.10	0.15
Severe	0.23	0.01	0.00	0.00	0.25	0.00	0.01	0.01	0.23	0.25
	0.40	0.15	0.05	0.40	1.00	0.40	0.15	0.05	0.40	1.00
	$\overline{U}^{TR} = 50.91$					$\overline{U}^{TR} = 38.56$				
	p^{acc}		n^{trt}			p^{acc}		n^{trt}		
U-Bayes with Dep	0.92		3.85			*0.02*		*3.03*		
U-Bayes with Ind	0.32		2.65			*1.00*		*4.00*		

Note: p^{acc} = P(declare E acceptable), and n^{trt} = mean number of cohorts treated. U-Bayes with two different statistical models are compared; (1) U-Bayes with Dep, U-Bayes with a model that assumes dependence between the outcomes and (2) U-Bayes with Ind: U-Bayes with a model that assumes independence between the outcomes. The utility function in Table 2 with $\underline{U}$ = **44.62** is used for both. Values for *truly unacceptable E* are given in *red italics*.

We considered five additional simulation scenarios, Scenarios 11–15, to examine how the performance of U-Bayes may change if the assumed inferential model does not account for potential dependence between the outcomes. These scenarios all have the same marginal probabilities, but the joint probability distributions differ in terms of their degrees of association. Specifically, we considered $\rho^{TR} = -0.9, -0.6, 0.0, 0.6$, or 0.9, respectively, for the five scenarios. Under the assumed simulation design in Section 4.1, a large negative value of ρ^{TR} leads to larger $\overline{U}^{TR}$, and $\rho^{TR} = 0.0$ represents the case of independence between the outcomes. The scenarios have respective $\overline{U}^{TR} = 50.91$, 48.54, 44.70, 40.88, and 38.56. Recall that $\underline{U} = 44.62$, which implies that E is acceptable when $\rho^{TR} = -0.9$ or -0.6 but unacceptable when $\rho^{TR} = 0.6$ or 0.9, while $\overline{U}^{TR}$ is very close to $\underline{U}$ when $\rho^{TR} = 0.0$. We call the model that assumes independence by fixing $\rho = 0.0$ 'U-Bayes with Ind,' and call U-Bayes with a model that assumes dependence by treating ρ as an unknown parameter 'U-Bayes with Dep.' The simulation results under Scenarios 11–15 are summarized in Table 2 in the Supporting Information. We illustrate the results for Scenarios 11 and 15 in Table 4. Although E is truly acceptable in Scenario 11 but truly unacceptable in Scenario 15, U-Bayes with Indep identifies E as acceptable with probabilities $p^{acc} = 0.32$ and 1.00, respectively, with $n^{trt} = 2.68$ and 4.00 cohorts treated on average for the two scenarios. In contrast, U-Bayes with Dep identifies E as acceptable with probabilities $p^{acc} = 0.92$ and 0.02, respectively. This shows that, by allowing the inferential model to learn potential dependencies between outcomes, U-Bayes achieves greater accuracy in decision-making. Table 3 in Supporting Information illustrates the results for different combinations of (L, N_{max}) in Scenarios 11–15. Consistent with earlier findings, U-Bayes with Dep performs better for smaller L and/or larger N_{max}. In contrast, the performance of U-Bayes with Indep does not improve with larger N_{max}.

5 | Discussion

We have proposed a new Bayesian single-arm phase II trial design, U-Bayes, that monitors a treatment's acceptability using one stopping rule based on a utility function for ordinal toxicity and response. Our simulations showed that U-Bayes is highly effective in evaluating a treatment acceptability, because it accounts for the joint distribution of Toxicity and Response through their mean utility. The simulations showed that the two marginal probability-based methods each may have very undesirable OCs because they reduce information by dichotomizing the two ordinal outcomes for decision making. Moreover, Prob-based I and Prob-based II may strongly disagree with each other because they use different cut-offs to define "toxicty" and "response." That is, whether E is found to be promising or not greatly depends on the choice of the cut-points used to determine the marginal events. U-Bayes does away with these problems by using the full bivariate distribution over the observed ordinal outcomes, and consequently it is likely to lead to more clinically justified decisions in practice.

Utility functions are inherently subjective because they reflect the desirability of each pair of outcomes from the clinical investigator's viewpoint, which reflects overall patient benefit numerically. Therefore, elicitation of $U(\mathbf{y})$ and calibration of the threshold $\overline{U}$ require close communication between the clinicians and statisticians planning a trial. This subjectivity is an advantage of U-Bayes, rather than a disadvantage, since efficacy-toxicity trade-offs are intrinsic to clinical decision making, and specifying the $U(\boldsymbol{y})$ values makes the desirabilities of all $\boldsymbol{y}$ values explicit.

In principle, the U-Bayes approach can be applied to more complex clinical settings. For example, U-Bayes is suitable for a basket trial where a targeted therapy is evaluated in multiple diseases with a common biomarker for a target that E is designed to attack. In such settings, if desired, a different utility function

can be elicited for each disease, providing a more tailored basis for disease-specific decision-making.

Acknowledgments

Juhee Lee's research was supported by NSF grant DMS-1662427. Peter F. Thall's research was supported by NIH/NCI grants 1R01CA261978 and 5 P30 CA016672 45.

Disclosure

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. M. Chang, T. Therneau, H. Wieand, and S. Cha, "Designs for Group Sequential Phase II Clinical Trials," *Biometrics* 43 (1987): 865–874.

2. R. M. Simon, "Optimal Two-Stage Designs for Phase II Clinical Trials," *Controlled Clinical Trials* 10 (1989): 1–10.

3. P. F. Thall and R. M. Simon, "Practical Bayesian Guidelines for Phase IIB Clinical Trials," *Biometrics* 50 (1994): 337–349.

4. J. Wathen, P. F. Thall, J. D. Cook, and E. Estey, "Accounting for Patient Heterogeneity in Phase II Clinical Trials," *Statistics in Medicine* 27 (2008): 2802–2815.

5. J. Park, G. Hsu, E. Siden, K. Thorlund, and E. Mills, "An Overview of Precision Oncology Basket and Umbrella Trials for Clinicians," *CA: A Cancer Journal for Clinicians* 70 (2020): 125–137.

6. P. Thall, J. K. Wathen, B. N. Bekele, R. E. Champlin, L. H. Baker, and R. S. Benjamin, "Hierarchical Bayesian Approaches to Phase II Trials in Diseases With Multiple Subtypes," *Statistics in Medicine* 22, no. 5 (2003): 763–780.

7. K. M. Cunanan, A. Iasonos, R. Shen, C. B. Begg, and M. Gonen, "An efficient basket trial design," *Statistics in Medicine* 36 (2017): 1568–1579.

8. Y. Chu and Y. Yuan, "BLAST: Bayesian Latent Subgroup Design for Basket Trials Accounting for Patient Heterogeneity," *Journal of the Royal Statistial Society, C* 67 (2018): 723–740.

9. M. LeBlanc, C. Rankin, and J. Crowley, "Multiple Histology Phase II Trials," *Clinical Cancer Research* 15 (2009): 4256–4262.

10. P. F. Thall, R. M. Simon, and E. H. Estey, "Bayesian Sequential Monitoring Designs for Single-Arm Clinical Trials With Multiple Outcomes," *Statistics in Medicine* 14 (1995): 357–379.

11. P. F. Thall, R. M. Simon, and E. H. Estey, "New Statistical Strategy for Monitoring Safety and Efficacy in Single-Arm Clinical Trials," *Journal of Clinical Oncology* 14 (1996): 296–303.

12. H. Zhou, J. Lee, and Y. Yuan, "BOP2: Bayesian Optimal Design for Phase II Clinical Trials With Simple and Complex Endpoints," *Statistics in Medicine* 36 (2017): 3302–3314.

13. V. Sambucini, "Bayesian Predictive Monitoring With Bivariate Binary Outcomes in Phase II Clinical Trials," *Computational Statistics and Data Analysis* 132 (2019): 18–30.

14. L. Jiang, F. Yan, P. Thall, and X. Huang, "Comparing Bayesian Early Stopping Boundaries for Phase II Clinical Trials," *Pharmaceutical Statistics* 19 (2020): 928–939.

15. P. F. Thall and K. Russell, "A Strategy for Dose-Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phaes I/II Cliical Trials," *Biometrics* 54 (1998): 251–264.

16. P. F. Thall and J. D. Cook, "Dose-Finding Based on Efficacy-Toxicity Trade-Offs," *Biometrics* 60 (2004): 684–693.

17. Y. Yuan, H. Nguyen, and P. Thall, *Bayesian Designs for Phase I-II Clinical Trials* (Boca Raton, FL: CRC Press, 2016).

18. P. Mozgunov and T. Jaki, "A Flexible Design for Advanced Phase I/II Clinical Trials With Continuous Efficacy Endpoints," *Biometrical Journal* 61 (2019): 1477–1492.

19. J. Lee, P. F. Thall, and P. Msaouel, "Precision Bayesian Phase I-II Dose-Finding Based on Utilities Tailored to Prognostic Subgroups," *Statistics in Medicine* 40, no. 24 (2021): 5199–5217.

20. B. Guo, Y. Zang, L. H. Lin, and R. Zhang, "A Bayesian Phase I/II Design to Determine Subgroup-Specific Optimal Dose for Immunotherapy Sequentially Combined With Radiotherapy," *Pharmaceutical Statistics* 22 (2022): 143–161.

21. P. F. Thall and H. Q. Nguyen, "Adaptive Randomization to Improve Utility-Based Dose-Finding With Bivariate Ordinal Outcomes," *Journal of Biopharmaceutical Statistics* 22, no. 4 (2012): 785–801.

22. P. F. Thall, H. Q. Nguyen, T. M. Braun, and M. H. Qazilbash, "Using Joint Utilities of the Times to Response and Toxicity to Adaptively Optimize Schedule–Dose Regimes," *Biometrics* 69, no. 3 (2013): 673–682.

23. T. A. Murray, P. F. Thall, and Y. Yuan, "Utility-Based Designs for Randomized Comparative Trials With Categorical Outcomes," *Statistics in Medicine* 35, no. 24 (2016): 4285–4305.

24. J. Lee, P. F. Thall, and K. Rezvani, "Optimizing Natural Killer Cell Doses for Heterogeneous Cancer Patients on the Basis of Multiple Event Times," *Journal of the Royal Statistical Society Series C: Applied Statistics* 68, no. 2 (2019): 461–474.

25. J. Lee, P. F. Thall, and P. Msaouel, "Bayesian Treatment Screening and Selection Using Subgroup-Specific Utilities of Response and Toxicity," *Biometrics* 79 (2023): 2458–2473.

26. T. Murray, Y. Yuan, P. Thall, and W. Hofstetter, "A Utility-Based Design for Randomized Comparative Trials With Ordinal Outcomes and Prognostic Subgroups," *Biometrics* 74 (2018): 1095–1103.

27. S. Chib and E. Greenberg, "Analysis of Multivariate Probit Models," *Biometrika* 85 (1998): 347–361.

28. C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (New York, NY: Springer, 2007).

Supporting Information

Additional supporting information can be found online in the Supporting Information section.