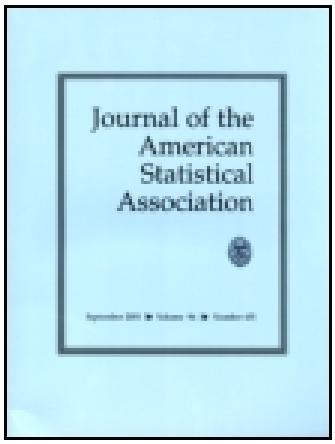
2014

Optimizing sedative dose in preterm infants undergoing treatment for respiratory distress syndrome

Peter F. Thall



Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Optimizing Sedative Dose in Preterm Infants Undergoing Treatment for Respiratory Distress Syndrome

Peter F. Thall, Hoang Q. Nguyen, Sarah Zohar & Pierre Maton

Accepted author version posted online: 01 Apr 2014.Published online: 02 Oct 2014.

To cite this article: Peter F. Thall, Hoang Q. Nguyen, Sarah Zohar & Pierre Maton (2014) Optimizing Sedative Dose in Preterm Infants Undergoing Treatment for Respiratory Distress Syndrome, Journal of the American Statistical Association, 109:507, 931-943, DOI: 10.1080/01621459.2014.904789

To link to this article: http://dx.doi.org/10.1080/01621459.2014.904789

Optimizing Sedative Dose in Preterm Infants Undergoing Treatment for Respiratory Distress Syndrome

Peter F. THALL, Hoang Q. NGUYEN, Sarah ZOHAR, and Pierre MATON

The intubation-surfactant-extubation (INSURE) procedure is used worldwide to treat preterm newborn infants suffering from respiratory distress syndrome, which is caused by an insufficient amount of the chemical surfactant in the lungs. With INSURE, the infant is intubated, surfactant is administered via the tube to the trachea, and at completion the infant is extubated. This improves the infant's ability to breathe and thus decreases the risk of long-term neurological or motor disabilities. To perform the intubation safely, the newborn infant first must be sedated. Despite extensive experience with INSURE, there is no consensus on what sedative dose is best. This article describes a Bayesian sequentially adaptive design for a multi-institution clinical trial to optimize the sedative dose given to preterm infants undergoing the INSURE procedure. The design is based on three clinical outcomes, two efficacy and one adverse, using elicited numerical utilities of the eight possible elementary outcomes. A flexible Bayesian parametric trivariate dose-outcome model is assumed, with the prior derived from elicited mean outcome probabilities. Doses are chosen adaptively for successive cohorts of infants using posterior mean utilities, subject to safety and efficacy constraints. A computer simulation study of the design is presented. Supplementary materials for this article are available online.

KEY WORDS: Adaptive design; Bayesian design; Clinical trial; Decision theory; Dose-finding; Neonatal; Phase I-II trial; Surfactant; Utility.

1. INTRODUCTION

Respiratory distress syndrome (RDS) in preterm newborn infants is characterized by an inability to breathe properly. RDS is associated with the facts that the infant's lungs have not developed fully and do not have a sufficient amount of surfactant, a compound normally produced in the lungs, which facilitates breathing. A relatively new but widely used procedure for preterm infants suffering from RDS is intubation-surfactant-extubation (INSURE), which is carried out when the infant is a few hours old. Once RDS has been diagnosed, the INSURE procedure is carried out as soon as possible to reduce the need for mechanical ventilation and risk of bronchopulmonary dysplasia. With INSURE, the infant is intubated, surfactant is administered via the tube to the trachea, and at completion the infant is extubated. The surfactant spreads from the trachea to the surface of the alveola, where it lowers alveolar surface tension and reduces alveolar collapse, thus improving lung aeration and decreasing respiratory effort. The aim is to improve the infant's ability to breathe and thus increase the probability of survival without long-term neurological or motor disabilities (Verder et al. 1994; Bohlin et al. 2007; Stevens et al. 2007). In most cases, the INSURE procedure takes no more than 1 hr, and ideally it is completed within 30 min. Because intubation is invasive, to allow it to be done safely and comfortably the infant first must be sedated. The drugs propofol (Ghanta et al. 2007) and remifentanyl (Welzing et al. 2009) are widely used for this purpose. Although the benefits of the INSURE procedure are well established, it also carries risks associated with intubation done while the infant is awake, and risks associated with the sedative. These include possible adverse behavioral and emotional effects if the infant is under-sedated as well as adverse hemodynamic effects associated with over-sedation. The goal in choosing a sedative dose is to sedate the infant sufficiently so that the procedure may be carried out, but avoid over-sedating. While it is clear that dose should be quantified in terms of amount per kilogram (kg) of the infant's body weight, little is known about what the optimal dose of any given sedative may be for the INSURE procedure. Propofol doses that are too high, or that are given recurrently or by continuous infusion, have been associated with serious adverse effects in the neonatal or pediatric populations (Murdoch and Cohen 1999; Sammartino et al. 2010; Vanderhaegen et al. 2010). Unfortunately, there is no broad consensus regarding the dose of any sedative in the community of neonatologists. The doses that actually are used vary widely, with each neonatologist using their preferred dose chosen based on personal clinical experience and consensus within their neonatal unit.

Pediatric clinical trials are challenging primarily due to ethical considerations, including informed consent, the fact that many pediatricians are hesitant to experiment with children, and the fact that adverse events may have lifelong consequences. These issues are especially difficult with newborn infants just a few hours old. While there is an extensive literature on adaptive dose-finding methods, these have been developed primarily for chemotherapy in oncology, which is a very different medical setting than sedation of neonates as described above. To date, no adaptive dose-finding design has been developed specifically for infants.

Peter F. Thall (E-mail: *rex@mdanderson.org*) is Professor and Hoang Q. Nguyen (E-mail: *honguyen@mdanderson.org*) is Senior Computational Scientist, Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX 77230-1402. Sarah Zohar is Research Associate, INSERM, UMR 1138, Centre de Recherche des Cordeliers, Universite Paris 5 and 6, Paris, France (E-mail: *sarah.zohar@inserm.fr*). Pierre Maton is Neonatologist and Head of Service Neonatal, CHC Saint Vincent, Rocourt, Belgium (E-mail: *pierre.maton@chc.be*). The authors thank the editor, an associate editor, and two referees for their detailed and constructive comments. This research was supported by NIH NCI grant 2RO1 CA083932.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Applications and Case Studies
DOI: 10.1080/01621459.2014.904789

The primary aim of the clinical trial described here is to optimize the dose of propofol given at the start of the INSURE procedure. Six possible doses are considered: 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 mg/kg body weight. Inherent difficulties in determining an optimal propofol dose are that there are both desirable and undesirable clinical outcomes related to dose, the probability of each outcome may vary as a complex, possibly nonmonotone function of dose, and the outcomes do not occur independently of each other. In any dose-finding clinical trial in humans, it is not ethical to randomize patients fairly among doses because, a priori, some doses are considered unsafe or ineffective, and as data are obtained some doses may turn out to be either unsafe or to have unacceptably low efficacy. These ethical considerations motivate the use of sequential, outcome-adaptive, "learn-as-you-go" dose-finding methods (see O'Quigley, Pepe, and Fisher 1990; Thall and Russell 1998; Chevret 2006; Cheung 2011). Such methods are especially important when treating newborn infants diagnosed with RDS, where sedative dose prior to intubation may have adverse hemodynamic effects and failure of the INSURE procedure may result in prolonged mechanical ventilation, a recognized risk factor for long-term adverse pulmonary outcomes (Stevens et al. 2007). Consequently, to optimize propofol dose in a reliable and ethical manner in the setting of the INSURE procedure, a clinical trial design must (1) account for unknown, potentially complex relationships between dose and key clinical outcomes, (2) account for inherent risk-benefit trade-offs between efficacy and adverse outcomes, (3) adaptively learn and make decisions using the accumulating dose-outcome data during the trial, and (4) reliably choose a final, "optimal" dose that can be recommended for future use worldwide with the INSURE procedure.

The clinical trial design described here satisfies all of these requirements. It uses a Bayesian sequentially outcome-adaptive method that relies on subjective utilities, elicited from neonatologists who perform the INSURE procedure, that account for the benefits of desirable outcomes and the risks of adverse outcomes. To characterize propofol dose effects in a realistic and practical way, we define three co-primary outcomes, including two desirable efficacy outcomes and one undesirable adverse outcome. The first efficacy outcome is that a "good sedation state," GSS, is achieved quickly. GSS is a composite event defined in terms of five established ordinal sedation assessment criteria variables scored within 5 min of the first sedative administration (Hummel et al. 2008). These five variables are $A_1 =$ Crying Irritability, $A_2 =$ Behavior State, $A_3 =$ Facial Expression, $A_4 =$ Extremities Tone, and $A_5 =$ Vital Signs. Each variable takes on an integer value in the set $-2, -1, 0, +1, +2$, with $A_j = -2$ corresponding to highest sedation and $A_j = +2$ to highest infant discomfort. The Vital Signs criterion score A_5 is defined in terms of heart rate (HR), respiration rate (RR), mean blood pressure (BP), and saturated oxygen in the circulating blood (SaO_2). Supplementary Table S1 gives detailed definitions of these five assessment variables.

The overall sedation assessment score is defined as $Z = \sum_{j=1}^{5} A_j$, and a good sedation score is defined as GSS $= \{-7 \le Z \le -3\}$. Because a GSS is required to intubate the infant, if it is not achieved with the initial propofol dose, then an additional fixed dose of 1.0 mg/kg propofol is given. If this still does not achieve a GSS, then use of another sedative is allowed at the discretion of the attending clinician. A nontrivial dimension reduction is performed in defining GSS, since Z is defined in terms of the variables $A_1, \ldots, A_4$ and A_5, which in turn is a function of three hemodynamic measurements. However, $A_1, \ldots, A_5$, Z, and GSS were defined by neonatologists who have extensive experience with the INSURE procedure.

Because it is desirable to complete the INSURE procedure as quickly as possible, the design also accounts for the efficacy event, EXT, that the infant is extubated within at most 30 min of intubation. This is motivated by the desire to sedate the infant sufficiently so that the INSURE procedure may be carried out, but not over-sedate. In addition to the efficacy events GSS and EXT, it is essential to monitor adverse events and include them in the dose-finding procedure. To do this, a third, composite adverse event was defined. The adverse hemodynamic event, HEM, is defined to have occurred if the baby's HR falls below 80 beats per minute, SaO_2 falls below 60%, or mean BP decreases by more than 5 mm Hg from a chosen inferior limit corresponding to the infant's gestational age. The time interval for monitoring the infant's HR, SaO_2, and BP values to score HEM includes both the period while the infant is intubated and the subsequent 3 hr following extubation. Thus, HEM is defined very conservatively.

Our proposed methodology is very different from adaptive dose-finding methods based on a single outcome. For example, a method based on GSS alone might choose a dose to maximize Pr(GSS | dose), maximize information using about this dose–response function using D-optimal or A-optimal designs, or possibly find the "minimum effective dose" for which it is likely that Pr(GSS | dose) $\ge p_G^*$ for some fixed target p_G^*. There is an extensive literature on such methods. Some useful references are Fedorov and Leonov (2001), Atkinson, Donev, and Tobias (2006), Dette et al. (2008), and Bornkamp et al. (2011). In the present setting, a method that is ethically acceptable must account for more than one outcome, and must quantify the trade-offs between the risk of HEM and the benefits of GSS and EXT. This requires specifying and estimating a trivariate dose-outcome probability distribution for these three events. Even if this function were known perfectly, however, some numerical representation of the desirabilities of the eight possible elementary outcomes still would be needed to decide which dose is best. We quantify this using elicited utilities, described in Section 3.

The propofol trial design uses a sequentially outcome-adaptive Bayesian dose-finding method based on a numerical utility of each of the eight possible combinations of the three outcomes GSS, EXT, and HEM. The numerical utilities, given in Table 1, were elicited from the neonatologists planning the trial, who are experienced with the INSURE procedure and have observed and dealt with these events in their clinical practice. Before the elicitation, the maximum numerical utility 100 was assigned to the best possible event (GSS = yes, EXT = yes, HEM = no), and the minimum numerical utility 0 was assigned to the worst possible event (GSS = no, EXT = no, HEM = yes). The six remaining intermediate values were elicited subject to the obvious constraints that the utility must increase as either GSS or EXT goes from "no" to "yes" and must decrease as HEM goes from "no" to "yes." The range 0–100 was chosen for convenience since it is easy to work with, although in general any numerical domain with which the area experts are

Table 1. Consensus elicited utilities and alternative utilities of the eight possible elementary outcomes

	GSS = Yes		GSS = No	
	EXT = Yes	EXT = No	EXT = Yes	EXT = No
a. Elicited consensus utilities.				
HEM = Yes	60	20	40	0
HEM = No	100	80	90	70
b. Alternative utilities 1, with *GSS* given greater importance compared with the consensus utility.				
HEM = Yes	80	60	20	0
HEM = No	100	90	45	35
c. Alternative utilities 2, with EXT given greater importance compared with the consensus utility.				
HEM = Yes	80	10	70	0
HEM = No	100	40	95	35
d. Alternative utilities 3, with HEM given greater importance compared with the consensus utility.				
HEM = Yes	30	10	20	0
HEM = No	100	90	95	85

NOTE: GSS = {good sedation score}, EXT = {extubation within 30 min}, HEM = {an adverse value of heartbeat, blood oxygen level, or blood pressure during the INSURE procedure or within 3 hr after extubation}.

comfortable could be used. By quantifying the desirability of each of the eight possible outcomes, the utility function formalizes the inherent trade-off between the INSURE procedure's risks and benefits, insofar as they are characterized by these three events. An essential property of the numerical utilities is that they quantify the subjective opinions of the area experts. This is an advantage of the methodology since, inevitably, any multidimensional criterion must be reduced to a one-dimensional object if decisions are to be made. However a dimension reduction is done, it is inherently subjective.

For trial conduct, the first cohort is treated at 1.0 mg/kg. The design chooses doses adaptively for all subsequent cohorts, subject to dose safety and efficacy constraints. Each decision is based on the dose-outcome data from all previously treated infants, using the posterior mean utilities of the six doses. To avoid getting stuck at a suboptimal dose, a well-known problem with "greedy" sequential algorithms that always maximize an objective function (see Sutton and Barto 1988), once a minimal sample is obtained at the current optimal dose, one version of the design randomizes adaptively among acceptable doses with posterior mean utility close to the maximum.

A variety of Bayesian decision theoretical methods have been proposed that are based on the utilities of making correct or incorrect decisions at the end of the trial. These include designs for phase II trials (see Stallard 1998; Stallard, Thall, and Whitehead 1999; Leung and Wang 2001; Stallard and Thall 2001; Chen and Smith 2009) and for randomized phase III trials (see Christen et al. 2004; Lewis et al. 2007; Wathen and Thall 2008). These methods optimize benefit to future patients. This is fundamentally different from the present approach, which assigns doses based on elicited joint utilities of the clinical outcomes, and at the end of the trial relies on the same criterion, posterior mean utility of each dose, to make a final recommendation. Bayesian clinical trial designs with similar sequentially adaptive Bayesian decision structures based on utilities have been proposed by Houede et al. (2010), Thall et al. (2011), and Thall and Nguyen (2012). The third design is the basis for a currently ongoing trial to optimize the dose of radiation therapy for pediatric brain tumors, based on bivariate ordinal efficacy and toxicity outcomes.

Section 2 describes the Bayesian multivariate dose-outcome model. The utility function and decision criteria used for trial conduct are presented in Section 3, and outcome-adaptive randomization criteria used in a modified version of the design are given in Section 4. An extensive simulation study of the design's behavior under a range of different possible scenarios is summarized in Section 5. We close with a brief discussion in Section 6.

2. PROBABILITY MODEL

2.1 Dose–Response Functions

Denote the outcome indicators $Y_G = I(\text{GSS}) = I\{-7 \le Z \le -3\}, Y_E = I(\text{EXT})$, $Y_H = I(\text{HEM})$. In the dose–response model, we will use the standardized doses obtained by dividing the raw doses by their mean, $x_1 = 0.5/1.75 = 0.286, \ldots, x_6 = 3.0/1.75 = 1.714$, with unsubscripted x denoting any given dose. The observed outcome vector is $\boldsymbol{O} = (Z, Y_E, Y_H)$. Because historical data of the form $(x, \boldsymbol{O})$ are not available, the following dose-outcome model was developed based on the collective experiences and prior beliefs of the neonatologists planning the propofol trial, and extensive computer simulations studying properties of various versions of the model and design.

Adaptive decisions in the trial are based on the behavior of $\boldsymbol{Y} = (Y_G, Y_E, Y_H)$ as a function of x. The distributions of the later outcomes, Y_E and Y_H, may depend quite strongly on the sedation score Z achieved at the start of the INSURE procedure. It is unlikely that Y_E and Y_H are conditionally independent given Z and x, and the definition of Z includes some of the hemodynamic events used to define HEM. To reflect these considerations, our joint model for $[\boldsymbol{O} \mid x]$ is based on the probability factorization

$$[Z, Y_E, Y_H \mid x, \boldsymbol{\theta}] = [Z \mid x, \boldsymbol{\theta}_Z][Y_E, Y_H \mid x, Z, \boldsymbol{\theta}_{E,H}], \quad (1)$$

where $\boldsymbol{\theta}_Z$ and $\boldsymbol{\theta}_{E,H}$ are subvectors of the model parameter vector $\boldsymbol{\theta}$. Expression (1) says that x may affect Z, while both x and Z may affect (Y_E, Y_H). To account for association between Y_E and Y_H, we first specify the conditional marginals of $[Y_E \mid x, Z]$ and $[Y_H \mid x, Z]$, and use a copula (Nelsen 1999) to obtain a bivariate distribution. Indexing $k = E, H$, we define these marginals using logistic regression models (McCullagh and Nelder 1989),

$$\pi_k(x, Z, \boldsymbol{\theta}_k) = \Pr(Y_k{=}1 \mid x, Z, \boldsymbol{\theta}_k) = \text{logit}^{-1}\{\eta_k(x, Z, \boldsymbol{\theta}_k)\}, \quad (2)$$

with linear terms taking the form

$$\eta_k(x, Z, \boldsymbol{\theta}_k) = \theta_{k,0} + \theta_{k,1}x^{\theta_{k,4}} + \theta_{k,2}\, f(Z) + \theta_{k,3}(1 - Y_G), \quad (3)$$

where $f(Z) = \{(Z + 5)/15\}^2$ and we denote $\boldsymbol{\theta}_k = (\theta_{k,0}, \theta_{k,1}, \theta_{k,2}, \theta_{k,3}, \theta_{k,4})$. For $k = E, H$, $\theta_{k,1}$ is the dose effect, $\theta_{k,2}$ is the sedation score effect, $\theta_{k,3}$ is the effect of not achieving a GSS, and x is exponentiated by $\theta_{k,4}$ to obtain flexible dose–response curves. We standardize Z in η_E and η_H so that

Table 2. Prior means, interval probabilities for Z and x = dose, and utilities

	Propofol dose (mg/kg)					
	0.5	1.0	1.5	2.0	2.5	3.0
a. Elicited prior interval probabilities for Z						
$-10 \le Z \le -8$	0.05	0.10	0.20	0.30	0.40	0.60
$-7 \le Z \le -3$	0.55	0.65	0.75	0.66	0.58	0.39
$-2 \le Z \le 10$	0.40	0.25	0.05	0.04	0.02	0.01
b. Elicited prior means of $\pi_E(z, x)$						
$Z = -10$	0.99	0.98	0.90	0.70	0.60	0.25
$Z = -5$	0.99	0.98	0.97	0.95	0.90	0.75
$Z = 0$	0.95	0.90	0.80	0.50	0.20	0.10
$Z = +10$	0.70	0.30	0.10	0.05	0.03	0.01
c. Elicited prior means of $\pi_H(z, x)$						
$Z = -10$	0.01	0.10	0.20	0.30	0.50	0.70
$Z = -5$	0.01	0.02	0.05	0.10	0.15	0.40
$Z = 0$	0.01	0.20	0.40	0.70	0.80	0.90
$Z = +10$	0.30	0.40	0.70	0.95	0.98	0.99
d. Prior mean utilities and probabilities, obtained by averaging over Z						
$\bar{U}(x \mid \boldsymbol{\theta})$	94.0	91.6	90.9	83.5	74.8	50.0
$\bar{\pi}_G(x \mid \boldsymbol{\theta})$	0.55	0.65	0.75	0.66	0.58	0.39
$\bar{\pi}_H(x \mid \boldsymbol{\theta})$	0.02	0.08	0.12	0.20	0.32	0.57
$\bar{\pi}_E(x \mid \boldsymbol{\theta})$	0.97	0.95	0.94	0.84	0.75	0.46
$\bar{\pi}_S(x \mid \boldsymbol{\theta})$	0.54	0.63	0.71	0.58	0.47	0.24

its numerical value does not have unduly large effects for values in the Z domain far away from -5, with $(Z + 5)/15$ squared to reflect the functional form of the elicited prior in Table 2. For example, the extreme score Z = +10 is represented by $f(Z) = 1$ rather than 225.

Specifying domains of the elements of $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_H$ requires careful consideration. The intercepts $\theta_{E,0}$ and $\theta_{H,0}$ are real-valued, with the exponents $\theta_{E,4}, \theta_{H,4} > 0$. Based on clinical experience with propofol and other sedatives used in the INSURE procedure, as reflected by the elicited prior means in Table 2, we assume that $\theta_{E,1}, \theta_{E,2} < 0$ while $\theta_{H,1}, \theta_{H,2} > 0$. This says that, given sedation score Z achieved initially, $\pi_E(x, Z, \boldsymbol{\theta})$ decreases and $\pi_H(x, Z, \boldsymbol{\theta})$ increases with dose. Similarly, failure to achieve a GSS can only increase the probability $\pi_H(x, Z, \boldsymbol{\theta})$ of an adverse hemodynamic event and decrease the probability $\pi_E(x, Z, \boldsymbol{\theta})$ of extubation within 30 min, so $\theta_{H,3} > 0$ while $\theta_{E,3} < 0$.

Denote the joint distribution $\pi_{E,H}(a, b \mid x, Z, \boldsymbol{\theta}_k) = \Pr(Y_E = a, Y_H = b \mid x, Z, \boldsymbol{\theta}_k)$, for $a, b \in \{0, 1\}$. Given the marginals $\pi_k(x, Z, \boldsymbol{\theta})$, $k = E, H$, temporarily suppressing $(x, Z, \boldsymbol{\theta})$ for brevity, the Gumbel–Morgenstern copula model is

$$\pi_{E,H}(a, b) = \pi_E^a(1 - \pi_E)^{1-a}\pi_H^b(1 - \pi_H)^{1-b} + \rho(-1)^{a+b}\pi_E(1 - \pi_E)\pi_H(1 - \pi_H) \quad (4)$$

with association parameter $-1 < \rho < +1$. The joint conditional distribution of $[Y_E, Y_H \mid x, Z]$ is parameterized by $\boldsymbol{\theta}_{E,H} = (\boldsymbol{\theta}_E, \boldsymbol{\theta}_H, \rho)$, which has dimension $5 + 5 + 1 = 11$, and $\boldsymbol{\theta}_Z$, which will be described below. Combining terms, and denoting $\pi_Z(z \mid x, \boldsymbol{\theta}_Z) = \Pr(Z = z \mid x, \boldsymbol{\theta}_Z)$, the joint distribution of $[Z, Y_E, Y_H \mid x]$ is

$$\Pr(Z = z, Y_E = a, Y_H = b \mid x, \boldsymbol{\theta}) = \pi_Z(z \mid x, \boldsymbol{\theta}_Z)\pi_{E,H}(a, b \mid x, z, \boldsymbol{\theta}_{E,H}) \quad (5)$$

for $z = -10, -9, \ldots, +9, +10$ and $a, b \in \{0, 1\}$.

An important property of the model is that the unconditional marginal distributions of the two later events, Y_E and Y_H, may be complex, nonmonotone functions of x. This is because their marginals first are defined in (2) conditional on the initial sedation score, Z, and their unconditional marginals are obtained by averaging over the distribution of Z,

$$\bar{\pi}_k(x, \boldsymbol{\theta}_k, \boldsymbol{\theta}_Z) = {}_{\text{def}} \Pr(Y_k = 1 \mid x, \boldsymbol{\theta}_k, \boldsymbol{\theta}_Z) = \sum_{z=-10}^{+10} \pi_k(x, z, \boldsymbol{\theta}_k)\, \pi_Z(z \mid x, \boldsymbol{\theta}_Z).$$

The unconditional joint distribution $\bar{\pi}_{E,H}(x, \boldsymbol{\theta}_k, \boldsymbol{\theta}_Z)$ is computed similarly, from (4) and (5). The probability $\bar{\pi}_H(x, \boldsymbol{\theta}_k, \boldsymbol{\theta}_Z)$ of HEM plays a key role in the design because it is used as a basis for deciding whether x is acceptably safe. Similarly, overall success is defined as S = (GSS and EXT) = ($-7 \le Z \le -3$ and $Y_E = 1$), which has probability $\pi_S(x, \boldsymbol{\theta})$ that depends on $\pi_Z(z \mid x, \boldsymbol{\theta}_Z)$. Thus, a key aspect of how the outcomes are observed that affects the statistical model and method is that, for an infant given propofol dose x, $\bar{\pi}_H(x, \boldsymbol{\theta}_k, \boldsymbol{\theta}_Z)$ and $\pi_S(x, \boldsymbol{\theta})$ are averages over the initial sedation score distribution, and thus these probabilities depend on $\boldsymbol{\theta}_Z$.

2.2 Extended Beta Regression Model for Sedation Score

To specify a flexible distribution of $[Z \mid x]$, we employ the technical device of first defining a beta regression model for a latent variable W having support [0, 1] with mean that is a decreasing function of x, and then defining the distribution of Z in terms of the distribution of W. We formulate the beta regression model for $[W \mid x]$ using the common reparameterization of the Be(a, b) model in terms of its mean $\mu = a/(a + b)$ and $\psi = a + b$, where $\mu = \mu_x$ varies with x and the pdf is

$$f_W(w \mid \boldsymbol{\theta}_Z, x) = \frac{\Gamma(\psi)}{\Gamma(\mu_x\psi)\Gamma((1 - \mu_x)\psi)} \times w^{\mu_x\psi-1}(1 - w)^{(1-\mu_x)\psi-1}, \quad \text{for } 0 < w < 1, \quad (6)$$

(see Williams 1982; Ferrari and Cribari-Neto 2004), and $\Gamma(\cdot)$ denotes the gamma function. Denote the indexes of the doses in increasing order by $j(x) = 1, \ldots, J$. We assume a saturated model for the mean of $[W \mid x]$,

$$\mu_x = \left\{1 + \sum_{r=1}^{j(x)} \alpha_r\right\}^{-1},$$

where $\alpha_1, \ldots, \alpha_J > 0$. Our preliminary simulations showed that assuming constant ψ in the beta regression model for $[W \mid x]$ results in a model for $[Z \mid x]$, shown below, that is not sufficiently flexible across a range of possible dose-outcome scenarios to facilitate reliable utility-based dose finding. To obtain a more flexible model, we explored the behavior of several parametric functions for ψ. We found that the function

$$\psi_x = \{\mu_x(1 - \mu_x)\}^{1-2\gamma_1}(2 + \gamma_2 x^{\gamma_3})^2 \quad (7)$$

with $\gamma_1, \gamma_2 > 0$ and γ_3 real-valued gives a model that does a good job of fitting a wide range of simulated data. The initial rationale for this particular functional form was to model the standard deviation as the function $\sigma_x = \{\mu_x(1-\mu_x)\}^{\nu}/(2+\zeta x^{\alpha})$, with $\nu > 0$. To ensure the usual beta distribution parameter constraints $\sigma_x < 0.50$ and $\psi_x > 0$, it was necessary to modify this so that $\sigma_x = [\{\mu_x(1-\mu_x)\}/(1+\psi_x)]^{1/2}$ with ψ_x given by (7). Modeling the effective sample size (ESS) parameter as a function of x and μ_x in this way, in addition to the more common practice of defining a regression model for the mean, is similar in spirit to the generalized beta regression model of Simas, Barreto-Souza, and Rocha (2010).

Denote the incomplete beta function $B(w, c, d) = \int_0^w u^{c-1}(1-u)^{d-1}dw$, for $0 < w < 1$ and $c, d > 0$. Using the continuous distribution of $[W \mid x]$ given in (6), we define the discrete distribution for $[Z \mid x]$ as

$$\begin{aligned}\pi_Z(z \mid x, \boldsymbol{\theta}_Z) &= \Pr(z - 0.5 \le 21\,W - 10.5 \le z + 0.5 \mid x, \boldsymbol{\theta}_Z)\\ &= \Pr\{(z+10)/21 \le W \le (z+11)/21 \mid x, \boldsymbol{\theta}_Z\}\\ &= B\left\{\frac{z+11}{21}, \mu_x\psi_x, (1-\mu_x)\psi_x\right\}\\ &\quad - B\left\{\frac{z+10}{21}, \mu_x\psi_x, (1-\mu_x)\psi_x\right\} \qquad (8)\end{aligned}$$

for $z = -10, -9, \ldots, +9, +10$, where $\boldsymbol{\theta}_Z = (\boldsymbol{\alpha}, \boldsymbol{\gamma}) = (\alpha_1, \ldots, \alpha_J, \gamma_1, \gamma_2, \gamma_3)$. Since $J = 6$ propofol doses will be studied, this model for the distribution of Z in terms of the generalized beta latent variable W expresses the probability of a GSS in terms of the incomplete beta function evaluated at arguments characterized by x, the six dose–response parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_6)$ of μ_x, and the three parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ of ψ_x. While this model for $[Z \mid x]$ may seem somewhat elaborate, it must be kept in mind that Z is a sum with 21 possible values and its distribution is a function of J possible doses, so for the propofol trial a $6 \times 20 = 120$ dimensional distribution is represented by a nine-parameter model.

It follows from (8) that the probability of GSS $= (-7 \le Z \le -3)$ is

$$\begin{aligned}\pi_G(x, \boldsymbol{\theta}_Z) = {}& B\{8/21, \mu_x\psi_x, (1-\mu_x)\psi_x\}\\ &- B\{3/21, \mu_x\psi_x, (1-\mu_x)\psi_x\}. \qquad (9)\end{aligned}$$

While the distribution of W is monotone in dose by construction, it should be clear from expressions (6)–(9) that $\pi_G(x, \boldsymbol{\theta}_Z)$ is a complex, possibly nonmonotone function of dose.

2.3 Prior, Likelihood, and Posterior Computation

Collecting terms, the model parameter vector is $\boldsymbol{\theta} = (\rho, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\theta}_E, \boldsymbol{\theta}_H)$, which has 20 elements. To establish a prior, we assumed $\rho \sim \text{Unif}[-1, +1]$, and for the remaining 19 parameters, $\boldsymbol{\theta} - \rho$, we used the following pseudo-sample-based approach, similar to that of Thall and Nguyen (2012). The pseudo-samples were obtained by treating the elicited means of the probabilities $\pi_E(z, x)$ and $\pi_H(z, x)$ and interval probabilities $\Pr(l \le Z \le u \mid x)$ in Table 2 as the true state of nature. For each dose x, we used these elicited probabilities to generate a pseudo-sample of 100 iid patient outcomes,

$$\tilde{\mathcal{D}}(x) = \left\{(\tilde{Z}^i(x), \tilde{Y}_E^i(x, (\tilde{Z}^i(x)), \tilde{Y}_H^i(x, (\tilde{Z}^i(x)),\ i = 1, \ldots, 100\right\}.$$

To generate each pseudo-sample, it first was necessary to specify $\pi_Z(z \mid x)$ for all combinations of x and $z = -10, \ldots, +10$. For each x, we did this by first fitting the three interval probabilities in the corresponding column of Table 2(a) to a beta(a_x, b_x), then partitioning $[0, 1]$ into 21 equal subintervals and setting each $\pi_Z(z \mid x)$ to be the fitted beta probability of the corresponding subinterval. To obtain $\pi_E(x, z)$ for all 21 values of z, we linearly interpolated the rows of Table 2(b), and we obtained $\pi_H(x, z)$ similarly from Table 2(c). Using these probabilities, for each i and x, we first simulated $\tilde{Z}^i(x)$ from $\pi_Z(z \mid x)$ and then simulated $\tilde{Y}_k^i(x, \tilde{Z}^i(x))$ from $\pi_k(x, \tilde{Z}^i(x))$ for $k = E$ and H. Given the combined pseudo-sample $\tilde{\mathcal{D}} = \cup_x \tilde{\mathcal{D}}(x)$, and assuming a highly noninformative pseudo-prior on $\boldsymbol{\theta} - \rho$, we computed a pseudo-posterior $p(\boldsymbol{\theta} - \rho \mid \tilde{\mathcal{D}})$. This entire process was repeated 3000 times, and the average of the 3000 pseudo-posterior means was used as the prior mean of $\boldsymbol{\theta} - \rho$. The pseudo-sample size 100 was chosen to be large enough to provide reasonably reliable pseudo-posteriors, but small enough so that the computations could be carried out feasibly. Pseudo-sampling provides a reliable alternative to nonlinear least squares, which often fails to converge in this type of setting.

For priors, we assumed that $\{\alpha_1, \ldots, \alpha_6, -\theta_{E,1}, -\theta_{E,2}, -\theta_{E,3}, \theta_{H,1}, \theta_{H,2}, \theta_{H,3}\}$ were normal truncated below at 0, $\{\gamma_1, \gamma_2, \theta_{E,4}, \theta_{H,4}\}$ were lognormal, and $\{\gamma_3, \theta_{E,0}, \theta_{H,0}\}$ were normal. Given the prior means established by the pseudo-sampling method, we calibrated the prior variances to be uninformative in the sense that ESS (Morita, Thall, and Mueller 2008, 2010) of the prior was 0.10. Numerical prior means and variances are given in supplementary Table S2.

Let N denote the maximum trial sample size. Index the patients enrolled in the trial by $i = 1, \ldots, N$, and denote the observed outcomes by $\mathbf{O}_i = (Z_i, Y_{i,E}, Y_{i,H})$, and the assigned dose by $x_{[i]}$ for the ith patient. Let $n = 1, \ldots, N$ denote an interim sample size where an adaptive decision is made during the trial, and $\mathcal{O}_n = (\mathbf{O}_1, \ldots, \mathbf{O}_n)$ denote the observed data from the first n patients. The likelihood for the first n patients in the trial is

$$\begin{aligned}\mathcal{L}(\mathcal{O}_n \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} f(\mathbf{O}_i \mid x_{[i]}, \boldsymbol{\theta})\\ &= \prod_{i=1}^{n} \pi_Z(Z_i \mid x_{[i]}, \boldsymbol{\theta}_Z)\pi_{E,H}(Y_{i,E}, Y_{i,H} \mid x_{[i]}, Z_i, \boldsymbol{\theta}_{E,H}).\end{aligned}$$

The posterior based on this interim sample is

$$p(\boldsymbol{\theta} \mid \mathcal{O}_n) \propto \mathcal{L}(\mathcal{O}_n \mid \boldsymbol{\theta})\ \text{prior}(\boldsymbol{\theta}).$$

All posterior quantities used for decision making by the trial design were computed using Markov chain Monte Carlo with Gibbs sampling (Robert and Casella 1999).

3. DECISION CRITERIA

3.1 Utilities

Denote the utility function by $U(\mathbf{y})$, where $\mathbf{y} = (y_G, y_E, y_H) \in \{0, 1\}^3$ is an elementary outcome. The numerical utilities for the propofol trial outcomes were obtained by first fixing the scores of the best and worst possible elementary outcomes to be $U(1, 1, 0) = 100$ and $U(0, 0, 1) = 0$, and eliciting the remaining six scores as values between 100 and 0 from

neonatologists familiar with the INSURE procedure. An admissible utility $U(y_G, y_E, y_H)$ must increase in y_G and y_E and decrease in y_H. While these admissibility requirements may seem obvious, they must be kept in mind during the elicitation process. Although we used the range [0, 100] for U, in general for a given application any convenient interval may be used, depending on what the area experts find intuitively appealing.

To construct dose-finding criteria from the utility function $U(\boldsymbol{y})$, we first define the *mean utility of dose* x *given* $\boldsymbol{\theta}$,

$$\bar{U}(x \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{y}} U(\boldsymbol{y})\, \pi_{G,E,H}(\boldsymbol{y} \mid x, \boldsymbol{\theta}), \tag{10}$$

where the joint distribution $\pi_{G,E,H}$ is as given earlier. This expression says that, if one knew the parameters $\boldsymbol{\theta}$, then the mean utility (10) is what one would expect to achieve by giving an infant dose x. Since $\boldsymbol{\theta}$ is not known, it must be estimated. Rather than computing a frequentist estimator $\hat{\boldsymbol{\theta}}$ and basing decisions on $\bar{U}(x \mid \hat{\boldsymbol{\theta}})$, we will exploit our Bayesian model to compute statistical decision criteria, as follows. Let data_n denote the observed dose-outcome data from n babies at any interim point in the trial, $1 \le n < N$. Let $p(\boldsymbol{\theta} \mid \text{data}_n)$ denote the current posterior of $\boldsymbol{\theta}$. The *posterior mean utility of dose* x *given* data_n is

$$u(x \mid \text{data}_n) = \int_{\boldsymbol{\theta}} \bar{U}(x \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta} \mid \text{data}_n) d\boldsymbol{\theta}. \tag{11}$$

In words, based on what has been learned from the observed data from n babies, the posterior mean utility $u(x \mid \text{data}_n)$ is what one would expect to achieve if the next baby were given dose x. An important point is that, with small sample sizes, some of the eight elementary events may not occur, and in this case $u(x \mid \text{data}_n)$ will be based partly on the prior. Note that (11) is obtained by averaging over the distribution of $[\boldsymbol{Y} \mid x, \boldsymbol{\theta}]$ in (10) to obtain $\bar{U}(x \mid \boldsymbol{\theta})$, and then averaging this mean utility over the posterior of $\boldsymbol{\theta}$. We denote by x_n^{opt} the dose having maximum $u(x \mid \text{data}_n)$ among the doses under study. For brevity, we denote $u_n^{\text{opt}} = u(x_n^{\text{opt}} \mid \text{data}_n)$. Subject to the restriction that an untried dose may not be skipped when escalating, the design U^{opt} chooses each successive cohort's dose to maximize $u(x \mid \text{data}_n)$ among all $x \in \{x_1, \ldots, x_6\}$.

It may seem appropriate to place a probability distribution on the utility function U to reflect uncertainty about what alternative utilities others may have. If a distribution $q(U)$ is assumed for U, using the elicited consensus utility as the mean U_q under q, then one would need to integrate over $q(U)$ as well as $\pi_{G,E,H}(\boldsymbol{y})$ and $p(\boldsymbol{\theta} \mid \text{data}_n)$ to obtain $u(x \mid \text{data}_n)$. This computation gives the original posterior mean utility (11), however, essentially because the trial data provide no new information about U. We will address this issue by sensitivity analyses to U, in Section 5.

3.2 Dose Acceptability Criteria

A critical issue is that a dose that is "optimal" in terms of the utility alone may be unacceptable in terms of either safety or overall success rate. To ensure that any administered dose has both an acceptably high success rate and an acceptably low adverse event rate, based on the current data, we define the following two posterior acceptability criteria. Given the fixed upper limit $\bar{\pi}_H^*$, we say that *a dose* x *is unsafe* if

$$\Pr\{\bar{\pi}_H(x, \boldsymbol{\theta}_H, \boldsymbol{\theta}_Z) > \bar{\pi}_H^* \mid \text{data}_n\} > p_{U,H} \tag{12}$$

for fixed upper probability cut-off $p_{U,H}$. Recall that the overall success event is $S = (Y_G = 1 \text{ and } Y_E = 1)$, that a GSS was achieved with the initial propofol administration and the INSURE procedure was completed with extubation within 30 min. Denoting $\pi_S(x, \boldsymbol{\theta}) = \Pr(S = 1 \mid x, \boldsymbol{\theta})$, the probability of this event is given by

$$\begin{aligned}\pi_S(x, \boldsymbol{\theta}) &= \Pr(Y_E = 1,\ Y_G = 1 \mid x, \boldsymbol{\theta}) \\ &= \Pr(Y_E = 1 \text{ and } -7 \le Z \le -3 \mid x, \boldsymbol{\theta}) \\ &= \sum_{z=-7}^{-3} \Pr(Y_E = 1 \mid x, Z = z, \boldsymbol{\theta}_E)\pi_Z(z \mid x, \boldsymbol{\theta}_Z),\end{aligned}$$

parameterized by $(\boldsymbol{\theta}_E, \boldsymbol{\theta}_Z)$. We say that *a dose* x *has unacceptably low overall success probability* if

$$\Pr\{\pi_S(x, \boldsymbol{\theta}_E, \boldsymbol{\theta}_Z) < \pi_S^* \mid \text{data}_n\} > p_{U,S} \tag{13}$$

for fixed upper probability cut-off $p_{U,S}$. We will refer to the subset of doses that do not satisfy either (12) or (13) as *acceptable doses*. We denote this subset by $\mathcal{A}_n$, and we denote the modification of design U^{opt} restricted to $\mathcal{A}_n$ by $U^{\text{opt}} + \text{Acc}$.

4. ADAPTIVE RANDOMIZATION

Intuitively, it may seem that the best dose is simply the one maximizing the posterior mean utility, possibly enforcing the additional acceptability criteria given above. However, it is well known in sequential decision making that a "greedy" algorithm that always chooses each successive action by optimizing some decision criterion risks getting stuck at a suboptimal action. A greedy algorithm may get stuck at a suboptimal action due to the fact that, because it repeatedly takes the suboptimal action, it fails to take and thus obtain enough data on an optimal action to determine, statistically, that it is truly optimal. This problem is sometimes known as the "optimization versus exploration" dilemma (see Robbins 1952; Gittins 1979; Sutton and Barto 1998). This fact has been recognized only recently in the context of dose-finding clinical trials (Azriel, Mandel, and Rinott 2011; Thall and Nguyen 2012; Oron and Hoff 2013). In the propofol trial, always choosing an "optimal" dose x by maximizing $u(x \mid \text{data}_n)$ is an example of a greedy algorithm, even if x is restricted to $\mathcal{A}_n$. A simple aspect of this problem is that the statistics $u(x_1 \mid \text{data}_n), \ldots, u(x_K \mid \text{data}_n)$ are actually quite variable for most values of n during the trial, and simply maximizing their means ignores this variability. This problem has both ethical and practical consequences, since maximizing the posterior mean utility for each cohort may lead to giving suboptimal doses to a substantial number of the infants in the trial, and it also may increase the risk of recommending a suboptimal dose at the end. To deal with this problem, we use adaptive randomization (AR) to improve this greedy algorithm and thus the reliability of the trial design. Our AR criterion is similar to that used by Thall and Nguyen (2012). One goal of the AR is to obtain a design that, on average, treats more patients at doses with higher actual utilities and is more likely to choose a dose with maximum or at least high utility at the end of the trial. At the same time, it must not allow an unacceptable risk for the two infants in each

cohort. Thus, while the AR is implemented using probabilities proportional to the posterior mean utilities, it is restricted to the set $\mathcal{A}_n$ of acceptable doses. Given current data$_n$, the next cohort is randomized to dose $x_j \in \mathcal{A}_n$ with probability

$$p_{j,n} = \frac{u(x_j \mid \text{data}_n)}{\sum_{r=1}^{K} u(x_r \mid \text{data}_n)\, I(x_r \in \mathcal{A}_n)}. \quad (14)$$

The following algorithm is a hybrid of utility maximization and AR. It chooses doses according to U^{opt} + Acc, unless the current optimal dose has at least δ more patients than any other acceptable dose. In this case, it applies the AR criterion (14) to choose a dose, as follows. Denote the sample size at dose x_j after n patients have been treated by $m_n(x_j)$, so that $m_n(x_1) + \ldots m_n(x_K) = n$. Among the doses in $\mathcal{A}_n$, if $m_n(x^{\text{opt}}) \geq m_n(x_j) + \delta$ for all $x_j \neq x^{\text{opt}}$, then assign x_j with probability $p_{j,n}$. Otherwise, assign x^{opt}.

For ethical reasons, AR must be applied carefully. Once enough data have been obtained to apply AR reliably, it is ethically inappropriate to randomize patients to a dose that is unlikely to be best. Formally, we say that *x is unlikely to be best* if

$$\Pr\{\bar{U}(x, \boldsymbol{\theta}) = \max_{x'} \bar{U}(x', \boldsymbol{\theta}) \mid \text{data}_n\} < p_L \quad (15)$$

for fixed lower probability cut-off p_L. Thus, AR is applied to the set of doses that not only are acceptable in terms of the safety and efficacy criteria (12) and (13), but that also do not satisfy (15), that is, that are not unlikely to be best. This restriction is most useful when larger sample sizes are available, later in the trial, and has the effect of reducing the numbers of patients treated at inferior doses. We denote this hybrid algorithm by U^{opt} + Acc + AR$_\delta$.

For each design U^{opt}, U^{opt} + Acc, and U^{opt} + Acc + AR$_\delta$, the first cohort is treated with 1.0 mg/kg, untried doses may not be skipped when escalating, but there is no constraint on de-escalation. Acc restricts doses to $\mathcal{A}_n$. For U^{opt} + Acc + AR$_\delta$, doses unlikely to be best also are excluded, and the AR criterion is used only if, within this subset of doses, x^{opt} has at least δ more patients than any other dose. For both U^{opt} + Acc, and U^{opt} + Acc + AR$_\delta$, if it is determined that $\mathcal{A}_n = \phi$, the trial is stopped and no dose is selected. For all three designs, if the trial is not stopped early, at the end of the trial, the dose x_{select} having maximum posterior mean utility, $u(x|\text{data}_N)$, is selected.

While the trial will be shut down if A_n is empty, that is, no dose is acceptable, we consider this very unlikely. If this happens, then for neonatologists performing the INSURE procedure using propofol, in practice a safe dose with HEM rate < 0.10 but a success rate lower than 0.60 would be used. This might motivate a subsequent trial to study the idea of titrating the dose in more than one administration for each infant. However, optimizing such a multistage procedure is a much more complex problem, and would require a very different design.

5. SIMULATION STUDY

In the simulations, the trial has maximum sample size $N = 60$, cohort size $c = 2$, and acceptability cut-offs $p_{U,H} = p_{U,S} = 0.95$, with $p_L = 0.05$ when AR$_\delta$ is used. In preliminary simulations, these design parameters were varied, along with the prior variances, to study their effects and obtain a design with desirable properties. The hybrid U^{opt} + Acc + AR$_\delta$ was studied for $\delta = 2$, 4, 6, 8, and 10. Since the results were insensitive to δ in this range, only the case $\delta = 2$ is reported.

We also included the following ad hoc nonmodel-based four-stage design suggested by a Referee as a comparator. Stage 1: Randomize 24 patients to each of the 6 doses (4 per dose). Select the 4 doses with the highest mean utility $\bar{U}(x)$ for evaluation in stage 2. Stage 2: Randomize 16 patients to each of the 4 selected doses (4 per dose), and select the 3 doses (from all 6) with highest $\bar{U}(x)$ for evaluation in stage 3. Stage 3: Randomize 12 patients to each of the 3 newly selected doses (4 per dose), and select the 2 doses (from all 6) with the highest $\bar{U}(x)$ for evaluation in stage 4. Stage 4: Randomize 8 patients to each of the 2 remaining doses (4 per dose), and select the best dose, having highest $\bar{U}(x)$ across all 6 doses. This design uses 60 patients, evaluates at least 4 patients per dose, and the selected dose has information on up to 16 patients. While it interimly selects (drops) doses with higher (lower) empirical mean utilities, it does not have rules that drop doses in terms of their empirical HEM or Success rates.

We used the following criteria to assess and compare the designs. The first is the proportion of the difference between the utilities of the best and worst possible doses achieved by x_{select}, scaled to the domain [0, 100],

$$R_{\text{select}} = 100\, \frac{u^{\text{true}}(x_{\text{select}}) - u_{\min}}{u_{\max} - u_{\min}}.$$

The second criterion quantifies how well a method assigns doses to patients in the trial,

$$R_{\text{treat}} = 100\, \frac{\frac{1}{N}\sum_{i=1}^{N} u^{\text{true}}(x_{[i]}) - u_{\min}}{u_{\max} - u_{\min}},$$

where $u^{\text{true}}(x_{[i]})$ is the true utility of the dose given to the ith patient. Larger values correspond to better design performance, with R_{select} quantifying benefit to future patients while R_{treat}, which may be regarded as an ethical criterion, quantifying benefit to the patients treated in the trial.

Table 3 compares the four designs, based on mean values across 3000 simulated trials under each of 9 different dose-outcome scenarios, given in supplementary Tables S3.1– S3.9. Scenario 1 is based on the elicited prior probabilities. The beta regression model was used to obtain all 21 true $\pi_Z(x)$ values from three interval probabilities, and linear interpolation was used to obtain true $\pi_E(x)$ and $\pi_H(x)$. Otherwise, none of the scenarios are model-based. The scenarios assume that a larger dose will shift the Z distribution toward -10, which is reasonable given the nature of the sedative drug. Given this, the interval probabilities for Z vary widely across the scenarios. The scenarios' true $\pi_E(x)$ and $\pi_H(x)$ have the same general trends as the prior in that $\pi_E(x)$ decreases and $\pi_H(x)$ increases with x given Z. To reflect the prior belief that Y_E and Y_H are slightly negatively correlated (Table 2), we set $\rho = -0.1$ when generating the true joint distributions of each scenario. Preliminary simulation results were insensitive to the assumed true ρ value. Both U^{opt} and four-stage have no early stopping rules, so these designs always treat 60 patients. Due to the much larger number adverse HEM events of four-stage in Scenarios 1, 2, and 8, the fact that it treats 60 patients in Scenarios 8 and 9 where no doses are acceptable, and the much lower R_{treat} values across all

Table 3. Comparison of alternative designs. Scenarios 8 and 9 have no acceptable dose, so R_{select} values are less relevant and thus have a gray background

Design		Scenario 1	2	3	4	5	6	7	8	9
U^{opt}	R_{select}	96	93	99	90	73	49	30	95	48
	R_{treat}	96	92	98	90	64	46	21	95	42
	% None	0	0	0	0	0	0	0	0	0
	# Pats	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0
	# HEM	4.1	2.7	2.4	2.8	2.2	2.3	2.1	19.3	2.0
	# Succ	36.7	40.8	39.6	33.0	25.7	20.2	9.2	37.1	11.0
U^{opt} + Acc	R_{select}	95	93	99	95	93	89	88	96	99
	R_{treat}	96	92	98	92	79	69	64	96	73
	% None	4	0	1	2	4	7	10	100	93
	# Pats	58.9	59.8	59.8	59.4	59.0	58.1	56.9	15.4	40.6
	# HEM	4.2	2.6	2.4	2.9	2.3	2.6	2.9	4.9	1.9
	# Succ	36.4	40.6	39.3	35.1	32.2	28.7	25.5	9.4	11.9
U^{opt} + Acc + AR_2	R_{select}	95	94	94	95	92	89	94	98	97
	R_{treat}	92	84	87	87	76	71	69	94	72
	% None	4	1	1	4	5	6	7	100	95
	# Pats	59.0	59.7	59.7	59.2	58.9	58.5	57.9	15.4	39.7
	# HEM	5.7	4.5	2.8	3.8	2.6	2.9	3.0	5.0	1.9
	# Succ	36.5	39.0	36.4	35.1	32.4	30.3	28.2	9.5	11.6
4-Stage	R_{select}	97	97	92	93	89	82	84	90	86
	R_{treat}	83	65	73	76	66	63	59	74	69
	% None	0	0	0	0	0	0	0	0	0
	# Pats	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0
	# HEM	8.8	8.9	3.2	5.0	2.7	2.9	2.7	22.8	2.7
	# Succ	34.5	35.4	33.3	31.9	29.8	28.4	23.1	36.4	16.8

NOTE: A dose x is unacceptable if either $\bar{\pi}_H(x, \boldsymbol{\theta}) > 0.10$ or $\pi_S(x, \boldsymbol{\theta}) < 0.60$ with posterior probability > 0.95.

scenarios, this design is unethical. Compared with U^{opt} + Acc and U^{opt} + Acc + AR_2, four-stage has R_{select} values that are slightly higher in Scenarios 1 and 2 with the price being many more occurrences of HEM, and in Scenarios 3–7 it has lower R_{select} values. Comparison of U^{opt} to U^{opt} + Acc shows the effects of including dose acceptability criteria in a sequentially adaptive utility-based design. While these two designs have similar values of R_{select} and R_{treat} for Scenarios 1–4, the importance of the acceptability rules is shown clearly by the other scenarios, where U^{opt} + Acc has greatly superior performance. Moreover, the mean of 19.2 adverse HEM events for U^{opt} in Scenario 8 illustrates the potential danger of using a design with a utility-based decision criterion without an early stopping rule for safety. The much higher values of R_{select} and R_{treat} for U^{opt} + Acc in Scenarios 5–7 show that it is both more reliable and more ethical in these cases compared to U^{opt}.

After excluding U^{opt} and four-stage as ethically unacceptable, comparison between U^{opt} + Acc and U^{opt} + Acc + AR_2 shows the effects of including AR. Recall that AR_2 randomizes patients among acceptable doses having $u(x \mid \text{data})$ close to $u(x^{opt} \mid \text{data})$, to better explore the dose domain. These designs have very similar R_{select} values for Scenarios 1–6, with U^{opt} + Acc + AR_2 showing a slight advantage in Scenario 7. As expected, U^{opt} + Acc has slightly larger R_{treat} values and slightly smaller mean numbers of HEM events in most scenarios. Consequently, for the propofol trial, U^{opt} + Acc is the better of the two ethical designs, but by a small margin.

Table 4 summarizes the simulations in more detail for U^{opt} + Acc. In each of Scenarios 1–7, the selection rates, subsample sizes, and success event rates for the six doses all follow the $u^{\text{true}}(x)$ values, and doses with comparatively low $u^{\text{true}}(x)$ are selected seldom or not at all. The design is very likely to stop the trial and select no dose in both Scenario 8, where all doses are unsafe with $\pi_H^{\text{true}}(x) \geq 0.29$, and Scenario 9, where all doses have a low success probability with $\pi_S^{\text{true}}(x) \leq 0.41$. In particular, U^{opt} + Acc does a good job of controlling the HEM event rate at very low values across all scenarios. Figure 1 illustrates properties of U^{opt} + Acc in four selected scenarios.

The numerical limits $\pi_H(x) \leq 0.10$ and $\pi_S(x) \geq 0.60$ in the propofol trial are very demanding, and they constrain the acceptable dose set severely. This is ethically appropriate for a trial where the patients are newborn infants and, although the optimal sedative dose is not known, the INSURE procedure has been very successful. Recall that adding AR to the design is motivated by the desire to reduce the chance of getting stuck at a suboptimal dose. In other structurally similar settings, different numerical values for the dose admissibility limits $\bar{\pi}_H^*$ and π_S^* may produce substantively different behavior of U^{opt} + Acc and U^{opt} + Acc + AR_δ. As a hypothetical but realistic example, consider an oncology trial of an anticancer agent where G is a desirable early biological effect, E is tumor response, and H is toxicity. Suppose that, based on what has been seen with standard chemotherapy, $\bar{\pi}_H^* = 0.25$ and $\pi_S^* = 0.40$ are appropriate numerical values for the dose acceptability criteria (12) and

Table 4. Simulation results using the U^{opt} + Acc design

Dose (mg/kg)		0.5	1.0	1.5	2.0	2.5	3.0	% None, sum
Scenario 1	u^{true}	94.0	**91.6**	90.9	83.5	74.7	49.9	
	% Sel	18	69	9	0	0	0	4
	# Pats	12.1	42.8	3.9	0.1	0.0	0.0	58.9
	# HEM	0.2	3.5	0.5	0.0	0.0	0.0	4.2
	# Succ	6.5	27.1	2.8	0.0	0.0	0.0	36.4
Scenario 2	u^{true}	**95.9**	92.3	84.3	79.9	75.0	68.7	
	% Sel	51	48	1	0	0	0	0
	# Pats	23.9	35.2	0.7	0.0	0.0	0.0	59.8
	# HEM	0.3	2.2	0.1	0.0	0.0	0	2.6
	# Succ	17.2	23.0	0.4	0.0	0.0	0.0	40.6
Scenario 3	u^{true}	93.0	**94.4**	92.2	88.7	86.0	80.6	
	% Sel	8	89	2	0	0	0	1
	# Pats	8.2	50.0	1.4	0.1	0.0	0.0	59.8
	# HEM	0.3	2.0	0.1	0.0	0.0	0.0	2.4
	# Succ	4.4	34.0	0.9	0.0	0.0	0.0	39.3
Scenario 4	u^{true}	88.2	91.7	**93.2**	91.3	82.1	75.3	
	% Sel	2	43	51	2	0	0	2
	# Pats	4.4	34.8	19.2	0.8	0.1	0.0	59.4
	# HEM	0.2	1.6	1.0	0.1	0.0	0.0	2.9
	# Succ	1.5	19.7	13.3	0.5	0.0	0.0	35.1
Scenario 5	u^{true}	80.6	85.7	90.9	**92.9**	90.4	84.4	
	% Sel	0	0	35	58	2	0	4
	# Pats	3.7	6.4	28.4	19.7	0.8	0.1	59.0
	# HEM	0.1	0.2	1.1	0.9	0.0	0.0	2.3
	# Succ	0.3	1.8	15.6	13.9	0.5	0.0	32.2
Scenario 6	u^{true}	83.6	87.0	88.8	90.7	**92.6**	89.6	
	% Sel	0	0	1	45	45	2	7
	# Pats	4.4	6.3	10.6	23.5	12.6	0.7	58.1
	# HEM	0.1	0.2	0.4	1.1	0.7	0.0	2.6
	# Succ	0.4	1.8	4.3	13.1	8.7	0.4	28.7
Scenario 7	u^{true}	87.5	83.4	82.1	87.5	89.8	**91.8**	
	% Sel	0	0	0	1	48	42	10
	# Pats	4.6	5.0	5.2	8.7	23.1	10.3	56.9
	# HEM	0.1	0.2	0.2	0.4	1.3	0.6	2.9
	# Succ	0.5	0.7	0.9	3.5	12.8	7.2	25.5
Scenario 8	u^{true}	82.1	80.5	78.5	75.2	69.7	59.9	
	% Sel	0	0	0	0	0	0	100
	# Pats	7.2	7.8	0.4	0.0	0.0	0.0	15.4
	# HEM	2.1	2.6	0.1	0.0	0.0	0.0	4.9
	# Succ	4.2	4.9	0.2	0.0	0.0	0.0	9.4
Scenario 9	u^{true}	79.9	82.2	83.9	85.1	86.0	85.9	
	% Sel	0	0	0	0	1	6	93
	# Pats	4.6	5.1	5.9	6.9	9.4	8.6	40.6
	# HEM	0.1	0.2	0.2	0.3	0.5	0.6	1.9
	# Succ	0.4	0.8	1.4	2.2	3.6	3.5	11.9

NOTES: A dose x is unacceptable if either $\bar{\pi}_H(x, \boldsymbol{\theta}) > 0.10$ or $\pi_S(x, \boldsymbol{\theta}) < 0.60$ with posterior probability > 0.95. Utilities of unacceptable doses have a gray background. The highest utility among acceptable doses is given in boldface.

(13). Changing only these two design parameters to reflect this hypothetical oncology setting, we resimulated U^{opt} + Acc and U^{opt} + Acc + AR_2 to assess the effect of including AR in the design, under Scenarios 1–7. Table 5 summarizes the results. In terms of both R_{select} and R_{treat}, the design U^{opt} + Acc performs slightly better in Scenarios 1–3, where the optimal dose is close to the starting dose, but U^{opt} + Acc + AR_2 is greatly superior in Scenarios 5–7, where the optimal dose is far away from the starting dose. The general message is that including AR may be regarded as an insurance policy against extremely poor behavior in some cases, with the price being a small drop in R_{select} and R_{treat} in other cases.

Table 5. Simulation study comparing U^{opt} + Acc and U^{opt} + Acc + AR_2 for a hypothetical trial where the acceptability limits $\pi_H(x) \le 0.25$ and $\pi_S(x) \ge 0.40$ are appropriate

		Scenario						
Design		1	2	3	4	5	6	7
U^{opt} + Acc	R_{select}	96	93	99	91	82	61	65
	R_{treat}	96	92	98	90	73	53	50
	% None	0	0	0	0	0	0	2
	# Pats	60.0	60.0	60.0	60.0	60.0	60.0	59.3
	# HEM	4.1	2.7	2.4	2.8	2.2	2.5	2.8
	# Succ	36.7	40.7	39.4	33.2	29.0	22.9	21.7
U^{opt} + Acc + AR_2	R_{select}	95	93	95	96	91	84	90
	R_{treat}	92	83	88	88	76	67	66
	% None	0	0	0	0	0	0	1
	# Pats	60.0	60.0	60.0	60.0	59.9	59.9	59.5
	# HEM	5.7	4.6	2.7	3.7	2.6	2.7	3.1
	# Succ	36.7	39.1	36.7	35.4	32.6	29.0	27.6

We also evaluated our design's performance under simpler versions of the model obtained by dropping $f(Z)$, Y_G, or both from the linear term (3). We found that dropping $f(Z)$ results in a design that escalates far too slowly or often fails to escalate when higher doses have higher utility. Dropping Y_G, so that neither $\pi_E(x, \boldsymbol{\theta})$ nor $\pi_H(x, \boldsymbol{\theta})$ depends on Y_G, causes the design to stop early far too often in cases where Y_E or Y_H actually are associated with Y_G. As a final comparator, we used the bivariate CRM (Braun 2002) with Success as "efficacy" and HEM as "toxicity," since this method is model-based but simpler than our method (supplementary Table 8). Because the bivariate CRM requires that the probability of efficacy must increase with dose, and our elicited prior has nonmonotone $\pi_S(x)$, to implement it, we adjusted the prior mean success probabilities to be nearly flat over the last four doses rather than decreasing. For the one stopping rule allowed by the available bivariate CRM software, we chose the toxicity rule with upper limit 0.10. The simulation results show that the bivariate CRM performs much worse than our method in six scenarios (2 through 7), and about the same in the other three.

To evaluate robustness to the model assumptions, we perturbed the scenarios' true probabilities in each of three ways: (1) mixing the true beta score distribution with a piecewise uniform score distribution in various proportions (supplementary Table 5), (2) changing the assumed optimal Z scores for Pr(EXT) and Pr(HEM) (supplementary Table 6), and (3) increasing the true risks by various amounts when GSS is not achieved (supplementary Table 7). We found that, when the model was misspecified by these perturbations, in most cases the early stopping probability tended to increase, but the dose selection performance (both R_{select} and R_{treat}) remained relatively high.

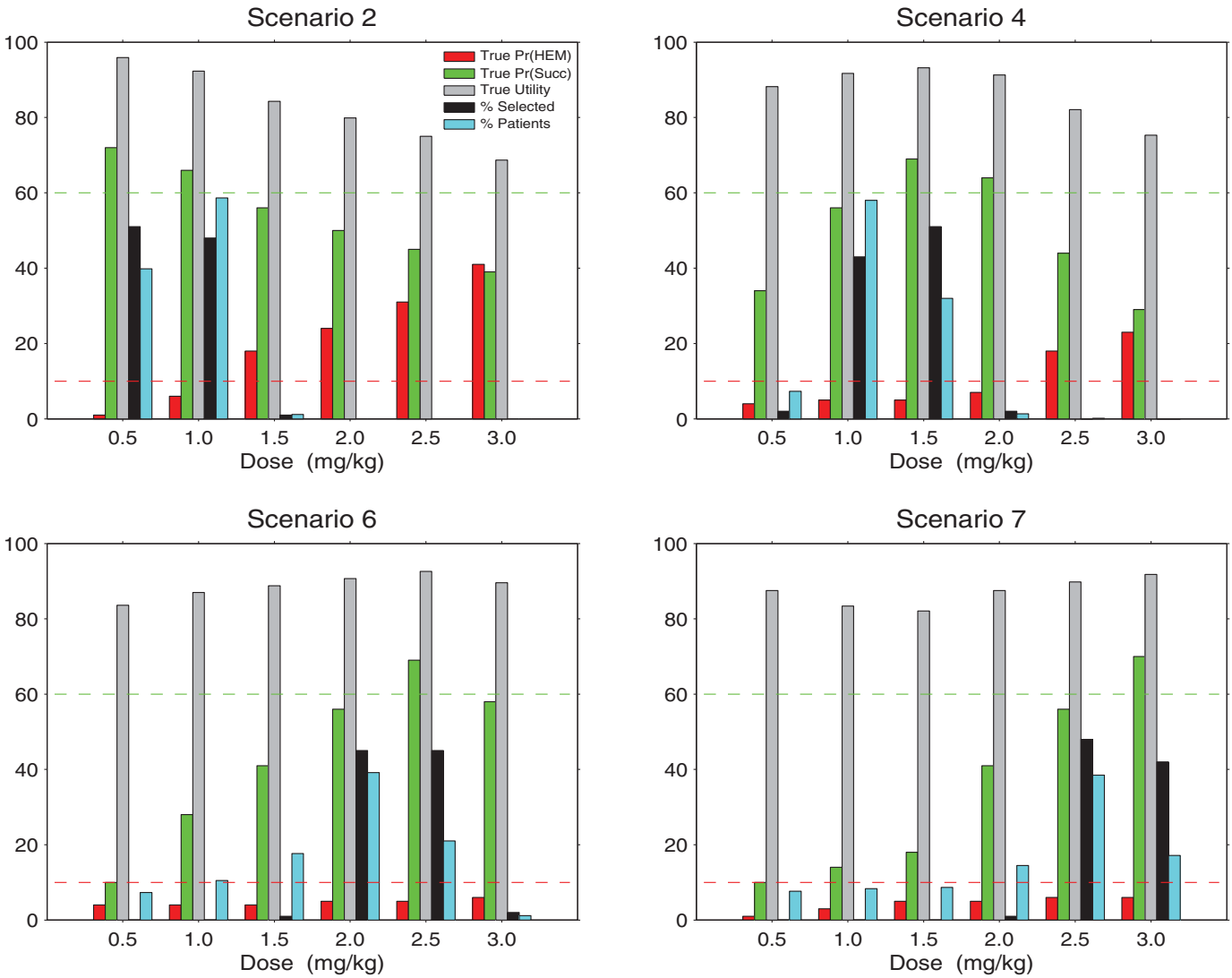


Figure 1. Simulation results for the design U^{opt} + Acc using the greedy algorithm with safety and efficacy acceptability rules, under four selected scenarios. For convenience, probabilities as percentages, utilities, and selection percentages are given together in the same plot. Horizontal dashed lines show the upper limit 10% for π_H(dose) and lower limit 60% for π_S(dose).

Table 6. Comparison of results obtained by conducting the trial using the consensus utility with the design U^{opt} + Acc, but analyzing the resulting data using each of the alternative utilities

		Scenario								
Utility used for analysis		1	2	3	4	5	6	7	8	9
Consensus	R_{select}	95	93	99	95	93	89	88	96	99
	R_{treat}	96	92	98	92	79	69	64	96	73
Alternative 1:	R_{select}	87	92	95	86	90	88	85	76	98
GSS more important	R_{treat}	87	90	92	77	73	66	55	75	64
Alternative 2:	R_{select}	96	93	99	97	95	90	90	97	98
EXT more important	R_{treat}	97	92	99	94	84	73	72	97	75
Alternative 3:	R_{select}	92	93	99	98	94	90	84	93	73
HEM more important	R_{treat}	93	92	99	97	81	73	68	93	73

NOTE: R_{select} values have a gray background in Scenarios 8 and 9 because these have no acceptable doses.

A key issue is that the elicited neonatologists' consensus utilities are subjective, and others may have different utilities. To address this, we carried out two sensitivity analyses. For the first, which addresses this concern by anticipating how the trial results may be interpreted by others after its completion, we evaluated the results of the trial conducted as before using the elicited consensus utility, but analyzed using each of the three alternative utilities given in Table 1. These alternative utilities numerically reflect the respective viewpoints that, compared to the consensus utility, GSS is more important, EXT is more important, or HEM is more important. Note that, for each alternative, several numerical values of $U(\boldsymbol{y})$ differ substantially from the corresponding values of the consensus utility. For the second sensitivity analysis, we simulated the trial conducted using each alternative utility in place of the consensus utility. The results, summarized in Tables 6 and 7, show that the design appears to be quite robust to changes in numerical utility values, either for

Table 7. Comparison of results if different alternative utilities are used to conduct the trial in place of the consensus utility, for the design U^{opt} + Acc

		Scenario								
Utility		1	2	3	4	5	6	7	8	9
Consensus	R_{select}	95	93	99	95	93	89	88	96	99
	R_{treat}	96	92	98	92	79	69	64	96	73
	% None	4	0	1	2	4	7	10	100	93
	# Pats	58.9	59.8	59.8	59.4	59.0	58.1	56.9	15.4	40.6
	# HEM	4.2	2.6	2.4	2.9	2.3	2.6	2.9	4.9	1.9
	# Succ	36.4	40.6	39.3	35.1	32.2	28.7	25.5	9.4	11.9
Alternative 1:	R_{select}	88	89	97	89	92	90	86	72	97
GSS more important	R_{treat}	87	89	92	78	75	68	55	75	64
	% None	4	1	1	2	4	6	9	99	94
	# Pats	59.1	59.8	59.7	59.4	59.0	58.3	57.2	15.4	40.3
	# HEM	4.4	2.8	2.4	2.9	2.4	2.7	2.9	4.8	1.9
	# Succ	36.8	40.4	39.3	35.5	32.6	29.4	26.0	9.4	11.8
Alternative 2:	R_{select}	96	94	99	96	95	89	90	98	96
EXT more important	R_{treat}	97	92	99	94	85	73	72	97	75
	% None	4	0	1	2	3	6	9	100	93
	# Pats	58.9	59.9	59.7	59.3	59.0	58.4	57.5	15.4	40.8
	# HEM	4.2	2.5	2.4	2.9	2.4	2.7	2.9	4.9	1.9
	# Succ	36.5	40.9	39.2	34.9	32.5	29.2	25.9	9.4	12.0
Alternative 3:	R_{select}	93	94	99	98	93	90	84	94	74
HEM more important	R_{treat}	93	92	98	97	80	73	67	93	73
	% None	4	1	1	2	3	6	9	100	94
	# Pats	59.0	59.9	59.7	59.2	59.1	58.3	57.3	15.4	40.7
	# HEM	4.1	2.2	2.4	2.8	2.3	2.6	2.9	4.8	1.9
	# Succ	36.4	41.3	39.3	34.8	32.0	28.6	25.6	9.4	11.9

NOTE: R_{select} values have a gray background in Scenarios 8 and 9 because these have no acceptable doses.

trial conduct or data analysis. Thus, the trial results based on the consensus utility should be acceptable for a wide audience of other neonatologists who may have differing opinions.

6. DISCUSSION

We have presented a Bayesian model and method for choosing sedative doses in a clinical trial involving newborn babies being treated for RDS with the INSURE procedure. The design is based on elicited utilities of three binary clinical outcome variables. The proposed method sequentially optimizes doses using posterior expected utilities, with additional restrictions to exclude doses that are likely to be either unsafe or inefficacious.

Using the utility function to reduce the three-dimensional outcome (Y_G, Y_E, Y_H) to a single quantity may be regarded as a technical device that is ethically desirable. Comparison of U^{opt} to U^{opt} + Acc clearly shows that use of the greedy utility-based algorithm per se gives a design that is ethically unacceptable, but that this can be fixed by adding dose admissibility criteria. As shown by the hypothetical example where the limits on $\pi_{U,H}$ and $\pi_{U,S}$ were replaced with different numerical values that might be more appropriate in an oncology trial (Table 5), in some settings using AR may be preferable.

Important caveats are that a particular utility function is setting-specific, and it may not be reasonable to attempt to include outcomes having dramatically different clinical importance in the utility function. For example, in cancer trials it may not be possible to construct a utility including both death and tumor response. This is a practical and ethical limitation of this type of utility-based methodology.

Application of a complex outcome-adaptive clinical trial design presents several important practical challenges. The first step, which has been our focus here, is to establish the design, write the necessary computer program, and obtain approval from the physicians who will treat patients enrolled in the trial. Key elements in implementation include (1) establishing a database and procedure for data entry in the clinic, (2) obtaining approval of the trial protocol by the Institutional Review Boards of all participating medical centers, and (3) implementing the design using the database and computer program as patients are enrolled, treated, and evaluated. Updating the database in real time, which is critically important for outcome-adaptive designs, is challenging since it requires research nurses or data managers to enter patient outcomes in a timely manner. The required data usually are simple, however. For example, the vector (x, Z, Y_E, Y_H) is all that is required by the propofol trial design. Computing each assigned dose is straightforward, since it requires only one run of the computer program using the updated database.

Upon completion of the trial, in addition to recommending an optimal dose, inferences from the final data will include summaries of the posterior distributions of the key outcome probabilities, including $\pi_G(x, \boldsymbol{\theta}_Z)$, $\bar{\pi}_E(x, \boldsymbol{\theta}_E, \boldsymbol{\theta}_Z)$, $\bar{\pi}_H(x, \boldsymbol{\theta}_H, \boldsymbol{\theta}_Z)$, and the success event probability, $\pi_S(x, \boldsymbol{\theta}_E, \boldsymbol{\theta}_Z)$. This will be

done by cross-tabulating posterior means and 95% credible intervals (cis) with dose x. This table also will include the posterior means $u(x \mid \text{data}_N)$ and 95% cis of the utilities $\bar{U}(x \mid \boldsymbol{\theta})$, which provide a set of natural summary statistics for evaluating and comparing the doses. Corresponding plots of the posteriors will provide a graphical illustration of what has been learned about each of these parametric quantities. As suggested in our sensitivity analyses, the summaries of $u(x \mid \text{data}_N)$ could be repeated for each of several reasonable alternative utilities, such as those in Table 1. Finally, it also will be important to include nonmodel-based summaries of the empirical distribution of the sedation score Z and the count of each of event G, E, H, and S for each dose.

The propofol trial design synthesizes ideas from several areas, including phase I-II dose finding, sequential optimization, decision analysis, Bayesian statistics, and intervention in preterm newborns. For future studies in neonatal care and similar medical settings, several potential extensions and improvements are worth mentioning. More general regimes might include multiple agents, two or more different administration schedules, or more than one cycle of therapy. Use of multicategory ordinal rather than binary outcomes would provide a more refined assessment of treatment or dose effects, and thus a more informed basis for decision making. Accounting for effects of known prognostic covariates to optimize so-called "individualized" therapies also is highly desirable, although such a design is likely to be complex and logistically difficult, since it would require rapid evaluation of the necessary covariates and adaptive computation of the dose in real time.

Designing clinical trials in children is challenging, both technically and ethically. Successful use of this type of statistical methodology in the propofol trial may serve as proof-of-concept, and possibly provide a bridge to future pediatric trials using similar approaches.

7. SUPPLEMENTARY MATERIALS

The Supplement contains the assessment criteria for determining sedation score (Table S1); prior means and variances (Table S2); elicited prior means and interval probabilities for sedation score, and mean utilities, under each simulation scenario (Tables S3.1S3.9); a graph of the sedation state distribution as a function of dose under each simulation scenario (Supplementary Figure 1); and plots of the true means for Pr(GSS), Pr(EXT), Pr(HEM), Pr(Success) and utility under each simulation scenario (Supplementary Figure 2). Additional simulation results are summarized for the design Uopt + Acc using the greedy algorithm with safety and efficacy dose acceptability rules (Supplementary Figure 3); sensitivity of the design to prior standard deviations (Table S4); effects of mixtures for the true sedation score distribution (Table S5); different true optimal sedation scores (Table S6); different amounts of increased risk when a good sedation score is not achieved (Table S7); and comparison to the bivariate Continual Reassessment Method (Table S8).

[Received May 2013. Revised March 2014.]

REFERENCES

Atkinson, A. C., Donev, A., and Tobias, R. (2006), *Optimal Experimental Designs, With SAS. Oxford Statistical Series, 34*, London: Oxford University Press. [932]

Azriel, D., Mandel, M., and Rinott, Y. (2011), "The Treatment Versus Experimentation Dilemma in Dose-Finding Studies," *Journal of Statistical Planning and Inference*, 141, 2759–2768. [936]

Bohlin, K., Gudmundsdottir, T., Katz-Salamon, M., Jonsson, B., and Blennow, M. (2007), "Implementation of Surfactant Treatment During Continuous Positive Airway Pressure," *Journal of Perinatology*, 27, 422–427. [931]

Bornkamp, B., Bretz, F., Dette, H., and Pinheiro, J. (2011), "Response-Adaptive Dose-Finding Under Model Uncertainty," *Annals of Applied Statistics*, 5, 1611–1631. [932]

Braun, T. M. (2002), "The Bivariate Continual Reassessment Method: Extending the CRM to Phase I Trials of Two Competing Outcomes," *Controlled Clinical Trials*, 23, 240–256. [939]

Chen, Y., and Smith, B. J. (2009), "Adaptive Group Sequential Design for Phase II Clinical Trials: A Bayesian Decision Theoretical Approach," *Statistics in Medicine*, 28, 3327–3362. [933]

Cheung, Y.-K. (2011), *Dose Finding by the Continual Reassessment Method*, New York: Chapman and Hall/CRC Press. [932]

Chevret, S. (ed.) (2006), *Statistical Methods for Dose-Finding Experiments*, West Sussex, UK: Wiley. [932]

Christen, J., Muller, P., Wathen, K., and Wolf, J. (2004), "Bayesian Randomized Clinical Trials: A Decision-Theoretical Sequential Design," *Canadian Journal of Statistics*, 32, 387–402. [933]

Dette, H., Bretz, F., Pepelyshev, A., and Pinhiero, J. (2008), "Optimal Designs for Dose-Finding Studies," *Journal of American Statistical Association*, 103, 1225–1237. [932]

Fedorov, V., and Leonov, S. L. (2001), "Optimal Design of Dose Response Experiments: A Model-Oriented Approach," *Drug Information Journal*, 35, 1373–1383. [932]

Ferrari, S. L. P., and Cribari-Neto, F. (2004), "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics*, 31, 799–815. [934]

Ghanta, S., Abdel-Latif, M. E., Lui, K., Ravindranathan, H., Awad, J., and Oei, J. (2007), "Propofol Compared With the Morphine, Atropine, and Suxamethonium Regimen as Induction Agents for Neonatal Endotracheal Intubation: A Randomized, Controlled Trial," *Pediatrics*, 119, 1248–1255. [931]

Gittins, J. C. (1979), "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society,* Series B, 41, 148–177. [936]

Houede, N., Thall, P. F., Nguyen, H., Paoletti, X., and Kramar, A. (2010), "Utility-Based Optimization of Combination Therapy Using Ordinal Toxicity and Efficacy in Phase I/II Trials," *Biometrics*, 66, 532–540. [933]

Hummel, P., Puchalski, M., Creech, S. D., and Weiss, M. G. (2008), "Clinical Reliability and Validity of the N-PASS: Neonatal Pain, Agitation and Sedation Scale With Prolonged Pain," *Journal of Perinatology*, 28, 55–60. [932]

Leung, D. H., and Wang, Y. G. (2001), "A Bayesian Decision Approach for Sample Size Determination in Phase II Trials," *Biometrics*, 57, 309–312. [933]

Lewis, R. J., Lipsky, A. M., and Berry, D. A. (2007), "Bayesian Decision-Theoretic Group Sequential Clinical Trial Design Based On a Quadratic Loss Function: A Frequentist Evaluation," *Clinical Trials*, 4, 5–14. [933]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed), New York: Chapman and Hall. [933]

Morita, S., Thall, P. F., and Mueller, P. (2008), "Determining the Effective Sample Size of a Parametric Prior," *Biometrics*, 64, 595–602. [935]

——— (2010), "Evaluating the impact of prior assumptions in Bayesian biostatistics," *Statistics in Biosciences*, 2, 1–17. [935]

Murdoch, S. D., and Cohen, A. T. (1999), "Propofol-Infusion Syndrome in Children," *Lancet*, 353, 2074–2075. [931]

Nelsen, R. B. (1999), *An Introduction to Copulas. Lecture Notes in Statistics* (Vol. 139), New York: Springer-Verlag. [933]

O'Quigley, J., Pepe, M., and Fisher, L. (1990), "Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer," *Biometrics*, 46, 33–48. [932]

Oron, A. P., and Hoff, P. D. (2013), "Small-Sample Behavior of Novel Phase I Cancer Trial Designs," *Clinical Trials*, 10, 63–80. [936]

Robbins, H. (1952), "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 58, 527–535. [936]

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer. [935]

Sammartino, M., Garra, R., Sbaraglia, F., and Papacci, P. (2010), "Propofol Overdose in a Preterm Baby: May Propofol Infusion Syndrome Arise in Two Hours?," *Paediatric Anaesthesia*, 20, 973–974. [931]

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010), "Improved Estimators for a General Class of Beta Regression Models," *Journal of Computational Statistics and Data Analysis*, 54, 348–366. [935]

Stallard, N. (1998), "Sample Size Determination for Phase II Clinical Trials Based On Bayesian Decision Theory," *Biometrics*, 54, 279–294. [933]

Stallard, N., and Thall, P. F. (2001), "Decision-Theoretical Designs for Pre-Phase II Screening Trials in Oncology," *Biometrics*, 57, 1089–1095. [933]

Stallard, N., Thall, P. F., and Whitehead, J. (1999), "Decision Theoretical Designs for Phase II Clinical Trials With Multiple Outcomes," *Biometrics*, 55, 971–977. [933]

Stevens, T. P., Harrington, E. W., Blennow, M., and Soll, R. F. (2007), "Early Surfactant Administration With Brief Ventilation vs. Selective Surfactant and Continued Mechanical Ventilation for Preterm Infants With or at Risk for Respiratory Distress Syndrome," *Cochrane Database of Systematic Reviews*, 4, CD003063. [931]

Sutton, R. S., and Barto, A. G. (1998), *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press. [936]

Thall, P. F., and Nguyen, H. Q. (2012), "Adaptive Randomization to Improve Utility-Based Dose-Finding With Bivariate Ordinal Outcomes," *Journal of Biopharmaceutical Statistics*, 22, 785–801. [933,935,936]

Thall, P. F., and Russell, K. T. (1998), "A Strategy for Dose Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials," *Biometrics*, 54, 251–264. [932]

Thall, P. F., Szabo, A., Nguyen, H. Q., Amlie-Lefond, C. M., and Zaidat, O. O. (2011), "Optimizing the Concentration and Bolus of a Drug Delivered by Continuous Infusion," *Biometrics*, 67, 1638–1646. [933]

Vanderhaegen, J., Naulaers, G., Van Huffel, S., Vanhole, C., and Allegaert, K. (2010), "Cerebral and Systemic Hemodynamic Effects of Intravenous Bolus Administration of Propofol in Neonates," *Neonatology*, 98, 57–63. [931]

Verder, H., Robertson, B., Greisen, G., Ebbesen, F., Albertsen, P., Lundstrom, K., and Jacobsen, T.; for the Danish-Swedish Multicenter Study Group, (1994), "Surfactant Therapy and Nasal Continuous Positive Airway Pressure for Newborns With Respiratory Distress Syndrome," *The New England Journal of Medicine*, 331, 1051–1055. [931]

Wathen, J. K., and Thall, P. F. (2008), "Bayesian Adaptive Model Selection for Optimizing Group Sequential Clinical Trials," *Statistics in Medicine*, 27, 5586–5604. [933]

Welzing, L., Kribs, A., Huenseler, C., Eifinger, F., Mehler, K., and Roth, B. (2009), "Remifentanil for INSURE in Preterm Infants: A Pilot Study for Evaluation of Efficacy and Safety Aspects," *Acta Paediatrica*, 98, 1416–1420. [931]

Williams, D. A. (1982), "Extra Binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148. [934]