

UC Davis

UC Davis Previously Published Works

Title

Advantages of Bayesian monitoring methods in deciding whether and when to stop a clinical trial: an example of a neonatal cooling trial

Permalink

<https://escholarship.org/uc/item/0rn8h4s0>

Journal

Trials, 17(1)

ISSN

1468-6708

Authors

Pedroza, Claudia
Tyson, Jon E
Das, Abhik
[et al.](#)

Publication Date

2016-12-01

DOI

10.1186/s13063-016-1480-4


Peer reviewed

RESEARCH

Open Access



Advantages of Bayesian monitoring methods in deciding whether and when to stop a clinical trial: an example of a neonatal cooling trial

Claudia Pedroza^{1*} , Jon E. Tyson¹, Abhik Das², Abbot Laptook³, Edward F. Bell⁴, Seetha Shankaran⁵ and for the Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network

Abstract

Background: Decisions to stop randomized trials are often based on traditional *P* value thresholds and are often unconvincing to clinicians. To familiarize clinical investigators with the application and advantages of Bayesian monitoring methods, we illustrate the steps of Bayesian interim analysis using a recent major trial that was stopped based on frequentist analysis of safety and futility.

Methods: We conducted Bayesian reanalysis of a factorial trial in newborn infants with hypoxic-ischemic encephalopathy that was designed to investigate whether outcomes would be improved by deeper (32 °C) or longer cooling (120 h), as compared with those achieved by standard whole body cooling (33.5 °C for 72 h). Using prior trial data, we developed neutral and enthusiastic prior probabilities for the effect on predischarge mortality, defined stopping guidelines for a clinically meaningful effect, and derived posterior probabilities for predischarge mortality.

Results: Bayesian relative risk estimates for predischarge mortality were closer to 1.0 than were frequentist estimates. Posterior probabilities suggested increased predischarge mortality (relative risk > 1.0) for the three intervention groups; two crossed the Bayesian futility threshold.

Conclusions: Bayesian analysis incorporating previous trial results and different pre-existing opinions can help interpret accruing data and facilitate informed stopping decisions that are likely to be meaningful and convincing to clinicians, meta-analysts, and guideline developers.

Trial registration: ClinicalTrials.gov NCT01192776. Registered on 31 August 2010.

Keywords: Bayesian methods, Factorial trial, Hypothermia, Phase III trial, Stopping rules, Trial monitoring

Background

Decisions to stop randomized trials are often based on traditional *P* values from sequential monitoring methods that diminish the possibility of making false positive claims of benefit or harm with repeated interim analyses [1, 2]. However, these decisions are often unconvincing to clinicians or guideline development panels [3]. The appropriate stopping guidelines are unclear [3–12], and the decisions

are often difficult for data and safety monitoring committee (DSMC) members, who bear the responsibility for both avoiding recommendations to continue trials too long once harm or futility is suspected and for stopping trials too soon because of misleading interim findings. Safety concerns for patients enrolled in the current trial are self-evident in the former case. In the latter case, an erroneous conclusion that a truly beneficial therapy is ineffective or harmful can also be detrimental to a very large number of future patients, particularly if the therapy would in fact reduce rates of major adverse outcomes.

Bayesian methods for monitoring efficacy, safety, and futility have been proposed [13–24]. Bayesian approaches

* Correspondence: claudia.pedroza@uth.tmc.edu

¹Center for Clinical Research and Evidence-Based Medicine, McGovern Medical School at The University of Texas Health Science Center at Houston, 6431 Fannin St, MSB 2.106, Houston, TX 77030, USA
Full list of author information is available at the end of the article

have potential advantages [25–27] that include incorporating the results from prior trials to better assess the likelihood of treatment benefit or harm and ensure that treatment recommendations are well justified, based on all relevant trials [16]. Another advantage of a Bayesian approach is that uncertainty from all parameter estimates is accounted for in reported summaries, which is particularly important with sparse data [28, 29]. For monitoring of trials, at a given interim analysis the posterior probability of the treatment effect is computed from the prior probability (referred to in this paper as “the prior”) and the interim data. Then DSMCs can weigh the current evidence for benefit, harm, or futility from the posterior probability and decide whether to continue recruitment, or to pause or terminate the trial.

A Bayesian approach can also incorporate a wide range of viewpoints and indicate the magnitude of the difference between treatment groups that would be needed at the end of the trial to convince those who are skeptical, as well as those who were enthusiastic about the value of the therapy prior to the trial [16, 30]. Decisions to stop a trial early based on the best available data from all relevant trials and on the identification of clinically meaningful differences are most likely to be convincing to meta-analysts, practice guideline developers, and clinicians.

Despite these advantages, only a small percentage of Phase III trials have adopted Bayesian monitoring methods, largely due to the seemingly daunting task of specifying priors [31], the perceived drawback of their subjective nature, computational burden [21], and lack of familiarity with implementation and interpretation of these methods [32, 33]. The objective of this report is to illustrate the use of a Bayesian approach and its advantages in trial monitoring with a concrete example of a major randomized trial of hypothermia in neonates with hypoxic-ischemic encephalopathy [34], a condition associated with a high risk of death or severe impairment. This trial used a frequentist monitoring plan and was stopped early for futility and safety concerns though stopping boundaries were not crossed. We present how Bayesian stopping guidelines can be specified using clinically meaningful treatment effects. We show how information from prior trials can be incorporated and utilized in addressing whether even an enthusiastic (about treatment benefit) clinician or investigator should be convinced by negative interim findings.

Methods

Optimizing Cooling Trial

Therapeutic hypothermia may be considered the most important advance in decades in the treatment of newborn infants with hypoxic-ischemic encephalopathy, a condition probably resulting from severe acute hypoxia-ischemia occurring within hours before birth [35, 36]. Whole body cooling to 33.5 °C for 72 h was shown to reduce the risk of

death or neurodevelopmental impairment at 18–22 months of age by 28 % (relative risk, 0.72; 95 % confidence interval, 0.54–0.95) among term infants in a randomized trial conducted by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Neonatal Research Network (NRN) [37]. However, even with cooling, only 56 % of the infants survived without severe or moderate impairment, and it was postulated that refined approaches to cooling would further improve the outcome. Based on evidence from studies on animals and neonates [38–42], the NRN launched a trial (Optimizing Cooling Trial) in 18 centers to assess whether the use of deeper cooling (to 32 °C), longer cooling (for 120 h), or both would further increase survival without impairment over that achieved with standard cooling [34–36].

Study design

The Optimizing Cooling Trial utilized a 2 × 2 factorial design to test depth and duration of cooling. Infants of 36 weeks gestational age or older with severe acidosis or need for resuscitation at birth with moderate or severe hypoxic-ischemic encephalopathy were randomized to four hypothermia groups: 33.5 °C for 72 h, 32.0 °C for 72 h, 33.5 °C for 120 h, or 32.0 °C for 120 h. Randomization was stratified by center and severity of hypoxic-ischemic encephalopathy (moderate or severe) with a dichotomous composite primary outcome of death or moderate or severe disability at 18–22 months of age. The sample size calculation was based on frequentist marginal analyses of the two cooling factors (comparing 33.5 °C with 32 °C and 72 h with 120 h) and assumed that there was no large interaction between duration and depth of cooling. The trial planned to enroll a total of 726 infants to detect the hypothesized relative risk (RR) of 0.73 in either factor with expected primary outcome rates of 37.5 % and 27.5 % for the two marginal groups with 80 % power and a two-sided α of 0.05 (Table 1). A completed CONSORT 2010 checklist is provided in Additional file 1.

Table 1 Hypothesized rates of primary outcome of death or moderate or severe impairment at 18–22 months: these rates were used for sample size calculation

	Depth of cooling		Margin
	33.5 °C	32.0 °C	
Duration of cooling			
72 h	45 %	30 %	37.5 %
120 h	30 %	25 %	27.5 %
Margin	37.5 %	27.5 %	

Ethical considerations

The research study was approved by the local institutional review board of the Women and Infants Hospital of Rhode Island, Case Western Reserve University, the Children’s Mercy Hospital, the University of Cincinnati Medical Center, the Cincinnati Children’s Hospital Medical Center, the Good Samaritan Hospital, the Duke University School of Medicine, the University of North Carolina at Chapel Hill, Emory University, Grady Memorial Hospital, Indiana University, the Nationwide Children’s Hospital, RTI International, Stanford University, the University of Alabama at Birmingham Health System, the University of California—Los Angeles, the University of Iowa, Mercy Medical Center, the University of New Mexico Health Sciences Center, the University of Pennsylvania, the University of Texas Southwestern Medical Center at Dallas, the University of Texas Health Science Center at Houston, Wayne State University, the University of Michigan Medical Center, the University of Rochester Medical Center, and the University of Buffalo Women’s and Children’s Hospital of Buffalo. Written informed consent was obtained from a parent or guardian for each enrolled infant.

Trial monitoring

The trial was monitored for safety outcomes of cardiac arrhythmia, persistent acidosis, major vessel thrombosis, alteration of skin integrity, major bleeding, and death using Pocock boundaries constructed to maintain an overall α of 0.05 for each outcome. Prespecified safety interim analyses were planned after the first 50 infants were enrolled and then for every 25 infants thereafter. The results of each interim analysis were reviewed by the independent DSMC. Though not specified in the study protocol, futility analyses of pre-discharge mortality were also performed at the request of the DSMC during the final interim check. These marginal analyses (comparing two cooling factors) were performed by calculating conditional power using the hypothesized treatment effect for the primary outcome as the alternative hypothesis.

Interim analysis using frequentist approaches

Enrollment started in October 2010. Following the recommendation of the DSMC, the NICHD Director stopped the trial for concerns about safety and futility on 27 November 2013, after the eighth DSMC review. A total of 364 infants had been enrolled (Additional file 2). The observed rates of pre-discharge mortality are shown in Table 2. The RR (95 % confidence interval) for pre-discharge mortality, adjusted for level of hypoxic-ischemic encephalopathy and center, was 1.37 (0.92–2.04) for the duration of cooling comparison, and 1.24 (0.69–2.25) for the depth of cooling comparison. Although the data did not cross the stopping boundaries for safety, the

Table 2 Observed rates of pre-discharge mortality for the Optimizing Cooling Trial

	Depth of cooling		Margin
	33.5 °C	32.0 °C	
Duration of cooling			
72 h	7 % (7/95)	14 % (13/90)	11 %
120 h	16 % (15/96)	17 % (14/83)	16 %
Margin	12 %	16 %	

conditional power of 2 % for both marginal comparisons indicated a low probability of finding a statistically significant reduction in pre-discharge deaths were the study to continue to completion.

Data on the primary outcome of death or moderate or severe disability at 18 to 22 months were available for only a few infants and hence did not play a role in the decision to stop the trial, a reason that some observers might question this decision. Moreover, the pre-discharge mortality with standard cooling (7 %) in the Optimizing Cooling Trial was less than half that in a prior NRN trial (19 %) using the same eligibility criteria [37]—and the mortality rates for all three experimental groups (longer cooling, deeper cooling, both) were less than 19 %. Thus, the very low mortality in the standard group might be a “random low” as in other trials when the interim findings after an equal or larger number of patients proved to be quite misleading [3, 4, 43, 44]. However, this low mortality might be partly or fully due to improvements in care or changes in the patient population between the two trials, especially considering the 7 year gap between them. Compared with cooled infants in the prior NRN trial, infants in the Optimizing Cooling Trial were less likely to have severe hypoxic-ischemic encephalopathy (23 % versus 32 %), to be intubated at birth (79 % versus 95 %), or to have seizures (29 % versus 43 %).

What could a Bayesian monitoring approach add to an interim analysis?

In developing stopping guidelines, data from the prior NRN trial could be used in identifying what negative interim findings for pre-discharge mortality would be convincing to enthusiasts as well as skeptics or neutral clinicians. While no prior data exist for longer or deeper cooling, the observed proportion of pre-discharge mortality for the cooled group in the prior trial can inform the expected rate for the standard group in the Optimizing Cooling Trial, as well as realistic treatment effects for this outcome.

Another advantage of a Bayesian approach is that it forces investigators to consider carefully what posterior probability of benefit or harm would justify stopping the trial—an exercise that requires close collaboration between clinical investigators and biostatisticians. The appropriate stopping probability threshold should arguably be lower

for treatment harms than benefits, meaning that less evidence might be required for presence of harm than absence of benefit to stop a trial. Moreover, a particularly high probability of benefit might be required for therapies that are invasive, hazardous, or extremely expensive. Once the threshold posterior probabilities have been selected, simulations can be performed if necessary to satisfy regulatory agencies or verify acceptable frequentist characteristics (type I error, power) [21, 26, 45].

While the *P* values required by frequentist stopping guidelines can be modified based on these considerations, this is rarely done in practice. Bayesian stopping guidelines based on these considerations may thus be more flexible as well as more easily explained and more meaningful to clinicians and developers of practice guidelines. They also have the added benefit of forcing difficult but productive upfront discussions among clinical investigators and biostatisticians that can enrich all aspects of trial design.

Bayesian monitoring of the Optimizing Cooling Trial

In addition to the necessary elements of an acceptable data monitoring plan [46], three main components need to be specified for a Bayesian monitoring plan: (1) prior evidence (in some circumstances, expert opinion might also be considered) of treatment effect; (2) clinically important treatment effect(s) for the primary outcome and any important outcomes, including death, to be monitored; and (3) probability thresholds for stopping a trial (Table 3). We

illustrate how these three components can be specified in practice using the Optimizing Cooling Trial as an example.

We performed a Bayesian reanalysis of predischage mortality data from the 364 infants enrolled in the Optimizing Cooling Trial. We present posterior probabilities for the comparisons of the three hypothermia groups with standard cooling partly because the study protocol stated the possibility of terminating one or more groups for safety or futility and continuing the trial with the remaining groups. This approach of simultaneously monitoring all groups for futility was found to be superior to or as good as an approach that first assesses an interaction term between the interventions and then examining the main effects if no significant interaction is found [47]. To compare with the frequentist results, we also provide posterior probabilities for the marginal comparisons of longer cooling or deeper cooling.

Bayesian model

We used a binomial model with a log link to estimate the RR of predischage mortality with longer cooling, deeper cooling, or both, compared with standard cooling. We included the main effects of duration and depth of cooling and the interaction term to assess its magnitude. Letting y_i be the outcome of predischage death (1 = yes, 0 = no), the model is expressed as:

Table 3 Summary of key components of a Bayesian monitoring plan

Component	Specification	Example
Prior distributions	<ul style="list-style-type: none"> • Previous studies on the control rate or treatment effect can be used as prior information • Prior beliefs about the treatment effect should be elicited from experts to inform the strength of the evidence needed to convince them 	Two-arm trial of Treatment A versus B (control): <ul style="list-style-type: none"> • Evidence from three previous trials on rate of outcome under treatment B: 17 %, 25 %, 30 % • Evidence from two studies on treatment effect for different population: RR 0.98 (95 % CI: 0.73–1.3); RR 0.75 (95 % CI: 0.56–1.0) Prior distributions, center (95 % CrI): <ul style="list-style-type: none"> • Control rate: 25 % (5–55 %) • Skeptical prior for treatment effect: RR 1.10 (0.7–2.0) • Enthusiastic prior for treatment effect: RR 0.85 (0.5–1.0)
Clinically important treatment effect	<ul style="list-style-type: none"> • Investigators should specify how big a treatment effect needs to be in order to stop a trial and recommend its use or advise against it 	<ul style="list-style-type: none"> • A relative risk reduction of 15 % or more is needed to recommend treatment A, RR < 0.85 • An absolute increase of 2 % in safety outcome would be unacceptable, RD > 0.02
Stopping thresholds	<ul style="list-style-type: none"> • For each type of monitoring, i.e., safety, efficacy, or futility, the level of confidence to stop the trial early needs to be specified • For most cases, it should be based on a clinically important effect • Efficacy: the trial will stop early if the likelihood of seeing a clinically important effect is very large, even under a skeptical prior • Futility: if the likelihood of a clinically important effect is small even under an enthusiastic prior, the trial would stop early • Safety: the trial would stop if the probability of increasing harm is large enough under an enthusiastic prior 	At any preplanned interim analysis, any of these occurrences would make the DSMC consider stopping the trial: <ul style="list-style-type: none"> • Efficacy under skeptical prior: Pr(RR < 0.85) > 0.99 • Futility under enthusiastic prior: Pr(RR < 0.85) < 0.10 • Safety under enthusiastic prior: Pr(RD > 0.02) > 0.70

DSMC, data and safety monitoring committee, CI, confidence interval, CrI credible interval, Pr, probability, RD risk difference, RR, relative risk

$$y_i \sim \text{Bernoulli}(p_i),$$

$$\log(p_i) = \beta_0 + \beta_1 \text{ depth}_i + \beta_2 \text{ duration}_i + \beta_3 \text{ depth}_i \times \text{duration}_i,$$

where p_i is the probability of pre-discharge death for infant i , depth and duration are coded as 1 for 32.0 °C and 120 h (experimental interventions) and 0 otherwise, and β are regression coefficients. In the log RR scale, the marginal effects of depth and duration are given by $\beta_1 + \beta_3/2$ and $\beta_2 + \beta_3/2$, respectively, with negative values of β indicating decreased mortality. β_0 is the log probability of pre-discharge death for the standard cooling group. The effects of each of the three intervention groups compared with standard cooling are β_1 for deeper cooling given alone; β_2 for longer cooling given alone; and $\beta_1 + \beta_2 + \beta_3$ when both therapies are given. Thus, this model easily allows for the estimation of the marginal treatment effect of the two factors as well as individual treatment group comparisons [48]. We did not include center or level of hypoxic-ischemic encephalopathy variables in our analysis.

While this model gives direct estimates of RRs, it is straightforward to derive estimates for other risk measures, such as absolute risk difference (RD) or its reciprocal, the number needed to treat (see Additional file 3).

Prior distributions

We assumed independent normal prior distributions for the β regression coefficients: $\beta_0 \sim \text{normal}(\mu_0, \tau_0^2)$, $\beta_1 \sim \text{normal}(\mu_1, \tau_1^2)$, $\beta_2 \sim \text{normal}(\mu_2, \tau_2^2)$, and $\beta_3 \sim \text{normal}(\mu_3, \tau_3^2)$, where the μ and τ are, respectively, the means and standard deviations of the distributions. We specified a set of neutral prior distributions and a set of enthusiastic priors for the two marginal interventions. With both sets of priors, we assumed an expected pre-discharge mortality rate of 19 % for the standard cooling group, the observed rate in the previous NRN trial. The prior for β_0 has a mean of $\log(0.19) = -1.66$ and a standard deviation of 0.565, which increases the uncertainty observed in the previous NRN trial (0.19 in log scale) by a factor of three (to account for population differences between the two cooling trials).

Neutral priors

We centered the RR at 1.0 (mean of 0 in the log RR scale), indicating no *a-priori* difference between the treatments being compared, and used a 95 % credible interval (CrI; this interval is interpreted as having a 95 % probability of containing the true RR) of 0.33–3.0. While empirical evidence from Cochrane systematic reviews of neonatal studies [49] indicates that the great majority of observed treatment effects on mortality are in the range of 0.5–2.0, there was very little previous information on the two interventions being tested, and we broadened this

interval to allow for the possibility of greater harm or benefit. The implied neutral priors for the β coefficients have means of $\mu_0 = -1.66$ and $\mu_1 = \mu_2 = \mu_3 = 0$ and standard deviations of $\tau_0 = \tau_1 = \tau_2 = 0.565$ and $\tau_3 = 0.14$. The prior standard deviation for β_3 indicates an *a-priori* probability of 0.025 of a qualitative interaction between longer and deeper cooling (meaning that the effect of longer cooling on the outcome changes direction in the presence or absence of deeper cooling). This corresponds to specifying that the likelihood of reducing the relative risk by 24 % with longer and deeper cooling (RR = 0.76) is only 0.025 when the treatment effect is zero (RR = 1) in the presence of only one of the interventions (i.e., $\Pr[\beta_3 < \log(0.76) | \beta_1, \beta_2 = 0] = 0.025$) [50]. For sensitivity analysis, we set $\tau_3 = 0.408$, which corresponds to a 0.25 *a-priori* probability of a qualitative interaction [48].

Enthusiastic priors

Pre-discharge death rates with cooling were available only from the prior multicenter NRN trial. Based largely on this trial, we chose to center the enthusiastic priors at a treatment effect half the size of the hypothesized reductions for the primary outcome (death or impairment at 18–22 months). We used the same standard deviations as for the neutral prior. For the two marginal effects, we centered the prior at a RR of 0.85 (assumed rates of 16 % for longer cooling or deeper cooling and 19 % for standard cooling). The implied priors for the β coefficients have means $\mu_0 = -1.66$, $\mu_1 = \mu_2 = -0.1625$, and $\mu_3 = 0$. The prior for β_3 is again centered at 0, since there is no prior expectation of an interaction, with a small *a-priori* probability (0.025) of a qualitative interaction between longer and deeper cooling.

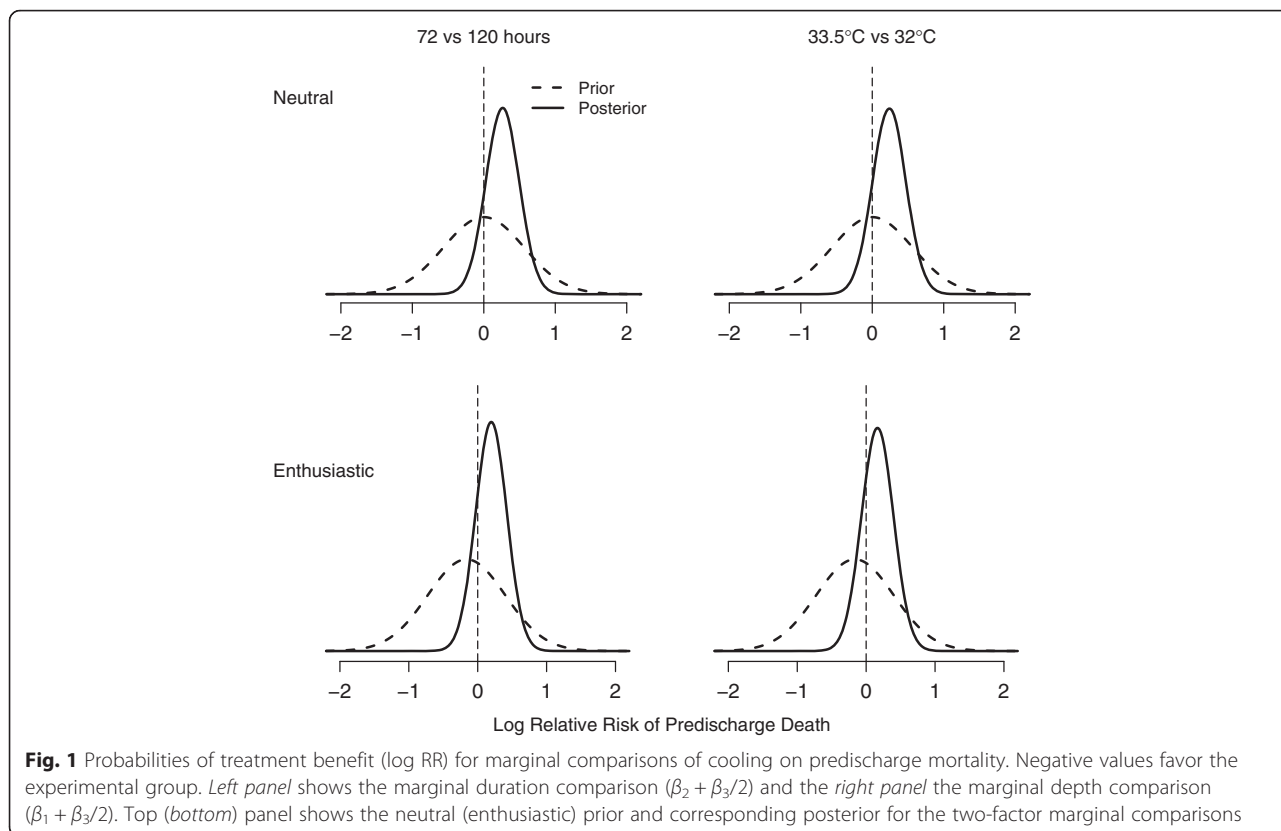
Implied priors for the marginal comparisons and for the three cooling groups compared with standard cooling are shown in Figs. 1 and 2, respectively.

Interim monitoring for futility

We defined a futility stopping guideline for pre-discharge mortality based on a clinically meaningful treatment effect size [22, 24]. If the posterior probability of this clinically meaningful mortality reduction dropped below a prespecified threshold, the DSMC would consider terminating the trial. Suppose the investigators had decided when designing the trial that the interventions had to reduce pre-discharge mortality by 10 % or more, meaning a RR < 0.90 (or an absolute RD of 2 %). If the probability of one of the cooling groups reducing pre-discharge mortality by at least 10 % fell below 0.10,

$$\Pr(\text{RR} < 0.90 | \text{Interim data}) < 0.10,$$

the cooling group would be stopped for futility. If all groups met this threshold, the trial would be stopped.



Suppose that instead of a reduction in RR, the trial investigators were interested in an absolute RD, defined as $p_{\text{control}} - p_{\text{treatment}}$ of at least 1 %. While a 1 % absolute difference in mortality has been considered a small effect in neonatal trials, differences of this magnitude for death or for composite outcomes of death or major cardiovascular events have been considered sufficiently important to justify treatment recommendations from adult trials [51–55]. With the number of disability free life years at stake in a neonatal trial, a 1 % difference in mortality or death or disability would likewise be important. The stopping guideline was that a hypothermia group would be stopped for futility if the probability of a RD of 1 % or more fell below 0.10,

$$\Pr(\text{RD} > 0.01 | \text{Interim data}) < 0.10.$$

Interim monitoring for safety

As another example of what the investigators might have selected as a stopping guideline in designing the trial, enrollment in any intervention group could be stopped if the probability of a 5 % absolute increase in mortality (the RD being less than negative 5 %) were 50 % or greater,

$$\Pr(\text{RD} < -0.05 | \text{Interim data}) > 0.50.$$

Acceptable operating characteristics would need to be verified before using these thresholds [26].

Implementation details

All models were fitted via Markov chain Monte Carlo (MCMC) methods, since the posterior distribution of the parameters is not in closed form [56]. We constrained all $p_i < 1$ in the models. All analyses were conducted using WinBUGS 1.4.3 [57]. For each model, we used three MCMC chains, each with 40,000 iterations, in addition to an initial burn-in of 4,000 chains. We examined trace plots of all the parameters to monitor convergence and calculated the Gelman–Rubin diagnostic \hat{R} . Additional file 3 gives the sample code for fitting this log binomial model in WinBUGS, and shows how to calculate RRs and RDs from the model parameters for both the two-factor marginal comparisons and the three intervention groups.

Results

For all models, trace plots showed good mixing of the 3 MCMC chains and the $\hat{R} < 1.01$ for all parameters, indicating convergence.

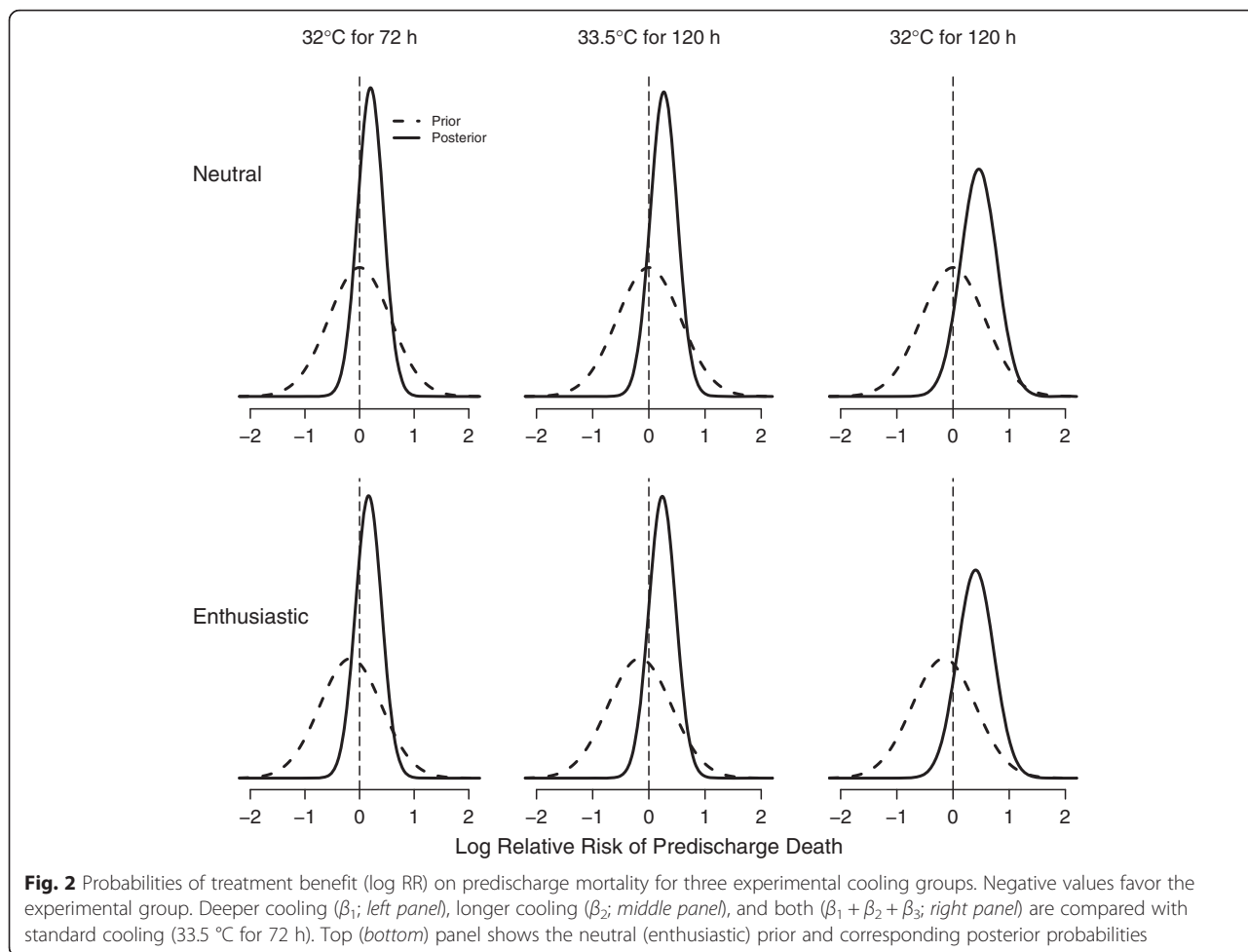


Fig. 2 Probabilities of treatment benefit (log RR) on pre-discharge mortality for three experimental cooling groups. Negative values favor the experimental group. Deeper cooling (β_1 ; left panel), longer cooling (β_2 ; middle panel), and both ($\beta_1 + \beta_2 + \beta_3$; right panel) are compared with standard cooling (33.5 °C for 72 h). Top (bottom) panel shows the neutral (enthusiastic) prior and corresponding posterior probabilities

Marginal comparisons of longer cooling and deeper cooling

Under neutral priors, the posterior distribution of the RR for longer cooling has a median of 1.30, (95 % CrI, 0.82–2.04) and an 87 % probability of increased pre-discharge mortality with longer duration (Fig. 1, top left panel). Deeper cooling has an 81 % posterior probability of increased pre-discharge death compared with standard cooling (RR: posterior median, 1.22; 95 % CrI, 0.77–1.87; Fig. 1). The likelihood of the $RR < 0.90$ given the interim trial data is 6 % and 9 % for longer cooling and deeper cooling, respectively, which are below the 10 % futility stopping threshold. Using the risk difference for monitoring, the probability of a $RD > 1$ % (indicating reduced mortality of 1 % or more with intervention) is 12 % for deeper cooling and 8 % for longer cooling, and longer cooling would be stopped.

Figure 1 (bottom panels) shows the posterior distributions calculated from the enthusiastic priors. The posterior median for the RR for duration of cooling is 1.27 (95 % CrI, 0.81–2.01) with 84 % posterior probability of increased pre-discharge death with longer cooling. Compared

with standard cooling, deeper cooling has 76 % probability of increased pre-discharge mortality (RR: posterior median, 1.18; 95 % CrI, 0.74–1.85). The probability of a 10 % or greater reduction in pre-discharge mortality ($RR < 0.90$) is 7 % (9 % for $RD > 1$ %) and 12 % (15 % for $RD > 1$ %) for the duration and depth interventions. Longer cooling again crosses the futility stopping threshold.

Posterior distributions for three intervention group comparisons

Table 4 and Fig. 2 give posterior summaries for the RRs of the three cooling groups (longer cooling, deeper cooling, both) compared with standard cooling. The posterior medians for all three groups are above 1.0, indicating increased mortality compared with standard cooling under both neutral and enthusiastic priors. Under the neutral prior, the probability of any reduced mortality is small for the 32 °C for the 120 h group. However, an enthusiastic prior gives a 25 % probability of reduced mortality for the 32 °C for 72 h group. The probabilities of treatment benefit of the three experimental groups achieving a clinically meaningful value of $RR < 0.90$ are

Table 4 Summaries of posterior probabilities of relative risk of predischarge mortality

	RR posterior median (95 % credible interval)		Evidence of any benefit Pr(RR < 1.0)		Futility monitoring Pr(RR < 0.90)	
	Neutral	Enthusiastic	Neutral	Enthusiastic	Neutral	Enthusiastic
32.0 °C for 72 h	1.23 (0.76–1.92)	1.19 (0.74–1.87)	20 %	25 %	10 %	13 %
33.5 °C for 120 h	1.31 (0.82–2.09)	1.27 (0.80–2.03)	13 %	16 %	6 %	8 %
32.0 °C for 120 h	1.60 (0.82–2.97)	1.50 (0.79–2.83)	8 %	11 %	4 %	6 %

The three experimental hypothermia groups are compared with standard cooling (33.5 °C for 72 h) under a neutral and enthusiastic prior. RR values less than 1.0 favor experimental groups
Pr probability, *RR* relative risk

13 %, 8 %, and 6 % under an enthusiastic prior and somewhat smaller under the neutral prior. Applying the same futility criterion as for the marginal interventions, only the 32 °C for 72 h group would not meet the stopping threshold under the enthusiastic prior. Similarly, using the futility stopping guideline based on a minimum RD of 1 %, the 33.5 °C for 120 h and 32 °C for 120 h groups would be stopped with the interim data (Table 5). Finally, only the 32 °C for 120 h group crosses the safety stopping threshold for predischarge death.

Sensitivity analyses using a larger standard deviation for the prior distribution of the interaction term resulted in similar posterior probabilities (not shown) with no differences in the crossing of stopping thresholds.

Discussion

The use of the best available prior information is one of the main advantages of the Bayesian approach, as it allows for formal evaluation of all available evidence. While no prior data exist for longer or deeper cooling, we explain how the data from a prior NRN trial of standard cooling [37] could be used to identify what negative interim findings for predischarge mortality should be convincing to enthusiasts as well as skeptics or neutral clinicians. The Bayesian analyses presented here incorporated this information into the prior probabilities while excluding large treatment effects, which are almost never observed with clinical interventions [58, 59]. The resulting posterior estimates of the RR being closer to 1.0 than the unadjusted frequentist estimates (e.g., 1.30 versus 1.50 for duration of cooling) even under a neutral prior.

These Bayesian interim analyses illustrate how the Bayesian approach can be used by DSMCs to evaluate the “totality of available evidence” [60] when deciding whether to stop a trial early. By using enthusiastic priors, a DSMC can judge whether the current evidence should be sufficient to convince an investigator with a strong prior belief in treatment benefit that there is little chance of benefit from the intervention. Under our proposed priors and stopping guidelines based on RRs and focusing on marginal comparisons, the trial data would not support the existence of treatment benefit for longer cooling but would not completely rule out the existence of a clinically important benefit (12 % probability) for deeper cooling.

We conducted a supplementary analysis using a neutral prior (centered at 0) for the intercept (and the same neutral priors for all other parameters) that essentially ignored the evidence on the rate of predischarge mortality from the previous NRN trial. The results are given in Additional file 3 and show RR estimates that are further away from 1.0 (indicating less shrinking towards a null effect). Under this prior, all three experimental groups cross the futility stopping threshold. Other priors not presented here, such as a skeptical prior (centered at RR > 1) or robust priors (e.g., Student’s *t* distributions) for sensitivity analyses, could also be presented to give DSMC members and investigators a complete picture of pre-existing expert views of a therapy’s benefits and harms. While the subjectivity of the prior is usually seen as the biggest drawback of a Bayesian approach, we see it as an advantage, since it formalizes how experts with

Table 5 Summaries of posterior probabilities of the absolute risk difference of predischarge mortality

	RD posterior mean (95 % credible interval)		Futility monitoring Pr(RD > 0.01) ^a		Safety monitoring Pr(RD < -0.05) ^b	
	Neutral	Enthusiastic	Neutral	Enthusiastic	Neutral	Enthusiastic
32.0 °C for 72 h	-0.02 (-0.08, 0.03)	-0.02 (-0.08, 0.04)	11 %	15 %	19 %	16 %
33.5 °C for 120 h	-0.03 (-0.09, 0.02)	-0.03 (-0.09, 0.03)	8 %	9 %	28 %	25 %
32.0 °C for 120 h	-0.06 (-0.15, 0.03)	-0.06 (-0.15, 0.03)	5 %	7 %	61 %	54 %

^aRD > 0.01 indicates 1 % or more reduced mortality

^bRD < -0.05 indicates a 5 % or more absolute increase in mortality

The three experimental groups are compared standard cooling (33.5 °C for 72 h) under a neutral and enthusiastic prior. Positive values of RD favor the experimental groups

Pr probability, *RD* risk difference

differing pre-existing opinions will view the results. A DSMC can then consider whether the results of a trial would be convincing to the whole community.

When factorial designs are used, the possibility of dropping one or two treatment groups and continuing the trial with the remaining groups should be discussed before starting the trial. Here we illustrated Bayesian monitoring for the three experimental cooling groups. The results show that the futility stopping threshold based on RRs and RDs would have been crossed for the 32.0 °C for 120 h and 33.5 °C for 120 h groups under both priors. However, the 32.0 °C for 72 h cooling group would not have crossed the threshold under either prior. Of course, different stopping thresholds would lead to different decisions. For example, if we instead used a RD of $\geq 5\%$ to monitor futility, the posterior probabilities of this treatment effect, $\Pr(\text{RD} > 0.05)$, would be less than 2% for all three groups, even with the enthusiastic prior. These results illustrate the need to fully explore at the planning stage the choice of priors and relevant quantities to monitor for efficacy, safety, and futility, as well as the probability thresholds for stopping the trial. All aspects of the trial design and monitoring plan can be evaluated with simulation studies (using different scenarios to represent a range of potential treatment effects) to ensure adequate type I and II errors [26, 45]. As with frequentist monitoring rules, planned Bayesian interim analyses and stopping guidelines should be pre-specified in the protocol [61].

A frequentist interim futility analysis using conditional power (the probability of obtaining a statistically significant benefit given current interim data) calculations led to the recommendation to stop this trial. A disadvantage of this approach is that it might put too much focus on an arbitrary level of significance [7] or a secondary outcome. A high probability of not showing a statistically significant benefit is not necessarily sufficient reason to stop a trial, since significance may not be needed to justify recommendation of a therapy with at least equivalent benefit that is less invasive, hazardous, more convenient, or less expensive than the comparison therapy. With the overriding importance of death and of death or disability, clinicians and their patients and families might consider a lower rate of death or of death or disability for the interventions assessed in the Optimizing Cooling Trial to justify their use, even if the differences did not reach statistical significance.

A Bayesian futility monitoring plan could alternatively use the predictive probability of a successful trial at the end of planned enrollment. This predictive probability is analogous to the frequentist conditional power but accounts for the uncertainty of the current parameter values rather than assuming fixed values. However, we would argue that it suffers from the same limitations as

conditional power. Although as Saville et al. [21] and Emerson et al. [62] point out, any stopping rule based on the posterior distribution can be converted into a stopping rule based on the predictive probability. For the Optimizing Cooling Trial, lack of primary outcome data also precluded us from adopting a predictive probability approach.

Challenges to the wider use of Bayesian monitoring methods include the perceived subjectivity of this approach, the difficulty of eliciting priors from investigators and previous studies, computational complexities, and reluctance from funding agencies and journals to embrace Bayesian methods in clinical research. Statisticians and clinical investigators must collaborate to find ways of overcoming these barriers to best inform decision makers. For the Optimizing Cooling Trial, we prespecified and planned to conduct a Bayesian final analysis for the primary outcome of death or disability.

Conclusions

To help familiarize clinical investigators with Bayesian monitoring methods, we reanalyzed the Optimizing Cooling Trial, a neonatal trial that was stopped early for safety and futility. We incorporated information on pre-discharge mortality rate from a previous multicenter neonatal cooling trial into the prior distribution. When we incorporate external data and take the view of an enthusiast, two of the three intervention groups would be stopped for futility.

Bayesian analyses incorporating previous trial results and different pre-existing opinions can help interpret accruing data and facilitate informed stopping decisions likely to be meaningful and convincing to clinicians, meta-analysts, and guideline developers. Given the advantages of Bayesian trial monitoring, investigators should consider the use of Bayesian methods in Phase III clinical trials.

Additional files

Additional file 1: CONSORT 2010 checklist. (PDF 677 kb)

Additional file 2: CONSORT flow diagram. (PDF 37 kb)

Additional file 3: Additional analysis results and sample WinBUGS code. (PDF 132 kb)

Abbreviations

CrI, credible interval; DSMC, data and safety monitoring committee; MCMC, Markov chain Monte Carlo; NICHD, Eunice Kennedy Shriver National Institute of Child Health and Human Development; NRN, Neonatal Research Network; RD, absolute risk difference; RR, relative risk

Acknowledgements

The National Institutes of Health, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, and the National Center for Advancing Translational Sciences provided grant support for the Neonatal Research Network's Optimizing Cooling Trial through cooperative agreements. While NICHD staff had input into the study design, conduct, analysis, and

manuscript drafting, the comments and views of the authors do not necessarily represent the views of the NICHD.

We are indebted to our medical and nursing colleagues and the infants and their parents who agreed to take part in this study. The following investigators, in addition to those listed as authors, participated in this study.

NRN Steering Committee Chairs: Michael S. Caplan, MD, University of Chicago, Pritzker School of Medicine (2006–2011); Richard A. Polin, MD, Division of Neonatology, College of Physicians and Surgeons, Columbia University, (2011–present).

Alpert Medical School of Brown University and Women & Infants Hospital of Rhode Island (U10 HD27904): Martin Keszler, MD; Angelita M. Hensman, MS RNC-NIC; Elisa Vieira, RN BSN; Emilee Little, RN BSN; Birju Shah, MD; Nicholas Guerina, MD; Joseph Bliss, MD PhD; Hussnain Mirza, MD; Ross Sommers, MD.

Case Western Reserve University, Rainbow Babies & Children's Hospital (U10 HD21364, M01 RR80): Michele C. Walsh, MD MS; Anna Maria Hibbs, MD; Nancy S. Newman, RN; Bonnie S. Siner, RN; Arlene Zadell, RN.

Children's Mercy Hospital and University of Missouri Kansas City School of Medicine (U10 HD68284): William E. Truog, MD; Eugenia K. Pallotto, MD; Howard W. Kilbride, MD; Cheri Gauldin, RN BSN CCRC Anne Holmes, RN MSN MBA-HCM CCRC; Kathy Johnson, RN CCRC.

Cincinnati Children's Hospital Medical Center, University of Cincinnati Medical Center, and Good Samaritan Hospital (U10 HD27853, UL1 TR77): Kurt Schibler, MD; Suhas G. Kallapur, MD; Barbara Alexander RN; Estelle E. Fischer MHA MBA; Teresa L. Gratton, PA; Cathy Grisby, BSN CCRC; Lenora Jackson, CRC; Jennifer Jennings, RN BSN; Kristin Kirker, CRC; Greg Muthig, BA; Sandra Wuertz, RN BSN CLC.

Duke University School of Medicine, University Hospital, University of North Carolina, and Duke Regional Hospital (U10 HD40492, UL1 RR24128): C. Michael Cotten, MD MHS; Ronald N. Goldberg, MD; Joanne Finkle, RN JD; Kimberley A. Fisher, PhD FNP-BC IBCLC; Sandra Grimes, RN BSN; Matthew M. Laughon, MD MPH; Carl L. Bose, MD; Janice Bernhardt, MS RN; Cindy Clark, RN.

Emory University, Children's Healthcare of Atlanta, Grady Memorial Hospital, and Emory University Hospital Midtown (U10 HD27851, UL1 TR454): Barbara J. Stoll, MD; David P. Carlton, MD; Shannon E. G. Hamrick, MD; Ellen C. Hale, RN BS CCRC; Yvonne Loggins, RN.

Eunice Kennedy Shriver National Institute of Child Health and Human Development: Rosemary D. Higgins, MD; Stephanie Wilson Archer, MA. Indiana University, Riley Hospital for Children at Indiana University Health and Methodist Hospital (U10 HD27856, UL1 TR6): Gregory M. Sokol, MD; Brenda B. Poindexter, MD MS; Leslie Dawn Wilson, BSN CCRC; Susan Gunn, NNP CCRC; Lucy Smiley, CCRC.

Nationwide Children's Hospital and The Ohio State University (U10 HD68278): Edward G. Shepherd, MD; Leif D. Nelin, MD; Sudarshan R. Jadcherla, MD; Pablo J. Sánchez, MD; Patricia A. Luzader, RN; Christine A. Fortney, PhD RN; Nehal A. Parikh, MD; Bronte Clifford; Julie Gontentag, BSN; Marissa E. Jones, RN MBA; Jodi A. Ulloa, MSN APRN NNP-BC; L. Yossef, MD; Erin Ferns; Tiffany Sharp; Jon Wispe, MD; Elizabeth Bonachea, MD; Jonathan Slaughter, MD, MPH; Louis G. Chicoine, MD; Brandon Hart, MD; Krista Haines, MD; Ish Gulati, MD; Michael Hokenson, MD; Roopali Bapat, MD; Nahla Zaghoul, MD; Ruth Seabrook, MD; Thomas Bartman, MD; Jennifer Fuller, MS RNC; Sarah McGregor, BSN RNC; Marliese Dion Nist, BSN; Tara Wolfe, BSN; Elizabeth Ann Rodgers, BSN.

RTI International (U10 HD36790): Dennis Wallace, PhD; Kristin M. Zaterka-Baxter, RN BSN CCRP; Margaret Crawford, BS CCRP; Barry Eggleston, MS; Jenna Gabrio, BS CCRP; Marie G. Gantz, PhD; Scott A. McDonald, BS; Jamie E. Newman, PhD MPH; Jeanette O'Donnell Auman, BS; Carolyn M. Petrie Huitema MS CCRP. Stanford University and Lucile Packard Children's Hospital (U10 HD27880, M01 RR70, UL1 TR93): Krisa P. Van Meurs, MD; David K. Stevenson, MD; M. Bethany Ball, BS CCRC; Melinda S. Proud, RCP.

University of Alabama at Birmingham Health System and Children's Hospital of Alabama (U10 HD34216, M01 RR32): Waldemar A. Carlo, MD; Namasivayam Ambalavanan, MD; Monica V. Collins, RN BSN MaEd; Shirley S. Cosby, RN BSN. University of California—Los Angeles, Mattel Children's Hospital, Santa Monica Hospital, Los Robles Hospital and Medical Center, and Olive View Medical Center (U10 HD68270): Uday Devaskar, MD; Meena Garg, MD; Teresa Chanlaw, MPH; Rachel Geller, RN BSN.

University of Iowa and Mercy Medical Center (U10 HD53109, UL1 TR442): Dan L. Ellsbury, MD; Tarah T. Colaizy, MD MPH; Jane E. Brumbaugh, MD; Karen J. Johnson, RN BSN; Donia B. Campbell, RNC-NIC; Jacky R. Walker, RN; Jonathan M. Klein, MD; Jeffrey L. Segar, MD; John M. Dagle, MD PhD; Julie B. Lindower, MD MPH; Steven J. McElroy, MD; Glenda K. Rabe, MD; Robert D. Roghair, MD; Lauritz R. Meyer, MD; Cary R. Murphy, MD; Vipinchandra Bhavsar, MB BS.

University of New Mexico Health Sciences Center (U10 HD53089, UL1 TR41): Kristi L. Watterberg, MD; Robin K. Ohls, MD; Conra Backstrom Lacy, RN; Sandra Beauman, MSN; Carol Hartenberger, BSN MPH.

University of Pennsylvania, Hospital of the University of Pennsylvania, Pennsylvania Hospital, and Children's Hospital of Philadelphia (U10 HD68244): Barbara Schmidt, MD MSc; Haresh Kirpalani, MB MSc; Kevin C. Dysart, MD; Sara B. DeMauro, MD MSCE; Aasma S. Chaudhary, BS RRT; Soraya Abbasi, MD; Toni Mancini, RN BSN CCRC; Dara M. Cucinotta, RN.

University of Rochester Medical Center, Golisano Children's Hospital, and the State University New York at Buffalo Women's and Children's Hospital of Buffalo (U10 HD68263, UL1 TR42): Nirupama Laroia, MD; Carl T. D'Angio, MD; Ronnie Guillet, MD PhD; Satyan Lakshminrusimha, MD; Karen Wynn, NNP RN; Holly I.M. Wadkins; Anne Marie Reynolds, MD MPH; Ann Marie Scorsone, MS; Patrick Conway, MS; Michael G. Sacilowski, BS; Stephanie Guilford, BS; Ashley Williams, MS Ed.

University of Texas Southwestern Medical Center, Parkland Health & Hospital System, and Children's Medical Center Dallas (U10 HD40689, M01 RR633): Myra Wyckoff, MD; Lina F. Chalak, MD MSCS; Pablo J. Sánchez, MD; Luc P. Brion, MD; Diana M. Vasil, RNC-NIC; Lijun Chen, PhD RN; Emma Ramon, RN. University of Texas Health Science Center at Houston Medical School and Children's Memorial Hermann Hospital (U10 HD21373): Kathleen A. Kennedy, MD MPH; Amir M. Khan, MD; Georgia E. McDavid, RN; Julie Arldt-McAlister, RN BSN; Katrina Burson, RN BSN; Carmen Garcia, RN CCRP; Karen Martin, RN; Sara C. Martin, RN BSN; Shawna Rodgers, RN; Patti L. Pierce Tate, RCP; Sharon L. Wright, MT (ASCP).

Wayne State University, University of Michigan, Hutzel Women's Hospital, and Children's Hospital of Michigan (U10 HD21385): Athina Pappas, MD; John Barks, MD; Rebecca Bara, RN BSN; Mary Christensen, RT; Stephanie A. Wiggins, MS; Diane F. White, RT.

Yale University – Richard A. Ehrenkranz, MD.

Authors' contributions

CP and JET conceived the study. CP conducted all analyses and wrote the first draft of the paper. JET, AD, AL, EFB, and SS made critical edits to earlier drafts. All authors read and approved the final version.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Center for Clinical Research and Evidence-Based Medicine, McGovern Medical School at The University of Texas Health Science Center at Houston, 6431 Fannin St, MSB 2.106, Houston, TX 77030, USA. ²Social, Statistical and Environmental Sciences Unit, RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852-3903, USA. ³Department of Pediatrics, Women & Infants Hospital of Rhode Island, The Warren Alpert Medical School of Brown University, 101 Dudley Street, Providence, RI 02905, USA. ⁴Department of Pediatrics, University of Iowa, 200 Hawkins Drive, Iowa City, IA 52240, USA. ⁵Department of Pediatrics, Neonatal-Perinatal Medicine, Wayne State University, Children's Hospital of Michigan, 3901 Beaubien Blvd., 4H46, Detroit, MI 48201, USA.

Received: 11 October 2015 Accepted: 21 June 2016

Published online: 22 July 2016

References

- Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*. 1982;38:153–62.
- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549–56.
- Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early. *BMJ*. 2012;344, e3863.
- Pocock SJ. When to stop a clinical trial. *BMJ*. 1992;305:235–40.
- Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics*. 1992;48:41–53.
- Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, et al. Issues in data monitoring and interim analysis of trials. *Health Technol Assess*. 2005;9:1–238. iii–iv.
- Pocock SJ. Current controversies in data monitoring for clinical trials. *Clin Trials*. 2006;3:513–21.

8. Fernandes RM, van der Lee JH, Offringa M. Data monitoring committees, interim analysis and early termination in paediatric trials. *Acta Paediatr.* 2011;100:1386–92.
9. Hey E. Clinical trials: when to start and when to stop. *Lancet.* 2002;359:1449.
10. Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. *J Clin Epidemiol.* 2008;61:241–6.
11. Mukherjee SD, Goffin JR, Taylor V, Anderson KK, Pond GR. Early stopping rules in oncology: considerations for clinicians. *Eur J Cancer.* 2011;47:2381–6.
12. Zannad F, Gattis Stough W, McMurray JVV, Remme WJ, Pitt B, Borer JS, et al. When to stop a clinical trial early for benefit: lessons learned and future approaches. *Circ Heart Fail.* 2012;5:294–302.
13. Parmar MKB, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. *Stat Med.* 1994;13:1297–312.
14. Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. *Stat Med.* 1993;12:1501–11. discussion 1513–7.
15. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. *J R Stat Soc Ser A.* 1994;157:357–416.
16. Fayers PM, Ashby D, Parmar MK. Tutorial in biostatistics Bayesian data monitoring in clinical trials. *Stat Med.* 1997;16:1413–30.
17. Robert E, Kass JBG. [Investigating therapies of potentially great benefit: ECMO]: comment: a Bayesian perspective. *Statist Sci.* 1989;4:310–17.
18. Berry DA. A case for Bayesianism in clinical trials. *Stat Med.* 1993;12:1377–93. discussion 1395–404.
19. Carlin BP, Sargent DJ. Robust Bayesian approaches for clinical trial monitoring. *Stat Med.* 1996;15:1093–106.
20. Dmitrienko A, Wang M-D. Bayesian predictive approach to interim monitoring in clinical trials. *Stat Med.* 2006;25:2178–95.
21. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin Trials.* 2014;11:485–93.
22. Thall PF, Simon R. Practical Bayesian guidelines for Phase IIb clinical trials. *Biometrics.* 1994;50:337–49.
23. Thall PF, Wooten LH, Tannir NM. Monitoring event times in early phase clinical trials: some practical issues. *Clin Trials.* 2005;2:467–78.
24. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for Phase II clinical trials. *Control Clin Trials.* 1994;15:463–81.
25. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. *BMJ.* 1999;319:508–12.
26. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov.* 2006;5:27–36.
27. Harrell FE, Shih YC. Using full probability models to compute probabilities of actual interest to decision makers. *Int J Technol Assess Health Care.* 2001;17:17–26.
28. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med.* 2005;24:2401–28.
29. Greenland S. Bayesian perspectives for epidemiological research. II Regression analysis. *Int J Epidemiol.* 2007;36:195–202.
30. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health care evaluation. Chichester: John Wiley & Sons, Ltd; 2004.
31. Pibouleau L, Chevreton S. Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices. *J Clin Epidemiol.* 2011;64:270–9.
32. Winkler RL. Why Bayesian analysis hasn't caught on in healthcare decision making. *Int J Technol Assess Health Care.* 2001;17:56–66.
33. Sheingold SH. Can Bayesian methods make data and analyses more relevant to decision makers? A perspective from Medicare. *Int J Technol Assess Health Care.* 2001;17:114–22.
34. Shankaran S, Laptook AR, Pappas A, McDonald SA, Das A, Tyson JE, et al. Effect of depth and duration of cooling on deaths in the NICU among neonates with hypoxic ischemic encephalopathy: a randomized clinical trial. *JAMA.* 2014;312:2629–39.
35. Jacobs SE, Berg M, Hunt R, Tarnow-Mordi WO, Inder TE, Davis PG. Cooling for newborns with hypoxic ischaemic encephalopathy. *Cochrane Database Syst Rev.* 2013;1:CD003311.
36. Papile L-A, Baley JE, Benitz W, Cummings J, Carlo WA, Eichenwald E, et al. Hypothermia and neonatal encephalopathy. *Pediatrics.* 2014;133:1146–50.
37. Shankaran S, Laptook AR, Ehrenkranz RA, Tyson JE, McDonald SA, Donovan EF, et al. Whole-body hypothermia for neonates with hypoxic-ischemic encephalopathy. *N Engl J Med.* 2005;353:1574–84.
38. Carroll M, Beek O. Protection against hippocampal CA1 cell loss by post-ischemic hypothermia is dependent on delay of initiation and duration. *Metab Brain Dis.* 1992;7:45–50.
39. Colbourne F, Corbett D. Delayed postischemic hypothermia: a six month survival study using behavioral and histological assessments of neuroprotection. *J Neurosci Off J Soc Neurosci.* 1995;15:7250–60.
40. Iwata O, Thornton JS, Sellwood MW, Iwata S, Sakata Y, Noone MA, et al. Depth of delayed cooling alters neuroprotection pattern after hypoxia-ischemia. *Ann Neurol.* 2005;58:75–87.
41. Perlman JM. Summary proceedings from the neurology group on hypoxic-ischemic encephalopathy. *Pediatrics.* 2006;117:S28–33.
42. Compagnoni G, Bottura C, Cavallaro G, Cristofori G, Lista G, Mosca F. Safety of deep hypothermia in treating neonatal asphyxia. *Neonatology.* 2008;93:230–5.
43. Pocock S, Wang D, Wilhelmssen L, Hennekens CH. The data monitoring experience in the Candesartan in Heart Failure Assessment of Reduction in Mortality and Morbidity (CHARM) program. *Am Heart J.* 2005;149:939–43.
44. Fleming TR, Neaton JD, Goldman A, DeMets DL, Launer C, Korvick J, et al. Insights from monitoring the CPCRA didanosine/zalcitabine trial. Terry Bein Community Programs for Clinical Research on AIDS. *J Acquir Immune Defic Syndr Hum Retrovirol.* 1995;10 Suppl 2:S9–18.
45. Food and Drug Administration. Guidance for the use of Bayesian statistics in medical device clinical trials. *J Biopharm Stat.* 2010; 5:2010.
46. Dixon DO, Freedman RS, Herson J, Hughes M, Kim K, Silverman MH, et al. Guidelines for data and safety monitoring for clinical trials not requiring traditional data monitoring committees. *Clin Trials.* 2006;3:314–9.
47. McClure LA, Coffey CS, Howard G. Monitoring futility in a two-by-two factorial design: the SPS3 experience. *Clin Trials.* 2013;10:250–6.
48. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. *Biometrics.* 1997;53:456–64.
49. Sinclair JC, Haughton DE, Bracken MB, Horbar JD, Soll RF. Cochrane neonatal systematic reviews: a survey of the evidence for neonatal therapies. *Clin Perinatol.* 2003;30:285–304.
50. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Stat Med.* 2002;21:2909–16.
51. Granger CB, Alexander JH, McMurray JVV, Lopes RD, Hylek EM, Hanna M, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2011;365:981–92.
52. Lagerqvist B, Fröbert O, Olivecrona GK, Gudnason T, Maeng M, Alström P, et al. Outcomes 1 year after thrombus aspiration for myocardial infarction. *N Engl J Med.* 2014;371:1111–20.
53. Mauri L, Kereiakes DJ, Yeh RW, Driscoll-Shempp P, Cutlip DE, Steg PG, et al. Twelve or 30 months of dual antiplatelet therapy after drug-eluting stents. *N Engl J Med.* 2014;371:2155–66.
54. Ikeda Y, Shimada K, Teramoto T, Uchiyama S, Yamazaki T, Oikawa S, et al. Low-dose aspirin for primary prevention of cardiovascular events in Japanese patients 60 years or older with atherosclerotic risk factors: a randomized clinical trial. *JAMA.* 2014;312:2510–20.
55. White HD, Held C, Stewart R, Tarka E, Brown R, Davies RY, et al. Darapladib for preventing ischemic events in stable coronary heart disease. *N Engl J Med.* 2014;370:1702–11.
56. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC; 2014.
57. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000;10:325–37.
58. Greenland S. Putting background information about relative risks into conjugate prior distributions. *Biometrics.* 2001;57:663–70.
59. Pedroza C, Han W, Truong VTT, Green C, Tyson JE. Performance of informative priors skeptical of large treatment effects in clinical trials: a simulation study. *Stat Methods Med Res.* 2015; doi:10.1177/0962280215620828. [Epub ahead of print]
60. Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA.* 2005;294:2228–30.
61. Tyson JE, Pedroza C, Wallace D, D'Angio C, Bell EF, Das A. Stopping guidelines for an effectiveness trial: what should the protocol specify? *Trials.* 2016;17:240.
62. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. <http://biostat.bepress.com/uwbiostat/paper243/> (2005). Accessed 5 Jun 2015.