Bayesian Analysis (2021) **16**, Number 3, pp. 825–844

# Robust Adaptive Incorporation of Historical Control Data in a Randomized Trial of External Cooling to Treat Septic Shock

Thomas A. Murray[*,††], Peter F. Thall[†,‡‡], Frederique Schortgen[‡], Pierre Asfar[§,¶], Sarah Zohar[‖,§§], and Sandrine Katsahian[‖,**,§§]

**Abstract.** This paper proposes randomized controlled clinical trial design to evaluate external cooling as a means to control fever and thereby reduce mortality in patients with septic shock. The trial will include concurrent external cooling and control arms while adaptively incorporating historical control arm data. Bayesian group sequential monitoring will be done using a posterior comparative test based on the 60-day survival distribution in each concurrent arm. Posterior inference will follow from a Bayesian discrete time survival model that facilitates adaptive incorporation of the historical control data through an innovative regression framework with a multivariate spike-and-slab prior distribution on the historical bias parameters. For each interim test, the amount of information borrowed from the historical control data will be determined adaptively in a manner that reflects the degree of agreement between historical and concurrent control arm data. Guidance is provided for selecting Bayesian posterior probability group-sequential monitoring boundaries. Simulation results elucidating how the proposed method borrows strength from the historical control data are reported. In the absence of historical control arm bias, the proposed design controls the type I error rate and provides substantially larger power than reasonable comparators, whereas in the presence bias of varying magnitude, type I error rate inflation is curbed.

**Keywords:** commensurate prior, conditional autoregressive model, evidence synthesis, intensive care unit, non-proportional hazards, restricted mean survival.

## 1 Background

The methodology described in this paper was motivated by the problem of designing a randomized controlled trial to evaluate the effectiveness of using external cooling to con-

[*]Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA, murra484@umn.edu
[†]Department of Biostatistics, M. D. Anderson Cancer Center, Houston, TX, USA
[‡]Service of Intensive Care Unit, Hôspital Intercommunal de Créteil, Créteil, France
[§]Service of medical Intensive care and hyperbaric oxygen therapy unit, Centre Hospitalier Universitaire Angers, Angers, France
[¶]Laboratoire de Biologie Neurovasculaire et Mitochondriale Intégrée, CNRS UMR 6214 - Inserm U1083, Université Angers, UBL, Angers, France
[‖]Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris, Paris, France
[**]CIC-EC 1418 Inserm, Hôpital Européen Georges-Pompidou, Paris, France
[††]Funded in part by NIH/NCI Grant P30-CA077598. Thanks to Medtronic Inc. for their support in the form of a Faculty Fellowship.
[‡‡]Funded in part by NIH/NCI Grant 5-R01-CA083932.
[§§]Katsahian S. and Zohar S. have equally contributed to this paper.

 https://doi.org/10.1214/20-BA1229

trol fever and thereby reduce mortality in patients with septic shock admitted to a hospital intensive care unit (ICU). Septic shock is a complex, life-threatening condition involving sepsis, which is an inflammatory immunological response to infection, combined with hypotension, which is dangerously low blood pressure (Angus and van der Poll, 2013; Singer et al., 2016). The 60-day mortality rate for patients with septic shock admitted to an ICU is about 40% (Annane et al., 2003; Caironi et al., 2014; Asfar et al., 2014).

Standard ICU treatment for septic shock may involve antibiotics or surgery for infection, infusion of fluids, and a vasopressor to maintain the patient's mean arterial pressure (MAP) at a minimal level to sustain organ function (Dellinger et al., 2013). The patient's vital signs are monitored continuously. The attending physician repeatedly adjusts the vasopressor dose to keep the patient's MAP in a targeted range, e.g., 65–70 mm Hg. The immediate goal is to achieve septic shock reversal defined as the patient sustaining a MAP within or above the targeted range for 24 hours without vasopressor administration. There is no consensus among intensive care physicians on the ideal MAP range, or whether this range should be higher for patients with a history of hypertension. A randomized controlled trial reported by Asfar et al. (2014) comparing the target ranges 80–85 versus 65–70 mm Hg for MAP in septic shock patients with a history of hypertension admitted to an ICU found no significant difference in either 28-day or 90-day mortality rates. While the target MAP range will not be the focus of the trial described in this paper, the trial reported by Asfar et al. (2014) is our source for historical control data.

Septic shock often is accompanied by high fever, which may be controlled by external cooling, done either by covering the patient with a blanket containing circulating cold water, or applying ice packs directly to the patient's body. A clinical trial ("Sepsiscool") that randomized 101 febrile septic shock patients to standard treatment augmented by external cooling and 99 to standard treatment alone reported by Schortgen et al. (2012) found statistically significant evidence at the 0.05 level that external cooling reduced vasopressor requirements. The external cooling arm had a higher rates of septic shock reversal (87% versus 72%), and 14-day survival (81% versus 66%). Statistically significant evidence was lacking, however, for a survival benefit at ICU discharge (65% versus 57%) and at hospital discharge (57% versus 52%). While fever control has circulatory benefits, fever is beneficial for fighting the underlying infection. Therefore, it is unclear whether fever control is beneficial or harmful in terms of survival, and the use of external cooling remains controversial (Shime et al., 2013). This background motivates the trial described here, which will compare $C$ = standard of care to $E$ = standard of care augmented with external cooling started immediately upon randomization. The primary goal is to obtain a confirmatory evaluation of $E$ versus $C$ on survival during the 60 day period following randomization. The choice of a 60 day follow up is motivated by the fact that, based on historical data and clinical experience, the great majority of deaths in each arm will occur within 30 days.

In this paper, we present a Bayesian group-sequential test to compare $E$ to $C$ in terms of their 60-day survival distributions. We use posterior probability stopping boundaries that correspond to established frequentist group sequential boundaries (Shi and Yin, 2019; Proschan et al., 2006). Statistical inference follows from a Bayesian hierarchical

discrete time survival regression model that facilitates robust adaptive incorporation of the historical control data from the trial reported by Asfar et al. (2014) into the group sequential posterior probabilities. Our model assumes first order intrinsic Gaussian Markov Random Field (IGMRF) priors for the hazard of death on each day of the observation period in each arm (Besag et al., 1991; Rue and Held, 2005). Compared to parametric survival regression models and Cox's proportional hazards model, our model provides greater flexibility for estimating the survival distribution on the domain [0, 60] days in each arm, and thus for estimating the difference in 60-day survival distributions between the two arms robustly. Our proposed comparator is more general than, but encompasses 60-day restricted mean survival time (RMS), which is the number of days that a patient is expected to survive during the 60 day period following randomization after admission to an ICU with septic shock (Royston and Parmar, 2013). We show that our proposed comparative testing and modeling framework, which neither requires proportional hazards treatment nor trial effects, provides greater power than the log-rank test under the targeted benefit of $E$ over $C$, even when not incorporating historical control data.

Our modeling approach for adaptively incorporating the historical control data into the group sequential tests is related to previous work on commensurate priors, see, e.g., (Hobbs et al., 2011, 2012; Murray et al., 2015). Power priors (Ibrahim and Chen, 2000) and multi-source exchangeability models (Kaizer et al., 2018) offer alternative approaches for incorporating historical data, although neither approach has been implemented with discrete time survival outcomes. The commensurate prior approach relies on hierarchically modeling data on the prospective and historical control arms while adaptively borrowing strength via priors that facilitate shrinking analogous parameters across data sources toward one another. Deviating from this strategy, we develop a regression model that explicitly parameterizes historical bias, and data-adaptively shrinks these bias parameters to zero using spike-and-slab priors. This framework facilitates robust adaptive incorporation of historical control data, based on the observed magnitude of historical bias, as the trial data accrue. Key innovations are the reformulation of commensurate priors through explicit parameterization of historical bias, development of a commensurate prior discrete time survival modeling framework, and formulation of a commensurate prior for a parameter following a GMRF prior.

In Section 2, we present our comparative test, which is based on a summary of the posterior 60-day survival distributions in the two arms. In Section 3, we describe the regression model to evaluate our comparative test. In Section 4, we discuss posterior estimation using a Gibbs sampler with Pólya-Gamma latent variables (Polson et al., 2013). In Section 5, we address important design considerations, including specification of the target benefit of $E$ relative to $C$ and the Bayesian posterior probability monitoring boundaries to ensure type I error rate control. In Section 6, we report the results of a computer simulation study that evaluates performance of various design implementations over a range of differences between the hazards of death for the historical and concurrent controls. In Section 7, we conclude with a brief discussion.

## 2 Comparative Test

For the primary analysis, we will follow patients who remain alive for 60 calendar days from randomization with primary outcome variable $Y^*$ being the follow up day of death. Thus, $Y^*$ has discrete support $\{1, \cdots, 60, > 60\}$ days during the follow up period, with $Y^* > 60$ if the patient is alive at the end of 60 day follow up. The discrete support arises from the fact that time of randomization and death each are recorded as a calendar date. If a patient dies during the calendar day on which they are randomized, then $Y^* = 1$; whereas, if a patient dies during the calendar day after they are randomized, then $Y^* = 2$; and so on. While $Y^*$ is unknown for any patient who is alive at the end of 60-day follow-up, it is partially observed with $Y^* > 60$. We use an asterisk to distinguish this from the observed time $Y$, which is subject to additional left-censoring mechanisms arising from loss to follow up and interim data analyses being carried out while some participants are in the midst of follow up. We will provide the formal connection between $Y^*$ and $Y$ later in Section 3, where the modeling framework is described.

The primary goal of the trial will be to evaluate the clinical effectiveness of $E$ compared to $C$ for reducing mortality. To do this, we compare the 60 day survival distributions in $E$ versus $C$ using posterior mean utilities, which in arm $a = E, C$ is defined as

$$\mu_a = \sum_{t=1}^{60} U(Y^* = t) \times \Pr(Y^* = t \,|\, A = a) + U(Y^* > 60) \times \Pr(Y^* > 60 \,|\, A = a), \quad (1)$$

where $U(\cdot)$ reflects the utility function. Formally, the trial will test the hypotheses

$$H_0 : \mu_E \leq \mu_C \qquad \text{versus} \qquad H_A : \mu_E > \mu_C.$$

The utility function $U(\cdot)$ in (1) should be defined to reflect the relative desirability of each possible realization of $Y^*$. In the sequel, we specify $U(Y^* = t) = t - 1$, for $t = 1, \ldots, 60$, but leave $R = U(Y^* > 60)$ unspecified subject to $R \geq 60$. There are a few notable specifications for $R$. Taking $R = 60$, $\mu_a$ in (1) reflects 60-day restricted mean survival time in arm $a = E, C$, which is the number of days a person is expected to survive during the 60 days following admission to the ICU with sepsis (Royston and Parmar, 2013). Taking $R = E[Y^* \,|\, Y^* > 60]$, i.e. 60-day conditional expected survival time, $\mu_a$ reflects mean survival time. In this case, $E[Y^* \,|\, Y^* > 60]$ could be estimated using a separate parametric model or external population data. Taking $R = \infty$ results in comparing arms on 60-day survival probability, since $\mu_E/\mu_C \to \Pr(Y^* > 60 \,|\, E)/\Pr(Y^* > 60 \,|\, C)$ as $R \to \infty$.

Sensitivity of the utility-based comparison in (1) to $R = U(Y^* \geq 60)$ may be assessed by varying $R$ from 60 to $\infty$ and identifying change points, if any, where the conclusion differs. When the 60-day survival distributions in each arm are identical, $\mu_E = \mu_C$ for all $R > 60$; and when both $\sum_{t=1}^{60} t \times \Pr(Y^* = t \,|\, A = a)$ and $\Pr(Y^* > 60 \,|\, A = a)$ are greater in one arm than the other, the conclusion will be invariant to $R$ as well. In contrast, when $\sum_{t=1}^{60} (t-1) \times \Pr(Y^* = t \,|\, A = a)$ and $\Pr(Y^* > 60 \,|\, A = a)$ favor opposite arms, the value of $R$ will determine which arm's 60-day survival distribution is clinically preferable. This

latter scenario includes crossing 60-day survival distributions in which one arm has a better 60-day survival probability but worse shorter-term survival probabilities.

Our comparative test will be based on the posterior mean utility in the two arms arising from our Bayesian hierarchical model described in Section 3. We will apply our testing criteria in a group-sequential manner and reject the null hypothesis as follows:

$$\text{If } \Pr\left(\mu_E > \mu_C \,|\, \text{Data}\right) \geq p, \text{ then stop the trial and recommend } E,$$

where 'Data' refers to the historical data and the currently available prospective trial data. Similarly, we will fail to reject the null hypothesis as follows:

$$\text{If } \Pr\left(\mu_E > \mu_C \,|\, \text{Data}\right) < q, \text{ then stop the trial and recommend } C.$$

In Section 5, we discuss timing for the group sequential tests, and a method for determining boundary values $p$ and $q$ that control the type I error rate, and a maximum sample size that controls the type II error rate under the targeted benefit of $E$ over $C$.

## 3 Probability Model

To develop a Bayesian model for learning about $\mu_E$ and $\mu_C$ from the data, we exploit the connection between these quantities and the hazard of death in each arm on each day during the 60-day follow-up period. We denote the hazard of death under regimen $A = a$ on day $t$ as $\pi_{a,t} = \Pr(Y^* = t \,|\, Y^* \geq t, A = a)$, for $a = E, C$ and $t = 1, \ldots, 60$. Due to the two identities,

$$\begin{aligned} \Pr(Y^* = t \,|\, A = a) &= \pi_{a,t} \times \prod_{s=1}^{t-1} \left[1 - \pi_{a,s}\right], \; t = 1, \ldots, 60, \\ \Pr(Y^* > 60 \,|\, A = a) &= \prod_{t=1}^{60} \left[1 - \pi_{a,t}\right], \end{aligned} \tag{2}$$

$\mu_a$ is a function of $\{\pi_{a,t} : t = 1, \ldots, 60\}$. We use this relationship to construct a Bayesian hierarchical model for learning about the hazard of death in each arm on each day during the 60 day follow up period, and thereby learning about $\mu_E$ and $\mu_C$, based on the historical control data reported by Asfar et al. (2014) and the prospective trial data.

To do this, we propose the following discrete time survival regression model. Denote the $i$-th patient's treatment arm by $a_i = E$, $C$ or $C_0$ for the historical control patients. We will assume that $Y_i^* \,|\, A = a_i; \boldsymbol{\beta}$, $i = 1, \ldots, n$, are independent, where $\boldsymbol{\theta}$ denotes the vector of model parameters, which we will describe in detail below. Let

$$\pi_{a,t}(\boldsymbol{\beta}) = \Pr(Y^* = t \,|\, Y^* \geq t, A = a; \boldsymbol{\beta})$$

denote the model-based hazard of death in arm $A = a$ on day $t$, and note that

$$\Pr(Y^* \leq t \,|\, A = a_i; \boldsymbol{\beta}) = 1 - \prod_{s=1}^{t} \left[1 - \pi_{a,s}(\boldsymbol{\beta})\right].$$

We assume the following logistic regression model for the hazard of death in arm $A = a$ on day $t$ during the follow up period:

$$\text{logit}\,\{\pi_{a,t}(\boldsymbol{\beta})\} = \gamma + g_t + 1_E(a)(\lambda + \ell_t) + 1_{C_0}(a)(\zeta + z_t), \tag{3}$$

where $\text{logit}\{x\} = \log\{x/(1-x)\}$ for $0 \le x \le 1$, $1_A(x)$ is the indicator function for $x \in A$, and $\boldsymbol{\beta} = (\gamma, g_1, \ldots, g_{60}, \lambda, \ell_1, \ldots, \ell_{60}, \zeta, z_1, \ldots, z_{60})^{\mathsf{T}}$. To ensure identifiability, we impose the restrictions $\sum_{t=1}^{60} g_t = 0$, $\sum_{t=1}^{60} \ell_t = 0$, and $\sum_{t=1}^{60} z_t = 0$ throughout.

To see how this model facilitates borrowing strength from the historical control data, note that

$$\text{logit}\,\{\pi_{C_0,t}(\boldsymbol{\beta})\} - \text{logit}\,\{\pi_{C,t}(\boldsymbol{\beta})\} = \zeta + z_t.$$

On follow-up day $t$, $\zeta + z_t$ is the historical-versus-trial control treatment bias defined as the log odds ratio of the hazard of death in the historical control arm $C_0$ versus the prospective control arm $C$. If $\zeta = 0$ and $z_t = 0$, $t = 1, \ldots, 60$, then the prospective and historical control patients' survival times are exchangeable, and it is appropriate to incorporate all data from $C_0$ when comparing $E$ to $C$. As the values of these historical bias parameters move away from 0, the difference between the prospective and historical control survival distributions increases, and it becomes increasingly inappropriate to incorporate $C_0$ data into the test statistic.

For patients randomized to the trial control arm $C$, our regression model in (3) allows the hazard of death to be different on each day of the 60-day observation period, with

$$\text{logit}\,\{\pi_{C,t}(\boldsymbol{\beta})\} = \gamma + g_t,$$

on day $t$. Thus, the parameters $\{g_t : t = 1, \ldots, 60\}$ quantify daily deviations from the shared parameter $\gamma$ for the hazard of death. Our regression model also allows the odds ratio for the hazard of death in arm $E$ versus $C$ to be different on each day of the 60-day observation period, with

$$\text{logit}\,\{\pi_{E,t}(\boldsymbol{\beta})\} - \text{logit}\,\{\pi_{C,t}(\boldsymbol{\beta})\} = \lambda + \ell_t.$$

The parameters $\{\ell_t : t = 1, \ldots, 60\}$ quantify daily deviations from the shared parameter $\lambda$, with $(\lambda + \ell_t)$ the $E$-versus-$C$ effect on the hazard of death on day $t$, for $t = 1, \ldots, 60$.

Leveraging insights from Ghosh et al. (2018) on prior specification for logistic regression and its similarity to the proposed model, we assume $\gamma \sim t_7(-5, 5)$ and $\lambda \sim t_7(0, 2.5)$, where $t_\nu(\mu, \sigma)$ denotes a generalized t-distribution with degrees of freedom $\nu$, location $\mu$ and scale $\sigma$. Our choice of $-5$ for the location of the prior distribution on $\gamma$ reflects the prior information that $\Pr(Y^* > 60 \,|\, C) = 0.58$ and the model identity $\Pr(Y^* > 60 \,|\, C; \boldsymbol{\theta}) = [1 + \exp(-\gamma)]^{-60}$. To partially pool information from adjacent days about the hazard of death in each arm, we use first-order IGMRF priors that assume independent increments in the time-varying log-hazard parameters, $(g_t - g_{t-1}) \,|\, \sigma_g^2 \overset{iid}{\sim} \text{Normal}(0, \sigma_g^2)$ and $(\ell_t - \ell_{t-1}) \,|\, \sigma_\ell^2 \overset{iid}{\sim} \text{Normal}(0, \sigma_\ell^2)$, for $t = 2, \ldots, 60$. Hennerfeind et al. (2006) also use this GMRF prior in a discrete time survival regression

model. Following Besag et al. (1991) and Rue and Held (2005), the independence of the increments follows from specifying improper priors for $\boldsymbol{g}$, and $\boldsymbol{\ell}$, as follows:

$$p_0\left(\boldsymbol{g}\,|\,\sigma_g^2\right) \propto \{\sigma_g^2\}^{-\frac{59}{2}} \exp\left\{-\frac{\boldsymbol{g}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{g}}{2\sigma_g^2}\right\},$$

where $\boldsymbol{Q}$ is a $60\times 60$ symmetric tridiagonal matrix with diagonal equal to $(1,2,\ldots,2,1)$ and first off-diagonals equal to $(-1,\ldots,-1)$, with $p_0\left(\boldsymbol{\ell}\,|\,\sigma_\ell^2\right)$ taking the same form. We specify $\sigma_g^2,\sigma_\ell^2 \overset{ind}{\sim}$ Scale-inv-$\chi^2\{1, 2.5^2/59\}$ which has a prior effective sample size of 1 and an expected value of $2.5^2/59$. Note that $\text{Var}\{\text{g}_1 - \text{g}_{60}\ |\,\sigma_{\text{g}}^2 = \text{E}[\sigma_{\text{g}}^2]\} = 2.5^2$ which reflects our prior expectation that the log-odds-ratio for the hazard of death on days 1 and 60 is very likely between $-5$ and 5. The same logic applies to $(\ell_1 - \ell_{60})$, which reflects the difference in the log-odds-ratio between arms $E$ and $C$ on days 1 and 60. In this way, these prior distributions are weakly informative.

To facilitate data-adaptive borrowing of strength from the historical control data reported by Asfar et al. (2014), we assume the following spike-and-slab prior for $\zeta$:

$$\zeta\,|\,\nu_\zeta \sim \nu_\zeta \times \text{Normal}(0, 1/R_\zeta) + (1-\nu_\zeta)\times t_7(0, 2.5^2) \ \text{ and } \ \nu_\zeta \sim \text{Bernoulli}(p_\zeta), \quad (4)$$

and the following multivariate spike-and-slab prior for $\boldsymbol{z}$:

$$p_0\left(\boldsymbol{z}\,|\,\sigma_z^2,\nu_z\right) \propto \nu_z\{1/R_z\}^{-\frac{59}{2}} \exp\left\{-\frac{\boldsymbol{z}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{z}}{2/R_z}\right\} + (1-\nu_z)\{\sigma_z^2\}^{-\frac{59}{2}} \exp\left\{-\frac{\boldsymbol{z}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{z}}{2\sigma_z^2}\right\}, \quad (5)$$
$$\nu_z \sim \text{Bernoulli}(p_z) \ \text{ and } \ \sigma_z^2 \sim \text{Scale-inv-}\chi^2\{1, 1/R_z\}.$$

By setting $R_\zeta = 4000$, conditional on $\nu_\zeta = 1$, $\zeta$ is effectively equal to zero. Similarly, by setting $R_z = 4000$, conditional on $\nu_z = 1$, $\boldsymbol{z}$ follows a IGMRF highly concentrated about $\boldsymbol{0}$, and thus $\boldsymbol{z}$ is effectively equal to $\boldsymbol{0}$. Recall that we constrain $\sum_{t=1}^{60} z_t = 0$. Therefore, conditional on $(\nu_z,\nu_\zeta) = (1,1)$, our probability model will effectively treat historical and prospective control patients as exchangeable, thereby facilitating full borrowing of strength from the historical control data. In contrast, conditional on $\nu_\zeta = 0$, $\zeta$ follows a weakly informative $t_7(0, 2.5^2)$ prior distribution, and conditional on $\nu_z = 0$ and $\sigma_z^2$, $\boldsymbol{z}$ follows a IGMRF prior with $(z_t - z_{t-1})\,|\,\sigma_z^2 \overset{iid}{\sim} \text{Normal}(0,\sigma_z^2)$, $t = 2,\ldots,60$ where $\sigma_z^2$ has a prior expected value equal to $1/R_z$ but effective sample size equal to 1. This prior specification for $\sigma_z^2$ facilitates posterior computation in that it is feasible to transition from $\nu_z = 1$ to $\nu_z = 0$, yet conditional on $\nu_z = 0$, the data will strongly influence the full conditional posterior for $\sigma_z^2$. Conditional $(\nu_z,\nu_\zeta) = (0,0)$, our probability model will treat historical and prospective control patients as non-exchangeable, thereby facilitating much less borrowing of strength from the historical control data. For $(\nu_z,\nu_\zeta) = (1,0)$, the historical bias satisfies proportional odds and our model will borrow some strength regarding the temporal variation in the hazards, whereas for $(\nu_z,\nu_\zeta) = (0,1)$, the average hazard across the 60 day follow up is similar in the two control arms, but the temporal variation differs. By setting $p_\zeta = 0.5$, and $p_z = 0.5$, *a priori* we assign the spike and slab distributions in (4) and (5) equal weight. Our model will learn about the values of $\nu_\zeta$ and $\nu_z$ adaptively in the usual Bayesian manner, by conditioning on the observed data, and this will be reflected in the group

sequential decision making during the trial. Because it is possible for $\nu_z = 0$ with $\sigma_z^2 < 1/R_z$, we suggest assessing posterior evidence for exchangeability with regard to $\boldsymbol{z}$ based upon $\nu_z^* = \nu_z + (1 - \nu_z)I(\sigma_z^2 \leq 1/R_z)$ rather than $\nu_z$. Our approach for incorporating historical data differs from existing commensurate prior approaches in several key ways. First, we use regression to explicitly parameterize a historical trial effect, or bias, which in this case is $\zeta + z_t$ on follow-up day $t = 1, \ldots, 60$. Second, we specify each spike-and-slab prior as a two-part mixture of a Gaussian distribution with its density highly concentrated about zero (the spike) and a weakly informative distribution (the slab) for the corresponding parameter (cf. Murray et al., 2015). Defining the spike in this way, rather than as a probability mass at zero, facilitates posterior computation (see, e.g., George and McCulloch, 1993). Third, we specify a multivariate spike-and-slab prior for $\boldsymbol{z}$ that is a two-part mixture of GMRF prior distributions, which simultaneously facilitates temporal smoothing across elements of $\boldsymbol{z}$ and shrinkage of all elements to zero.

We will refer to the above model formulation as the "Comm NPO" model, i.e. the commensurate non-proportional-odds prior model. In the sequel, for comparison, we consider three modifications to the Comm NPO model. The Comm PO model omits $\boldsymbol{z}$ altogether and thus assumes proportional odds for the hazard of death in the concurrent and historical control arms. The "Trad NPO" model, where Trad indicates traditional, ignores the historical control data altogether, as in a traditional clinical trial, and thus includes neither $\zeta$ nor $\boldsymbol{z}$, since historical data needed for estimating these parameters is not incorporated. The "Trad PO" model additionally omits the $\boldsymbol{\ell}$ term, and thus assumes that the hazards for death in the concurrent arms satisfy the proportional odds assumption.

## 4 Posterior Estimation

To estimate the posterior distribution arising from the Comm NPO model, we develop a blocked Gibbs sampling algorithm. Let $n_a$ denote the number of patients in arm $a = E, C$, or $C_0$, with $n_{C_0} = 761$ and $n = n_E + n_C + n_{C_0}$. Indexing patients by $i$, for patients who die during follow up we will observe $Y_i^*$, and otherwise only know that $Y_i^* \geq U_i^*$, where $U_i^*$ denotes the patient's administrative right-censoring time. As usual, we define the observable random variables $Y_i = \min\{Y_i^*, U_i^*\}$ and $\delta_i = I\{Y_i = Y_i^*\}$. At a given test, we will have sample observations $\boldsymbol{D}_n = \{(y_i, \delta_i, a_i) : i = 1, \ldots, n\}$. Under the proposed probability model described in Section 3, the likelihood function takes the form

$$L(\boldsymbol{\beta} \mid \boldsymbol{D}_n) = \prod_{i=1}^{n} \prod_{t=1}^{y_i} \{\pi_{a_i,t}(\boldsymbol{\beta})\}^{\delta_{i,t}} \{1 - \pi_{a_i,t}(\boldsymbol{\beta})\}^{1-\delta_{i,t}} = \prod_a \prod_t \frac{\{\exp(\boldsymbol{x}_{a,t}^\mathsf{T}\boldsymbol{\beta})\}^{d_{a,t}}}{\{1 + \exp(\boldsymbol{x}_{a,t}^\mathsf{T}\boldsymbol{\beta})\}^{r_{a,t}}},$$

denoting the linear terms by $\boldsymbol{x}_{a,t}^\mathsf{T}\boldsymbol{\beta}$, with $\delta_{i,t} = \delta_i 1_{\{t,\ldots,60\}}(y_i)$, $r_{a,t} = \sum_{i=1}^n 1_{\{t,\ldots,60\}}(y_i) 1_a(a_i)$, and $d_{a,t} = \sum_{i=1}^n \delta_{i,t} 1_a(a_i)$. That is, $\delta_{i,t}$ is the binary indicator that patient $i$ died on day $t$, $r_{a,t}$ is the number of patients in arm $a$ who were alive at the start of day $t$ and thus were at risk of dying that day, and $d_{a,t}$ is the number of patients in arm $a$ who died on day $t$, $a = E, C, C_0$ and $t = 1, \ldots, 60$.

The likelihood function may be factored into one contribution from the historical control data and another from the prospective trial data as follows:

$$L(\boldsymbol{\beta} \mid \boldsymbol{D}_n) = \left[ \prod_{a=E,C} \prod_t \frac{\{\exp(\boldsymbol{x}_{a,t}^{\mathsf{T}}\boldsymbol{\beta})\}^{d_{a,t}}}{\{1+\exp(\boldsymbol{x}_{a,t}^{\mathsf{T}}\boldsymbol{\beta})\}^{r_{a,t}}} \right] \times \left[ \prod_t \frac{\{\exp(\boldsymbol{x}_{C_0,t}^{\mathsf{T}}\boldsymbol{\beta})\}^{d_{C_0,t}}}{\{1+\exp(\boldsymbol{x}_{C_0,t}^{\mathsf{T}}\boldsymbol{\beta})\}^{r_{C_0,t}}} \right].$$

Only the likelihood contribution of historical control data depends on the historical bias parameters $\zeta$ and $\boldsymbol{z}$. In this way, the proposed approach augments the prospective data likelihood with a historical control data contribution that includes historical bias parameters. This factorization illuminates two things: (i) multiple historical data sources could augment the prospective likelihood function each with their own contribution and historical bias parameters, (ii) individual-level historical data are not required for inference, rather sufficient statistics are all that is needed, which in this case are the number at risk and the number who died on each day during the 60-day follow up period. Recognizing that $L(\boldsymbol{\beta} \mid \boldsymbol{D}_n)$ has the structure of a logistic regression model and following a similar strategy to (Polson et al., 2013), we developed an efficient and conditionally conjugate Pólya-Gamma Gibbs sampler. We provide additional details in the Supplement (Murray et al., 2020), and freely-available, user-friendly `R` software for implementation of all models under consideration (see Supplementary Materials).

# 5 Design Considerations

## 5.1 Preliminaries

For the septic shock ICU trial, we anticipate an accrual rate of 34 patients per month across 10–15 sites. Patients will be randomized in permuted blocks by site, so that throughout the trial at each site there will be approximately an equal number of patients assigned to arms $E$ and $C$. We aim to enroll up to 956 patients, i.e. 478 patients in each concurrent arm, which is expected to take 28 months. We provide justification for this sample size below. We plan to carry out up to three tests at 10 month intervals. To ensure that each patient who remains alive will be followed for at least 60 days, we plan to complete enrollment at 28 months and carry out the final test at 30 months.

Because the goal is to determine whether $E$ confers a survival benefit relative to $C$, but not in demonstrating that $E$ is harmful, we are concerned with controlling the probabilities of incorrectly recommending $E$ when $\mu_E = \mu_C$, i.e. making a type I error, and of incorrectly recommending $C$ when $E$ achieves a target benefit with $\mu_E > \mu_C$, i.e. type II error. A key complication, which is a central issue in our testing procedure, is that we do not know the degree to which the historical and prospective control data will agree. As we will demonstrate later, the degree of similarity between these data sources will affect the type I and II error rates of our design, since posterior inference will be based on the proposed Comm NPO model that incorporates the historical control data adaptively. To deal with this, we choose monitoring boundaries and a sample size

that control the type I and II error rates when posterior estimation is based on the Trad NPO model, which is fit only to the prospective data, as done conventionally. Specifically, omitting the historical control data $C_0$, we design the trial to have type I error probability 0.025 and type II error probability 0.10, equivalently, power 0.90. This strategy leads to increased power and reduced expected sample size when the historical data are congruous with the prospective data, while providing the desired power when the historical data are so incongruous that the Comm NPO model learns to borrow zero strength. For intermediate cases, the operating characteristics of the proposed design with the Comm NPO model are less clear, but may be evaluated using computer simulation.

## 5.2 Determining an Alternative

Under our utility-based comparative test, the meaning of the word "power" is not entirely obvious, since the alternative $H_A: \mu_E > \mu_C$ is determined by the two time-varying hazard functions and a shared utility function. Below, we explain how elicited relative risk reduction (RRR) values were used to determine the two hazard functions and resulting different numerical values for $\mu_E$ and $\mu_C$ used to compute power and select a sample size. The two survival distributions under this alternative are given in Figure 1.

Our approach will ensure that the design is adequately powered even when the historical control data are substantially incongruous with the prospective control data. This is because, in this case, the proposed Comm NPO model will tend to borrow little strength from the historical control data and result in posterior inference that is similar to that from the Trad NPO model fit only to the prospective data. In contrast, when the historical and prospective control data are congruous, the proposed Comm NPO model will tend to borrow substantially from the historical control data, leading to a design with larger power and smaller mean sample size. In Section 6, we will report computer simulation results that illustrate how the magnitude of the difference between the historical and prospective control data will affect the operating characteristics of the proposed design.

To obtain numerical values of $\mu_C$ and target $\mu_E$ for constructing a design, we set $R = 60$ so these parameters reflect 60-day RMS, and elicited information from FS and SK, who will be running the trial and are co-authors of this paper. To do this, we used the RRR in mortality through follow up day $t$, which was familiar to FS and SK, defined as

$$RRR_t = 1 - \frac{\Pr(Y^* \le t \,|\, A = E)}{\Pr(Y^* \le t \,|\, A = C)}.$$

Motivated by the survival benefits of external cooling observed in the trial reported by Schortgen et al. (2012), Drs. Schortgen (FS) and Katsahian (SK) hypothesized a 40% RRR in mortality for $E$ versus $C$ through calendar day 15, but a smaller 20% reduction through day 30, denoted by $RRR_{15} = 0.40$ and $RRR_{30} = 0.20$. To fully specify the target alternative scenario, we defined hazards of death in each concurrent arm on each day during the 60-day observation period, subject to the hypothesized $RRR_{15} = 0.40$

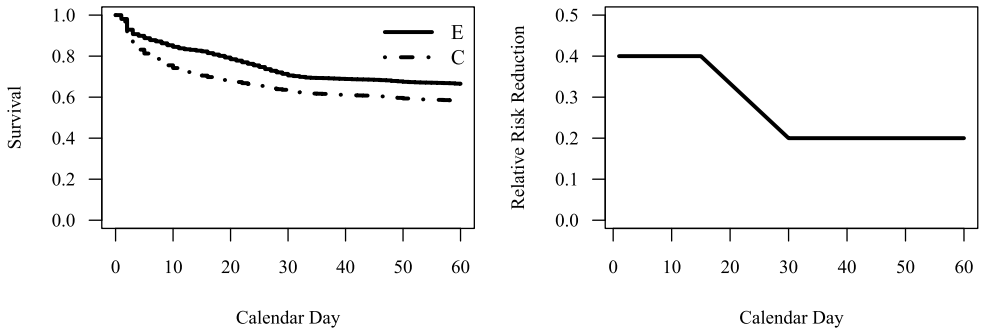


Figure 1: Survival distributions for the two prospective arms under the target alternative scenario (left panel) and the corresponding relative risk reduction of mortality for $E$ versus $C$ (right panel).

and $RRR_{30} = 0.20$, as follows. First, we set $\pi_{C,t} = \widehat{\pi}_{C_0,t}$, for $t = 1, \ldots, 60$, where $\widehat{\pi}_{C_0,t}$ is the posterior mean from a discrete time survival model with the GMRF prior fit to the historical control data reported by Asfar et al. (2014). The survival distribution in arm $C$ that corresponds to these hazards is displayed as the dot-dash line in the left panel of Figure 1. Second, relying on the information provided by FS and SK, we assumed that $RRR_1 = \cdots = RRR_{15} = 0.40$, $RRR_{30} = \cdots = RRR_{60} = 0.20$, and used linear interpolation to obtain values for $RRR_{16}$ through $RRR_{29}$. The resulting RRR curve is displayed in the right panel of Figure 1. Third, given $\pi_{C,t}$ and $RRR_t$, we solved for $\pi_{E,t}$, for each $t = 1, \ldots, 60$. This gave the null RMS $\mu_C = \widehat{\mu}_{C_0} = 39.8$ days and target $\mu_E = 44.9$ days. The survival distribution in arm $E$ that corresponds to the elicited hazards is displayed as the solid line in the left panel of Figure 1.

## 5.3 Controlling the Type I and II Error Rates

We will allow the monitoring boundary values for our comparative test to differ across the three planned data analyses. Let $p_j$ and $q_j$ denote the boundary values for analysis $j = 1, 2, 3$. To ensure that the design recommends either $E$ or $C$ at the trial's conclusion, we require $p_3 = q_3$. Following the theoretical method of Shi and Yin (2019), for $j = 1, 2, 3$, we set $p_j = \Phi(a_j)$ and $q_j = \Phi(b_j)$ where $a_j$ and $b_j$ reflect upper and lower frequentest group-sequential Z-score boundaries that provide the desired type I error rate and $\Phi(x)$ denotes the standard normal distribution function. Specifically, we set $p_1 = 0.999$, $q_1 = 0.145$, $p_2 = 0.995$, $q_2 = 0.408$ and $p_3 = q_3 = 0.977$, which were obtained from the `gsDesign` package in `R` and correspond to an asymmetric group-sequential design based on the Hwang-Shih-DeCani spending function (Hwang et al., 1990). For this design, asymptotically the probability of crossing the upper boundary at or before analysis $j$ under the null is given by $0.025[1 - \exp\{4(j/3)\}]/[1 - \exp(4)]$, $j = 1, 2, 3$. Similarly, asymptotically the probability of crossing the lower boundary without crossing the upper boundary at or before analysis $j$ under the null is given by $(1 - 0.025)[1 - \exp\{2(j/3)\}]/[1 - \exp(2)]$, $j = 1, 2, 3$. We used an iterative algorithm

to select a sample size that controls type II error rate when using the Trad PO model that ignores the historical control data so that incorporating these data provides the potential for greater power and shorter a trial. Details about this iterative algorithm are provided in the Web Supplement. We validated the type I and II error rates of our design for the planned sample size and probability models using computer simulation, as reported in Section 6.

# 6 Simulation Study

Our computer simulation study has two major aims: (i) validate the type I and II error rates of our planned group sequential design, when posterior estimation is based on the Trad NPO model fit only to prospective data, and (ii) assess how the degree of agreement between the prospective control and historical control data affects the proposed design's operating characteristics when posterior estimation is based on the proposed Comm NPO model that adaptively incorporates the historical control data. For comparison, we included a group-sequential log-rank test with monitoring boundaries analogous to the proposed design, which we refer to as the "LRT" approach. We also evaluated the Comm PO model that assumes historical control arm bias follows a proportional-odds assumption, and the Trad PO model that does not incorporate the historical control arm data and also assumes the concurrent arms follow a proportional-odds assumption. For the Bayesian methods, after 200 warm-up iterations, we ran our Gibbs sampler for 2,000 iterations and used the resulting samples for posterior calculations. Each Gibbs sampler run for the Comm NPO model took about six minutes. The samplers for other models required fewer steps and were slightly faster, requiring one to five minutes per run. To speed up computations, we simulated trials in parallel on a HP Linux distributed cluster.

We simulated 2,000 trials each under a null scenario and a target alternative scenario. For aim (i), prospective control data were generated from a survival distribution with $\pi_{C,t} = \widehat{\pi}_{C_0,t}$, for $t = 1, \ldots, 60$, i.e. $\zeta = 0$ and $\boldsymbol{z} = \mathbf{0}$, so that $C$ was exactly congruous with the observed historical control data $C_0$. For the null scenario, the prospective external cooling data were generated from this same survival distribution. For the alternative scenario, data were generated from the target survival distribution displayed in Figure 1. This investigation facilitates achieving aim (i) as well as evaluating the benefit of using our proposed Comm NPO model when the prospective and historical control data tend to be congruous. To achieve aim (ii), we simulated an additional 10,000 trials with varying differences between $\mu_C$ and $\widehat{\mu}_{C_0} = 38.9$ under both the null and the target alternative, which we always defined in terms of the elicited RRR curve displayed in Figure 1. For each trial, we sampled $\zeta \sim \text{Uniform}(-0.5, 0.5)$, but kept $\boldsymbol{z} = \mathbf{0}$, and generated prospective control data from a survival distribution with $\text{logit}\{\pi_{C_0,t}\} = \text{logit}\{\widehat{\pi}_{C,t}\} + \zeta$, for $t = 1, \ldots, 60$. Recall that $\zeta$ quantifies the magnitude of disagreement between the true survival distribution of the $C$ data and the observed historical $C_0$ data. Since $\boldsymbol{z} = \mathbf{0}$, the shape of the temporal variation in the hazards for death remain similar across sources, however. Since $\{\pi_{a,t}, t = 1, \cdots, 60\}$ determine $\mu_a$ through (1), $\mu_C$ is a one-to-one function of $\zeta$ in our computer simulation study. We determined that $\zeta \in [-0.5,\ 0.5]$, which corresponds to $\mu_C \in [31.0,\ 46.6]$ days, provides a suitably wide

| Method | Analysis 1 | 2 | 3 | Overall | mSS |
|---|---|---|---|---|---|
| *Null Scenario (Type I Error Probabilities)* | | | | | |
| LRT | 0.000 | 0.002 | 0.017 | 0.020 | 955.1 |
| Trad PO | 0.002 | 0.004 | 0.018 | 0.024 | 953.7 |
| Trad NPO | 0.001 | 0.004 | 0.016 | 0.020 | 954.3 |
| Comm PO | 0.002 | 0.004 | 0.010 | 0.016 | 953.8 |
| Comm NPO | 0.002 | 0.003 | 0.009 | 0.014 | 954.3 |
| *Target Alternative Scenario (Power Figures)* | | | | | |
| LRT | 0.105 | 0.406 | 0.329 | 0.840 | 780.2 |
| Trad PO | 0.142 | 0.387 | 0.320 | 0.849 | 763.0 |
| Trad NPO | 0.152 | 0.447 | 0.292 | 0.890 | 740.4 |
| Comm PO | 0.321 | 0.486 | 0.163 | 0.970 | 625.3 |
| Comm NPO | 0.284 | 0.484 | 0.192 | 0.960 | 648.9 |

Table 1: Probabilities of recommending $E$ over $C$ under the null and target alternative scenario with $\zeta = 0$ and $\boldsymbol{z} = \mathbf{0}$. Reported values reflect the proportion of 2,000 simulated trials that recommended $E$ at analysis 1, 2 and 3, and overall, respectively. mSS = mean sample size when ignoring the non-binding lower stopping boundary.

range of values for $\mu_C$ around $\widehat{\mu}_{C_0} = 38.9$ to evaluate the operating characteristics of our proposed design under varying magnitudes of disagreement between the prospective and historical control arms. Under the null, we used $\pi_{E,t} = \pi_{C,t}$, $t = 1, \ldots, 60$. Under the alternative, given $\zeta$ and $\{\pi_{C,t} : t = 1, \ldots, 60\}$, we used $\{\pi_{E,t} : t = 1, \ldots, 60\}$ values corresponding to the elicited RRR. We generated observations for the $E$ arm from the survival distribution with the resulting hazards.

For each simulated trial, we assigned each of the 956 subjects a random enrollment day drawn uniformly between calendar days 1 and 855, and we carried out data analyses at calendar days 305, 610 and 915. Based on the enrollment day for each simulated subject, we were able to determine the observed data at each of the three analyses. For the LRT, at each data analysis we calculated and stored the log-rank test statistic. For the Bayesian designs, at each data analysis we calculated and stored $\Pr(\mu_E > \mu_C \,|\, \boldsymbol{D}_n)$, $\widehat{\mu}_E = \mathrm{E}[\mu_E \,|\, \boldsymbol{D}_n]$, $\widehat{\mu}_C = \mathrm{E}[\mu_C \,|\, \boldsymbol{D}_n]$ and $\mathrm{Var}[\mu_E - \mu_C \,|\, \boldsymbol{D}_n]$. For the Comm NPO and PO models, we also calculated and stored the posterior probability of exchangeability between the current and historical control arms with respect to $\zeta$ and $\boldsymbol{z}$, e.g., $\mathrm{PPE} = \Pr(\nu_\zeta = 1 \,|\, \boldsymbol{D}_n)$. To quantify the amount of borrowing from the historical data at each analysis, we defined, calculated and stored the relative information gain (RIG) as follows,

$$\mathrm{RIG} = \frac{1/\mathrm{Var}(\mu_E - \mu_C \,|\, \boldsymbol{D}_n; \text{Comm NPO})}{1/\mathrm{Var}(\mu_E - \mu_C \,|\, \boldsymbol{D}_n; \text{Trad NPO})} - 1,$$

and similarly for the Comm NPO model. RIG measures the increase in posterior precision for the treatment effect of interest, $\mu_E - \mu_C$, due to incorporation of the historical control data. Compared to effective historical sample size (Hobbs et al., 2011), RIG is a more general metric for the amount of strength borrowed from the historical data.

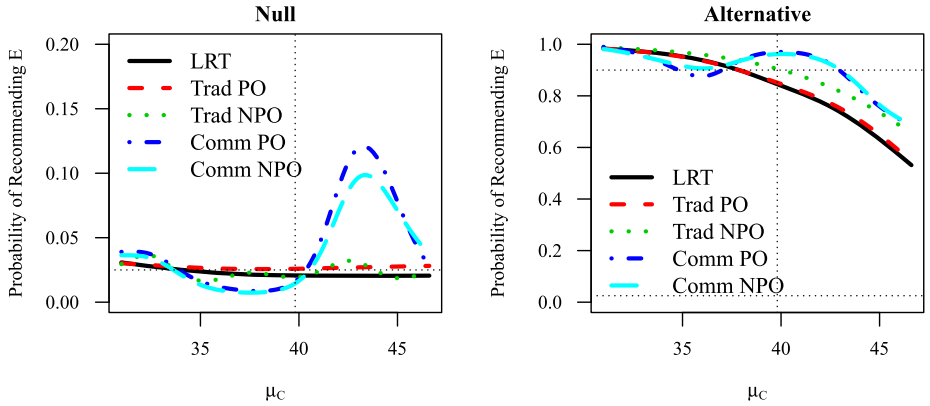


Figure 2: Probability of recommending external cooling as a function of $\mu_C$ under the null (left panel) and target alternative (right panel). The vertical dotted line reflects the point of zero historical bias with $\widehat{\mu}_{C_0} = 39.8$ as estimated from the historical control data. The horizontal dotted lines reflect the desired 0.025 one-sided type I error rate and 0.90 power. For visual clarity, the y-axis range differs across the two panels.

Table 1 gives the proportions of simulated trials that recommended $E$ over $C$ under the null and alternative hypotheses, in the case with $\zeta = 0$ and $\boldsymbol{z} = 0$ where $C$ and $C_0$ are fully congruous. The results in Table 1 reflect a non-binding lower boundary for recommending $C$, i.e. the trial may continue when the lower boundary is crossed. When adhering to the lower boundary, for all three methods and under both the null and alternative, the proportion of trials recommending $E$ is unchanged, whereas mSS = mean sample size is much smaller. Because stopping early to recommend $E$ is rare under the null, mSS is similar for all methods. Under the target alternative, LRT has the largest mSS, followed by Trad PO, Trad NPO, Comm NPO, and then Comm PO, which stopped to recommend $E$ at the first analysis in nearly one third of the simulated trials.

The LRT and Trad methods exhibited one-sided type I error rates near the intended nominal 0.025 level, whereas the Comm methods exhibited slightly lower rates. Under the target alternative scenario, the LRT and Trad PO methods exhibit similar power near 0.85, whereas the Trad NPO method exhibits power near 0.90, as intended. The lower power of these methods is likely due to the fact that LRT is most powerful for proportional-hazards effects and the Trad PO model assumes proportional-odds, neither of which is the case here. In contrast, the Comm PO and NPO methods exhibit similarly higher power of 0.97 and 0.96, respectively. This substantial increase in power is due to incorporating the historical control data when the prospective control data follow a survival distribution that is congruous with the observed historical control data. In particular, there is little power loss when modeling the historical bias with both $\zeta$ and $\boldsymbol{z}$, rather than $\zeta$ alone.

Figure 2 displays an estimate of the probability of recommending $E$ under the null and target alternative as a function of $\mu_C$, corresponding to each method. The probabil-

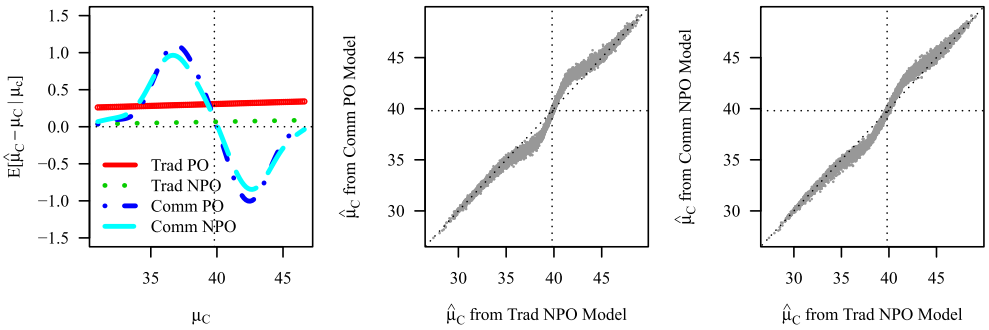


Figure 3: Empirical estimate for control arm bias $E[\widehat{\mu}_C - \mu_C \,|\, \mu_C]$ in $\mu_C$ (left panel), and scatter plot of $\widehat{\mu}_C$ from Trad NPO versus Comm PO (middle panel) and Comm NPO (right panel). Results displayed are at the third analysis under the target alternative. Each grey dot corresponds to one simulated trial.

ity estimates are based on a generalized additive model fit to the relevant simulated data. The estimates are similar when based, instead, on LOESS (locally weighted smoothing) (Cleveland, 1979). Under the null, LRT and the two Trad models have one-sided type I error rates near the 0.025 nominal rate for all $\mu_C$, whereas the type I error rate for the Comm models vary with $\mu_C$. In particular, the type I error rate inflates when $\mu_C$ is slightly above $\widehat{\mu}_{C_0} = 39.8$, deflates when $\mu_C$ is slightly below $\widehat{\mu}_{C_0}$, and moves back toward the nominal rate as $\mu_C$ deviates further from $\widehat{\mu}_{C_0}$. The Comm NPO model exhibits less variability than the Comm PO model in this regard. Under the target alternative, LRT and Trad PO exhibit lower power than Trad NPO for all $\mu_C$, whereas the power of Curr-Hist follows the same trend as its type I error rate, achieving the most power when $\mu_C$ is equal to or slightly greater than $\widehat{\mu}_{C_0}$. For the Comm models, both the null and alternative curves for Pr(Recommending $E$) as a function of $\mu_C$ in Figure 2 have non-monotone shapes, with a temporary dip in each curve for values near and below $\widehat{\mu}_{C_0}$ and a temporary rise for values near and above $\widehat{\mu}_{C_0}$. This is due to the fact that, when $\mu_C$ is in these intermediate ranges below and above $\widehat{\mu}_{C_0}$, $\widehat{\mu}_C$ based on the Comm models tends to be shrunk toward $\widehat{\mu}_{C_0}$. Thus, the type I error is inflated for values of $\mu_C$ slightly above $\widehat{\mu}_{C_0}$ and the power is reduced for values of $\mu_C$ slightly below $\widehat{\mu}_{C_0}$. This estimation bias may be considered the price paid for using a Comm model to borrow strength from the historical data. Additional simulation results in the Web Supplement (see Supplementary Materials) demonstrate that specifying a lower prior probability on the spike distributions (i.e. $p_\zeta$ and $p_z$) will reduce type I error inflation, but at the cost of lower power when the two sources are congruous.

Figure 3 displays estimates of $E[\widehat{\mu}_C - \mu_C \,|\, \mu_C]$ as a function of $\mu_C$ for each Bayesian model, and scatter plots of $\widehat{\mu}$ based on each Comm model versus the Trad NPO model. Both Trad models are biased upward for all $\mu_C$, through the Trad PO model much more so, likely because the PO assumption is violated under the target alternative. Under the null, which satisfies the PO assumption, neither Trad model exhibits noticeable bias for any $\mu_C$ (results not shown here). Both Comm models tend to over-estimate $\mu_C$ when

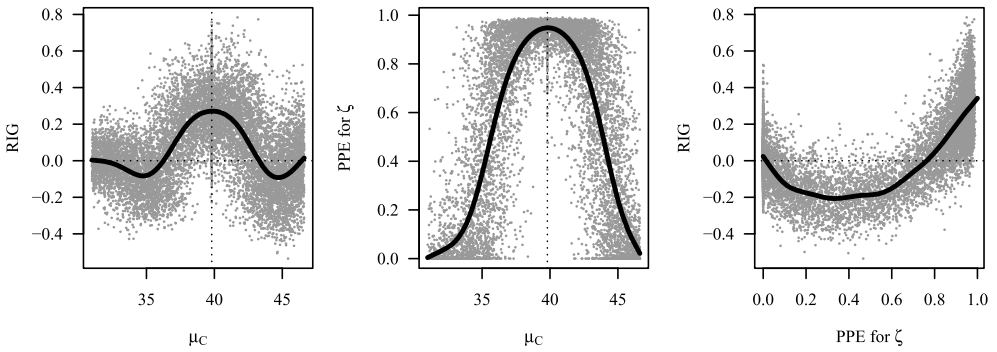


Figure 4: Borrowing properties of the Comm NPO model under the null (results are similar under the alternative). Scatter plots of PPE = Posterior probability of exchangeability for $\zeta$ (left panel), RIG = Relative information gain for $\widehat{\mu}_C$ versus Trad NPO (middle panel), and RIG vs PPE for $\zeta$ (right panel). Each grey dot corresponds to one simulated trial.

$\mu_C$ is below $\widehat{\mu}_{C_0}$ and under-estimate $\mu_C$ when $\mu_C$ is slightly above $\widehat{\mu}_{C_0}$. As shown by the middle and right panels, posterior mean estimates of $\mu_C$ from each Trad model are shrunk toward the value of $\widehat{\mu}_{C_0}$ from the Comm NPO model when this estimate is between 35 and 45.

Figure 4 further elucidates the borrowing properties of the Comm NPO model, with scatter plots of PPE = posterior probability of exchangeability for $\zeta$ (left panel) and RIG (middle panel) versus $\widehat{\mu}_C$, as well as RIG versus PPE for $\zeta$ (right panel). The Comm PO model exhibits similar patterns in this regard. Recall that the Comm NPO model allows for temporal variation in the historical control arm bias through $\boldsymbol{z}$ as well. Because this simulation study only varied the value of $\zeta$, the PPE for $\boldsymbol{z}$ was relatively invariant in $\mu_C$ at around 0.85 and thus is not illustrated here. RIG provides an alternative means to capture the strength of borrowing across all bias parameters, regardless of the Comm model's complexity. When $\widehat{\mu}_C$ from Trad NPO is near $\widehat{\mu}_{C_0}$, Comm NPO tends to have PPE near 1, which leads to RIG $> 0$ and thus greater posterior precision for the treatment effect of interest. In contrast, as $\widehat{\mu}_C$ from Trad NPO deviates far from $\widehat{\mu}_{C_0}$, PPE and RIG tend toward 0, indicating that incorporation of historical control data that is incongruous with the new control data leads to little gain in posterior precision for the treatment effect of interest. When $\widehat{\mu}_C$ from Trad NPO is near 35 or 45, Comm NPO tends to have intermediate values for PPE and RIG $< 0$, i.e. it suffers some reduction in posterior precision for the treatment effect of interest.

# 7 Discussion

We have proposed and investigated the operating characteristics of a Bayesian group sequential design for a randomized controlled trial to assess the effectiveness of using external cooling to control fever in patients with septic shock admitted to an ICU.

The proposed design will use a group sequential comparative test based on the difference between 60-day survival distributions in the external cooling and control arms. Statistical inference will be based on a Bayesian discrete time survival model that facilitates borrowing strength from the historical control data using an intuitive regression framework.

Our computer simulations showed that, if the historical control data are unbiased, then the proposed approach will control the type I error rate and provide substantially greater power than similar Bayesian designs that do not incorporate the historical control data as well as a more traditional design based on a group-sequential log-rank test. In contrast, when the historical data are biased, regardless of the magnitude, the proposed approach provides reasonable power and avoids run-away type I error inflation and bias in the estimate of the treatment effect of interest. Increasing the prior probability on the spike distribution for each historical bias parameter will reduce the maximum type I error inflation and estimate bias at the cost of lower power when the data sources are congruous.

If desired, one may modify the regression model's linear predictor to obtain covariate-adjusted treatment effects. For example, the proposed model may be extended to include enrollment center, patient prognostic covariates, or possibly high dimensional biologic/genomic variables. While doing this is conceptually straightforward, technically it may be much more complicated. In part, this is because such an extension may involve treatment-covariate interactions, thus requiring a more complex regression model and a covariate-specific group sequential decision structure.

## Supplementary Material

Web Supplement for "Robust Adaptive Incorporation of Historical Control Data in a Randomized Trial of External Cooling to Treat Septic Shock"
(DOI: 10.1214/20-BA1229SUPP; .pdf). Additional Gibbs sampling algorithm and sample size calculation details, as well as additional simulation results are available as a web supplement. The extra simulation results illustrate the impact of hyperparameter specification for the spike-and-slab prior distributions. Freely available `R` programs, and the required historical control data, to fit the proposed Bayesian models and reproduce the simulation results in the article may be obtained from the first author's Github page `https://8tmurray.github.io/`.

## References

Angus, D. C. and T. van der Poll (2013). Severe sepsis and septic shock. *New England Journal of Medicine 369*(9), 840–851. PMID: 23984731. 826

Annane, D., P. Aegerter, M. C. Jars-Guincestre, and B. Guidet (2003). Current epidemiology of septic shock. *American Journal of Respiratory and Critical Care Medicine 168*(2), 165–172. 826

Asfar, P., F. Meziani, J.-F. Hamel, F. Grelon, B. Megarbane, N. Anguel, J.-P. Mira, P.-F. Dequin, S. Gergaud, N. Weiss, F. Legay, Y. Le Tulzo, M. Conrad, R. Robert, F. Gonzalez, C. Guitton, F. Tamion, J.-M. Tonnelier, P. Guezennec, T. Van Der Linden, A. Vieillard-Baron, E. Mariotte, G. Pradel, O. Lesieur, J.-D. Ricard, F. Hervé, D. du Cheyron, C. Guerin, A. Mercat, J.-L. Teboul, and P. Radermacher (2014). High versus low blood-pressure target in patients with septic shock. *New England Journal of Medicine 370*(17), 1583–1593. PMID: 24635770. 826, 827, 829, 831, 835

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*(1), 1–20. MR1105822. doi: https://doi.org/10.1007/BF00116466. 827, 831

Caironi, P., G. Tognoni, S. Masson, R. Fumagalli, A. Pesenti, M. Romero, C. Fanizza, L. Caspani, S. Faenza, G. Grasselli, G. Iapichino, M. Antonelli, V. Parrini, G. Fiore, R. Latini, and L. Gattinoni (2014). Albumin replacement in patients with severe sepsis or septic shock. *New England Journal of Medicine 370*(15), 1412–1421. PMID: 24635772. 826

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*(368), 829–836. MR0556476. 839

Dellinger, R. P., M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, T. M. Osborn, M. E. Nunnally, S. R. Townsend, K. Reinhart, R. M. Kleinpell, D. C. Angus, C. S. Deutschman, F. R. Machado, G. D. Rubenfeld, S. A. Webb, R. J. Beale, J.-L. Vincent, R. Moreno, and S. S. C. Guidelin (2013, Feb). Surviving Sepsis Campaign: International guidelines for management of severe sepsis and septic shock: 2012. *Critical Care Medicine 41*(2), 580–637. 826

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889. 832

Ghosh, J., Y. Li, and R. Mitra (2018, 06). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis 13*(2), 359–383. MR3780427. doi: https://doi.org/10.1214/17-BA1051. 830

Hennerfeind, A., A. Brezger, and L. Fahrmeir (2006). Geoadditive survival models. *Journal of the American Statistical Association 101*(475), 1065–1075. MR2324146. doi: https://doi.org/10.1198/016214506000000348. 830

Hobbs, B. P., B. P. Carlin, S. J. Mandrekar, and D. J. Sargent (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics 67*(3), 1047–1056. MR2829239. doi: https://doi.org/10.1111/j.1541-0420.2011.01564.x. 827, 837

Hobbs, B. P., D. J. Sargent, and B. P. Carlin (2012, 09). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis 7*(3), 639–674. MR2981631. doi: https://doi.org/10.1214/12-BA722. 827

Hwang, I. K., W. J. Shih, and J. S. De Cani (1990). Group sequential designs using

a family of type i error probability spending functions. *Statistics in Medicine 9*(12), 1439–1445. 835

Ibrahim, J. G. and M.-H. Chen (2000, 02). Power prior distributions for regression models. *Statistical Science 15*(1), 46–60. MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 827

Kaizer, A. M., J. S. Koopmeiners, and B. P. Hobbs (2018). Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics 19*(2), 169–184. MR3799610. doi: https://doi.org/10.1093/biostatistics/kxx031. 827

Murray, T. A., B. P. Hobbs, and B. P. Carlin (2015, 09). Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *The Annals of Applied Statistics 9*(3), 1549–1570. MR3418735. doi: https://doi.org/10.1214/15-AOAS840. 827, 832

Murray, T. A., Thall, P. F., Schortgen, F., Zohar, S., and Katsahian, S. (2020). "Web Supplement for "Robust Adaptive Incorporation of Historical Control Data in a Randomized Trial of External Cooling to Treat Septic Shock"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1229SUPP. 833

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 827, 833

Proschan, M. A., K. K. G. Lan, and J. T. Wittes (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer, New York, NY. 826

Royston, P. and M. K. Parmar (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology 13*(1), 1–15. 827, 828

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca-Raton, FL: Chapman & Hall/CRC Press. MR2130347. doi: https://doi.org/10.1201/9780203492024. 827, 831

Schortgen, F., K. Clabault, S. Katsahian, J. Devaquet, A. Mercat, N. Deye, J. Dellamonica, L. Bouadma, F. Cook, O. Beji, C. Brun-Buisson, F. Lemaire, and L. Brochard (2012). Fever control using external cooling in septic shock. *American Journal of Respiratory and Critical Care Medicine 185*(10), 1088–1095. 826, 834

Shi, H. and G. Yin (2019). Control of type I error rates in Bayesian sequential designs. *Bayesian Analysis 14*(2), 399–425. MR3934091. doi: https://doi.org/10.1214/18-BA1109. 826, 835

Shime, N., K. Hosokawa, and G. MacLaren (2013). Does cooling really improve outcomes in patients with septic shock? *American Journal of Respiratory and Critical Care Medicine 187*(11), 1274–1275. PMID: 23725626. 826

Singer, M., C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, R. Bauer, Michael amd Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S.

Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, and D. C. Angus (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *The Journal of the American Medical Association 315*(8), 801–810. 826