Incorporating Historical Models with Adaptive Bayesian Updates

Philip S. Boonstra[1]* Ryan P. Barbaro[2,3]

[1]Department of Biostatistics
[2]Division of Pediatric Critical Care
[3]Child Health Evaluation and Research Unit
University of Michigan, Ann Arbor, MI

Version: March 14, 2018

Abstract

This paper considers Bayesian approaches for incorporating information from a historical model into a current analysis when the historical model includes only a subset of covariates currently of interest. The statistical challenge is two-fold. First, the parameters in the nested historical model are not generally equal to their counterparts in the larger current model, neither in value nor interpretation. Second, because the historical information will not be equally informative for all parameters in the current analysis, additional regularization may be required beyond that provided by the historical information. We propose several novel extensions of the so-called power prior that adaptively combine a prior based upon the historical information with a variance-reducing prior that shrinks parameter values toward zero. The ideas are directly motivated by our work building mortality risk prediction models for pediatric patients receiving extracorporeal membrane oxygenation, or ECMO. We have developed a model on a registry-based cohort of ECMO patients and now seek to expand this model with additional biometric measurements, not available in the registry, collected on a small auxiliary cohort. Our adaptive priors are able to leverage the efficiency of the original model and identify novel mortality risk factors. We support this with a simulation study, which demonstrates the potential for efficiency gains in estimation under a variety of scenarios. Bias-Variance Tradeoff; Combining Information; Hierarchical Shrinkage; Power Prior; Regularized Horseshoe Prior

1 Introduction

When a statistical model is published, there are often already models for the same outcome. Although the new model and the existing models may each differ in their target populations,

*Correspondence to: Philip S. Boonstra, PhD, SPH II, 1415 Washington Hts, Ann Arbor, MI, 48109; philb@umich.edu

underlying sets of predictors, or in other ways [e.g., Becker and Wu, 2007], there is usually some historical information available when the new model was built. In that sense, this framework of sequential but independent model development is not fully utilizing available historical information. In this paper, we propose Bayesian approaches that incorporate the posterior distribution from the historical model into a prior for the new model when the set of historical covariates is strictly nested within the set of the new covariates.

Our motivation for this work is a short-term mortality risk prediction model ("Ped-RESCUERS") for pediatric patients receiving extracorporeal membrane oxygenation (ECMO) support using information on 1611 pediatric patients treated between the years 2009-2012 [Barbaro et al., 2016]. The source population was the Extracorporeal Life Support Organization (ELSO), an international registry of ECMO patients, and the pertinent data are limited to patient clinical characteristics (weight, age, sex, primary diagnosis, co-morbidities, complications, pre-ECMO supportive therapies) and ECMO-specific measurements (blood gas measurements and ventilator settings). In total, Ped-RESCUERS uses eleven predictors. Patient-specific biometric measurements of renal, hepatic, neurologic and hematologic dysfunction that may be associated with mortality on ECMO are not generally collected in the ELSO registry. We posited a list of eleven such additional potential risk factors and collected a cohort of 178 non-overlapping patients at three ECMO-providing centers. The data consist of both the eleven risk factors in PED-RESCUERS and eleven biometric measurements not in the registry. We require a model that potentially includes all 22 predictors. We report on these new data in Barbaro et al. [2018]. However, given the ratio of sample size to number of predictors and the subsequent variability in estimates, it is more statistically incumbent to make use of the information from the original large cohort of patients with unmeasured biometric measurements to gain efficiency. Yet, the process of doing so may introduce bias, due to the different predictors included in each model. It is these competing objectives we seek to balance.

In other cases, there may only be very limited historical information, meaning that the number of historical predictors is *much* less than the total number of potential predictors under study. It is reasonable to expect that incorporating even such limited historical information should result in a model that is non-inferior to a modeling approach ignoring the historical information entirely. For example, one alternative to using a literal historical prior would be to regularize estimation and prediction with a prior that shrinks parameters toward zero, as in the Bayesian Lasso [Park and Casella, 2008] and others [Griffin and Brown, 2005, Armagan et al., 2013]. Based on this logic, an ideal strategy combines these approaches: incorporating whatever historical information is available and, for those parameters about which it is not informative, controlling variability by shrinking them to zero in the 'usual' way.

We achieve this here through an extension of the power prior, which uses the historical data likelihood as the current prior, and this prior density is further raised to a power $0 \leq \phi \leq 1$ [Ibrahim and Chen, 2000, Ibrahim et al., 2015]. Setting $\phi = 0$ or $\phi = 1$ corresponds, respectively, to ignoring the historical likelihood entirely or a fully Bayesian update. Using $0 < \phi < 1$ allows for partial borrowing of the historical likelihood in the presence of heterogeneity, and this can be made adaptive by considering a hyperprior on ϕ itself [Duan et al., 2006, Neuenschwander et al., 2009]. The novel idea in our approach is to

use ϕ to vary the relative contributions of the historical prior ($\phi = 1$) and a variance-reducing prior that shrinks to zero ($\phi = 0$).

In the classical power prior, the historical and current models include the same predictors. In one extension, Chen et al. [1999] approach a related problem by constructing a second, artificial historical likelihood that uses a constant for the outcome and copies of the added covariates from the current data. This has the effect of shrinking the corresponding regression coefficients toward zero and so is actually more similar to a typical shrinkage-to-zero prior. Ibrahim et al. [2002] consider the general setting of fitting GLMs using power priors when there are missing covariates in the historical and/or current datasets. Crucially, both the historical and current likelihoods condition on the same set of covariates, and missingness is ancillary to the main statistical problem.

Beyond the power prior, there exist alternative approaches for incorporating historical information. A meta-analysis statistically combines univariable or multivariable associations from multiple studies based upon each analysis' variance or covariance matrix [Walker et al., 2008, Chen et al., 2012, Jackson and Riley, 2014]. The main advantage of a meta-analysis is its simplicity, even when combining more than two models, because only summaries statistics are required. Among other assumptions, however, all models to be combined must include the same predictors. Recently, several authors have proposed strategies for incorporating summary-level historical information via constraints on the likelihood [Chatterjee et al., 2016, Grill et al., 2017, Cheng et al., 2018]. Antonelli et al. [2017] propose Bayesian approaches for borrowing information from the dataset with additional covariates to improve estimation of the average causal effect in the dataset with fewer covariates. Relative to previous important work in this area, we highlight two distinctive features of our approach. First, we account for the underlying uncertainty in the historical information by using the posterior variance from the historical model as the prior variance for the current model. Second, we explicitly ignore any historical information on the intercept parameter, which, in the case of a binary outcome, borrows information while still allowing for differences in the underlying true prevalence between the historical and the current models.

The proceeding sections develop the required ingredients of our approach that combines historical-based prior shrinkage and shrinkage to zero. Section 2 reviews the 'regularized horseshoe prior' [Carvalho et al., 2009, 2010, Piironen and Vehtari, 2015, 2016, 2017], which is the shrinkage-to-zero prior that we use. Section 3 outlines the construction two historical-based priors derived from models fit to a subset of the current set of covariates under consideration. Section 4 proposes how to adaptively combine the historical information with the shrinkage prior. Section 5 and 6 demonstrate our methods with a simulation study and analysis of the motivating ECMO mortality risk prediction model, respectively. Section 7 concludes with a discussion.

2 Shrinkage-to-Zero Priors

Let $g(\cdot)$ denote the link function of a generalized linear model (GLM) and $\pi(\cdot|\cdot)$ and $\pi(\cdot)$ denote conditional and marginal distributions, respectively. We will use the prior/posterior

nomenclature to indicate whether conditioning is on data. Capital and lowercase letters, respectively, indicate random data and observed data; all types of Greek letters will be reserved for parameters. Standard font will be used for scalar or vector valued quantities, and boldface font will be reserved for matrix-valued quantities.

A GLM for an outcome Y is fit to a length-p+q vector of covariates X, $g(E[Y|X=x]) = x^\top\beta$, using n datapoints, $\{\boldsymbol{y}, \boldsymbol{x}\}$. The covariates $\boldsymbol{x}$ are standardized to their empirical mean and empirical standard deviation. We do not distinguish between the first p and the final q elements of β yet (these identify the historical and current elements, respectively) but will do so in subsequent sections. The vector $\beta = \{\beta_1, \ldots, \beta_{p+q}\}$ is of primary interest. Given the likelihood $\pi(\boldsymbol{y}|\beta)$ and prior $\pi(\beta|\theta)\pi(\theta)$, with θ a vector of hyperparameters that is conditionally independent of $\boldsymbol{y}$ given β, the posterior is $\pi(\beta, \theta|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\beta)\pi(\beta|\theta)\pi(\theta)$. Often, the prior $\pi(\beta|\theta)\pi(\theta)$ is selected to regularize parameters by shrinking estimates toward zero, thereby reducing variance and increasing efficiency, as discussed in the Introduction. Many such shrinkage priors can be written as products of conditionally independent normal priors on β_j: e.g., $\theta = \{\theta_1, \ldots, \theta_{p+q}\}$ and $\pi(\beta|\theta) = \prod_j N(\beta_j|0, \theta_j^2)$. For example, if each θ_j^2 is independently inverse-gamma-distributed with a common shape and scale parameter equal to $k/2$, each β_j is marginally student-t-distributed with k degrees of freedom [e.g. Gelman et al., 2014]. A different choice of $\pi(\theta)$ conferring more adaptive shrinkage properties is the 'regularized horseshoe' [Carvalho et al., 2009, 2010, Piironen and Vehtari, 2015, 2016, 2017]. Given constants c, d and hyperparameters τ, $\lambda = \{\lambda_1, \ldots, \lambda_{p+q}\}$, the hyperprior is $\pi(\theta) \equiv \pi(\tau)\prod_{j=1}^{p+q}\pi(\lambda_j)$, where

$$\begin{aligned}
&\pi(\tau) = C^+(\tau|0,1); \quad \pi(\lambda_j) = C^+(\lambda_j|0,1), j = 1, \ldots, p+q \\
&\theta_j \equiv \theta(\tau, \lambda_j) = \left(1/d^2 + 1/[c^2\tau^2\lambda_j^2]\right)^{-1/2} \\
\Rightarrow &\pi_{\text{SZ}}(\beta|\theta) = \prod_j^{p+q} N(\beta_j|0, \theta_j^2). \qquad (1)
\end{aligned}$$

C^+ indicates the positive half-Cauchy distribution. The SZ subscript indicates 'shrinkage to zero'. The hyperparameter τ globally shrinks all parameters, while the λ_js multiplicatively offset τ and thus admit large individual variance components. The original horseshoe [Carvalho et al., 2009] implicitly used $c = 1$ and $d = \infty$, i.e. $\theta_j = \tau\lambda_j$, and others have since generalized it. First, Piironen and Vehtari [2016] suggested considering alternative values of c, which scales the global shrinkage, by linking its value to an implicit assumption about the a priori effective number of non-zero parameters in the model, say ξ_{eff}. The relationship is given by $\xi_{\text{eff}} \approx \sum_j \left(1 + [\sigma/\sqrt{n}]^2\theta_j^{-2}\right)^{-1}$, where σ is the dispersion. So, for example, if $p + q = 20$, $n = 500$ and $\sigma = 2$, under the original horseshoe, the prior mean of ξ_{eff} is $E[\xi_{\text{eff}}] \approx 17.1$. If instead $c = 0.01$, then $E[\xi_{\text{eff}}] \approx 3.3$. Typically, using $c = 1$ codifies a prior belief that most of the $p + q$ parameters are non-zero, which is unlikely when $p + q$ itself is large. Further, the expression for ξ_{eff} highlights that the choice of c ought to scale with $\sigma/\sqrt{n}$, all other things being equal. Larger sample sizes warrant smaller values of c. Based on this, Piironen and Vehtari recommend selecting $\tilde{\xi}_{\text{eff}} = E[\xi_{\text{eff}}]$ and then, assuming σ^{-2} is fixed, numerically solving $\tilde{\xi}_{\text{eff}} = E\left[\sum_j \left(1 + [\sigma/\sqrt{n}]^2\theta_j^{-2}\right)^{-1}\right]$ for c (θ_j is a function of c), where the expectation is taken with respect to $\pi(\tau, \lambda_j)$. The result will usually be

$c \ll 1$.

Subsequent work by Piironen and Vehtari [2017] argued that the original horseshoe tends to *under*-shrink large elements of β. A numerical consequence of this is that the original horseshoe may encounter challenges in its stochastic search through heavy tails [Piironen and Vehtari, 2015]. As a solution to both of these problems, they suggest to soft-truncate the tails of the horseshoe prior by including a diffuse normal prior with variance d^2. Choosing a finite-valued d results in the regularized horseshoe prior. Our strategy for choosing the hyperparameters c and d in this paper is to set d equal to a large value, $d = 15$, and then numerically solve $\tilde{\xi}_{\text{eff}} = E[\sum_j (n\sigma^{-2}\theta_j^2)/(1 + n\sigma^{-2}\theta_j^2)]$ for c, as before. A large d has minimal effect in the middle of the horseshoe prior but effectively thins out the heavy tails. We further discuss $\tilde{\xi}_{\text{eff}}$ in Section 5. The prior in (1) is the hierarchical shrinkage prior that we will extend in Section 4 to adaptively incorporate historical information. But first, Section 3 considers the prerequisite non-adaptive prior using the historical information alone.

3 Historical Shrinkage Prior

Separate the covariate vector into X^o and X^a, of length p and q, respectively. The *original* covariates X^o were measured in the historical analysis, and the *added* covariates X^a were not. We are interested in modeling $E[Y|X^o = x^o, X^a = x^a]$, but the historical model only estimates the smoothed version $E[Y|X^o = x^o] = E[E[Y|X^o = x^o, X^a]|X^o = x^o]$. The historical analysis conveys information about

$$g(E[Y|X^o = x^o]) = \mu_{\text{hist}} + (x^o)^\top \alpha, \tag{2}$$

This knowledge is quantified by the posterior distribution of α given the historical data, about which one likely only has access to summary statistics, e.g. the mean and covariance matrix. Our interest is not in model (2) but rather the embiggened model

$$g(E[Y|X^o = x^o, X^a = x^a]) = \mu + (x^o)^\top \beta^o + (x^a)^\top \beta^a. \tag{3}$$

We have a dataset of n observations, $\{\boldsymbol{y}, \boldsymbol{x}^o, \boldsymbol{x}^a\}$ and a likelihood $\pi(\boldsymbol{y}|\beta^o, \beta^a)$. A standard analysis of $\{\boldsymbol{y}, \boldsymbol{x}^o, \boldsymbol{x}^a\}$ alone might employ a shrinkage-to-zero prior as described in Section 2; that prior does not distinguish between historical and current covariates. Keeping in mind our ultimate goal of incorporating the historical information we have about α, this section lays out an alternative prior formulation based upon the historical analysis. We will then combine these priors in Section 4.

3.1 Naive Bayesian Update

A naive Bayesian (NB) update would directly apply the historical posterior on α as a prior on β^o, since these parameters correspond to the same set of covariates, namely X^o. More formally, one might use $m_\alpha \equiv E[\alpha]$ and $\boldsymbol{S}_\alpha \equiv \text{Var}[\alpha]$ as the respective prior mean and

variance for β^o. Then, given an optional scaling hyperparameter η, a conditional prior might be

$$\pi_{\text{NB}}(\beta^o|\eta) = N(\beta^o|m_\alpha, \eta \boldsymbol{S}_\alpha) \tag{4}$$

However, model (2) cannot hold for all patterns x^o if model (3) is the true generating model unless $\beta^a = 0$ (or, if for some fixed $q \times p$ matrix of weights $\boldsymbol{B}$, $X^a = \boldsymbol{B}X^o$ almost surely. In the special case that g is the identity link, this condition may be relaxed to equality in expectation, i.e. $E[X^a|X^o = x^o] = Bx^o$). Thus, in general, the naive Bayesian is implicitly assuming that $\beta^a \approx 0$, so as to be able to equate α and β^o in (4). To be consistent with this assumption, $\pi_{\text{NB}}(\beta^o|\eta)$ should be accompanied by a prior on β^a that strongly shrinks to zero. We discuss this further in Section 4.

3.2 Sensible Bayesian Update

Although the naive Bayesian update may improve efficiency, by construction it assumes $\alpha \approx \beta^o$ and $\beta^a \approx 0$. In general α and β^o are not equal, neither in value nor interpretation. Further, it is illogical to begin with a strong prior assumption that β^a – the novel set of covariates of interest – is approximately zero. The naive Bayesian update will introduce bias when β^a is far from zero. A more sensible Bayesian update would approximate the many-to-few mapping from $\{\beta^o, \beta^a\}$ to α and place a prior on that mapping. That mapping naturally arises from iterating the conditional expectation of Y given $X^o = x^o$:

$$E[Y|X^o = x^o] = E[E[Y|X^o = x^o, X^a]|X^o = x^o] = E[g^{-1}\left(\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a\right)|X^o = x^o].$$

Applying Model (2), i.e. taking $g(\cdot)$ of both sides, which – because the true(r) model is (3) – will only be an approximation of the conditional mean of Y given X^o, we obtain the following:

$$\mu_{\text{hist}} + (x^o)^\top\alpha \approx g\, E[g^{-1}\left(\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a\right)|X = x^o] \tag{5}$$

$$\mu_{\text{hist}} \approx g\, E[g^{-1}\left(\mu + (X^a)^\top\beta^a\right)|X^o = 0]. \tag{6}$$

This relates the available historical model with the current model. In particular, (6) obtains an approximation for the historical intercept parameter μ_{hist} by plugging in $x^o = 0$, which is predicated on $x^o = 0$ falling within the observed support of X^o and achieved by centering the covariates. This is useful because taking the difference between (5) and (6) completely removes μ_{hist} from the equation:

$$(x^o)^\top\alpha \approx g\, E[g^{-1}\left(\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a\right)|X^o = x^o] - g\, E[g^{-1}\left(\mu + (X^a)^\top\beta^a\right)|X^o = 0]. \tag{7}$$

Equation (7) is the basis of the sensible Bayesian update: it links Model (3) to a linear function of the parameters from Model (2), about which there is historical information. Furthermore, like the naive Bayesian update, the sensible Bayesian update avoids borrowing information on the historical intercept μ_{hist}. This relaxes a critical assumption: we do not

need to assume that the historical and current data generating models are identical but rather that the regression coefficients, that is, the underlying generating values of $\{\beta^o, \beta^a\}$, are identical. We discuss this more in the concluding section.

When the link function g is non-linear, constructing a prior based upon (7) would necessitate a Jacobian adjustment, and the adaptive priors that we subsequently develop in Section 4 would require numerically integrating over this Jacobian *at each iteration* of the Markov Chain, rendering such an approach computationally intractable. Practically, then, we must further linearize the mapping to obviate the Jacobian adjustment. Moving g across the integrals,

$$\begin{aligned}(x^o)^\top \alpha &\approx E[\mu + (x^o)^\top \beta^o + (X^a)^\top \beta^a | X^o = x^o] - E[\mu + (X^a)^\top \beta^a | X^o = 0] \\ &= (x^o)^\top \beta^o + \left(E[X^a | X^o = x^o]^\top - E[X^a | X^o = 0]^\top\right) \beta^a. \end{aligned} \tag{8}$$

For a given p-length vector x^o, we can use Equation (8) to link a $(p+q)$-dimensional set of parameter values $\{\beta^o, \beta^a\}$ to a linear combination of α, capturing one dimension of information about α. With a linearly independent set of p vectors x^o, we can create the desired $(p+q) \to p$ mapping and capture the available p dimensions of information.

In theory, (8) holds for any arbitrary vector x^o. However, as a consequence of the derivations in this section, we intuitively understand x^o to correspond to a vector of the original covariates. This is important because we need to be able to calculate or approximate the expectations in the mapping. Let $\boldsymbol{V}^o$ denote a $p \times p$ matrix of linearly independent columns, with the jth row representing a hypothetical pattern of the original covariates. Analogously, let $\boldsymbol{V}^a$ denote a $p \times q$ matrix of p hypothetical patterns of the added covariates. Then, the length-p vectorized mapping is

$$\begin{aligned} &\boldsymbol{v}^o \alpha \approx \boldsymbol{v}^o \beta^o + \left(E[\boldsymbol{V}^a | \boldsymbol{V}^o = \boldsymbol{v}^o] - E[\boldsymbol{V}^a | \boldsymbol{V}^o = \boldsymbol{0}_{p\times p}]\right) \beta^a \\ &\Rightarrow \alpha \approx \beta^o + \boldsymbol{P}\beta^a, \end{aligned} \tag{9}$$

where $\boldsymbol{P} \equiv (\boldsymbol{v}^o)^{-1} \left(E[\boldsymbol{V}^a | \boldsymbol{V}^o = \boldsymbol{v}^o] - E[\boldsymbol{V}^a | \boldsymbol{V}^o = \boldsymbol{0}_{p\times p}]\right)$. Analogous to the naive Bayesian update,

$$\pi_{\text{SB}}(\beta^o + \boldsymbol{P}\beta^a) = N(\{\beta^o + \boldsymbol{P}\beta^a\} | m_\alpha, \eta \boldsymbol{S}_\alpha). \tag{10}$$

Here $\boldsymbol{v}^o$ is fixed and known. Contrasting (4) and (10), the latter incorporates a linear offset to account for the differences in Model (2) and (3). The sensible Bayesian update thus approximates and accounts for the difference between α and β^o. However, the prior in (10) will still be insufficient on its own, as it only informs p dimensions of a $p+q$ parameter space. We return to this point in Section 4.

The rows of $\boldsymbol{v}^o$ must form a linearly independent set, i.e. $\boldsymbol{v}^o$ must be invertible, so that information on α is not lost in the transformation. Additional minimal requirements come from noting that the integral in the expectation of (9) can be numerically estimated by repeatedly sampling from the conditional distribution of the augmented covariates given the original covariates and averaging over these samples. This is symbolically written as

$$\boldsymbol{P} \approx (\boldsymbol{v}^o)^{-1} \frac{1}{M} \sum_{m=1}^{M} \left(\boldsymbol{V}^a_{(m)} - \tilde{\boldsymbol{V}}^a_{(m)}\right), \tag{11}$$

where $\boldsymbol{V}^a_{(m)} \sim F(\boldsymbol{V}^a|\boldsymbol{V}^o = \boldsymbol{v}^o)$ and $\tilde{\boldsymbol{V}}^a_{(m)} \sim F(\boldsymbol{V}^a|\boldsymbol{V}^o = \boldsymbol{0}_{p\times p})$. With this in mind, the rows of $\boldsymbol{v}^o$, which represent hypothetical patterns of the original covariate, should ideally be uncorrelated with each row in $\boldsymbol{x}^o$, but, at the same time, not too "far" from the distribution of $\boldsymbol{x}^o$. We therefore selected $\boldsymbol{v}^o$ to be the set of eigenvectors from $\boldsymbol{S}_\alpha$.

After setting up the matrix $\boldsymbol{v}^o$ appropriately, multiple imputation with chained equations (MICE) provides a fast and flexible approach for calculating (11). We need to draw M samples of $V^a_j \sim F(V^a|V^o = v^o_j)$, for each j, and M samples of $\tilde{V}^a \sim F(V^a|V^o = 0)$. We will use the $n \times (p+q)$ matrix $\{\boldsymbol{x}^o, \boldsymbol{x}^a\}$ to infer the joint distribution of $\{X^o, X^a\}$. Operationally, first stack this matrix on top of a $p\times(p+q)$ matrix, with the jth row being the length-$(p+q)$ vector $\{v^o_j, \text{NA}, \ldots, \text{NA}\}$, that is the jth row of $\boldsymbol{v}^o$ followed by a length-q vector of "missing" values. These are the components to sample $\boldsymbol{V}^a_{(m)}$ in (11). Next, stack this $(n+p)\times(p+q)$ matrix onto a single additional row vector $\{0, \ldots, 0, \text{NA}, \ldots, \text{NA}\}$, which will be used to sample $\tilde{\boldsymbol{V}}^a_{(m)}$ in (11). Feed the resulting $(n+p+1)\times(p+1)$ matrix into the imputation software to obtain the needed multiple imputations. Because a monotone missingness pattern is satisfied by construction, one iteration of MICE is sufficient for convergence. Note that, in contrast to typical uses of MICE, we do not condition on the outcome Y because it does not appear in Equation (7).

Both priors here further consideration: the sensible Bayesian is intuitively preferable by adjusting for model misspecification, and the naive Bayesian avoids modeling the distribution of X^a given X^o.

4 Adaptive Weighting

Alone, neither type of prior from Section 2 or 3 would be acceptable in the context of this paper: the shrinkage-to-zero prior in Section 2 ignores the historical data, and the priors in Section 3 may be incomplete, particularly when the historical information is limited to a small number of covariates. In this section, we develop combined versions of the historical priors that continuously and adaptively vary between the priors in Section 2 and Section 3. Intuitively, these adaptive versions, called 'naive adaptive Bayes' (NAB) and 'sensible adaptive Bayes' (SAB), should be able to incorporate the historical information without sacrificing potentially large efficiency gains coming from shrinking to zero. We intuitively describe the two adaptive priors before formally defining them. Both share the following commonalities. Similar to the power prior, a hyperparameter $\phi \in [0, 1]$ weights the historical information by inversely scaling the variance $\boldsymbol{S}_\alpha$; larger (smaller) values of ϕ reflect greater (less) incorporation of the historical information. When ϕ is equal to zero, both NAB and SAB reduce to the shrinkage-to-zero prior in (1).

4.1 Naive Adaptive Bayes

Where NAB and SAB diverge is at $\phi = 1$, which corresponds to full use of the historical information. NAB extends the $\alpha \approx \beta^o$ assumption of its non-adaptive counterpart in Section

2.1. Therefore, the historical prior on β^o in (4) is fully used, and any additional shrinkage of β^o is very weak, since prior information on β^o is already available. Moreover, β^a should be strongly shrunk to zero, since that is generally the only parameterization in which $\alpha \approx \beta^o$. Mathematically, the NAB conditional prior given real-valued hyperparameters ϕ and τ and vector-valued hyperparameters λ and $\tilde{\lambda}$ is

$$\pi_{\text{NAB}}(\beta^o, \beta^a | \phi, \eta, \tau, \lambda, \tilde{\lambda}) = N(\beta^o | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p+q} N(\beta_j | 0, \tilde{\theta}_j^2) Z_{\text{NAB}}(\phi, \eta, \tau, \lambda), \tag{12}$$

$$\tilde{\theta}_j(\phi, \tau, \lambda, \tilde{\lambda}) = \begin{cases} \left(\frac{1}{d^2} + \frac{1-\phi}{c^2 \tau^2 \lambda_j^2} \right)^{-1/2}, & j = 1, \ldots, p \\ \left(\frac{1}{d^2} + \frac{1-\phi}{c^2 \tau^2 \lambda_j^2} + \frac{\phi}{\tilde{c}^2 \tilde{\lambda}_j^2} \right)^{-1/2}, & j = p+1, \ldots, p+q \end{cases},$$

$$Z_{\text{NAB}}(\phi, \eta, \tau, \lambda) = \left(\int_{\beta^o} N(\beta^o | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p} N(\beta_j | 0, \tilde{\theta}_j^2) d\beta^o \right)^{-1} \tag{13}$$

As desired, the impact of the shrinkage-to-zero prior decreases with ϕ. τ and λ are the same as in Section 2.1, and the constants c and d are selected as previously described. We set the other constant, $\tilde{c}$, equal to 0.05, i.e. a small but non-zero number to reflect the assumption that $\beta^a \approx 0$ when $\phi = 1$; however, we introduce an auxiliary hyperparameter vector $\tilde{\lambda}$, which allow for exceptionally large elements of β^a if warranted by the data. The NAB prior uses the constant d, which is not scaled by ϕ, to guarantee propriety of the posterior for any $\phi \in [0, 1]$. To summarize, NAB varies between standard shrinkage to zero ($\phi = 0$) and a Bayesian update under the assumption that $\alpha \approx \beta^o$ and $\beta^a \approx 0$ ($\phi = 1$).

Remark 1 The normalizing constant $Z_{\text{NAB}}(\phi, \eta, \tau, \lambda)$ in (13) ensures that the prior is a proper density for any configuration of the hyperparameters and must be calculated when any of the hyperparameters are themselves random. Its analytic expression is derived in the Supplement. The integral, which is calculated at each step of the Markov Chain, would become computationally intractable in the presence of a Jacobian, which is why we linearized the mapping, as in Equation (8).

4.2 Sensible Adaptive Bayes

For SAB, the modified prior in Equation (10) is fully employed when $\phi = 1$, and any additional shrinkage of β^o to zero is weak. However, because the sensible Bayesian update adjusts for the difference between β^o and α, it is not necessary to assume that $\beta^a \approx 0$. Thus, in SAB, the value of ϕ does not affect the contribution of the variance-reducing prior on β^a. The SAB conditional prior given real-valued hyperparameters ϕ and τ and vector-valued

hyperparameter λ is

$$\pi_{\text{SAB}}(\beta^o, \beta^a | \phi, \eta, \tau, \lambda) = N(\{\beta^o + \boldsymbol{P}\beta^a\} | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p+q} N(\beta_j | 0, \tilde{\theta}_j^2) Z_{\text{SAB}}(\phi, \eta, \tau, \lambda), \quad (14)$$

$$\tilde{\theta}_j(\phi, \tau, \lambda) = \begin{cases} \left(\frac{1}{d^2} + \frac{1-\phi}{c^2 \tau^2 \lambda_j^2} \right)^{-1/2}, & j = 1, \ldots, p \\ \left(\frac{1}{d^2} + \frac{1}{c^2 \tau^2 \lambda_j^2} \right)^{-1/2}, & j = p+1, \ldots, p+q \end{cases}$$

$$Z_{\text{SAB}}(\phi, \eta, \tau, \lambda) = \left(\iint_{\beta^o, \beta^a} N(\{\beta^o + \boldsymbol{P}\beta^a\} | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p+q} N(\beta_j | 0, \tilde{\theta}_j^2) d\beta^o d\beta^a \right)^{-1}$$

An expression for $Z_{\text{SAB}}(\phi, \eta, \tau, \lambda)$ is derived in the Supplement. Table 1 compares the values of the hyperparameter θ_j for the two adaptive Bayesian approaches. Like the NAB, the SAB prior is also proper for any ϕ as a consequence of d being finite.

4.3 Hyperpriors

We describe here our choices of hyperprior for the hyperparameters ϕ, η, and, for NAB, $\tilde{\lambda}$. The hyperpriors on the global and local shrinkage components, τ and λ, remain as given in Section 2.

The hyperparameter ϕ is critical because it distributes prior weight between shrinkage to zero (ϕ close to zero) and historical shrinkage (ϕ close to one). Thus, we consider two options. The first, which we call *agnostic*, is uniform over the unit interval. The second is a truncated normal distribution with mean and standard deviation of 1 and 0.25, respectively. This is an *optimistic* hyperprior in the sense that the mode is $\phi = 1$, encouraging full use of the historical information.

The hyperparameter η independently controls the historical prior shrinkage. This could simply be set to 1; we used an inverse-gamma distribution with shape and scale equal to 2.5, although our findings were generally insensitive to multiple different choices that we considered.

Finally, the hyperparameter vector $\tilde{\lambda}$, used by the NAB prior, controls the prior scale of β^a when $\phi = 1$. As with η, this could be set to 1, which would give that β^a is normal with standard deviation $(1/d^2 + 1/\tilde{c}^2)^{-1/2} = (1/15^2 + 1/0.05^2)^{-1/2} \approx 0.05$ when $\phi = 1$. We instead model $\tilde{\lambda}$ as an inverse-gamma with shape and scale equal to 0.5, which allows for some elements of $\tilde{\lambda}$ to be exceptionally large.

5 Simulation Study

We conducted a simulation study of logistic regression to evaluate our proposed methodology against a variety of data generating scenarios. All analyses were conducted in the R statistical

environment [R Core Team, 2016, Wickham, 2009, van Buuren and Groothuis-Oudshoorn, 2011] and its interface with Stan [Carpenter, 2017, Stan Development Team, 2017, 2018], which numerically characterizes posterior distributions using Hamiltonian Monte Carlo. The Stan scripts implementing the NAB and SAB priors are in Supplement S3.

Varying between each scenario were the fixed, unknown values of $\{\beta^o, \beta^a\}$ to be estimated (ten possibilities described in Table 3, ranging from $p+q=6$ to 100 predictors), the sample size of the historical data analyses ($n_{\text{hist}} \in \{100, 400, 1600\}$), and the sample size of the current data analyses ($n \equiv n_{\text{curr}} \in \{100, 200\}$). For each unique data generating scenario, we independently sampled 80 'historical' and 'current' datasets of size n_{hist} and n_{curr}, respectively. To generate the data, the covariates $\{X^o, X^a\}$ were sampled from multivariable normal distributions with constant correlation equal to 0.2. Then, given $\{X^o, X^a\}$, Y was sampled according to a logistic regression with regression coefficients $\{\beta^o, \beta^a\}$ fixed at one of the values in the third column of Table 3. The true value of the intercept parameter in the historical data ($\mu_{\text{hist}} = -1$) was larger than that of the current data ($\mu = -2$), yielding different marginal prevalences of the outcome. Each historical dataset, $\boldsymbol{y}_{\text{hist}}$, consisted of independent draws of $\{Y, X^o\}$, whereas each current dataset, $\boldsymbol{y} \equiv \boldsymbol{y}_{\text{curr}}$, consisted of independent draws of $\{Y, X^o, X^a\}$. In summary, the historical and current *generating models* differ in the true value of the intercept; the historical and current *datasets* structurally differ in that the former does not use X^a.

The fourth column of Table 3 gives the asymptotic coefficients from the misspecified logistic regression of Y on X^o, which the historical data analysis estimates. To emulate the historical analysis, an initial Bayesian logistic regression was fit to $\boldsymbol{y}_{\text{hist}}$ to estimate model (2). We applied a regularized horseshoe prior on α using Equation (1) with $d = 15$ and $\tilde{\xi}_{\text{eff}} = p^{1/3} - 0.5$, $n = n_{\text{hist}}$, and $\sigma = 2$ to determine the value of c. So, for example, when $p = 20$, the assumed effective number of non-zero parameters was about 2.21, and when $n_{\text{hist}} = 400$, solving $2.21 = E[\sum_j (n_{\text{hist}}\sigma^{-2}\theta_j^2)/(1 + n_{\text{hist}}\sigma^{-2}\theta_j^2)]$ yields $c \approx 0.0060$. Fixing $\sigma = 2$ in this equation corresponds to the largest dispersion in a logistic GLM and usually results in slightly less than $\tilde{\xi}_{\text{eff}}$ effective covariates compared to $\sigma < 2$ [Piironen and Vehtari, 2016]. We obtained samples from the historical posterior distribution $\pi(\alpha|\boldsymbol{y}_{\text{hist}})$. From this, we obtained estimates of m_α and $\boldsymbol{S}_\alpha$, the ingredients for the adaptive priors in the current data analysis: (12) and (14). We then conducted the current analysis, which consists of fitting another Bayesian logistic regression, this time to estimate the larger model in (3), using $\boldsymbol{y}_{\text{curr}}$. Each of five priors in the third column of Table 2 was paired with the likelihood of $\boldsymbol{y}_{\text{curr}}$, yielding five posterior distributions to be compared. Four of these priors are variants of the adaptive Bayesian update priors outlined in Section 4: two adaptive priors times two choices of hyperpriors on ϕ. The other prior, namely the regularized horseshoe prior in (1), was used as a reference; both of the adaptive priors would reduce to the regularized horseshoe prior if $\phi \equiv 0$. This is the 'Standard' posterior. We used $d = 15$ and $\tilde{\xi}_{\text{eff}} = (p+q)^{1/3} - 0.5$, $n = n_{\text{curr}}$, and $\sigma = 2$ to solve for c. We measured performance using root mean-squared error (RMSE), defined by

$$\text{RMSE} = \sqrt{E_{\pi(\beta|\boldsymbol{y})}(\beta - b)^\top(\beta - b)}, \tag{15}$$

where b is the fixed, true value of the regression coefficient vector, and the expectation is

taken over both the original and added covariates. For each of the historical-based adaptive priors, we calculated the RMSE ratio compared to the standard approach, such that ratios less than one indicate relatively better performance of the historical-based prior. For each unique data generating mechanism, we report the distribution of 80 RMSE ratios for each adaptive prior. The top panel of Figure 1 plots the RMSE ratios from the first 5 rows of Table 3, for which $p = 4$ and $q = 2$, and the bottom panel plots the RMSE ratios from the final 5 rows, for which $p + q$ ranged from 22 to 100. Figures S1 and S2 in the supplement plot the RMSE ratios separately for the original and added covariates, respectively.

In general, more historical data, i.e. larger n_{hist}, improved the relative performance of the adaptive priors, whereas more current data, i.e. larger n_{curr}, decreased the relative performance of the adaptive priors. Both of the adaptive priors were relatively *less* useful when $p + q$ was small: across all datasets in the top panel of Figure 1, the middle quartiles of the RMSE ratios for NAB(agnostic) was $\{0.68, 0.86, 1.04\}$, and for SAB(agnostic) it was $\{0.63, 0.76, 0.89\}$; across all datasets in the bottom panel, these were $\{0.21, 0.51, 0.69\}$ and $\{0.25, 0.56, 0.76\}$, respectively. One likely reason for this is the inherent variability from estimating many regression coefficients. Comparing the adaptive priors, NAB outperformed SAB in scenarios for which integral required by the latter ((9) and (10)) is difficult to estimate well through multiple imputation, e.g. $p \ll q$.

6 Application: Mortality Risk Prediction in Pediatric ECMO Patients

We demonstrate our methods on the data example discussed in the introduction. Ped-RESCUERS was fit to $n_{\text{hist}} = 1611$ historical patients, and $p = 11$ risk factors for short-term mortality were included. Our current data consists of $n_{\text{curr}} = 178$ patients, on which we have measured both the $p = 11$ original and the $q = 11$ added risk factors. The overall mortality rate in the Ped-RESCUERS cohort was 40.8%; in the current cohort it was 26.4%. Thus, ignoring historical information on the intercept μ_{hist}, as both types of prior do, is prudent.

We fit the following seven Bayesian logistic regression models. Ped-RESC is the model of the eleven original risk factors from Barbaro et al. [2016], using the 1611 patients. Ped-RESC2 fits this same model using the current 178 patients, using weakly informative Cauchy priors on the regression coefficients; we include this model so as to be able to assess differences due to study populations. The other five priors are as considered in the simulation study: a regularized horseshoe prior on all 22 risk factors ('Standard'), and agnostic and optimistic versions for each of the SAB and NAB priors. For all five priors, we used $\tilde{\xi}_{\text{eff}} = 11$ to reflect an optimistic prior assumption that 11/22 of the coefficients are non-zero. To account for sporadic missingness in the covariates of the current dataset (about 4% across all covariates), we used a pseudo-Bayesian strategy proposed by Zhou and Reiter [2010]. We first imputed 100 datasets using MICE. Then, for each completed dataset and each prior, we sampled 400 draws from the posterior distribution of the parameters conditional upon that completed

dataset, concatenating these across imputations to construct a sample of $400 \times 100 = 40,000$ posterior draws "averaged" over the imputations.

All odds ratios were standardized with respect to the observed distribution of the 178 current patients, allowing for a comparison of magnitudes both between and within all priors. Table 4 gives the posterior medians of the standardized odds ratios. Also included is the larger of (i) $\Pr(e^{\beta_k} > 1)$ and (ii) $\Pr(e^{\beta_k} < 1)$. Bolded results correspond to those with a $> 75\%$ probability of falling above or below 1, a simple binary indicator of variable importance. The first two blocks of rows correspond to the original risk factors, and the second two blocks of rows correspond to the added risk factors. Figure 2 and 3 give boxplots of the posterior distributions for the original and added covariates, respectively.

Comparing the PED-RESC and PED-RESC2 rows, the direction and magnitude of the observed associations in the sets of original risk factors were consistent between the two cohorts. One exception was in PED-RESC2, in which no patients with a primary diagnosis of asthma died, i.e. quasi-complete separation. All variable importance probabilities were generally closer to 1 in PED-RESC, a consequence of its larger sample size. The Standard approach (a shrinkage-to-zero prior), shrinks nearly all odds ratios, both original and added, close to one. This is one consequence of the size of n_{curr} relative to $p + q$. In contrast, all of the adaptive priors recover some of the original associations from PED-RESC.

Using the NAB priors, the variable importance probabilities of the original risk factors were all greater than 75%, as well as those of added risk factors of ALT and lactate; these were also important according to the Standard approach. The SAB priors did not find $PaCO_2$, malignancy, or preECMO milrinone to be important and, among the added risk factors, identified bilirubin, ALT, lactate, and PF ratio as important. The posterior means of the tuning parameter ϕ were 0.61 and 0.58, for NAB and SAB when $\pi(\phi) = \text{Unif}(\phi|0, 1)$ ('agnostic') and 0.84 and 0.83, respectively when $\pi(\phi) = N(\phi|1, 0.25^2)$ ('optimistic').

From Figure 2, there are two general differences between the Standard and the Adaptive priors. For the two original covariates that Standard identified as important (primary diagnosis of pertussis and number of hours intubated prior to ECMO), the posterior variability of the adaptive priors is smaller than the Standard prior. Among the remaining original coefficients, the Adaptive priors have larger posterior variability than the Standard prior; this is a consequence of the shrinkage-to-zero prior, which yields small posterior variance for coefficients that it identifies as likely to be zero-valued. The Standard prior is not necessarily preferred, because some of this shrinkage likely reflects an inability to reliably estimate coefficients rather than confidence that the coefficients are truly close to zero. In general, the SAB priors deviate more from PED-RESC than the NAB priors: NAB shrinks β^o directly towards the PED-RESC estimates, whereas SAB shrinks a linear combination of β^o and β^a toward the PED-RESC estimates.

7 Discussion

We have proposed novel adaptive Bayesian updates of a GLM when the historical model only includes a subset of the covariates of interest. The priors, with acronyms 'NAB' or 'SAB', adaptively combine literal prior information from the historical model, including underlying statistical uncertainty, with variance-reducing shrinkage to zero. Thus, they are flexible enough for use in many contexts, ranging from the historical information being highly informative about a few coefficients to being weakly informative about most coefficients, as evidenced by our simulation study. They generally outperformed or matched in performance a standard approach that ignores historical information. We demonstrated these ideas in our motivating case study for predicting short-term mortality risk in pediatric ECMO patients.

Specifically, we combined a registry-based mortality risk prediction model (the historical model) fit to 1611 patients' data with a broader model that includes biometric measurements (the current model) recorded on 178 patients. A standard shrinkage-to-zero prior shrunk most covariates, both original and added, to zero, which is typical behavior for such priors in the presence of substantial uncertainty. Taking into account a clinical perspective, it seems unlikely that only two of the eleven original variables identified by Ped-RESCUERS remain as risk factors for mortality after including the added biometric measurements. Agreeing more with our clinical expectation, the adaptive priors re-confirmed most of the original Ped-RESCUERS risk factors and also identified two (NAB) or four (SAB) of the added covariates as being potential risk factors. Lack of importance in some of the original factors may be due to correlation in the predictors: $PaCO_2$ and lactate are negatively correlated, both associated with the degree of acidosis in the body. Similarly, bilirubin and ALT both measure liver damage, which may explain that the NAB priors focused on ALT alone whereas SAB estimated both bilirubin and ALT as important. One finding that is counter to clinical intuition is SAB's failure to identify malignancy as a meaningful risk factor; NAB found it to be more associated with mortality, which is consistent with the Ped-RESCUERS model as well as our clinical prior.

As outlined in Section 3, the SAB prior requires specification of an imputation model for X^a given X^o, c.f. (8). In our simulation study, this approach worked well in scenarios for which that model was readily estimated, namely $p > q$. Its advantage over NAB was most evident in scenario 5; in this case, the true value of β^o differed from the misspecified true value of α, and so NAB was substantially *worse* than the Standard approach because the historical prior was biased. In contrast, in scenarios 9 and 10, for which $p \ll q$ and NAB outperformed SAB, the need for imputation was likely to the detriment of SAB (although it still outperformed the Standard approach). Such $p \ll q$ scenarios would correspond, for example, to the exploration of the utility of adding a panel of biomarkers, X^a, to an established risk prediction model, $Y|X^o$. Furthermore, there may be differences in the true underlying imputation model between the current and historical populations that no amount of current data could recover. NAB is free of this particular assumption and therefore not automatically inferior to SAB in all scenarios, despite the implied value judgment in our nomenclature of 'naive' versus 'sensible'. The only difference between the historical priors

in (4) and (10) being the offset $\boldsymbol{P}\beta^a$ in (10), replacing it with $\gamma\boldsymbol{P}\beta^a$, γ a random variable in $[0, 1]$, may be one way to leverage the advantages of both adaptive priors. Importantly, both NAB and SAB ignore information on the historical intercept μ_{hist}. Thus, rather than assuming that the underlying data-generating mechanism $[Y|X^o, X^a]$ is identical between the historical and current models, they both make a less restrictive assumption that the regression coefficients β^o and β^a are the same.

Shrinkage methods classically make a bias-variance tradeoff to improve overall performance: bias in the direction of zero in exchange for a reduction in variance. In contrast, the adaptive Bayesian updates we propose, which balance between historical-based shrinkage and shrinkage to zero, are making a bias-bias tradeoff. Both extremes of the adaptive priors ($\phi = 0$ and $\phi = 1$) reduce variance, and the question is rather one of determining which type of shrinkage is less biased.

Acknowledgments

All numerical analyses were conducted in the R statistical environment [R Core Team, 2016]. This work was partially supported by the National Institutes of Health [P30 CA046592] and the Extracorporeal Life Support Organization.

References

Joseph Antonelli, Corwin Zigler, and Francesca Dominici. Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics*, 18(3):553–568, 2017.

A Armagan, DB Dunson, and J Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

RP Barbaro, PS Boonstra, ML Paden, LA Roberts, GM Annich, RH Bartlett, FW Moler, and MM Davis. Development and validation of the pediatric risk estimate score for children using extracorporeal respiratory support (Ped-RESCUERS). *Intensive Care Medicine*, 42 (5):879–888, 2016.

RP Barbaro, PS Boonstra, KW Kuo, DT Selewski, DK Bailly, CL Stone, J Chow, GM Annich, FW Moler, and ML Paden. Evaluating pediatric mortality risk prediction among children receiving extracorporeal respiratory support. *Submitted*, 2018.

Betsy Jane Becker and Meng-Jia Wu. The synthesis of regression slopes in meta-analysis. *Statistical Science*, pages 414–429, 2007.

B Carpenter. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009.

CM Carvalho, NG Polson, and JG Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

Han Chen, Alisa K Manning, and Josée Dupuis. A method of moments estimator for random effect multivariate meta-analysis. *Biometrics*, 68(4):1278–1284, 2012.

M-H Chen, Joseph G Ibrahim, and Constantin Yiannoutsos. Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):223–242, 1999.

Wenting Cheng, Jeremy M. G. Taylor, Pantel S. Vokonas, Sung Kyun Park, and Bhramar Mukherjee. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine*, page doi:10.1002/sim.7600, 2018. ISSN 1097-0258. URL `http://dx.doi.org/10.1002/sim.7600`.

Yuyan Duan, Keying Ye, and Eric P Smith. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106, 2006.

A Gelman, HB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, 3rd edition, 2014.

JE Griffin and PJ Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, 2005.

Sonja Grill, Donna P. Ankerst, Mitchell H. Gail, Nilanjan Chatterjee, and Ruth M. Pfeiffer. Comparison of approaches for incorporating new information into existing risk prediction models. *Statistics in Medicine*, 36(7):1134–1156, 2017. ISSN 1097-0258. doi: 10.1002/sim.7190. URL `http://dx.doi.org/10.1002/sim.7190`. sim.7190.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Joseph G Ibrahim, Ming-Hui Chen, and Stuart R Lipsitz. Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, 30(1): 55–78, 2002.

Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015.

Daniel Jackson and Richard D Riley. A refined method for multivariate meta-analysis and meta-regression. *Statistics in Medicine*, 33(4):541–554, 2014.

Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, 2009.

T Park and G Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

J Piironen and A Vehtari. Projection predictive variable selection using Stan+R. 2015. arXiv preprint arXiv:1508.02502.

Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. 2016. arXiv preprint arXiv:1610.05559.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/.

Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 2.17.0*, 2017. http://mc-stan.org/.

Stan Development Team. RStan: the R interface to Stan, 2018. URL `http://mc-stan.org/`. R package version 2.17.3.

S van Buuren and K Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Esteban Walker, Adrian V Hernandez, and Michael W Kattan. Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine*, 75(6):431–439, 2008.

H Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, 2009.

Xiang Zhou and Jerome P. Reiter. A note on Bayesian inference after multiple imputation. *The American Statistician*, 64(2):159–163, 2010.

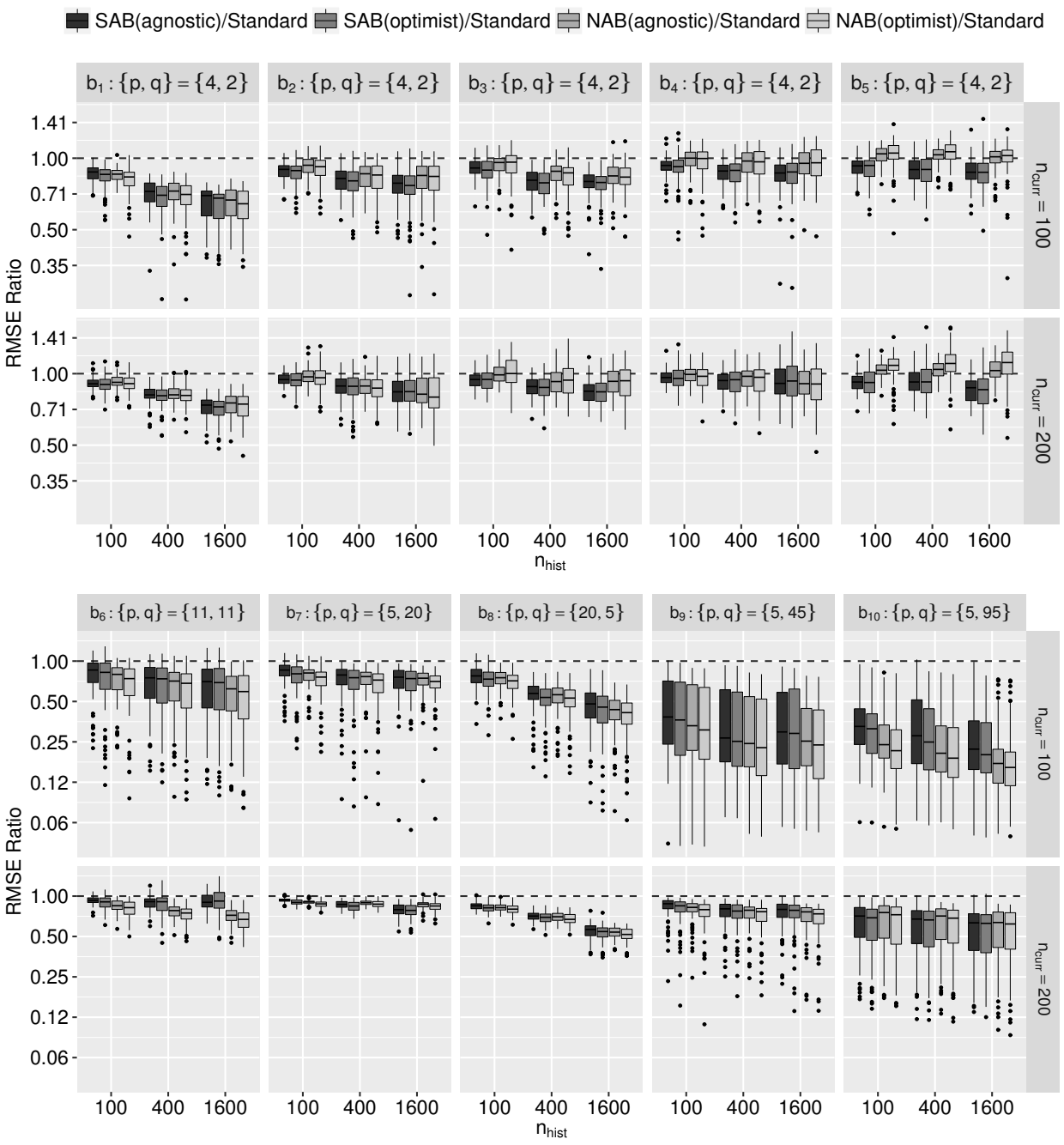

Figure 1: Boxplots of RMSE ratios (y-axis, on the $\log_2$-scale) comparing four adaptive priors that make use of the historical information against a standard hierarchical shrinkage prior against varying sample sizes of the historical data (n_{hist}; x-axis) for ten true values of the regression coefficients (b_k, $k = 1, \ldots, 10$; columns) taken from Table 3 and varying sample sizes of the current data (n_{curr}; rows). Each boxplot compares the posteriors across 50 independent datasets. Smaller ratios indicate that better performance of the corresponding adaptive prior.

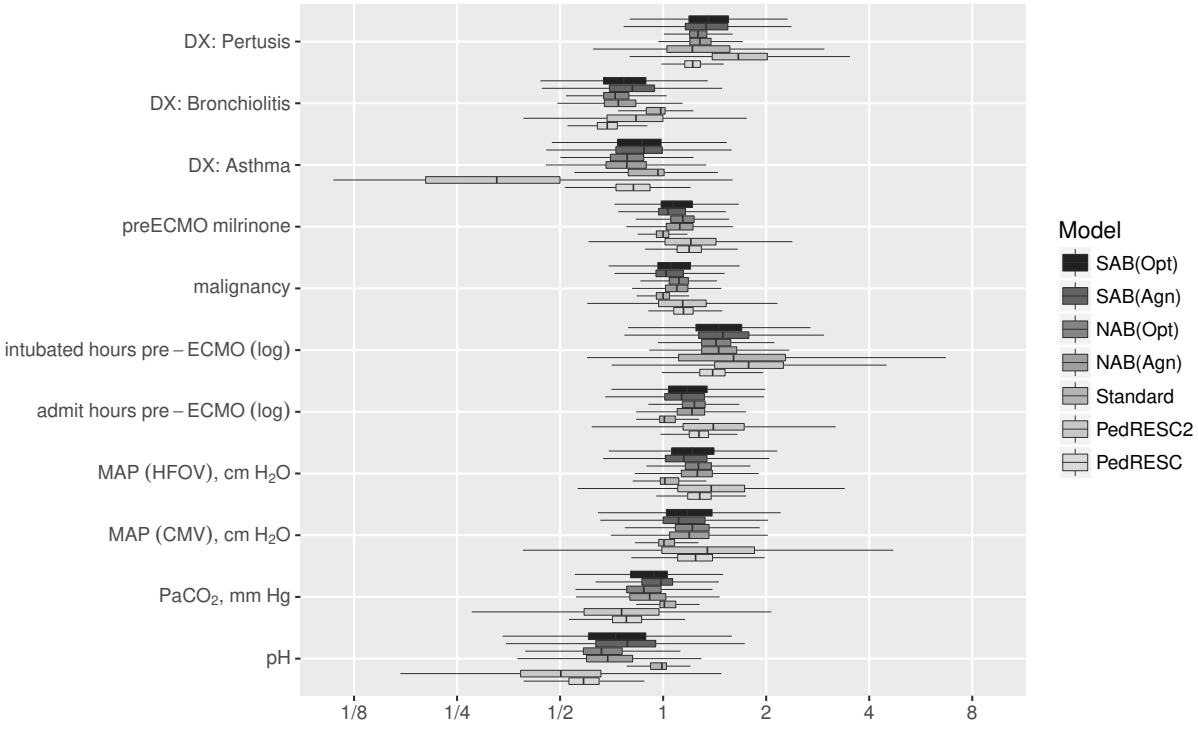

Figure 2: Boxplots of posterior draws for the *original* risk factors from seven Bayesian logistic models using different priors. All models except 'PedRESC' and 'PedRESC2' also include the *added* risk factors as given in Figure 3.

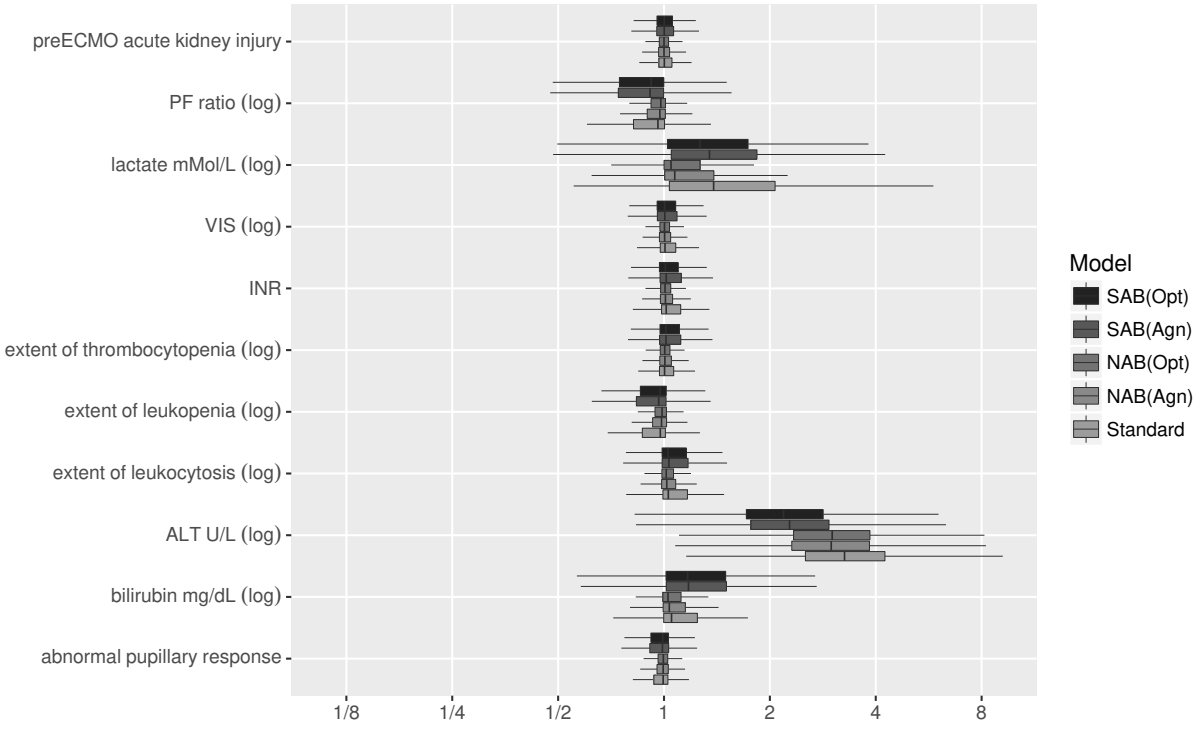

Figure 3: Boxplots of posterior draws for the *added* risk factors from five Bayesian logistic models using different priors. All models also include the *original* risk factors as given in Figure 2.

Table 1: Expressions for $\tilde{\theta}_j$ in the adaptive Bayesian updates.

	$j = 1, \ldots, p$, i.e. β^o	$j = (p+1), \ldots, (p+q)$, i.e. β^a	
	NAB and SAB	NAB	SAB
$\phi = 0$	$\left(\frac{1}{d^2} + \frac{1}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$	$\left(\frac{1}{d^2} + \frac{1}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$	$\left(\frac{1}{d^2} + \frac{1}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$
$\phi \in (0,1)$	$\left(\frac{1}{d^2} + \frac{1-\phi}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$	$\left(\frac{1}{d^2} + \frac{1-\phi}{c^2\tau^2\lambda_j^2} + \frac{\phi}{\tilde{c}^2\tilde{\lambda}_j^2}\right)^{-1/2}$	$\left(\frac{1}{d^2} + \frac{1}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$
$\phi = 1$	d	$\left(\frac{1}{d^2} + \frac{1}{\tilde{c}^2\tilde{\lambda}_j^2}\right)^{-1/2}$	$\left(\frac{1}{d^2} + \frac{1}{c^2\tau^2\lambda_j^2}\right)^{-1/2}$

Table 2: Summary of posterior distributions evaluated in the simulation study. Because the same likelihood is used for all methods, any differences are due to priors used. The 'Standard' approach is precisely the regularized horseshoe prior and used as a benchmark for comparing performance.

Label	Likelihood	Prior	Prior Eqn.	Hyperprior: $\pi(\phi) =$
Standard	$\pi(\boldsymbol{y}\|\beta^o, \beta^a)$	$\pi_{\text{SZ}}(\beta^o, \beta^a\|\tau, \lambda)$	Eqn. (1)	–
NAB(agnostic)	$\pi(\boldsymbol{y}\|\beta^o, \beta^a)$	$\pi_{\text{NAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda, \tilde{\lambda})$	Eqn. (12)	Unif$(\phi\|0, 1)$
NAB(optimist)	$\pi(\boldsymbol{y}\|\beta^o, \beta^a)$	$\pi_{\text{NAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda, \tilde{\lambda})$	Eqn. (12)	$N(\phi\|1, 0.25^2)1_{\phi\in[0,1]}$
SAB(agnostic)	$\pi(\boldsymbol{y}\|\beta^o, \beta^a)$	$\pi_{\text{SAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda)$	Eqn. (14)	Unif$(\phi\|0, 1)$
SAB(optimist)	$\pi(\boldsymbol{y}\|\beta^o, \beta^a)$	$\pi_{\text{SAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda)$	Eqn. (14)	$N(\phi\|1, 0.25^2)1_{\phi\in[0,1]}$

Table 3: Summary of fixed, true values of the regression coefficients $\{\beta^o, \beta^a\}$ from the generating logistic regression model, $[Y|X^o, X^a]$ used in the simulation study as well as the asymptotic true values of α from the misspecified reduced model, $[Y|X^o]$, when the covariates $\{X^o, X^a\}$ are jointly normally distributed with constant correlation 0.2.

Label	$\{p, q\}$	$\{\beta^o \vert\vert \beta^a\}$	α
b_1	$\{4, 2\}$	$\{0.5, 0.5, 0.5, 0.5 \vert\vert 0.5, 0.5\}$	$\{0.58, 0.58, 0.58, 0.58\}$
b_2	$\{4, 2\}$	$\{1, 0.5, 0, 0 \vert\vert 0.5, 1\}$	$\{0.99, 0.58, 0.16, 0.16\}$
b_3	$\{4, 2\}$	$\{1, -0.5, 0, 0 \vert\vert -0.5, -1\}$	$\{0.68, -0.58, -0.16, -0.16\}$
b_4	$\{4, 2\}$	$\{0.5, 0.5, 0, 0 \vert\vert 1, 1\}$	$\{0.58, 0.58, 0.19, 0.19\}$
b_5	$\{4, 2\}$	$\{0.5, 0.5, 0, 0 \vert\vert -1, -1\}$	$\{0.19, 0.19, -0.19, -0.19\}$
b_6	$\{11, 11\}$	$\{\underbrace{0.5, \ldots, 0.5}_{4}, \underbrace{0.25, \ldots, 0.25}_{7} \vert\vert 2, 1, 1, \underbrace{0, \ldots, 0}_{8}\}$	$\{\underbrace{0.45, \ldots, 0.45}_{4}, \underbrace{0.30, \ldots, 0.30}_{7}\}$
b_7	$\{5, 20\}$	$\{\underbrace{0.2, \ldots, 0.2}_{5} \vert\vert \underbrace{0.2, \ldots, 0.2}_{20}\}$	$\{\underbrace{0.49, \ldots, 0.49}_{5}\}$
b_8	$\{20, 5\}$	$\{\underbrace{0.2, \ldots, 0.2}_{20} \vert\vert \underbrace{0.2, \ldots, 0.2}_{5}\}$	$\{\underbrace{0.23, \ldots, 0.23}_{20}\}$
b_9	$\{5, 45\}$	$\{1, 1, 1, 0, 0 \vert\vert \underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0, \ldots, 0}_{35}\}$	$\{1, 1, 1, 0.35, 0.35\}$
b_{10}	$\{5, 95\}$	$\{1, 1, 1, 0, 0 \vert\vert \underbrace{0.25, \ldots, 0.25}_{20}, \underbrace{0, \ldots, 0}_{75}\}$	$\{1.08, 1.08, 1.08, 0.38, 0.38\}$

Table 4: Posterior medians of standardized odds ratios and, in parentheses, variable importance probabilities given as percentages, defined as the larger of (i) the posterior probability that an odds ratio exceeds one and (ii) the posterior probability that an odds ratio falls below one. Those in **bold** exceed 75.0%.

Original	pH	$PaCO_2$,mmHg	MAP(CMV), cmH_2O	MAP(HFOV), cmH_2O	admit hrs preECMO (log)	intubated hrs preECMO (log
PedRESC	**0.46(100.0%)**	**0.70(96.3%)**	**1.37(93.9%)**	**1.43(98.6%)**	**1.42(99.3%)**	**1.62(99.5%)**
PedRESC2	**0.37(96.1%)**	**0.67(77.4%)**	1.54(74.4%)	**1.59(83.5%)**	**1.63(87.1%)**	**2.29(95.5%)**
Standard	0.99(58.9%)	1.01(59.1%)	1.01(56.0%)	1.02(61.1%)	1.01(58.7%)	**1.98(90.4%)**
NAB(Agn)	**0.58(94.8%)**	0.88(70.3%)	**1.29(83.8%)**	**1.39(93.2%)**	**1.32(91.4%)**	**1.71(98.4%)**
NAB(Opt)	**0.55(97.5%)**	**0.83(78.3%)**	**1.33(89.1%)**	**1.41(96.2%)**	**1.35(95.6%)**	**1.67(99.0%)**
SAB(Agn)	**0.71(86.7%)**	0.98(56.6%)	1.16(74.8%)	**1.22(80.5%)**	**1.20(79.5%)**	**1.79(96.5%)**
SAB(Opt)	**0.63(91.7%)**	0.91(66.5%)	**1.26(80.8%)**	**1.33(86.9%)**	**1.26(84.9%)**	**1.71(96.6%)**

Original	malignancy	preECMO milrinone	DX:Asthma	DX:Bronchiolitis	DX:Pertussis
PedRESC	**1.22(95.1%)**	**1.29(95.8%)**	**0.75(93.9%)**	**0.58(100.0%)**	**1.33(99.5%)**
PedRESC2	1.21(70.8%)	**1.31(76.6%)**	**0.05(99.2%)**	**0.77(75.2%)**	**2.07(98.0%)**
Standard	1.00(50.8%)	1.00(51.3%)	0.94(69.9%)	0.98(63.5%)	**1.33(85.3%)**
NAB(Agn)	**1.14(80.6%)**	**1.18(81.5%)**	**0.70(91.2%)**	**0.65(95.7%)**	**1.43(98.9%)**
NAB(Opt)	**1.17(86.9%)**	**1.21(88.0%)**	**0.71(93.7%)**	**0.63(98.1%)**	**1.40(99.4%)**
SAB(Agn)	1.03(59.1%)	1.05(63.0%)	**0.83(77.8%)**	**0.74(88.7%)**	**1.52(94.5%)**
SAB(Opt)	1.07(64.7%)	1.11(70.3%)	**0.82(79.1%)**	**0.68(93.2%)**	**1.55(95.7%)**

Added	abnormal pupillary resp.	bilirubin mg/dL (log)	ALT U/L (log)	extent of leukocyt. (log)	extent of leukopen. (log)	extent of thrombocytopen. (
PedRESC	–	–	–	–	–	–
PedRESC2	–	–	–	–	–	–
Standard	0.99(56.8%)	1.07(72.6%)	**5.52(99.9%)**	1.04(67.9%)	0.97(66.7%)	1.01(55.1%)
NAB(Agn)	0.99(54.8%)	1.05(70.9%)	**4.86(99.5%)**	1.03(63.8%)	0.98(62.6%)	1.01(55.3%)
NAB(Opt)	0.99(54.6%)	1.04(69.1%)	**4.90(99.6%)**	1.02(63.4%)	0.98(61.7%)	1.01(54.7%)
SAB(Agn)	0.99(57.9%)	**1.26(81.5%)**	**3.27(99.1%)**	1.05(67.0%)	0.95(66.6%)	1.02(59.3%)
SAB(Opt)	0.99(57.5%)	**1.25(81.3%)**	**3.09(98.9%)**	1.04(65.8%)	0.96(64.6%)	1.02(59.9%)

Added	INR	VIS (log)	lactate mMol/L (log)	PF ratio (log)	preECMO acute kidney injury
PedRESC	–	–	–	–	–
PedRESC2	–	–	–	–	–
Standard	1.02(62.2%)	1.01(57.6%)	**1.60(86.1%)**	0.94(71.1%)	1.00(53.4%)
NAB(Agn)	1.01(58.0%)	1.01(53.7%)	**1.11(78.0%)**	0.96(68.5%)	1.00(50.1%)
NAB(Opt)	1.01(56.9%)	1.00(53.2%)	**1.07(75.0%)**	0.97(67.0%)	1.00(50.0%)
SAB(Agn)	1.02(60.6%)	1.01(54.9%)	**1.54(86.1%)**	**0.88(76.7%)**	1.00(52.2%)
SAB(Opt)	1.01(58.2%)	1.01(53.9%)	**1.41(83.8%)**	**0.89(76.0%)**	1.00(51.9%)