Biostatistics (2020) **21**, 2, *pp.* e47–e64
doi:10.1093/biostatistics/kxy053
Advance Access publication on September 21, 2018

Incorporating historical models with adaptive Bayesian updates

PHILIP S. BOONSTRA*

Department of Biostatistics, University of Michigan, 1415 Washington Hts,
SPHII, Ann Arbor, MI 48109, USA

philb@umich.edu

RYAN P. BARBARO

Division of Pediatric Critical Care and Child Health Evaluation and Research Unit, University of
Michigan, 1500 East Medical Center Drive, Mott F-6790/Box 5243, Ann Arbor, MI 48109, USA

Summary

This article 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 2-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 (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 use the information in 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.

Keywords: 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, 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 article, we propose Bayesian approaches that

*To whom correspondence should be addressed.

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 and 2012 (Barbaro *and others*, 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, and 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, neurological, and hematological 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 them on 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 covariates as predictors. We report on these new data in Barbaro *and others* (2018). However, given the ratio of sample size to number of covariates 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. 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 so as to increase overall efficiency, i.e. decrease root mean squared error (RMSE).

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 a prior based solely on historical data 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 *and others*, 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 includes the historical data likelihood, raised to a power $0 \leq \phi \leq 1$ (Ibrahim and Chen, 2000; Ibrahim *and others*, 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 *and others*, 2006; Neuenschwander *and others*, 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 *and others* (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 parameters toward zero and so is actually more similar to a typical shrinkage-to-zero prior. Ibrahim *and others* (2002) consider the general setting of fitting generalized linear models (GLMs) using power priors when values are missing in covariates of 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 *and others*, 2008; Chen *and others*, 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 *and others*, 2016; Grill *and others*, 2017; Cheng *and others*, 2018). Antonelli *and others* (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 about the intercept, 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 ingredients of our approach that combines historical-based prior shrinkage and shrinkage to zero. Section 2 reviews the "regularized horseshoe prior" (Carvalho *and others*, 2009, 2010; Piironen and Vehtari, 2015, 2017a, 2017b), which is the shrinkage-to-zero prior that we use. Section 3 outlines the construction of 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. Sections 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 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, $\{y, \boldsymbol{x}\}$. 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 original and added 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 $\pi(y|\beta)$ and prior $\pi(\beta|\theta)\pi(\theta)$, with θ a vector of hyperparameters that is conditionally independent of y given β, the posterior is $\pi(\beta, \theta|y) \propto \pi(y|\beta)\pi(\beta|\theta)\pi(\theta)$. Often, the prior $\pi(\beta|\theta)\pi(\theta)$ is selected to regularize parameters by shrinking estimates toward zero, reducing variance at a cost of some bias, thereby increasing efficiency. 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 *and others*, 2014). A different choice of $\pi(\theta)$ conferring more adaptive shrinkage properties is the "regularized horseshoe" (Carvalho *and others*, 2009, 2010; Piironen and Vehtari, 2015, 2017a, 2017b). 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 $\pi(\tau) = C^+(\tau|0,1)$, $\pi(\lambda_j) = C^+(\lambda_j|0,1)$ for $j = 1, \ldots, p+q$, and C^+ indicates the positive half-Cauchy distribution. Letting $\theta_j \equiv \left(1/d^2 + 1/[c^2\tau^2\lambda_j^2]\right)^{-1/2}$, the conditional prior for β is then

$$\pi_{\text{SZ}}(\beta|\theta) = \prod_{j=1}^{p+q} N(\beta_j|0, \theta_j^2), \tag{2.1}$$

The SZ subscript indicates “shrinkage to zero”. The hyperparameter τ globally shrinks β, while the λ_js multiplicatively offset τ and thus admit large individual variance components. The original horseshoe (Carvalho *and others*, 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 (2017a) 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, the authors 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 $\pi(\theta)$. The result will usually be $c \ll 1$.

Subsequent work by Piironen and Vehtari (2017b) argued that the original horseshoe tends to *under*-shrink large elements of β, which can also result in numerical difficulties, known as “divergent transitions”, as a result of a stochastic search through heavy tails (Piironen and Vehtari, 2015). As a solution, they suggest to soft-truncate its tails with 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 (2.1) 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 X 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

$$g(E[Y|X^o = x^o]) = \mu^o + (x^o)^\top\alpha. \tag{3.1}$$

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 (3.1) but rather the embiggened model

$$g(E[Y|X^o = x^o, X^a = x^a]) = \mu + (x^o)^\top\beta^o + (x^a)^\top\beta^a. \tag{3.2}$$

We have a dataset of n observations, $\{y, \boldsymbol{x}^o, \boldsymbol{x}^a\}$ and a likelihood function $\pi(y|\beta^o, \beta^a)$. A standard analysis of $\{y, \boldsymbol{x}^o, \boldsymbol{x}^a\}$ 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

$$\pi_{\text{NB}}(\beta^o|\eta) = N(\beta^o|m_\alpha, \eta\boldsymbol{S}_\alpha) \tag{3.3}$$

However, Model (3.1) cannot hold for all patterns x^o if Model (3.2) 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. $E[X^a|X^o = x^o] = \boldsymbol{B}x^o$). 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 (3.3). 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- corresponding to 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 place the prior on the many-to-few mapping from $\{\beta^o, \beta^a\}$ to α. Based on derivations below, we show that this mapping can be approximated by $\alpha \approx \beta^o + \boldsymbol{P}\beta^a$, where $\boldsymbol{P}$ is a certain $p \times q$ projection matrix. To see this, begin by 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 (3.1), i.e. taking $g(\cdot)$ of both sides, which—because the true(r) model is (3.2)—will only be an approximation of the conditional mean of Y given X^o, we obtain the following:

$$\mu^o + (x^o)^\top\alpha \approx g\,E[g^{-1}\left(\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a\right)|X^o = x^o] \tag{3.4}$$

$$\mu^o \approx g\,E[g^{-1}\left(\mu + (X^a)^\top\beta^a\right)|X^o = 0]. \tag{3.5}$$

This relates the available historical model with the current model. In particular, (3.5) obtains an approximation for the historical (and misspecified) intercept μ^o 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 (3.4) and (3.5) completely removes μ^o from the equation:

$$(x^o)^\top\alpha \approx g\,E[g^{-1}\left(\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a\right)|X^o = x^o] - g\,E[g^{-1}\left(\mu + (X^a)^\top\beta^a\right)|X^o = 0]. \tag{3.6}$$

Equation (3.6) is the basis of the sensible Bayesian update: it links Model (3.2) to a function of the parameters from Model (3.1), about which there is historical information. Furthermore, like the naive Bayesian update, the sensible Bayesian update avoids borrowing information on the historical intercept μ^o. This relaxes a critical assumption: we do not require that the historical and current data generating models are identical but, less restrictively, that the underlying true values of $\{\beta^o, \beta^a\}$ are equal.

When the link function g is non-linear, constructing a prior based upon (3.6) 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 approximate the mapping to obviate the Jacobian adjustment. Moving g across the integrals,

$$\begin{aligned}(x^o)^\top\alpha &\approx E[\mu + (x^o)^\top\beta^o + (X^a)^\top\beta^a|X^o = x^o] - E[\mu + (X^a)^\top\beta^a|X^o = 0]\\ &= (x^o)^\top\beta^o + \left(E[X^a|X^o = x^o]^\top - E[X^a|X^o = 0]^\top\right)\beta^a. \end{aligned} \tag{3.7}$$

For a given p-length vector x^o, we can use Equation (3.7) 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, (3.7) 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

$$\begin{aligned}&\boldsymbol{v}^o\alpha \approx \boldsymbol{v}^o\beta^o + \left(E[\boldsymbol{V}^a|\boldsymbol{V}^o = \boldsymbol{v}^o] - E[\boldsymbol{V}^a|\boldsymbol{V}^o = \boldsymbol{0}_{p\times p}]\right)\beta^a\\ &\Rightarrow \alpha \approx \beta^o + \boldsymbol{P}\beta^a,\end{aligned} \tag{3.8}$$

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,

$$\pi_{\text{SB}}(\beta^o + \boldsymbol{P}\beta^a) = N(\{\beta^o + \boldsymbol{P}\beta^a\}|m_\alpha, \eta\boldsymbol{S}_\alpha). \tag{3.9}$$

In calculating the posterior, $\boldsymbol{v}^o$ and $\boldsymbol{P}$ are treated as fixed and known constants. Section S1 of supplementary material available at *Biostatistics* online describes in detail how to construct $\boldsymbol{v}^o$ and how to use multiple imputation with chained equations (MICE) to calculate a Monte Carlo estimate of the integral in (3.8).

Contrasting the distributions in (3.3) and (3.9), the latter incorporates a linear offset to account for the differences between Models (3.1) and (3.2). The sensible Bayesian update thus approximates and adjusts for the difference between α and β^o. However, the prior in (3.9) 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. In summary, both ideas merit 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 article: 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 adaptively vary between the priors in Sections 2 and 3. Called "naive adaptive Bayes" (NAB) and "sensible adaptive Bayes" (SAB), these seek to incorporate the historical information without sacrificing potential efficiency gains coming from shrinking to zero. We 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 (2.1).

4.1. *Naive adaptive Bayes*

Where NAB and SAB differ 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 (3.3) is fully used when $\phi = 1$, and additional shrinkage of β^o is unnecessary. Moreover, β^a is strongly shrunk to zero, because that is generally the only parameterization for which $\alpha \approx \beta^o$. The hyperparameters of NAB are $\{\phi, \eta, \tau\}$ (scalars) and $\{\lambda, \tilde{\lambda}\}$ (vectors). The pre-specified constants are scalars c, d, and $\tilde{c}$. For notational simplicity, define $\tilde{\theta}_j$ to be

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

Then, the NAB conditional prior is

$$\pi_{\text{NAB}}(\beta^o, \beta^a | \phi, \eta, \tau, \lambda, \tilde{\lambda}) = N(\beta^o | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p+q} N(\beta_j | 0, \tilde{\theta}_j^2) Z_{\text{NAB}}(\phi, \eta, \tau, \lambda), \tag{4.1}$$

$$Z_{\text{NAB}}(\phi, \eta, \tau, \lambda) = \left(\int_{\beta^o} N(\beta^o | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p} N(\beta_j | 0, \tilde{\theta}_j^2) d\beta^o \right)^{-1}. \tag{4.2}$$

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 $\tilde{c}$ equal to 0.05, i.e. a small number, reflecting the assumption that $\beta^a \approx 0$ when $\phi = 1$; however, we introduce an auxiliary hyperparameter $\tilde{\lambda}$, allowing for non-zero elements of β^a if warranted by the data. The hyperparameter η separately controls the historical prior shrinkage. Hyperpriors for $\tilde{\lambda}$ and η are discussed below. The constant d guarantees 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 (4.2) ensures that the prior is proper for any configuration of the hyperparameters and must be calculated when any of the hyperparameters are random. Its analytic expression is derived in Section S2 of supplementary material available at *Biostatistics* online. The integral calculation must be updated at each step of the Markov Chain. This would become computationally intractable in the presence of a Jacobian from a non-linear transformation and is why we approximated the mapping as in (3.7).

Remark 2 A reviewer observed the similarity between NAB and the class of "penalized complexity" priors of Simpson *and others* (2017), both of which use the extreme end of the hyperparameter support ($\phi = 1$, in our notation) to essentially recapitulate some pre-specified simple model and prevent overfitting. Penalized complexity priors are defined for a much broader context and thus allow for more general base models, as opposed to our specific objective of incorporating historical information.

4.2. *Sensible adaptive Bayes*

For SAB, the modified prior in (3.9) 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. Defining the SAB hyperparameter $\tilde{\theta}_j$ to be

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

the SAB conditional prior is

$$\pi_{\text{SAB}}(\beta^o, \beta^a | \phi, \eta, \tau, \lambda) = N(\{\beta^o + \boldsymbol{P}\beta^a\} | m_\alpha, \eta \boldsymbol{S}_\alpha / \phi) \prod_{j=1}^{p+q} N(\beta_j | 0, \tilde{\theta}_j^2) Z_{\text{SAB}}(\phi, \eta, \tau, \lambda), \tag{4.3}$$

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

An expression for $Z_{\text{SAB}}(\phi, \eta, \tau, \lambda)$ is derived in Section S3 of supplementary material available at *Biostatistics* online. As with NAB, a finite-valued d ensures that $\pi_{\text{SAB}}(\beta^o, \beta^a | \phi, \eta, \tau, \lambda)$ is proper for any $\phi \in [0, 1]$.

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 distributes prior weight between shrinkage to zero ($\phi \approx 0$) and historical shrinkage ($\phi \approx 1$). We consider two hyperprior options. The first, called *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 *optimistic* because 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.

Finally, the hyperparameter vector $\tilde{\lambda}$, used by NAB, controls the prior scale of β^a when $\phi = 1$. As with η, each element of $\tilde{\lambda}$ 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 the components of $\tilde{\lambda}$ as inverse-gamma, each with shape and scale equal to 0.5, allowing for some elements of $\tilde{\lambda}$ to be 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. Code to reproduce the simulation study is available at https://github.com/psboonstra/AdaptiveBayesianUpdates.

Varying between each scenario were the fixed, unknown values of $\{\beta^o, \beta^a\}$ to be estimated (ten possibilities described in Table 2, 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 of the 60 unique data generating scenarios, we independently sampled 128 "historical" and "current" datasets of size n_{hist} and n_{curr}, respectively. To generate the data, preliminary values of $\{X^o, X^a\}$ were sampled from multivariable normal distributions with constant correlation equal to 0.2. After this, half of the elements of each of X^o and X^a were transformed using the sign function, i.e. $1_{[x>0]} - 1_{[x<0]}$. Then, given the resulting $\{X^o, X^a\}$ vector, Y was sampled from a logistic regression with parameters $\{\beta^o, \beta^a\}$ fixed at one of the values in the third column of Table 2. The true generating value of the intercept in the historical data model ($\mu_{\text{hist}} = -1$) was larger than that of the current data model ($\mu = -2$), yielding different marginal prevalences of the outcome. Each historical dataset consisted of n_{hist} independent draws of $\{Y, X^o\}$, whereas each current dataset consisted of n_{curr} 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.

Remark 3 The intercept μ^o in Model (3.1) is distinct from μ_{hist} described in the preceding paragraph: the former denotes the intercept from the asymptotically misspecified sub-model, and the latter is the true intercept from Model (3.2) that actually generated the historical data.

The fourth column of Table 2 gives the asymptotic parameter values 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 the historical outcomes y_{hist} to estimate Model (3.1). We applied a regularized horseshoe prior on α using (2.1) 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 $20^{1/3} - 0.5 \approx 2.21$, 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 $\sigma = 2$ corresponds to the largest dispersion in a logistic GLM and usually results in slightly less than $\tilde{\xi}_{\text{eff}}$ effective parameters compared with $\sigma < 2$ (Piironen and Vehtari, 2017a). We obtained samples from the historical posterior distribution $\pi(\alpha|y_{\text{hist}})$ and estimated m_α and $\boldsymbol{S}_\alpha$, the ingredients for the adaptive priors in the current data analysis. Then, the "current" analysis was conducted: a second Bayesian logistic regression to estimate the larger model in (3.2), using the current outcomes y_{curr}. Each of the five priors in the third column of Table 1 was paired with the likelihood of y_{curr}, yielding five posterior distributions to be compared. Four of these were adaptive Bayesian updates from in Section 4: two adaptive priors times two distinct hyperpriors on ϕ. The other was the regularized horseshoe in (2.1) and was used as a reference; we call this approach 'Standard'. 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. All of the adaptive priors are equivalent to Standard when $\phi \equiv 0$. For SAB, we estimated $\boldsymbol{P}$ (3.9) using Monte Carlo methods based upon 100 independent draws from MICE. We measured performance using RMSE $\equiv \sqrt{E_{\pi(\beta|y)}(\beta - b)^\top(\beta - b)}$,

Table 1. *Summary of posterior distributions evaluated in the simulation study*

Labels	Likelihood	Prior	Prior equation	Hyperprior: $\pi(\phi) =$
Standard	$\pi(y\|\beta^o, \beta^a)$	$\pi_{\text{SZ}}(\beta^o, \beta^a\|\tau, \lambda)$	Equation (2.1)	—
NAB(agnostic)	$\pi(y\|\beta^o, \beta^a)$	$\pi_{\text{NAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda, \tilde{\lambda})$	Equation (4.1)	Unif($\phi\|0, 1$)
NAB(optimist)	$\pi(y\|\beta^o, \beta^a)$	$\pi_{\text{NAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda, \tilde{\lambda})$	Equation (4.1)	$N(\phi\|1, 0.25^2)1_{\phi\in[0,1]}$
SAB(agnostic)	$\pi(y\|\beta^o, \beta^a)$	$\pi_{\text{SAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda)$	Equation (4.3)	Unif($\phi\|0, 1$)
SAB(optimist)	$\pi(y\|\beta^o, \beta^a)$	$\pi_{\text{SAB}}(\beta^o, \beta^a\|\phi, \eta, \tau, \lambda)$	Equation (4.3)	$N(\phi\|1, 0.25^2)1_{\phi\in[0,1]}$

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.

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 adaptive Bayesian updates, we calculated its RMSE ratio with respect to Standard, such that ratios less than one indicate relatively better performance of the adaptive update. For each unique data generating mechanism, we report the distribution of 128 RMSE ratios. The top panel of Figure 1 plots the RMSE ratios from the first 5 rows of Table 2, for which $p = 4$ and $q = 2$, and the bottom panel plots the ratios from the final 5 rows, for which $p+q \in \{22, 25, 50, 100\}$.

More historical data, i.e. larger n_{hist}, improved the relative performance of the adaptive updates. For example, across coefficient settings b_1–b_5, with $n_{\text{curr}} = 100$, the middle quartiles of the SAB(agnostic)/Standard RMSE ratio were $\{0.71, 0.80, 0.90\}$ when $n_{\text{hist}} = 100$ versus $\{0.51, 0.62, 0.75\}$ when $n_{\text{hist}} = 1600$. The NAB-type updates also improved with increasing n_{hist}, but, in absolute terms, they did not always improve upon Standard, e.g. for b_5 the quartiles of the RMSE ratio were $\{0.96, 1.09, 1.19\}$ when $n_{\text{hist}} = 100$ and $\{0.88, 1.01, 1.10\}$ when $n_{\text{hist}} = 1600$.

More current data, i.e. larger n_{curr}, hurt the typical relative performance of the adaptive updates, all other aspects being fixed, but also decreased the variability between datasets. This can be seen in Figure 1: when $n_{\text{curr}} = 200$ (the second and fourth rows), the boxplots move closer to 1.00 and with less spread relative to $n_{\text{curr}} = 100$ (the first and third rows, respectively). A fixed amount of historical data becomes relatively less valuable in the presence of more current data.

The adaptive priors were relatively less useful when $p + q$ was small: across all datasets in the top panel of Figure 1, for which $p + q = 6$, the middle quartiles of the RMSE ratios for NAB(agnostic) was $\{0.68, 0.84, 1.02\}$, and for SAB(agnostic) it was $\{0.61, 0.74, 0.86\}$; across all datasets in the bottom panel, for which $p + q \in \{22, 25, 50, 100\}$, these were $\{0.26, 0.54, 0.72\}$ and $\{0.28, 0.58, 0.78\}$, respectively. Comparing the adaptive priors, NAB outperformed SAB in scenarios for which the integral required by the latter, i.e. (3.8), was difficult to estimate well, e.g. $p \ll q$.

Figures S1 and S2 of supplementary material available at *Biostatistics* online plot the separate RMSE ratios for β^o and β^a, respectively. There is a clear trend of improving performance for estimating β^o, which is to be expected given its relationship to α. For the small $p + q$ settings, the SAB-type updates were about equivalent to the standard approach in estimating β^a, and the NAB-type updates were worse. Interestingly, in the larger $p + q$ settings, the adaptive priors were marginally better at estimating β^a than the standard approach. Our findings were relatively unchanged upon increasing the shape and scale of η, which scales $\boldsymbol{S}_\alpha$, to 25 (Figure S3 of supplementary material available at *Biostatistics* online). Finally, Table S1 of supplementary material available at *Biostatistics* online summarizes the running time of each prior considered. The SAB-type updates were slowest to run, owing in part to the required imputation step.

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 covariates, all of which are defined in Table S2 of supplementary material available at *Biostatistics* online. The overall mortality rate in the Ped-RESCUERS cohort was 40.8%; in the current cohort it was 26.4%.

We fit the following seven Bayesian logistic regression models. Ped-RESC is the model from Barbaro *and others* (2016) based upon the original covariates fit to the 1611 patients. Ped-RESC2 is this same model fit to 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 parameters ("Standard"), and agnostic and optimistic versions for each of the SAB and NAB priors. For all five priors, we used $\tilde{\xi}_{\text{eff}} = 11$

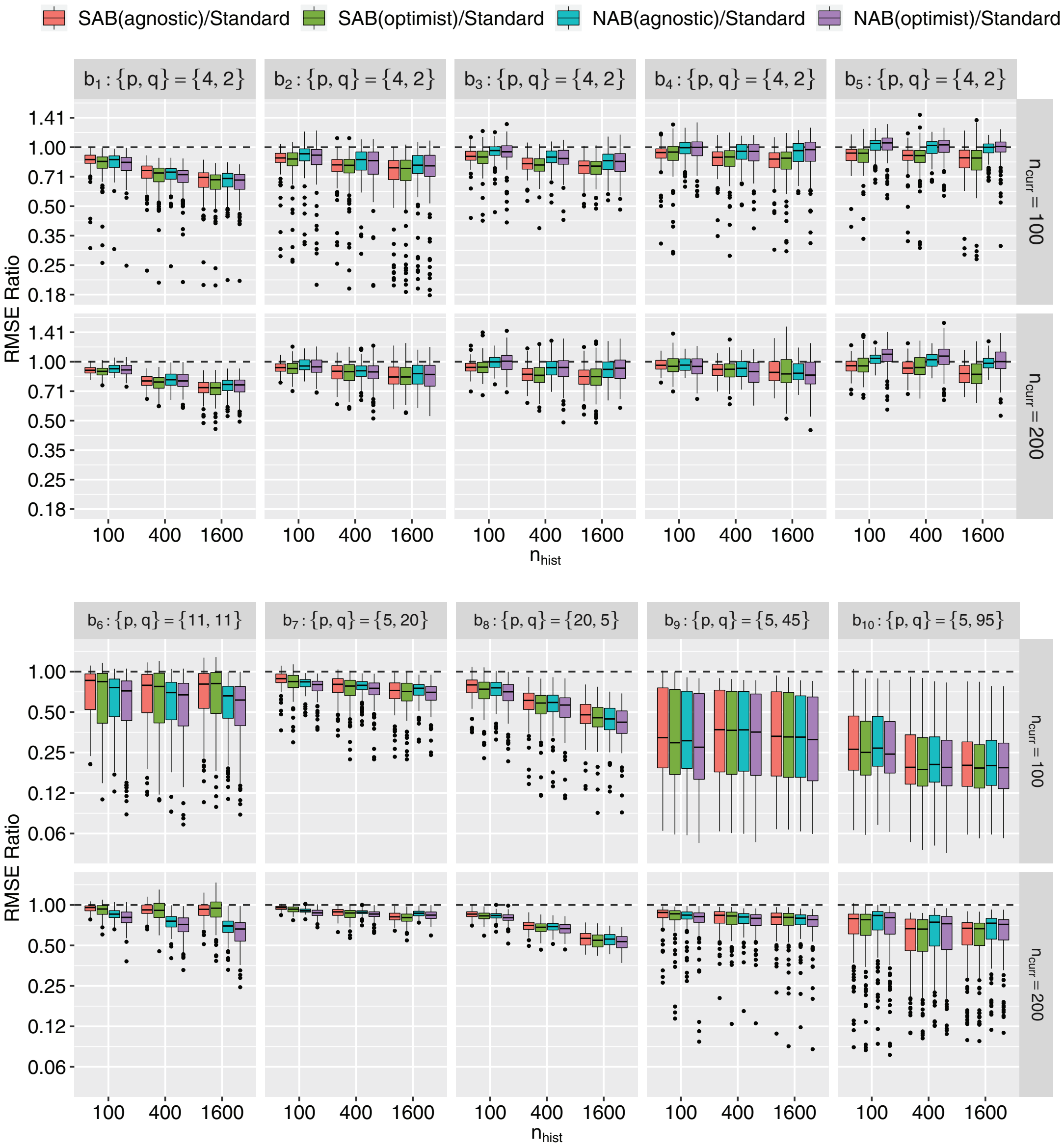

Fig. 1. Boxplots of RMSE ratios (y-axis, on the $\log_2$-scale) comparing four adaptive priors that make use of the historical information against a standard hierarchical shrinkage prior, plotted against varying sample sizes of the historical data (n_{hist}; x-axis) for ten true values of the regression coefficients (b_k, $k = 1, \ldots, 10$; columns) taken from Table 2 and varying sample sizes of the current data (n_{curr}; rows). Each boxplot compares the posteriors across 128 independent datasets. Smaller ratios indicate better performance of the corresponding adaptive prior.

to reflect an optimistic prior assumption that 11 of the 22 parameters are non-zero. To account for sporadic missingness in the current data (about 4% in total), 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 $400 \times 100 = 40,000$

Table 2. *Summary of fixed, true values of* $\{\beta^o, \beta^a\}$ *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,* $[Y|X^o]$

Label	$\{p, q\}$	$\{\beta^o \vert\vert \beta^a\}$	α
b_1	$\{4, 2\}$	$\{0.5, 0.5, 0.5, 0.5 \vert\vert 0.5, 0.5\}$	$\{0.55, 0.55, 0.58, 0.58\}$
b_2	$\{4, 2\}$	$\{1, 0.5, 0, 0 \vert\vert 0.5, 1\}$	$\{0.97, 0.54, 0.16, 0.16\}$
b_3	$\{4, 2\}$	$\{1,\text{-}0.5, 0, 0 \vert\vert \text{-}0.5,\text{-}1\}$	$\{0.73,\text{-}0.54,\text{-}0.16,\text{-}0.16\}$
b_4	$\{4, 2\}$	$\{0.5, 0.5, 0, 0 \vert\vert 1, 1\}$	$\{0.53, 0.53, 0.19, 0.19\}$
b_5	$\{4, 2\}$	$\{0.5, 0.5, 0, 0 \vert\vert \text{-}1,\text{-}1\}$	$\{0.25, 0.25,\text{-}0.19,\text{-}0.19\}$
b_6	$\{11, 11\}$	$\{\underbrace{0.5, \ldots, 0.5}_{4}, \underbrace{0.25, \ldots, 0.25}_{7} \vert\vert$ $2, 1, 1, \underbrace{0, \ldots, 0}_{8}\}$	$\{\underbrace{0.37, \ldots, 0.37}_{4}, 0.24, 0.24, \underbrace{0.29, \ldots, 0.29}_{5}\}$
b_7	$\{5, 20\}$	$\{\underbrace{0.2, \ldots, 0.2}_{5} \vert\vert \underbrace{0.2, \ldots, 0.2}_{20}\}$	$\{\underbrace{0.41, \ldots, 0.41}_{3}, 0.51, 0.51\}$
b_8	$\{20, 5\}$	$\{\underbrace{0.2, \ldots, 0.2}_{20} \vert\vert \underbrace{0.2, \ldots, 0.2}_{5}\}$	$\{\underbrace{0.23, \ldots, 0.23}_{10}, \underbrace{0.24, \ldots, 0.24}_{10}\}$
b_9	$\{5, 45\}$	$\{1, 1, 1, 0, 0 \vert\vert$ $\underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0, \ldots, 0}_{35}\}$	$\{0.89, 0.89, 0.89, 0.35, 0.35\}$
b_{10}	$\{5, 95\}$	$\{1, 1, 1, 0, 0 \vert\vert$ $\underbrace{0.25, \ldots, 0.25}_{20}, \underbrace{0, \ldots, 0}_{75}\}$	$\{0.97, 0.97, 0.97, 0.38, 0.38\}$

The "||" symbol is used to distinguish between elements of β^o and β^a.

posterior draws "averaged" over the imputations. The log-odds ratios (ORs), i.e. elements of β, were standardized with respect to the observed distribution of the 178 current patients, allowing for a comparison of magnitudes both between and within priors.

Table 3 gives the posterior medians of the ORs. 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 covariates, and the second two blocks of rows correspond to the added covariates. Figure 2 presents boxplots of the posterior distributions of all ORs.

Comparing PED-RESC and PED-RESC2, the direction and magnitude of the observed associations in the sets of original covariates were consistent between the two cohorts, with one exception: in PED-RESC2, no patients with a primary diagnosis of asthma died, i.e. the data were quasi-complete separated. All variable importance probabilities were generally closer to 1 in PED-RESC, due to its larger sample size. The regularized horseshoe (Standard) shrinks nearly all ORs, 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-type priors, the importance probabilities for the original risk factors were all greater than 75%, as well as those of added covariates of ALT and lactate. These two were also important according to Standard. Neither of the SAB priors found PaCO$_2$, MAP(CMV), admit hours pre-ECMO, malignancy, preECMO milrinone, or DX:Asthma to be important and, among the added covariates, identified bilirubin,

Table 3. *Posterior medians of standardized odds ratios (ORs) and, in parentheses, variable importance probabilities given as percentages, defined as the larger of (i) the posterior probability that an OR exceeds one and (ii) the posterior probability that an OR falls below one*

Model	pH	PaCO$_2$ (mmHg)	MAP(CMV) (cmH$_2$O)	MAP(HFOV) (cmH$_2$O)	Admit hours pre-ECMO (log)	Intubated hours pre-ECMO (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(96.0%)**	**0.67(76.9%)**	1.54(74.1%)	**1.59(83.4%)**	**1.64(86.8%)**	**2.30(95.7%)**
Standard	0.99(58.8%)	1.01(58.9%)	1.01(55.6%)	1.02(60.8%)	1.01(58.4%)	**1.96(90.6%)**
NAB(Agn)	**0.58(94.5%)**	0.88(70.1%)	**1.29(84.3%)**	**1.39(93.0%)**	**1.33(91.5%)**	**1.71(98.3%)**
NAB(Opt)	**0.55(97.5%)**	**0.82(78.4%)**	**1.33(88.7%)**	**1.41(96.2%)**	**1.36(95.7%)**	**1.67(99.0%)**
SAB(Agn)	0.90(73.7%)	1.02(56.9%)	1.10(69.2%)	**1.14(77.4%)**	1.03(59.2%)	**2.25(99.0%)**
SAB(Opt)	**0.86(76.9%)**	1.02(56.1%)	1.13(71.4%)	**1.15(78.4%)**	1.02(58.1%)	**2.25(99.4%)**

Model	Malignancy	Pre-ECMO 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(71.1%)	**1.31(76.6%)**	**0.05(99.2%)**	**0.76(75.1%)**	**2.07(97.9%)**
Standard	1.00(50.8%)	1.00(51.4%)	0.94(69.8%)	0.98(63.6%)	**1.33(85.2%)**
NAB(Agn)	**1.15(80.7%)**	**1.17(81.5%)**	**0.70(91.4%)**	**0.65(95.4%)**	**1.43(98.8%)**
NAB(Opt)	**1.17(87.0%)**	**1.21(87.8%)**	**0.71(93.6%)**	**0.63(98.2%)**	**1.40(99.4%)**
SAB(Agn)	1.03(59.2%)	0.99(54.9%)	0.96(60.6%)	**0.61(96.4%)**	**1.50(96.8%)**
SAB(Opt)	1.03(59.5%)	0.99(52.6%)	0.97(60.0%)	**0.57(97.9%)**	**1.49(96.9%)**

Model	Abnormal pupillary resp.	Bilirubin mg/dL (log)	ALT U/L (log)	Extent of leukocyt. (log)	Extent of leukopen. (log)	Extent of thrombocytopen. (log)
PedRESC	—	—	—	—	—	—
PedRESC2	—	—	—	—	—	—
Standard	0.99(56.4%)	1.07(72.5%)	**5.60(99.8%)**	1.04(68.7%)	0.97(66.6%)	1.01(55.1%)
NAB(Agn)	0.99(55.1%)	1.05(71.5%)	**4.88(99.6%)**	1.03(63.8%)	0.98(62.5%)	1.01(54.3%)
NAB(Opt)	0.99(54.6%)	1.04(69.6%)	**4.92(99.5%)**	1.02(63.1%)	0.98(61.5%)	1.01(54.1%)
SAB(Agn)	0.99(57.5%)	**1.13(75.5%)**	**4.87(99.8%)**	1.05(66.4%)	0.95(66.5%)	1.01(55.3%)
SAB(Opt)	0.99(57.8%)	**1.12(75.2%)**	**4.60(99.7%)**	1.04(64.6%)	0.96(65.1%)	1.01(55.3%)

Model	INR	VIS (log)	Lactate mMol/L (log)	PF ratio (log)	Pre-ECMO acute kidney injury
PedRESC	—	—	—	—	—
PedRESC2	—	—	—	—	—
Standard	1.02(62.0%)	1.01(57.9%)	**1.57(85.6%)**	0.94(71.4%)	1.00(53.7%)
NAB(Agn)	1.01(57.9%)	1.01(53.2%)	**1.10(77.6%)**	0.96(68.8%)	1.00(50.2%)
NAB(Opt)	1.01(56.7%)	1.00(53.0%)	**1.07(75.1%)**	0.97(67.6%)	1.00(50.0%)
SAB(Agn)	1.05(66.8%)	1.02(59.1%)	**1.94(90.7%)**	**0.87(77.2%)**	1.00(52.8%)
SAB(Opt)	1.05(66.8%)	1.02(60.1%)	**1.97(90.6%)**	**0.88(76.8%)**	1.00(51.9%)

Results in bold have percentages exceeding 75.0%.

ALT, lactate, and PF ratio as important. The posterior means of ϕ were 0.61 and 0.64, respectively, for the agnostic versions of NAB and SAB, and 0.84 and 0.82, respectively, for the optimistic versions.

From Figure 2, there are two general differences between Standard and the adaptive priors. For the two original covariates that Standard also identified as important (DX:Pertussis and intubated hours pre-ECMO), the posterior variability of the adaptive priors is smaller than Standard. Among the remaining original covariates, the adaptive priors have larger posterior variability than Standard; 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. This does not mean Standard is automatically preferred, because some of this

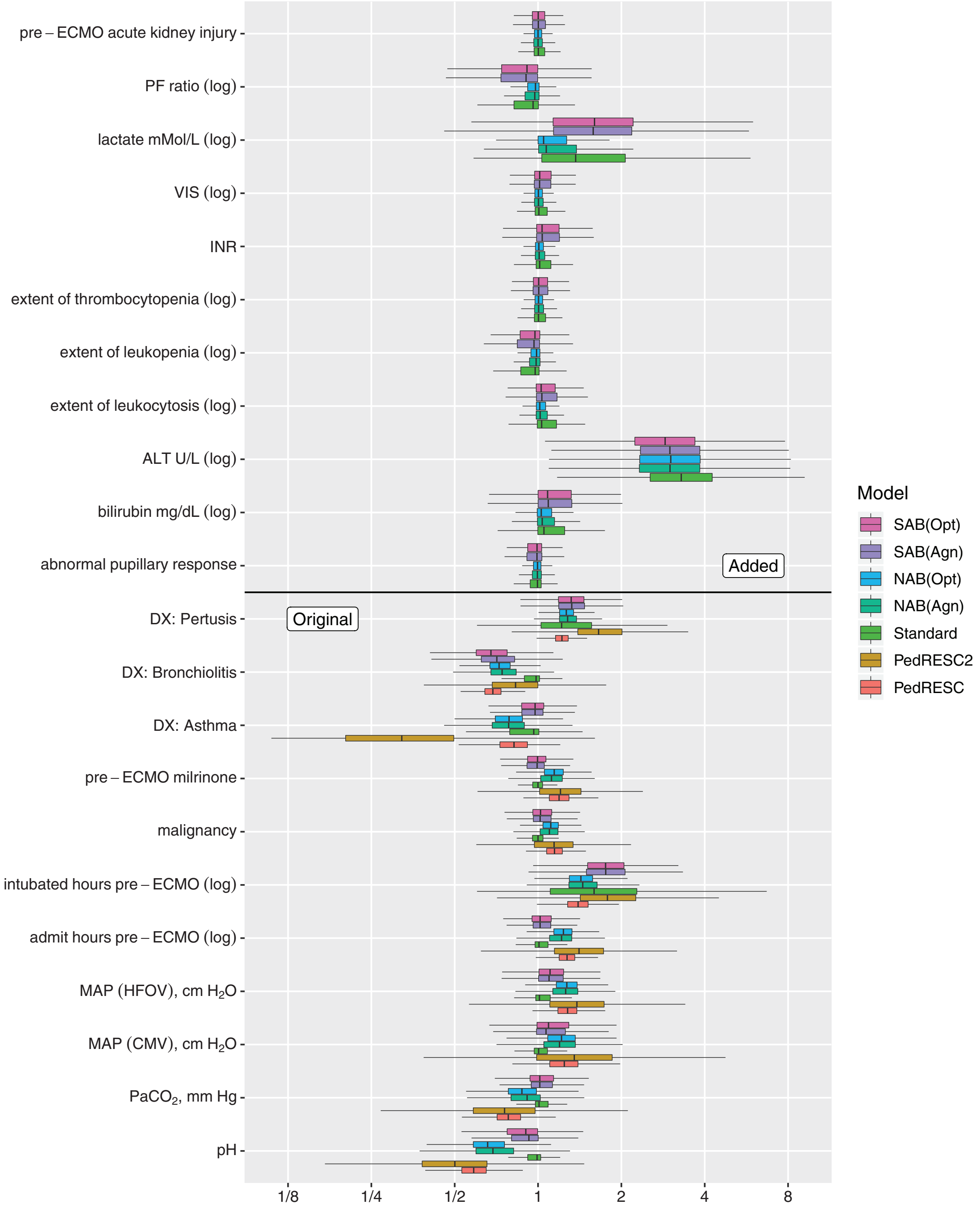

Fig. 2. Boxplots of posterior draws of odds ratios (ORs) from seven Bayesian logistic models using different priors. All models estimate ORs corresponding to the *original covariates* (the bottom eleven rows). Five models (all except “PedRESC” and “PedRESC2”) estimate ORs corresponding to the *added covariates* (the top eleven rows).

shrinkage likely reflects an inability to reliably estimate parameters rather than confidence that they are truly zero. In general, the SAB-type priors deviate more from PED-RESC than the NAB-type 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 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 parameters to being weakly informative about most parameters, 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 prior shrunk most ORs, both original and added, to one, 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 covariates identified by Ped-RESCUERS remain as risk factors for mortality after including the added biometric measurements. Agreeing more with our clinical expectation, the adaptive updates included between five and all eleven of the original Ped-RESCUERS risk factors and also identified two (NAB) or four (SAB) of the added covariates as likely risk factors. Some of this "loss" of importance may be due to correlation between the original and added covariates: $PaCO_2$ 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 found both bilirubin and ALT to be important. One surprising non-finding was SAB's failure to identify malignancy as a meaningful risk factor. NAB did identify malignancy as important, and this is more consistent with the Ped-RESCUERS model as well as our expectation based upon clinical experience and intuition.

The SAB prior depends upon both a simplifying functional approximation (3.6) and an imputation model for X^a given $X^o = x^o$ (3.7). In simulated scenarios for which the imputation model was readily estimated and had sufficient predictive ability, namely $p \approx q$, these approximations yielded considerable efficiency gains. Its advantage over NAB was most evident in coefficient setting b_5, in which the true value of β^o differed from the misspecified true value of α, and so NAB was substantially *worse* than Standard because the former was biased. Subsequent investigations have also suggested that SAB improves as μ is closer to zero. In contrast, in settings b_9 and b_{10}, for which $p \ll q$ and NAB slightly outperformed SAB, the need for imputation was likely to the detriment of SAB (although it still outperformed Standard). Such $p \ll q$ and small n_{curr} scenarios would correspond, for example, to an exploration of adding a panel of biomarkers, X^a, which are available on just a few subjects, to an established risk prediction model, $Y|X^o$. In that setting, the covariates X^o are probably weak predictors of X^a. Empirically, we have found that estimating $\boldsymbol{P}$ with 100 imputations worked as a good rule of thumb. Of course, there may be underlying differences between the current and historical populations in the true $[X^a|X^o]$ distribution that no amount of current data or number of imputations could recover. NAB is free of this particular distributional 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 (3.3) and (3.9) being the additive offset $\boldsymbol{P}\beta^a$ in (3.9), replacing it with $\gamma\boldsymbol{P}\beta^a$, $\gamma \in [0, 1]$, may be one way to leverage the advantages of both adaptive priors. Importantly, both NAB and SAB ignore information on the historical intercept, meaning that they assume that the historical and current models share $\{\beta^o, \beta^a\}$ in common but not necessarily the full data-generating mechanism $[Y|X^o, X^a]$.

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 to improve efficiency. 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.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank two referees, the Associate Editor and the Coordinating Editor for their thorough reviews and helpful comments. *Conflict of Interest:* None declared.

Funding

This work was partially supported by the National Institutes of Health [P30 CA046592; R01 CA129102] and the Extracorporeal Life Support Organization.

References

Antonelli, J., Zigler, C. and Dominici, F. (2017). Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics* **18**, 553–568.

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

Barbaro, R. P., Boonstra, P. S., Kuo, K. W., Selewski, D. T., Bailly, D. K., Stone, C. L., Chow, J., Annich, G. M., Moler, F. W. and Paden, M. L. (2018). Evaluating pediatric mortality risk prediction among children receiving extracorporeal respiratory support. *ASAIO Journal* doi:10.1097/mat.0000000000000813.

Barbaro, R. P., Boonstra, P. S., Paden, M. L., Roberts, L. A., Annich, G. M., Bartlett, R. H., Moler, F. W. and Davis, M. M. (2016). Development and validation of the pediatric risk estimate score for children using extracorporeal respiratory support (Ped-RESCUERS). *Intensive Care Medicine* **42**, 879–888.

Becker, B. J. and Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science* **22**, 414–429.

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

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In: van Dyk, D. and Welling, M. (editors), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, Volume 5. Proceedings of Machine Learning Research, Clearwater Beach, Florida USA: PMLR, pp. 73–80.

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**, 107–117.

CHEN, H., MANNING, A. K. AND DUPUIS, J. (2012). A method of moments estimator for random effect multivariate meta-analysis. *Biometrics* **68**, 1278–1284.

CHEN, M.-H., IBRAHIM, J. G. AND YIANNOUTSOS, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 223–242.

CHENG, W., TAYLOR, J. M. G., VOKONAS, P. S., PARK, S. K. AND MUKHERJEE, B. (2018). Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine* **37**, 1515–1530.

DUAN, Y., YE, K. AND SMITH, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics* **17**, 95–106.

GELMAN, A., CARLIN, H. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. AND RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd edition. Boca Raton, FL: CRC Press.

GRIFFIN, J. E. AND BROWN, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *Working Paper* 10, University of Warwick. Centre for Research in Statistical Methodology.

GRILL, S., ANKERST, D. P., GAIL, M. H., CHATTERJEE, N. AND PFEIFFER, R. M. (2017). Comparison of approaches for incorporating new information into existing risk prediction models. *Statistics in Medicine* **36**, 1134–1156.

IBRAHIM, J. G. AND CHEN, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

IBRAHIM, J. G., CHEN, M.-H., GWON, Y. AND CHEN, F. (2015). The power prior: theory and applications. *Statistics in Medicine* **34**, 3724–3749.

IBRAHIM, J. G., CHEN, M.-H. AND LIPSITZ, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics* **30**, 55–78.

JACKSON, D. AND RILEY, R. D. (2014). A refined method for multivariate meta-analysis and meta-regression. *Statistics in Medicine* **33**, 541–554.

NEUENSCHWANDER, B., BRANSON, M. AND SPIEGELHALTER, D. J. (2009). A note on the power prior. *Statistics in Medicine* **28**, 3562–3566.

PARK, T. AND CASELLA, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

PIIRONEN, J. AND VEHTARI, A. (2015). Projection predictive variable selection using Stan+R. arXiv preprint arXiv:1508.02502. https://arxiv.org/abs/1508.02502v1.

PIIRONEN, J. AND VEHTARI, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In: Singh, A. and Zhu, J. (editors), *Proceedings of the 20th International Conference on Articial Intelligence and Statistics*, Volume 54, Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 905–913.

PIIRONEN, J. AND VEHTARI, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**, 5018–5051.

R CORE TEAM. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. AND SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* **32**, 1–28.

STAN DEVELOPMENT TEAM (2017). *Stan Modeling Language User's Guide and Reference Manual, Version 2.17.0*. http://mc-stan.org/.

STAN DEVELOPMENT TEAM (2018). RStan: the R interface to Stan. R package version 2.17.3.

VAN BUUREN, S. AND GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**, 1–67.

WALKER, E., HERNANDEZ, A. V. AND KATTAN, M. W. (2008). Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine* **75**, 431–439.

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

ZHOU, X. AND REITER, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician* **64**, 159–163.

[*Received March 19, 2018; revised August 13, 2018; accepted for publication August 15, 2018*]