The Annals of Applied Statistics
2011, Vol. 5, No. 1, 201–231
DOI: 10.1214/10-AOAS373

BAYESIAN METHODS TO OVERCOME THE WINNER'S CURSE IN GENETIC STUDIES[1]

BY LIZHEN XU, RADU V. CRAIU[2] AND LEI SUN[2]

University of Toronto

Parameter estimates for associated genetic variants, report ed in the initial discovery samples, are often grossly inflated compared to the values observed in the follow-up replication samples. This type of bias is a consequence of the sequential procedure in which the estimated effect of an associated genetic marker must first pass a stringent significance threshold. We propose a hierarchical Bayes method in which a spike-and-slab prior is used to account for the possibility that the significant test result may be due to chance. We examine the robustness of the method using different priors corresponding to different degrees of confidence in the testing results and propose a Bayesian model averaging procedure to combine estimates produced by different models. The Bayesian estimators yield smaller variance compared to the conditional likelihood estimator and outperform the latter in studies with low power. We investigate the performance of the method with simulations and applications to four real data examples.

1. Introduction. Parameter estimates such as odds ratios (OR) for an associated genetic variant (e.g., SNP, Single-Nucleotide Polymorphism), reported from the same discovery samples that were initially used to declare statistical significance, are often grossly inflated compared to the values observed in the follow-up replication samples [e.g., Nair, Duffin and Helms (2009)]. This type of bias is a consequence of using the same data for both model selection and parameter estimation, because a declared associated variant must pass a stringent significance threshold. This phenomenon is also known as the Beavis effect [Xu (2003)] or the winner's curse [Zöllner and Pritchard (2007)] in the biostatistics literature.

The winner's curse has recently gained much attention in genetic studies, because it has been recognized as one of the major contributing factors to the failures of many attempted replication studies [e.g., Ioannidis, Thomas and Daly (2009)]. For example, five *Nature Genetic* publications in the first three months of 2009 acknowledged the effect of the winner's curse [e.g., Nair, Duffin and Helms (2009)]. In their recent *Nature Review* paper, Ioannidis, Thomas and Daly (2009) dedicated a section to the winner's curse and emphasized that "*the magnitude of the*

Received September 2009; revised May 2010.

[1]Supported by a research grant from the Canadian Institute of Health Research (CIHR).

[2]Supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Key words and phrases. Association study, Bayesian model averaging, hierarchical Bayes model, spike-and-slab prior, winner's curse.

winner's curse is inversely related to the power of the study. In typical circumstances, for 10% power, the inflation of an additive effect could be approximately 60%. . . . For small effects [anticipated for susceptibility loci associated with complex diseases/traits], even large meta-analyses could be grossly under-powered and emerging associations could be considerably inflated. For rare variants, the power can be <1%."

Some authors [e.g., Göring, Terwilliger and Blangero (2001)] have argued that reliable parameter estimates can be obtained only from an independent sample. However, collecting additional samples could be undesirable due to, for example, time and budget constraints as well as concerns over population heterogeneity and sampling differences. Two categories of methods were subsequently proposed to correct for the selection bias using the original samples only: the model-free resampling based methods [Sun and Bull (2005); Wu, Sun and Bull (2006); Yu et al. (2007); Jefferies (2007)] and the likelihood based methods [Zöllner and Pritchard (2007); Ghosh, Zou and Wright (2008); Zhong and Prentice (2008); Xiao and Boehnke (2009)]. Both types of approaches were shown to substantially reduce the estimation bias in relatively small samples, and comparable performances were observed by Faye et al. (2009). However, one caveat is that the variances of the proposed estimators in both categories are considerably higher than the original naïve estimator and lead to highly variable estimates of the sample size needed for replication studies. Although the increased variability is expected, due to the bias-variance trade-off, it may be too high to provide practical design recommendations. For example, Figure 4 of Zöllner and Pritchard (2007) shows that the bias-adjusted sample size estimates range from $\sim$500 to $\sim$100,000 compared to the actual required sample size of 1,261 for a successful replication study ($\alpha = 10^{-6}$, power $= 80\%$).

Motivated by the above observations and the fact that some form of prior information is often available in genetic studies, we propose here a Bayesian framework to further reduce the bias and decrease the variability in the estimates. In particular, we focus on the OR estimates from genome-wide association studies (GWAS) via logistic regression analyses of case-control disease status, because most of the current genetic mapping studies adopt the case-control GWAS design. We first describe the statistical model in Section 2. We prove in Section 3 that, conditional on statistical significance, there are no unbiased estimators for the log OR. We present the Bayesian methodology in Section 4 with detailed discussions on the prior specifications and the advantages of model averaging. We assess the performance of the proposed methods in Section 5 via extensive simulation studies under a general normal model and specific genetic models. We demonstrate the utility of our methods in Section 6 with applications to four different association studies, including a candidate gene study and three GWAS of either binary case-control or quantitative outcomes. Our concluding remarks are in Section 7.

2. The statistical model. Let β refer to the true log Odds Ratio (OR), the parameter of interest, for the risk allele of an associated SNP, and Z the statistic of the corresponding association test. Following Ghosh, Zou and Wright (2008), we assume that Z is asymptotically normally distributed and has the form

$$Z = \frac{\widehat{\beta}}{\widehat{SE}(\widehat{\beta})} \sim N\Big(\frac{\beta}{SE(\widehat{\beta})}, 1\Big),$$

where $\widehat{\beta}$ is the estimate for β from the logistic regression, $\text{logit}(E[Y]) = \alpha + \beta X$, in which the response variable Y is the affection status of a sample ($0 =$ unaffected and $1 =$ affected by the disease of interest) and the predictor $X \in \{0, 1, 2\}$ is the SNP genotype coded additively (X represents the number of copies of the risk allele). Other covariates may be also included in the model. Without loss of generality, we assume that the minor allele is the risk allele and the alternative of interest is one-sided, that is, $H_0 : \beta = 0$ vs. $H_1 : \beta > 0$. The association test in this case is based on the Wald test, and if the null hypothesis is rejected, the standard practice is to directly use the $\widehat{\beta}$ from the logistic regression as the estimate for β.

The above estimation procedure is essentially the same as the familiar practice of population mean estimation in the following more general statistical setup. Assuming that n i.i.d. samples, $\{X_1, \ldots, X_n\}$, were collected from a normal population with mean μ and variance σ^2, a significance test is first conducted for $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$ based on the statistic, $T_n = \frac{\overline{X}}{S/\sqrt{n}}$, which follows $N(\frac{\mu}{\sigma/\sqrt{n}}, 1)$, where $\overline{X}$ and S are the sample mean and standard deviation. The sample mean $\overline{X}$, calculated from the *same* sample, is subsequently used as an estimate for μ, without adjusting for the fact that the null hypothesis was rejected (i.e., $T_n > c$, where c is the critical value corresponding to type I error rate α) and that estimation is performed for samples with positive findings only. Note that, in our simplified model, although $E[\overline{X}] = \mu$, the conditional mean $E[\overline{X}|\overline{X} > (cS/\sqrt{n})]$ is strictly greater than μ, unless the power of the test is 100%. Thus, this naïve estimate, $\overline{X}$, is upward biased. The amount of bias is inversely proportional to the power as was first demonstrated by Göring, Terwilliger and Blangero (2001) in genome-wide linkage analyses and later by Garner (2007) for genome-wide association studies. The likelihood based methods proposed by Ghosh, Zou and Wright (2008) and others propose to correct for this selection bias by calculating the maximum likelihood estimate (MLE) of μ from the correct conditional likelihood. In this setting,

$$P(\mathbf{X}|\mu, \sigma^2, T_n > c) = \prod_{i=1}^{n} \frac{(1/\sqrt{2\pi\sigma^2}) \exp[-(X_i - \mu)^2/2\sigma^2]}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))}, \tag{2.1}$$

where Φ is the cumulative distribution function (c.d.f.) of the standard normal distribution.

Although the above normal model is a conceptual one, it connects directly with the logistic model used for case-control association studies. Specifically, β (the true log OR) corresponds to μ (the normal population mean), $\widehat{\beta}$ (the naïve estimate) corresponds to the statistic $\overline{X}$, and $\widehat{SE}(\widehat{\beta})$ corresponds to $S/\sqrt{n}$. In the following development of the bias correction Bayesian methods, we choose to focus on the normal model for a number of reasons. The key factor that influences the selection bias is the power of the association test, which depends on the noncentrality parameter, $\beta/SE(\hat{\beta})$. In practice, β is the true log OR, but $SE(\hat{\beta})$ is a complex function of multiple components including the prevalence of the disease in the population, the disease model (e.g., additive, dominant or others), the minor allele frequency of the SNP, the sample size and the significance threshold used [Slager and Schaid (2001)]. The normal model allows us to concisely control the main factor of interest, the power of the association test, in the simulation studies, by fixing the normal population mean ($\mu \leftrightarrow \beta$, the log OR) and considering practically meaningful ranges of significance threshold value, power and sample size (n), which in turn determine the normal population variance [σ, and $\sigma/\sqrt{n} \leftrightarrow SE(\hat{\beta})$]. Moreover, this conceptual normal model also covers association analyses of quantitative outcome, Y, for which a linear regression model is typically used, for example, $E[Y] = \alpha + \beta X$. In that case, the population mean μ in the conceptual normal model represents the regression coefficient β. In Section 6 we show how our Bayesian methods built upon this conceptual normal model can be applied to published association studies for which only the OR (or the regression coefficient), the association p-value, the sample size and the significance threshold were available.

In the following, we first show that there are no unbiased estimators for the population mean conditionally on the significance of the corresponding hypothesis test. We then proceed with the development of a catalogue of Bayesian estimators and the evaluation of their performance via simulation and application studies.

3. Lack of unbiased estimators for μ. Ghosh, Zou and Wright (2008) and other authors have demonstrated that the MLE from the correct conditional likelihood could substantially reduce the bias. However, they also observed via simulation studies that the conditional MLE tends to over-correct for large μ and under-correct for small μ. Stallard, Todd and Whitehead (2008) showed that there is no conditional unbiased estimators for the effect of treatment A from a sample that was first used to select treatment A over B, that is, conditioning on the fact that the sample effect of treatment A was larger than that of treatment B. Although previous authors [Zhong and Prentice (2008); Bowden and Dudbridge (2009)] discussed that a similar argument can be used in the case considered here, below we provide a formal proof to show that there are no unbiased conditional estimators for the population mean μ even when the population variance σ^2 is known.

Because T_n is a sufficient statistic for μ when σ is known, the completeness of the normal family of distributions implies that we can restrict the search for

unbiased estimators of $\frac{\mu}{\sigma/\sqrt{n}}$ to functions of T_n. Now suppose that some function $h(T_n)$ is an unbiased estimator of $\frac{\mu}{\sigma/\sqrt{n}}$ conditional on the statistical significance, that is, $T_n > c$. Let $g(T_n) = \{T_n - h(T_n)\}$, then

$$\begin{aligned}
E[g(T_n)|T_n > c] &= E[T_n|T_n > c] - E[h(T_n)|T_n > c] \\
&= \int_c^\infty T_n \frac{\phi(T_n - \mu/(\sigma/\sqrt{n}))}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))}\, d(T_n) - \frac{\mu}{\sigma/\sqrt{n}} \\
&= \frac{1}{B}\int_{c-\mu/(\sigma/\sqrt{n})}^\infty \left(z + \frac{\mu}{\sigma/\sqrt{n}}\right)\phi(z)\, dz - \frac{\mu}{\sigma/\sqrt{n}} \\
&= \frac{1}{B}\left[\int_{c-\mu/(\sigma/\sqrt{n})}^\infty z \cdot e^{-z^2/2}\, dz + B \cdot \frac{\mu}{\sigma/\sqrt{n}}\right] - \frac{\mu}{\sigma/\sqrt{n}} \\
&= \frac{1}{B}\left[\phi\left(c - \frac{\mu}{\sigma/\sqrt{n}}\right) + B \cdot \frac{\mu}{\sigma/\sqrt{n}}\right] - \frac{\mu}{\sigma/\sqrt{n}} \\
&= \frac{\phi(c - \mu/(\sigma/\sqrt{n}))}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))},
\end{aligned}$$

where $B = 1 - \Phi(c - \frac{\mu}{\sigma/\sqrt{n}})$.

Thus, we have

$$\int_c^\infty g(T_n) \frac{\phi(T_n - \mu/(\sigma/\sqrt{n}))}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))}\, dT_n = \frac{\phi(c - \mu/(\sigma/\sqrt{n}))}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))}, \tag{3.1}$$

which implies

$$\int_c^\infty g(T_n)\phi\left(T_n - \frac{\mu}{\sigma/\sqrt{n}}\right) dT_n = \phi\left(c - \frac{\mu}{\sigma/\sqrt{n}}\right). \tag{3.2}$$

Now, let $\delta_c(y)$ be the Dirac delta function defined for $y \geq c$ such that it is equal to 0 for all y greater than c and $\int_c^\epsilon \delta_c(y)\, dy = 1$ for all $\epsilon > 0$. It is easy to see that a solution to equation (3.2) is $g(T_n) = \delta_c(T_n)$. By the completeness of the normal distribution, the solution $g(T_n) \cdot \mathbf{1}_{\{T_n > c\}}$ is unique almost everywhere. Thus, $h(T_n) \cdot \mathbf{1}_{\{T_n > c\}} = T_n \cdot \mathbf{1}_{\{T_n > c\}}$ holds almost everywhere. Hence, T_n is also an unbiased estimator for $\frac{\mu}{\sigma/\sqrt{n}}$. However, $T_n \cdot \mathbf{1}_{\{T_n > c\}}$ has an upward bias equal to $\frac{\phi(c-\mu/(\sigma/\sqrt{n}))}{1-\Phi(c-\mu/(\sigma/\sqrt{n}))}$. Therefore, we conclude that there are no unbiased estimators of $\frac{\mu}{\sigma/\sqrt{n}}$ and hence no unbiased estimators of μ.

4. Bayesian bias correction.

4.1. *Prior specification.* The possible available prior information for genome-wide association studies (GWAS) is diverse due to, for example, results from previous genome-wide linkage analyses or candidate studies, or biological evidence on the SNPs. One common theme, however, is the anticipated low power of the

GWAS and the well-acknowledged fact that an apparent significantly associated SNP could be a false positive [Ioannidis, Thomas and Daly (2009)]. Thus, the performance of the proposed Bayesian methods is assessed in this context, although the practical implementation of the methods could be study specific depending on the type of the available prior.

The Bayesian paradigm allows us to incorporate in our model the prior belief that the significance of the effect observed *may be due to chance*. Mathematically, this belief can be modeled using a spike-and-slab prior which is essentially a mixture between a discrete probability with mass at zero and a continuous density f with support on the positive real line

$$p(\mu|\xi) = \xi\delta_{\{0\}}(\mu) + (1-\xi)f(\mu),$$

where ξ is either constant or a hyperparameter in the model.

The spike-and-slab priors have a long history in the Bayesian literature on variable selection and shrinkage estimation, for example, Box and Meyer (1986), Mitchell and Beauchamp (1988), George and McCulloch (1993), Chipman (1996), Clyde, DeSimone and Parmigiani (1996), Geweke (1996), and Kuo and Mallick (1998). A recent theoretical study by Ishwaran and Rao (2005) discusses the similarities between Bayesian procedures using the spike-and-slab priors and frequentist procedures.

We treat ξ as a hyperparameter with a Beta distribution, $\xi \sim \text{Beta}(a, b)$. The parameters a, b reflect our degree of prior belief in $\mu = 0$ (false positive) versus $\mu > 0$ (true positive). If we set $a = b = 1$, then $p(\xi|a = 1, b = 1)$ is the Uniform(0, 1) density, which implies that we do not favor, a priori, any region of (0, 1). This could be considered the "noninformative" prior for ξ. The choice $a = 2/3$ and $b = 2/3$ corresponds to our belief in two extreme outcomes: ξ is either close to 0 (believing in true positive, $\mu > 0$) or close to 1 (believing in false positive, $\mu = 0$). Smaller values for a and larger values for b, say, $a = 0.5$ and $b = 8$, lead to a higher prior confidence that the signal is real. Similarly, larger values for a and smaller values for b, say, $a = 8$ and $b = 0.5$, correspond to prior skepticism regarding the observed association between the significant SNP and the trait of interest. Figure 1 shows the Beta distribution of ξ for different values of a and b.

Although we focus on Beta(0.5, 8), and Beta(8, 0.5) in evaluating the performance of the proposed Bayesian methods, we conducted additional simulations to study the model's robustness to the choice of priors. Simulation results included in the supplementary material indicate that other values for a and b [e.g., Beta(0.5, 16) or Beta(4, 0.5)] that preserve the L-shaped or the "inverse" L-shaped density, as seen in Figure 1, produce very similar inferences.

In the existing likelihood approaches the sample variance, S^2, is typically used to estimate σ^2 [Ghosh, Zou and Wright (2008)]. Although the variance estimator has relatively high precision in large samples, it could be subject to the selection

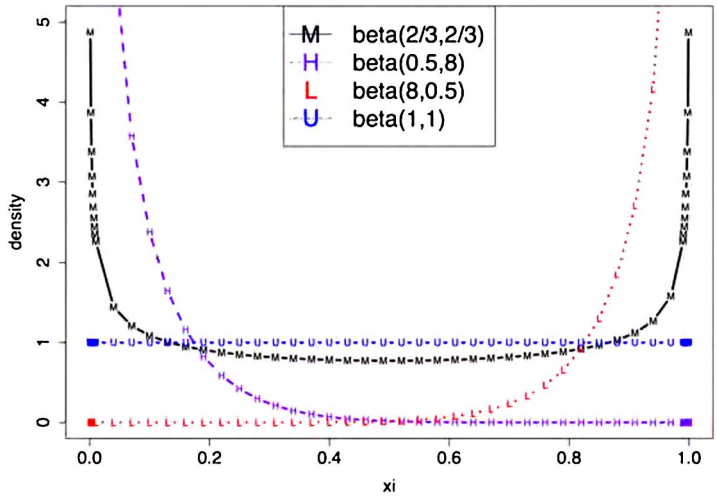

FIG. 1. *Density of the prior* Beta(a, b) *for* ξ *with different choices of* a *and* b.

bias in small samples [Faye et al. (2009)]. Therefore, we adopt an empirical Bayes prior for σ^2 in which the hyperparameters of the inverse gamma distribution, α_1 and α_2, are chosen so that the a priori mean of σ^2 is equal to S^2, the sample variance, but the prior variance of σ^2 is equal to 200. We note that additional simulations with more certainty about σ^2 (prior variance of σ^2 as small as 10) or less certainty (as large as 1000) produce very similar results.

We use Uniform$(0, A)$ to specify $f(\mu)$, the density function for the continuous component of the prior for μ, the log OR, where A represents the upper bound of log OR. However, in this parametrization the estimator is very sensitive to the choice of A. To show this, let Z be the latent mixture indicator so that $Z = 0$ if the significant SNP is a false positive ($\mu = 0$) and $Z = 1$ for a true positive ($\mu > 0$). It is not difficult to see that

$$Z|\vec{X}, \xi, \mu, \sigma^2 = \begin{cases} 0, & \text{with probability } \dfrac{p_0}{p_0 + p_1}, \\ 1, & \text{with probability } \dfrac{p_1}{p_0 + p_1}, \end{cases}$$

where $\vec{X} = \{X_1, \ldots, X_n\}$ and

$$p_0 = \frac{\xi}{1 - \Phi(c)},$$

$$p_1 = \frac{1}{A} \times \frac{(1 - \xi)\exp\{-(1/(2\sigma^2))(n\mu^2 - 2\mu\sum_{i=1}^{n} X_i)\}}{1 - \Phi(c - \mu/(\sigma/\sqrt{n}))}.$$

Thus, depending on the value of A, p_1 can be made arbitrarily small regardless of the data available. This can influence dramatically (even for $A = 2$) the performance of the computational algorithm used to obtain the posterior distribution of

interest (described in Section 4.3). One simple method to circumvent this problem is to use the reparametrization $\theta = \mu/A$ which dissolves the influence of A on p_1. Therefore, the proposed Bayesian method has the following hierarchical prior structure:

$$
\begin{aligned}
p(\theta|\xi) &= \xi g_0(\theta) + (1-\xi) g_1(\theta), \\
\xi &\sim \text{Beta}(a, b), \\
\sigma^2 &\sim \text{Inv-Gamma}(\alpha_1, \alpha_2),
\end{aligned} \tag{4.1}
$$

where $\alpha_1 = S^4/200 + 2$, and $\alpha_2 = S^6/200 + S^2$, S is the sample standard deviation, $g_0(\theta) = \delta_{\{0\}}(\theta)$ and $g_1(\theta)$ is the density of Uniform(0, 1).

In the actual implementation, we use $A = 2$ to reflect the known maximum log OR of SNPs identified for complex diseases and traits. For example, the truly associated SNP in the well-known major histocompatibility complex (MHC) region has perhaps the highest genetic effect observed to date, with a log OR of $\log(5.49) = 1.7$ [WTCCC (2007)]. We note that additional simulations showed that, as long as the reparametrization $\theta = \mu/A$ is used, results remain largely the same for higher upper bounds (e.g., $A = 6$ corresponding to a maximum OR ≈ 400). Applications in Section 6 also demonstrate the robustness of the model when it was applied not only to case-control data but also to an association study of a quantitative outcome.

4.2. *Posterior distribution.* The joint prior distribution for (θ, ξ) is

$$
\begin{aligned}
p(\theta, \xi) &= p(\theta|\xi) p(\xi) \\
&= \xi g_0(\theta) \xi^{a-1} (1-\xi)^{b-1} + (1-\xi) g_1(\theta) \xi^{a-1} (1-\xi)^{b-1}.
\end{aligned} \tag{4.2}
$$

Conditional on Z, the sampling distribution is

$$
\begin{aligned}
&P(\vec{X}|\theta, \sigma^2, Z, T_n > c) \\
&\quad \propto (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n X_i^2/(2\sigma^2)\}}{1 - \Phi(c)} \right)^{1-Z} \\
&\qquad \times \left(\frac{\exp\{-\sum_{i=1}^n (X_i - 2\theta)^2/(2\sigma^2)\}}{1 - \Phi(c - 2\theta/(\sigma/\sqrt{n}))} \right)^{Z}.
\end{aligned}
$$

If Z were observed, the posterior distribution for the vector (θ, ξ, σ^2) would be

$$
\begin{aligned}
&p(\theta, \xi, \sigma^2|\vec{X}, Z, T_n > c) \\
&\quad \propto p(\vec{X}, Z|\theta, \sigma^2, T_n > c) p(\theta|\xi) p(\xi) p(\sigma^2) \\
&\quad \propto (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n X_i^2/(2\sigma^2)\}\xi}{1 - \Phi(c)} \right)^{1-Z}
\end{aligned} \tag{4.3}
$$

$$\times \left(\frac{\exp\{-\sum_{i=1}^{n}(X_i-2\theta)^2/(2\sigma^2)\}(1-\xi)}{1-\Phi(c-2\theta/(\sigma/\sqrt{n}))}\right)^{Z}$$
$$\times \xi^{a-1}(1-\xi)^{b-1}\left(\frac{1}{\sigma^2}\right)^{\alpha_1+1}\exp\{-\alpha_2/\sigma^2\}$$

for $\theta, \xi \in [0,1]$, $\sigma > 0$ (detailed derivation provided in the Supplementary material). We note that the posterior distribution specified in equation (4.3) depends on the data only through the sufficient statistics for (μ, σ^2), $D_n = (\sum X_i, \sum X_i^2)$. This is particularly useful in practice when the original sample-specific data $\vec{X}$ are not available, but the sufficient statistics are provided or could be inferred from typically reported quantities such as the sample size, the observed OR and association p-value, and the significance threshold used.

4.3. *Sampling from the posterior distribution.* The latent variable Z is unobservable in practice, so equation (4.3) cannot be used directly to study the characteristics of the posterior distribution, $\pi(\theta, \xi, \sigma^2) = p(\theta, \xi, \sigma^2 | D_n, T_n > c)$. The traditional approach in this type of situation is to use Markov chain Monte Carlo (MCMC) techniques to sample from π. The posterior distribution has a mixture form for which the Data Augmentation algorithm of Tanner and Wong (1987) has been proven extremely efficient [see also van Dyk and Meng (2001)]. The algorithm relies on sampling alternatively from the distribution of $Z|D_n, \theta, \xi, \sigma^2$ and $\theta, \xi, \sigma^2|Z, D_n$. More precisely, at iteration t we carry out the following steps:

Step 1. Sample $Z_t \in \{0,1\}$ given ξ_{t-1}, θ_{t-1} and σ_{t-1}^2 from the conditional distribution

$$Z_t|\xi_{t-1}, \theta_{t-1}, \sigma_{t-1}^2 = \begin{cases} 0, & \text{with probability } \dfrac{p_0}{p_0+p_1}, \\ 1, & \text{with probability } \dfrac{p_1}{p_0+p_1}, \end{cases}$$

where

$$p_0 = \frac{\xi_{t-1}}{1-\Phi(c)},$$
$$p_1 = \frac{(1-\xi_{t-1})\exp\{-(1/(2\sigma_{t-1}^2))(4n\theta_{t-1}^2 - 4\theta_{t-1}\sum_{i=1}^{n}X_i)\}}{1-\Phi(c-2\theta_{t-1}/(\sigma_{t-1}/\sqrt{n}))}.$$

Step 2. (i) If $Z_t = 0$, sample

$$\xi_t \sim \text{Beta}(a+1, b),$$
$$\sigma_t^2 \sim p(\sigma^2|D_n) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+\alpha_1+1}\exp\left\{-\frac{1}{\sigma^2}\left(\alpha_2 + \frac{\sum_{i=1}^{n}X_i^2}{2}\right)\right\},$$

which is the inverse gamma distribution with shape parameter equal to $\frac{n}{2}+\alpha_1$, and scale parameter equal to $\alpha_2 + \frac{\sum_{i=1}^{n}X_i^2}{2}$. We also set $\mu_t = \theta_t = 0$.

(ii) If $Z_t = 1$, sample

$$\xi_t \sim \text{Beta}(a, b+1),$$

$$\theta_t \sim p(\theta|D_n, \sigma_{t-1}) \propto \frac{\exp\{-2n\theta^2/\sigma_{t-1}^2 - 2\theta\sum_{i=1}^n X_i/\sigma_{t-1}^2\}}{1-\Phi(c-2\theta/(\sigma_{t-1}/\sqrt{n}))}\mathbf{1}_{(0,1)}(\theta),$$

$$\sigma_t^2 \sim p(\sigma^2|(D_n, \xi_t, \theta_t) \propto \frac{\exp\{-1/(2\sigma^2)(\sum_{i=1}^n X_i^2 + 4n\theta_t^2 - 4\theta_t\sum_{i=1}^n X_i)\}}{(1-\Phi(c-2\theta_t/\sqrt{\sigma^2/n}))}$$
$$\times (\sigma^2)^{n/2+\alpha_1+1}\exp\{-\alpha_2/\sigma^2\}.$$

The sampling of θ_t and σ_t^2 at step 2(ii) cannot be carried out directly, so we apply a Metropolis–Hasting algorithm [Metropolis et al. (1953)]. We use 20,000 iterations to obtain 15,000 posterior samples, discarding the first 5000 "burn-in" samples. The sample mean of the above 15,000 posterior samples, $\overline{\theta}$, is used to estimate the posterior mean $E[\mu|D_n, T_n > c]$. That is, $\widehat{\mu}_B = 2\overline{\theta}$, where the factor 2 is due to the initial reparametrization $\theta = \mu/A$ and $A = 2$. (Additional simulations presented in the Supplementary material show that running the chain longer or discarding more "burn-in" samples provide similar results.)

4.4. *Bayesian Model Averaging* (*BMA*). The Bayesian model averaging (BMA) is a coherent and conceptually simple method devised to take into account the model uncertainty [see Hoeting et al. (1999) and references therein]. For the problem discussed here, the uncertainty is related to our lack of information regarding the power of the test performed in the first stage. If we knew, say, that the power of the test is high, then we would be more confident that the signal detected is a true signal and this would be reflected in our choice of the prior. In the absence of such information, one could adopt the BMA methodology to increase the robustness of the Bayesian estimator.

In the BMA paradigm, assume that Δ is the quantity of inferential interest for which a number of candidate models, say, $M_1, \ldots, M_K$, are available. Given the prior probability for each candidate model, $p(M_i)$, $1 \le i \le K$, the traditional BMA method assigns the posterior distribution given data D for Δ

$$p(\Delta|D) = \sum_{k=1}^K p(\Delta|M_k, D)p(M_k|D), \tag{4.4}$$

where

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=1}^K p(D|M_l)p(M_l)}$$

and

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)\,d\theta_k.$$

In our setting, $K = 2$ because only two models are considered. Let M_1 be the model with prior $p(\xi) = \text{Beta}(8, 0.5)$ (a priori favors the belief that the initial discovery is a false positive) and M_2 for $p(\xi) = \text{Beta}(0.5, 8)$ (a priori favors the belief that the initial discovery is a true positive). To specify the values for $p(M_1)$ and $p(M_2)$, we utilize the threshold value c in the following fashion, $p(M_1) = e^{(-c/2)}$ and $p(M_2) = 1 - e^{(-c/2)}$. Thus, our prior belief in model M_1 (with higher density for false positive) decreases as the testing threshold value increases at an exponential rate. The posterior probabilities for the two models can be derived as

$$p(M_i|D_n) = \frac{p(D_n|M_i)p(M_i)}{p(D_n|M_1)p(M_1) + p(D_n|M_2)p(M_2)}, \qquad i = 1, 2.$$

Thus,

$$\frac{p(M_1|D_n)}{p(M_2|D_n)} = \frac{p(D_n|M_1)}{p(D_n|M_2)} \cdot \frac{e^{(-c/2)}}{(1 - e^{(-c/2)})}. \tag{4.5}$$

The direct computation, however, is difficult because the integral

$$p(D_n|M) = \int\int_{(\mu,\xi,\sigma^2)} p(D_n|\mu, \xi, \sigma^2, M)p(\mu|\xi, M)p(\xi|M)p(\sigma^2|M)\, d\mu\, d\xi$$

cannot be calculated in a closed form. Note that

$$p(\mu, \xi, \sigma^2|D_n, M) = \frac{p(D_n|M, \mu, \xi, \sigma^2)p(\mu|\xi, M)p(\xi|M)p(\sigma^2|M)}{p(D_n|M)}, \tag{4.6}$$

thus $p(D_n|M)$ can be viewed as the normalizing constant of the posterior distribution $p(\mu, \xi, \sigma^2|D_n, M)$. Therefore, the first ratio in (4.5) is a ratio of two normalizing constants for two densities from which we can sample. The problem of estimating ratios of two normalizing constants has been discussed by, among others, Meng and Wong (1996) and Gelman and Meng (1998). We use the bridge sampling method proposed by Meng and Wong (1996) to compute the ratio in (4.5).

To compute (4.5), let $r = p(D_n|M_1)/p(D_n|M_2)$, $\omega = (\mu, \xi, \sigma^2)$, $\pi_i = p(\mu, \xi, \sigma^2|D_n, M_i)$ and $q_i(\mu, \xi, \sigma^2) = p(D_n|M_i, \mu, \xi, \sigma^2)p(\mu|\xi, M_i)p(\xi|M_i)p(\sigma^2|M_i)$, for $1 \le i \le 2$. Given $m = 10{,}000$ samples $\{(\mu_{i1}, \xi_{i1}, \sigma_{i1}^2), \ldots, (\mu_{in_i}, \xi_{in_i}, \sigma_{i1}^2)\}$ from each density π_i, we can approximate r using the iterative procedure of Meng and Wong (1996). Specifically, after starting with an initial estimate $\hat{r}^{(0)}$, at the $(t + 1)$st iteration, we compute

$$\begin{aligned}\hat{r}^{(t+1)} &= \frac{(1/m)\sum_{j=1}^{m}[q_1(\omega_{2j})/(s_1q_1(\omega_{2j}) + s_2\hat{r}^{(t)}q_2(\omega_{2j}))]}{(1/m)\sum_{j=1}^{m}[q_2(\omega_{1j})/(s_1q_1(\omega_{1j}) + s_2\hat{r}^{(t)}q_2(\omega_{1j}))]} \\ &\equiv \frac{(1/m)\sum_{j=1}^{n_2}[l_{2j}/(s_1l_{2j} + s_2\hat{r}^{(t)})]}{(1/m)\sum_{j=1}^{m}[1/(s_1l_{1j} + s_2\hat{r}^{(t)})]},\end{aligned} \tag{4.7}$$

where $s_i = 0.5$, and $l_{ij} = \frac{q_1(\omega_{ij})}{q_2(\omega_{ij})}$, for $1 \le j \le m$, $1 \le i \le 2$. Note that l_{ij} needs to be computed only once at the beginning of the algorithm. The convergent value of $\hat{r}^{(t)}$ is the one we choose to estimate r.

In the current setting l_{ij} is easy to compute since

$$l_{ij} = \frac{p(D_n|M_1, \mu_{ij}, \xi_{ij}, \sigma_{ij}^2)p(\mu_{ij}|\xi_{ij}, M_1)p(\xi_{ij}|M_1)p(\sigma_{ij}^2|M_1)}{p(D_n|M_2, \mu_{ij}, \xi_{ij}, \sigma_{ij}^2)p(\mu_{ij}|\xi_{ij}, M_2)p(\xi_{ij}|M_2)p(\sigma_{ij}^2|M_2)}$$
$$= \frac{p(\xi_{ij}|M_1)}{p(\xi_{ij}|M_2)} = \xi_{ij}^{7.5}(1-\xi_{ij})^{-7.5}.$$

From equations (4.4) and (4.5), we obtain the BMA estimator of μ,

$$\hat{\mu}_{BMA} = \frac{\hat{r}e^{(-c/2)}}{\hat{r}e^{(-c/2)} + 1 - e^{(-c/2)}}\hat{\mu}_1 + \frac{1 - e^{(-c/2)}}{\hat{r}e^{(-c/2)} + 1 - e^{(-c/2)}}\hat{\mu}_2, \tag{4.8}$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the posterior means of μ obtained under models M_1 and M_2, respectively.

5. Simulation study. We carried out two sets of simulations to examine the performances of the Bayesian methods and compared the results with those from the likelihood-based estimators of Ghosh, Zou and Wright (2008). The first set of simulations used data generated from the normal model that was used to outline and develop the Bayesian methods, and the second set used data simulated from a case-control genetic model. The nine estimators examined are as follows:

N: The naïve estimator ($\overline{X}$, the *unconditional* MLE).

MLE: The *conditional* MLE estimator based on equation (2.1), that is the β_1 estimator in Ghosh, Zou and Wright (2008).

NMLE: The mean of the Normalized Conditional Likelihood estimator, that is, the β_2 estimator of Ghosh, Zou and Wright (2008).

Ghosh: The average estimator of MLE and NMLE, that is, the β_3 estimator recommended by Ghosh, Zou and Wright (2008).

B.L: The Bayesian estimator based on equation (4.3) when the prior for ξ is Beta$(8, 0.5)$ (the prior belief is low power of the initial discovery study).

B.H: The Bayesian estimator based on equation (4.3) when the prior for ξ is Beta$(0.5, 8)$ (the prior belief is high power of the initial discovery study).

B.BMA: The BMA estimator obtained by averaging the B.L and B.H models, based on equation (4.8).

B.M: The Bayesian estimator based on equation (4.3) when the prior for ξ is Beta$(2/3, 2/3)$ (the prior belief is either low or high power).

B.Unif: The Bayesian estimator based on equation (4.3) when the prior for ξ is Uniform$(0, 1)$ (the "noninformative" prior).

Whenever an obtained estimate was negative, it was truncated to be zero following the standard practice of interpreting the "flip–flop" phenomenon occurring at the same SNP in the same population [Lin et al. (2007)]. That is, a SNP is found to be associated with the disease of interest in two independent studies, but the risk allele is reversed (i.e., the allele that increases the risk in one study is the protective allele that decreases the risk in another study).

5.1. *Simulation set* 1*—normal model*. We considered a factorial design in which the factors are the power of the association test, the type 1 error rate and the sample size. The power levels are {5%, 10%, 20%, 50%, 99%}, of which 99% allows us to investigate the asymptotic behavior of the methods while 20% or lower reflect the low power anticipated for genome-wide association studies (GWAS). The type 1 error rates, α, are $\{0.05, 10^{-4}, 10^{-6}\}$, of which 0.05 is the typical choice for a single SNP study, while the other two are suitable for high-throughput GWAS depending on the density of the SNPs being genotyped. The corresponding threshold values for the test statistics, c, are $\{1.645, 3.719, 4.753\}$. The true population mean is fixed at $\mu = 0.095 = \log(1.1)$, and the sample size ranges from $n = 100$ to over 10,000 depending on the combination of α and power. The values of the these parameters then uniquely determine the corresponding population variance, σ^2. The details of each simulation scenario are shown in Table 1.

Under each simulation scenario, we began by generating 200 significant data sets, that is, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, such that the value of the test statistic, $T_n = \frac{\overline{X}}{S/\sqrt{n}}$, is greater than c. We then computed the nine estimates, **N**, **MLE**, **NMLE**, **Ghosh**, **B.L**, **B.H**, **B.BMA**, **B.M** and **B.Unif**, for each significant data set.

Figure 2 provides detailed results when the type 1 error rate is 0.05 and the simulating parameter values are those in row 1 of Table 1. These plots confirm that, in the case of low power of the initial association study (e.g., 10%), the naïve estimator has a large upward bias. Even in the moderately powered studies (e.g., 20%), the naïve estimator could considerably overestimate the true effect size. Note that the two priors with opposite degrees of belief in the significance of the effect, **B.L** and **B.H**, produce quite different results. The **B.L** estimator conservatively shrinks the effect and, therefore, it is more reliable in those cases when the effect is small or zero. (See additional figures in Supplement for the case of no genetic effect, i.e., the apparent association is a false positive.) When the power of the test is relatively high (e.g., 50%), **B.H** outperforms the other estimators considered. While it is clear that **B.L** and **B.H** are complementing each other, **B.BMA**, designed to balance between **B.L** and **B.H**, performs well in a variety of settings. The performances of the other two estimators, **B.M** and **B.Unif**, are similar to one another but inferior to **B.BMA**. The natural implication is that putting equal prior weight on (0, 1) is equivalent to putting equal weight on ξ close to zero or close to 1. As expected, when the power is very high (e.g., 99%) there is little bias in the naïve estimate; the other estimates also converge to the true value with **B.L** lagging behind. This is due to the strong skepticism embedded in the **B.L** model about the finding.

In most of the cases, the Bayesian estimators achieve the anticipated reduction in bias as well as variance compared to the likelihood based estimators, **MLE**, **NMLE** and **Ghosh**. Of the three, we observed that **Ghosh** (i.e., the average of **MLE** and **NMLE**) performs the best, confirming the conclusion of Ghosh, Zou and

TABLE 1
Simulation scenarios for the normal model

	5%			**10%**			**20%**			**50%**			**99%**		
$\alpha\backslash$power	n	σ	$\sigma/\sqrt{n}$	n	σ	$\sigma/\sqrt{n}$	n	σ	$\sigma/\sqrt{n}$	n	σ	$\sigma/\sqrt{n}$	n	σ	$\sigma/\sqrt{n}$
0.05	–	–	–	100	2.623	0.262	200	1.678	0.119	1000	1.832	0.058	5000	1.697	0.024
10^{-4}	1000	1.453	0.046	2000	1.749	0.039	3000	1.814	0.033	5000	1.812	0.026	10,000	1.577	0.016
10^{-6}	2000	1.371	0.031	4000	1.736	0.027	5000	1.723	0.024	8000	1.793	0.020	16,000	1.702	0.013

Notes: Sample size (n) and population standard error (σ) needed to obtain the desired power at the prespecified type 1 error rate (α) when population mean $\mu = 0.0953 = \log(1.1)$.

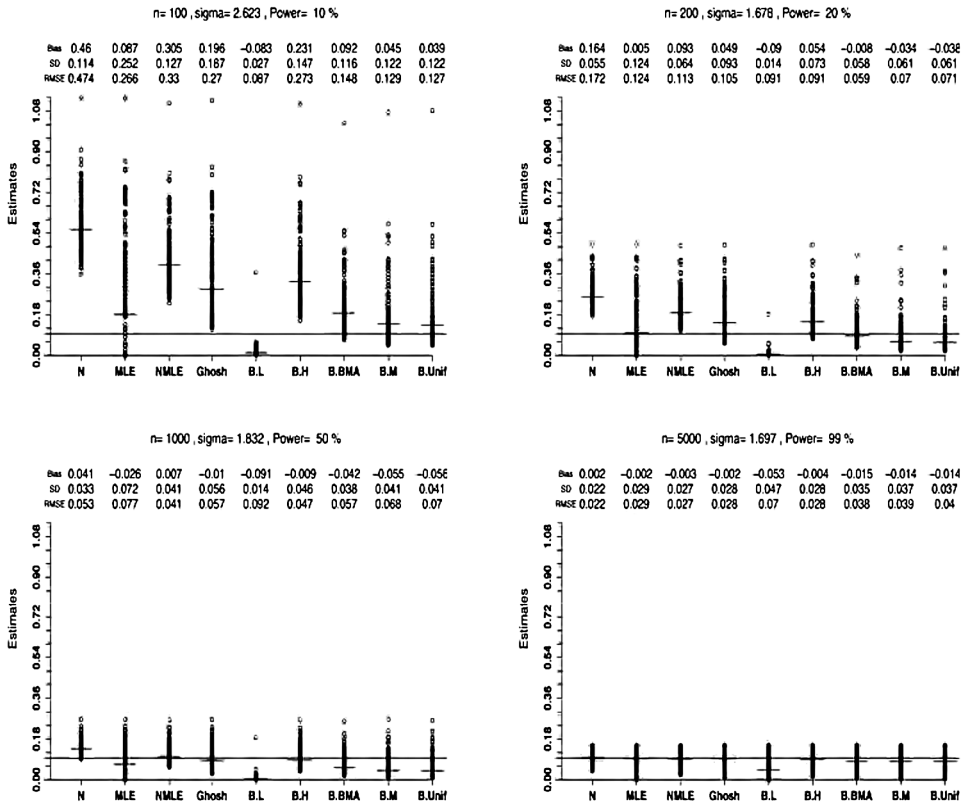

FIG. 2. *Performance of the nine estimators under the normal model with a type* 1 *error rate of* 0.05. *The population mean* $\mu = \log(1.1) = 0.0953$ *and power ranging from* 10%, 20%, 50% *to* 99%. *Details of the simulating parameters are given in row* 1 *of Table* 1. *Each circle represents an estimate, the horizontal is the averaged estimate over* 200 *simulated data sets, and the long horizontal line represents the true value of* μ. *The Bias, sample Standard Deviation* (*SD*) *and Root Mean Squared Error* (*RMSE*) *are also provided for each estimator.*

Wright (2008). Therefore, in what follows we focus on the comparison between **B.BMA** and **Ghosh**.

The advantage of **B.BMA** over **Ghosh** is especially obvious in the low power studies. For example, when the power of the test is 10%, the bias of **Ghosh** is 0.196, almost twice as big as 0.092 for **B.BMA**. The sample standard deviation of the **Ghosh** estimate is 0.186 compared to 0.116 for the **B.BMA** estimate. The Root Mean Squared Error (RMSE) for **B.BMA** is almost half that for **Ghosh** (0.148 vs. 0.273). To formally assess the significance of the difference between **Ghosh** and **B.BMA**, we performed a matched-pair t-test based on 50 simulation runs, and we obtained a t-statistic of -117.47 showing that the difference is significant. As expected, the advantage dissipates and the two perform similarly when the power of the initial association study increases.

As discussed by Ghosh, Zou and Wright (2008) and detailed in Section 2, the main factor that influences the estimation bias is the power of the association test which depends on the noncentrality parameter, $\mu/(\sigma/\sqrt{n})$. Thus, although μ has the interpretation of $\beta = \log \mathrm{OR}$ and was fixed at $\log(1.1)$, the results are qualitatively similar for larger OR with smaller sample size or smaller OR with larger sample size, as long as the ratio, $\mu/(\sigma/\sqrt{n})$, and the significance threshold value, α, stay the same.

Figure 3 shows the performance of the estimators when the type 1 error rate is 10^{-6} and the parameter values are from row 3 of Table 1. We found that all the bias correction estimators are showing a slight overcorrection. (Note that the scale in the y-axis differs between Figures 2 and 3.) In this setting, the results of **B.BMA** and **Ghosh** are very similar with **B.BMA** having a smaller variance. The difference between Figures 2 and 3 is due to the fact that the significance threshold used is drastically different, $\alpha = 0.05$ for Figure 2 and $\alpha = 10^{-6}$ for Figure 3, while the power of the association study of the same SNP is kept comparable by increasing the required sample size, n. As a result, the noncentrality parameter values, $\mu/(\sigma/\sqrt{n})$, are not directly comparable between the two cases.

5.2. *Simulation set* 2*—genetic model*. Following the setup of the simulations conducted by Ghosh, Zou and Wright (2008), we generated data for 500 cases and 500 controls from an additive genetic model with disease prevalence of 1%, minor allele frequency of 0.25, and the log OR, β, ranging from $\log(1.1)$ to $\log(2)$. The threshold value is $c = 5.0$, leading to the significance level $\alpha = 2.87 \times 10^{-7}$. For each log OR value, we began by generating 200 significant data sets such that the association test statistic, $\hat{\beta}/\widehat{SE}(\hat{\beta})$, is greater than c, where $\hat{\beta}$ is the log OR estimate obtained from the logistic regression model, and $\widehat{SE}(\hat{\beta})$ is the estimate of the standard error of $\hat{\beta}$. Using the summary statistics, $\hat{\beta}$ and $\widehat{SE}(\hat{\beta})$, the auxiliary information such as the sample size (we used $n = 1000$) and the threshold value of the test, we applied the Bayesian methods by letting $\hat{\mu} = \hat{\beta}$, and $S = \hat{\sigma} = \widehat{SE}(\hat{\beta}) \times \sqrt{n}$.

Figure 4 illustrates the results for log OR values equal to $\{\log(1.2), \log(1.3), \log(1.4), \log(1.8)\}$, corresponding to the power of detecting the associated SNP in the range $\{0.345\%, 4.515\%, 21.897\%, 99.5\%\}$. (Results for other log OR values are qualitatively similar.) The results obtained from the simulated genetic models confirm that the **B.BMA** has a smaller RMSE than **Ghosh** when the power of the association test is low. Although the variance reduction on the log OR scale is small, the implication on study design is practically important. Figure 5 shows the sample size estimation for a replication study with 80% power at the 0.05 significance level using the naïve log OR estimate, the **Ghosh** estimate and the **B.BMA** estimate obtained from the original discovery samples, as reported in Figure 4. Results show that the standard error in sample size estimation based on **Ghosh** is almost twice as big as that based on **B.BMA** when the power of the original association study is low (e.g., 20% or lower). In the low power case, we also note

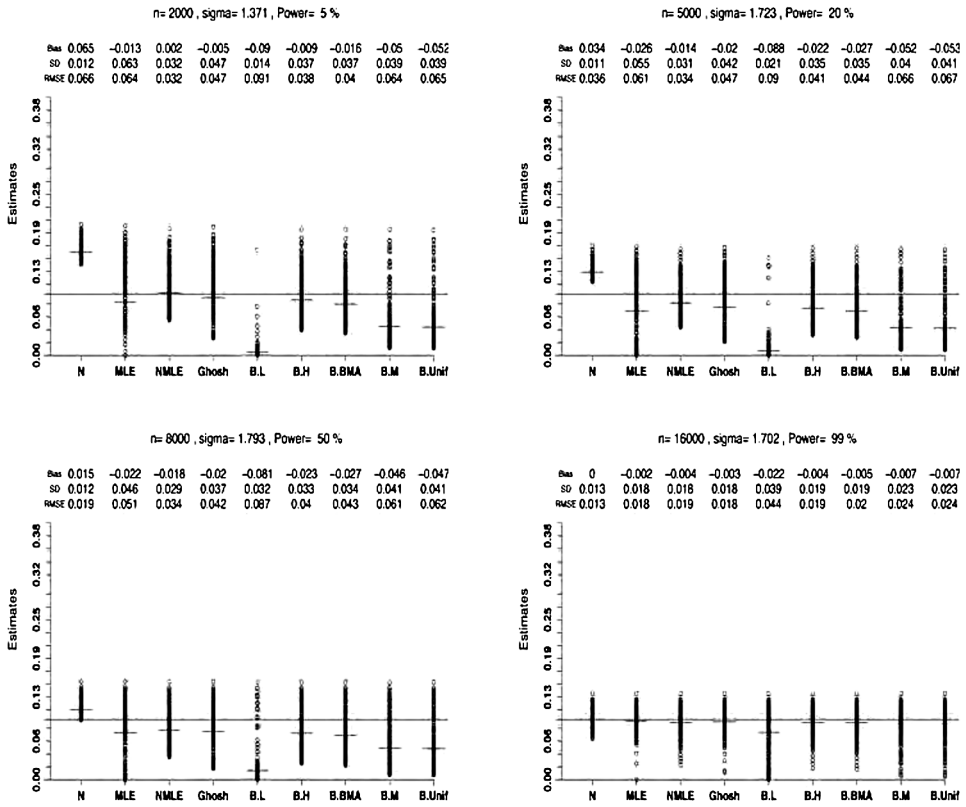

FIG. 3. *Performance of the nine estimators under the normal model with a type* 1 *error rate of* 10^{-6}. *The population mean* $\mu = \log(1.1) = 0.0953$ *and power ranging from* 5%, 20%, 50% *to* 99%. *Details of the simulating parameters are given in row* 3 *of Table* 1. *Each circle represents an estimate, the horizontal bar is the averaged estimate over* 200 *simulated data sets, and the long horizontal line represents the true value of* μ. *The Bias, sample Standard Deviation* (*SD*) *and Root Mean Squared Error* (*RMSE*) *are also provided for each estimator.*

that the sample size predicted based on **N**, the naïve estimate, is never sufficient. For example, for a SNP with log(OR) of log(1.2), the naïve sample size estimate centers around 222 with a maximum predicted size of 247, while the true expected required sample size is 1170. Although both **Ghosh** and **B.BMA** overestimate the necessary sample size for replication due to the overcorrection of effect size, we believe that a conservative sample size estimate is practically useful because it guards against sampling variation.

We also examined different effect levels when the type I error level is equal to 0.05 or 0.001, and we drew similar conclusions based on the results reported in Supplement. The additional simulation studies also include a null case where the apparent discovery is a false positive. In that case, **B.BMA** outperforms **Ghosh**, but **B.L** performs the best, as expected.

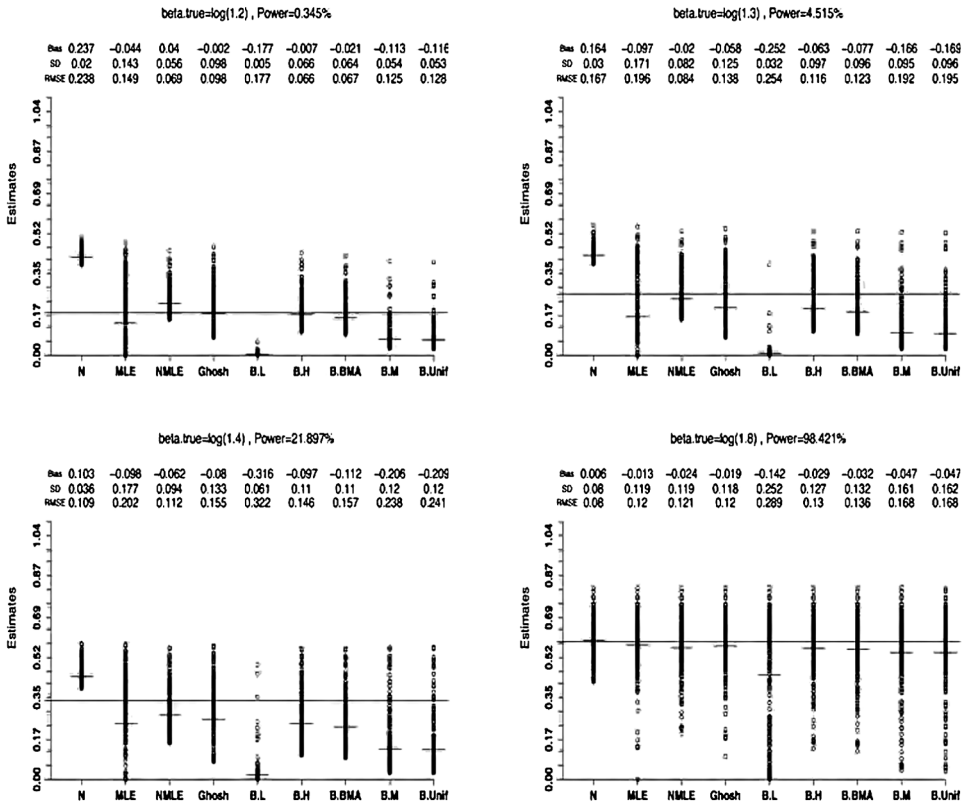

FIG. 4. *Performance of the nine estimators under an additive genetic model with a type* 1 *error rate of* $\alpha = 2.87 \times 10^{-7}(c = 5)$. *The sample size is* 1000 (500 *cases and* 500 *controls*), *the minor allele frequency of the causal SNP is* 0.25. *The effect of the SNP on the* log *OR scale ranging from* $\mu = \beta = \log(1.2), \log(1.3), \log(1.4)$ *to* $\log(1.8)$ *corresponding to power* $<1\%$, $\approx 5\%$, $\approx 20\%$ *and* $>95\%$ *to detect the association. Each circle represents an estimate, the horizontal bar is the average estimate over* 200 *simulated data sets, and the long horizontal line represents the true value of* μ. *The Bias, sample Standard Deviation* (*SD*) *and Root Mean Squared Error* (*RMSE*) *are also provided for each estimator.*

6. Application study. We applied the proposed Bayesian estimation methods to four data sets of which one is a candidate gene study and the other three are genome-wide association studies (GWAS) of either binary or quantitative outcomes. Specifically, the four studies are as follows:

(I) the candidate gene association study of Lymphoma by Wang et al. (2006),
(II) the GWAS of type 1 diabetes (T1D) by WTCCC (2007),
(III) the GWAS of psoriasis by Nair, Duffin and Helms (2009),
(IV) the GWAS of complications of T1D by Paterson et al. (2010).

The Lymphoma and WTCCC T1D data sets were chosen because they were previously analyzed by Ghosh, Zou and Wright (2008) via the likelihood-based ap-

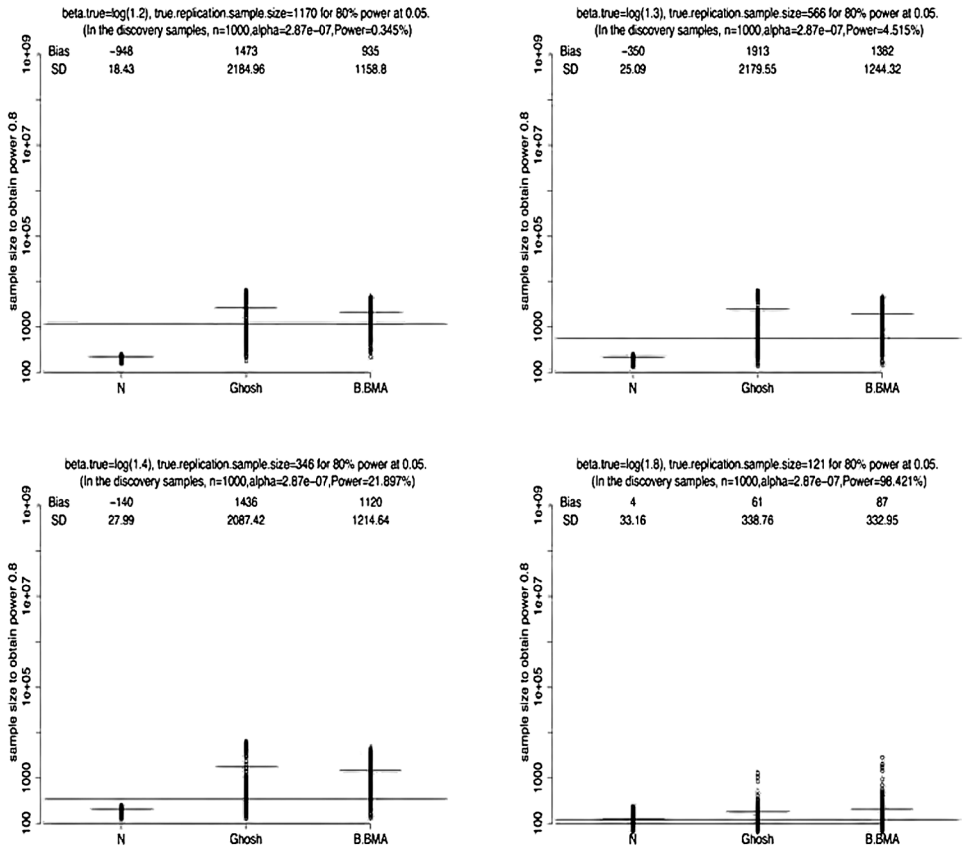

FIG. 5. *Performance of sample size estimation for replication studies under an additive genetic model. The initial discovery samples are the same as those in Figure* 4. *The replication sample size is calculated assuming a type* 1 *error rate of* 0.05 *and power of* 80%, *and it is calculated based on the estimate of the log OR by* ***N***, *the naïve estimation method*, ***Ghosh***, *the likelihood method, or* ***B.BMA***, *the Bayesian method applied to the simulated significant discovery samples. Each circle represents an estimate, the horizontal bar is the average estimate over* 200 *simulated data sets, and the long horizontal line represents the true expected required sample size.*

proach, and the other two studies were chosen because the genetic effect estimates from independent replication samples were reported by the study authors. In addition, the T1D complication data set allows us to demonstrate that the proposed methods can be easily and robustly applied to association studies of quantitative outcomes.

In each case, the results are summarized in a table containing the original reported genetic effect (i.e., the naïve estimate, **N**), the five different Bayesian estimators, **B.L**, **B.H**, **B.BMA**, **B.Unif** and **B.M**, and three likelihood methods, **MLE**, **NMLE** and **Ghosh**, as described in Section 5. The estimates produced by each method are compared with the estimates obtained from the independent replica-

tion samples reported in the literature. We note that the anticipated power for each study differs due to the apparent differences in study design [e.g., higher power for the candidate gene study of Wang et al. (2006) compared to the GWAS], the sample size [e.g., higher power for the GWAS of T1D by WTCCC (2007) with $n \approx 5000$ compared to the GWAS of T1D complication by Paterson et al. (2010) with $n = 667$], and the prior knowledge of a SNP (e.g., higher power for rs12191877 from chromosome 6 in the well-known MHC region that is strongly associated with Psoriasis compared to other novel SNPs). However, we report estimates from all five Bayesian estimators for a more complete comparison. The estimate from the replication samples serves as the benchmark, but the value itself should not be viewed as the true parameter value because of the sampling variation and the potential subpopulation and ascertainment differences between the original discovery and the follow-up replication studies.

We also report the corresponding confidence interval (CI) or the highest posterior density region/interval (HpdI), but it should be noted that the statistical interpretations of CI and HpdI are different and, therefore, these regions are not directly comparable. Although the HpdI with posterior mass $1 - \eta$ may be estimated using samples from the posterior under model M_1 for **B.L** or M_2 for **B.H**, there is no direct way to construct a HPD region for **B.BMA**, the model averaging estimator for the two models. However, a credible interval (CrdI) can be constructed using the normal approximation based on the model averaging estimator and its variance estimate [see equation (7) in Viallefont, Raftery and Richardson (2001)]. For the likelihood-based methods, we construct the CI following the method proposed by Ghosh, Zou and Wright (2008) that was shown to outperform the standard CI procedure. Specifically, the Ghosh $1 - \eta$ CI is the interval between the $\eta/2$ and $1 - \eta/2$ quantiles of the conditional density $p(T_n | T_n > c)$. Ghosh, Zou and Wright (2008) noted that, although they proposed three competing point estimates, **MLE**, **NMLE** and **Ghosh**, their procedure provided only a single CI.

6.1. *Application I—A candidate-gene study of lymphoma.* Wang et al. (2006) performed a candidate gene study of Lymphoma using a total of 48 SNPs genotyped on 318 cases and 766 controls, and they reported two significant SNPs using a p-value threshold of $\alpha = 0.002$. The naïve log OR estimate is log(1.54) for rs1800629 and log(1.40) for rs909253, however, the follow-up estimates obtained from a larger independent study are reduced considerably to log(1.29) for rs1800629 and log(1.16) for rs909253 [Rothman et al. (2006); Ghosh, Zou and Wright (2008)]. For each of the two SNPs, we applied the likelihood estimation methods as well as the Bayesian methods, using the naïve log OR estimates, $\hat{\mu} = \hat{\beta}$, and $S = \hat{\sigma} = \widehat{SE}(\hat{\beta}) \times \sqrt{n}$ inferred from the observed association p-value [p-value $= 1 - \Phi(|\hat{\beta}/\widehat{SE}(\hat{\beta})|)$], $n = 318 + 766 = 1084$ and $c = 2.878$ corresponding to $\alpha = 0.002$ (Table 2).

Results in Table 2 are consistent with simulation results of power 50% in Figure 2. Because of the anticipated high power of a candidate gene study, both

TABLE 2
Application I—the candidate gene study of Lymphoma by Wang et al. (2006)

SNPs of interest	**rs1800629**	**rs909253**
Discovery samples		
Association p-value	5.7×10^{-4}	7.4×10^{-4}
Reported effect	0.432	0.337
Likelihood estimates		
MLE (CI)	0.116 (0.000, 0.645)	0.010 (0.000, 0.498)
NMLE (CI)	0.247 (0.000, 0.645)	0.184 (0.000, 0.498)
Ghosh (CI)	0.182 (0.000, 0.645)	0.097 (0.000, 0.498)
Bayesian estimates		
B.L (HpdI)	0.005 (0.000, 0.013)	0.004 (0.000, 0.005)
B.H (HpdI)	0.196 (0.000, 0.508)	0.142 (0.000, 0.382)
B.BMA (CrdI)	0.150 (0.000, 0.428)	0.115 (0.000, 0.324)
B.Unif (HpdI)	0.068 (0.000, 0.377)	0.045 (0.000, 0.277)
B.M (HpdI)	0.074 (0.000, 0.397)	0.049 (0.000, 0.281)
Follow-up samples		
Follow-up estimate	0.255	0.148

Notes: The Reported Effect is naïve log OR estimate obtained from the original discovery samples (318 cases and 766 controls) of Wang et al. (2006), in which the association tests of these two SNPs were significant at the $\alpha = 0.002$ level. The follow-up estimate was obtained from a larger pooled analysis by Rothman et al. (2006). The other eight estimates were based on either the likelihood approach, **MLE**, **NMLE** and **Ghosh**, or the proposed Bayesian approach, **B.L**, **B.H**, **B.BMA**, **B.Unif** and **B.M** as summarized in Section 5. CI is the 95% confidence interval for the likelihood estimates, HpdI is the highest posterior density interval with posterior mass 95% and CrdI is the credible interval for the Bayesian estimates.

B.BMA and **Ghosh** overcorrect slightly with similar performance. We observe that the CrdI of **B.BMA** is smaller than the CI of **Ghosh**, although we noted before that the interpretation of the two intervals is different. Results suggest that **B.H** performs best among all the Bayesian methods, which is not surprising for a study with putative high power.

6.2. *Application II—A GWAS of Type* 1 *Diabetes*. The Type 1 Diabetes (T1D) GWAS from the WTCCC included approximatively 2000 cases and 3000 controls and the samples were genotyped on the Affymetrix 500K chip3 [WTCCC (2007)]. After a set of quality control criterions (e.g., the minor allele frequency of a SNP $> 5\%$, the genotyping missing rate $< 5\%$ and the p-value of the Hardy–Weinberg Equilibrium test $> 5.7 \times 10^{-7}$), the authors reported six significant loci at the 5×10^{-7} level. We focused on the four SNPs analyzed by Ghosh, Zou and Wright (2008) because the replication results are available from the study of Todd et al. (2007). For each SNP of interest, we applied the proposed estimation methods using the reported log OR estimates obtained from the WTCCC discovery sam-

ples, $\hat{\beta} = \hat{\mu}$, and $S = \hat{\sigma} = \widehat{SE}(\hat{\beta}) \times \sqrt{n}$ inferred from the observed association p-value, and $c = 4.892$ corresponding to $\alpha = 5 \times 10^{-7}$ (Table 3). In this application, the actual number of cases is $1963 - 37 = 1926$ and the number of controls is $(1480 - 24) + (1458 - 42) = 2872$, where the 37, 24 and 42 samples were deleted due to quality control issues, based on the information provided in the supplementary Tables 1 and 4 of WTCCC (2007). Thus, $n = 1926 + 2872 = 4798$ in this application.

Results in Table 3 show that if the original association result is extreme in that the p-value is considerably smaller than the threshold considered (i.e., rs17696736), then the prior influences the result only minimally. Similarly, the likelihood-based estimates are only slightly reduced from the published estimated log ORs. However, the follow-up estimate is considerably lower than the bias reduced estimates. As noted by Ghosh, Zou and Wright (2008), this suggests possible heterogeneity between the discovery and replication samples. A subtle but important explanation for the results in the last three columns of Table 3 where the replicated values are larger in absolute value than the estimates produced by each method is that the follow-up estimates here are also subject to the winner's curse, albeit less severe, because only estimates of successfully replicated SNPs were reported.

6.3. *Application III—A GWAS of Psoriasis*. Nair, Duffin and Helms (2009) conducted a two-stage association of Psoriasis, a chronic skin disease characterized by circumscribed red patches covered with white scales. The first stage is a GWAS with 438,670 SNPs genotyped on 1359 cases and 1400 controls, and the second stage is a replication study following up on 21 promising SNPs using a set of independent 5048 cases and 5051 controls. *"Owing to the winner's curse, odds ratios estimated in the discovery sample were larger than those estimated in the follow-up samples"* [Table 2 of Nair, Duffin and Helms (2009)]. The SNP selection criterion was mainly based on the ranking of the GWAS p-value, roughly corresponding to a p-value threshold of $\alpha = 10^{-4}$. For each SNP of interest, we applied the estimation methods using the reported log OR estimates obtained from the discovery samples, $\hat{\beta} = \hat{\mu}$, and $S = \hat{\sigma} = \widehat{SE}(\hat{\beta}) \times \sqrt{n}$ inferred from the observed association p-value, $n = 1359 + 1400 = 2759$ and $c = 3.719$ corresponding to $\alpha = 10^{-4}$ (Table 4).

When the results are as extreme as rs12191877 with $p = 4 \times 10^{-53}$ or as rs2082412 with $p = 5 \times 10^{-10}$, indicating high power at the chosen threshold level, all the bias correction estimators results in little change from the published estimate, including **B.L** despite its inherent prior skepticism of a finding. For the other less significant SNPs in the table, both **B.BMA** and **Ghosh** achieve substantial bias reduction. In general, **B.BMA** has a noticeably smaller variance for lower power cases, which in turn can produce more reliable sample size estimates for replication studies.

TABLE 3
Application II—the GWAS of T1D by WTCCC (2007)

SNPs of interest	**rs17696736**	**rs2292239**	**rs12708716**	**rs2542151**
Discovery samples				
Association p-value	7.27×10^{-14}	1.49×10^{-9}	1.28×10^{-8}	8.4×10^{-8}
Reported effect (CI)	0.315 (0.239, 0.399)	0.262 (0.182, 0.351)	−0.261 (−0.357, −0.174)	0.285 (0.182, 0.399)
Likelihood estimates				
MLE (CI)	0.314 (0.224, 0.397)	0.241 (0.095, 0.346)	−0.212 (−0.348, 0.000)	0.140 (0.000, 0.375)
NMLE (CI)	0.310 (0.224, 0.397)	0.217 (0.095, 0.346)	−0.182 (−0.348, 0.000)	0.154 (0.000, 0.375)
Ghosh (CI)	0.312 (0.224, 0.397)	0.229 (0.095, 0.346)	−0.197 (−0.348, 0.000)	0.147 (0.000, 0.375)
Bayesian estimates				
B.L (HpdI)	0.311 (0.221, 0.399)	0.019 (0.000, 0.210)	−0.006 (−0.008, 0.000)	0.004 (0.000, 0.010)
B.H (HpdI)	0.309 (0.221, 0.403)	0.212 (0.063, 0.345)	−0.170 (−0.306, 0.000)	0.126 (0.000, 0.294)
B.BMA (CrdI)	0.309 (0.234, 0.385)	0.207 (0.079, 0.336)	−0.161 (−0.318, −0.004)	0.117 (0.000, 0.280)
B.Unif (HpdI)	0.311 (0.220, 0.398)	0.172 (0.000, 0.312)	−0.087 (−0.283, 0.000)	0.045 (0.000, 0.240)
B.M (HpdI)	0.309 (0.211, 0.391)	0.173 (0.000, 0.310)	−0.092 (−0.286, 0.000)	0.046 (0.000, 0.249)
Follow-up samples				
Follow-up estimate (CI)	0.148 (0.086, 0.207)	0.247 (0.182, 0.308)	−0.186 (−0.248, −0.116)	0.254 (0.174, 0.337)

Notes: The reported effect is naïve log OR estimate obtained from the original discovery samples (1926 cases and 2872 controls) of WTCCC (2007), in which the association tests of these SNPs were significant at the $\alpha = 5 \times 10^{-7}$ level. The Follow-up Estimate was obtained from the replication study by Todd et al. (2007). The other eight estimates were based on either the likelihood approach, **MLE**, **NMLE** and **Ghosh**, or the proposed Bayesian approach, **B.L**, **B.H**, **B.BMA**, **B.Unif** and **B.M** as summarized in Section 5. CI is the 95% confidence interval for the likelihood estimates, HpdI is the highest posterior density interval with posterior mass 95% and CrdI is the credible interval for the Bayesian estimates.

TABLE 4
Application III—the GWAS of Psoriasis by Nair, Duffin and Helms (2009)

SNPs of interest	**rs12191877**	**rs2082412**	**rs17728338**	**rs20541**	**rs610604**
Discovery samples					
p-value	4×10^{-53}	5×10^{-10}	2×10^{-7}	6×10^{-6}	1×10^{-5}
Reported effect	1.026	0.445	0.542	0.315	0.247
Likelihood estimate					
MLE (CI)	1.026 (0.895, 1.157)	0.443 (0.287, 0.585)	0.514 (0.214, 0.746)	0.234 (0.000, 0.445)	0.162 (0.000, 0.349)
NMLE (CI)	1.026 (0.895, 1.157)	0.435 (0.287, 0.585)	0.476 (0.214, 0.746)	0.210 (0.000, 0.445)	0.154 (0.000, 0.349)
Ghosh (CI)	1.026 (0.895, 1.157)	0.439 (0.287, 0.585)	0.495 (0.214, 0.746)	0.222 (0.000, 0.445)	0.158 (0.000, 0.349)
Bayesian estimate					
B.L (hpdI)	1.026 (0.887, 1.153)	0.400 (0.000, 0.556)	0.049 (0.000, 0.494)	0.007 (0.000, 0.010)	0.005 (0.000, 0.009)
B.H (hpdI)	1.024 (0.891, 1.150)	0.436 (0.276, 0.587)	0.468 (0.170, 0.754)	0.197 (0.000, 0.377)	0.136 (0.000, 0.288)
B.BMA (CrdI)	1.024 (0.915, 1.132)	0.436 (0.304, 0.568)	0.444 (0.151, 0.738)	0.172 (0.000, 0.379)	0.122 (0.000, 0.279)
B.Unif (hpdI)	1.026 (0.898, 1.163)	0.437 (0.283, 0.592)	0.405 (0.000, 0.681)	0.094 (0.000, 0.339)	0.062 (0.000, 0.252)
B.M (hpdI)	1.026 (0.887, 1.146)	0.436 (0.268, 0.580)	0.402 (0.000, 0.687)	0.096 (0.000, 0.341)	0.063 (0.000, 0.253)
Follow-up samples					
Follow-up estimate	0.971	0.365	0.464	0.239	0.174

TABLE 4
(*Continued*)

SNPs of interest	**rs2066808**	**rs2201841**	**rs1076160**	**rs12983316**
Discovery samples				
Association p-value	2×10^{-5}	3×10^{-7}	2×10^{-5}	2×10^{-5}
Reported effect	0.519	0.300	0.231	0.308
Likelihood estimates				
MLE (CI)	0.231 (0.000, 0.728)	0.281 (0.107, 0.414)	0.103 (0.000, 0.324)	0.137 (0.000, 0.432)
NMLE (CI)	0.293 (0.000, 0.728)	0.258 (0.107, 0.414)	0.129 (0.000, 0.324)	0.173 (0.000, 0.432)
Gho0sh (CI)	0.262 (0.000, 0.728)	0.270 (0.107, 0.414)	0.116 (0.000, 0.324)	0.155 (0.000, 0.432)
Bayesian estimates				
B.L (HpdI)	0.008 (0.000, 0.011)	0.021 (0.000, 0.228)	0.003 (0.000, 0.005)	0.004 (0.000, 0.010)
B.H (HpdI)	0.247 (0.000, 0.571)	0.253 (0.076, 0.422)	0.110 (0.000, 0.257)	0.147 (0.000, 0.340)
B.BMA (CrdI)	0.221 (0.000, 0.54)	0.240 (0.074, 0.407)	0.097 (0.000, 0.239)	0.127 (0.000, 0.316)
B.Unif (HpdI)	0.097 (0.000, 0.472)	0.207 (0.000, 0.381)	0.042 (0.000, 0.209)	0.056 (0.000, 0.275)
B.M (HpdI)	0.099 (0.000, 0.482)	0.210 (0.000, 0.376)	0.044 (0.000, 0.213)	0.057 (0.000, 0.273)
Follow-up samples				
Follow-up estimate	0.293	0.122	0.086	0.086

Notes: The reported effect is naïve log OR estimate obtained from the original discovery samples (1359 cases and 1400 controls) of Nair, Duffin and Helms (2009), in which these SNPs were among the top 2000 SNPs based on the p-values of the association tests, corresponding to $\alpha = 10^{-4}$ level. The Follow-up estimate was obtained from the replication study by Nair, Duffin and Helms (2009). The other eight estimates were based on either the likelihood approach, **MLE**, **NMLE** and **Ghosh**, or the proposed Bayesian approach, **B.L**, **B.H**, **B.BMA**, **B.Unif** and **B.M** as summarized in Section 5. CI is the 95% confidence interval for the likelihood estimates, HpdI is the highest posterior density interval with posterior mass 95% and CrdI is the credible interval for the Bayesian estimates.

6.4. *Application IV—A GWAS of quantitative measures of T1D complications*. In the fourth setting of the GWA study of longitudinal repeated quantitative measures of phenotype HbA1c in the Diabetes Control and Complications Trial (DCCT) samples, a significant locus (at $\alpha = 5 \times 10^{-8}$) was identified in the conventional treatment group with 667 samples near SORCS1 (rs1358030 with p-value $= 4.66 \times 10^{-9}$). The association statistic was obtained via regression analysis of the average log (HbA1c) value vs. SNP with an additive genotype coding. The GWAS was performed on 841,342 SNPs, genotyped by the Illumina 1M BeadArray assay, that passed a set of quality control criteria [details in Paterson et al. (2010)].

The naïve estimate of the regression coefficient for rs1358030 is 0.045. However, the estimate obtained from the intensive treatment group with 637 samples is 0.005 (Table 5). Note that for the intensive treatment group, only the measures at

TABLE 5
Application IV—the GWAS of HbA1c in Type 1 Diabetes patients, by Paterson et al. (2010)

SNP of interest	**rs1358030**
Discovery samples	
Association p-value	4.66×10^{-9}
Reported effect	0.045
Likelihood estimates	
MLE (CI)	0.029 (0.000, 0.056)
NMLE (CI)	0.024 (0.000, 0.056)
Ghosh (CI)	0.027 (0.000, 0.056)
Bayesian estimates	
B.L (HpdI)	0.001 (0.000, 0.002)
B.H (HpdI)	0.021 (0.000, 0.048)
B.BMA (CrdI)	0.020 (0.000, 0.047)
B.Unif (HpdI)	0.007 (0.000, 0.040)
B.M (HpdI)	0.008 (0.000, 0.040)
Follow-up samples	
Follow-up estimate	0.005

Notes: The reported effect is the naïve estimate of the regression coefficient obtained from the 667 discovery samples, in which the association test of the SNP was significant at the $\alpha = 5 \times 10^{-8}$ level. The Follow-up estimate was obtained from 637 independent samples. The other eight estimates were based on either the likelihood approach, **MLE**, **NMLE** and **Ghosh**, or the proposed Bayesian approach, **B.L**, **B.H**, **B.BMA**, **B.Unif** and **B.M** as summarized in Section 5. CI is the 95% confidence interval for the likelihood estimates, HpdI is the highest posterior density interval with posterior mass 95% and CrdI is the credible interval for the Bayesian estimates.

the eligibility time-point (i.e., before the starting of the two different treatments) were used for the regression analysis so that the two groups are comparable and the intensive treatment group could be used as a replication data set.

Unlike the case control studies with binary response (diseased or not) considered previously, of interest here is a quantitative outcome, HbA1c, that measures the amount of glycated hemoglobin in blood. Therefore, the μ no longer represents the log OR but the corresponding coefficient in the linear regression model. Although we could consider choosing a more suitable prior, we adopted the same Uniform$(0, 2)$ density for $f(\mu)$ as for the case-control data to test the robustness of the Bayesian methods. (Results from other prior choices are discussed in Section 7.) To apply the Bayesian methods, we let $\hat{\mu} = 0.045$, $n = 667$, $c = 5.328$ (corresponding to the threshold used, the significance level is $\alpha = 5 \times 10^{-8}$), and the observed association p-value 4.66×10^{-9} (corresponding to a test statistic of 5.743) allows us to infer the standard error $S = \mu * \sqrt{n}/5.743 = 0.202$ (Table 5). As expected for the low power case, both **B.BMA** and **Ghosh** reduce the estimation bias but not sufficiently enough, and **B.L** performs better. However, in this case the estimates from **B.Unif** or **B.M** are closest to the one obtained from the follow-up study.

7. Conclusions and future work. We propose hierarchical Bayes methods to reduce selection bias in genetic association studies. The basis of the approach is a spike-and-slab prior which essentially allows for the possibility that the signal detected may be a false positive. The prior permits the researchers to quantify their belief in the strength of the signal. Depending on the prior, inference based on the posterior distribution may be different from model to model and, therefore, the researcher faces a (sometimes difficult) choice. To alleviate this dilemma, we consider a Bayesian model averaging strategy, **B.BMA**, in which we use the data to weigh in on the more appropriate model.

Simulation and application studies demonstrated that the **B.BMA** estimator performs well across different settings, and we recommend **B.BMA** when there is little information on the putative power of the initial discovery study. However, we also emphasize that model averaging is not necessarily the best approach for a given study. Factors such as study design and sample size should be taken into account in the decision of using a more conservative model like **B.L** or an anticonservative one like **B.H**. In general, **B.H** is suitable for candidate gene studies with putative high power as demonstrated in application I, and **B.L** is preferred for GWAS with putative low power as shown in application IV. Knowledge about the SNP of interest is also a factor. For example, little bias is expected for a SNP in a well-known associated region or with p-value significantly smaller than the chosen threshold as demonstrated by the first SNP (rs12191877) in Table 4 of application III, while substantial bias is expected for a SNP with p-value just below the threshold as shown by the last SNP (rs12983316) in the table.

We have carried out additional simulation studies to investigate the robustness of the Bayesian estimators. Results provided in Supplement show that the proposed methods are robust to the choice of prior for ξ, the hyperparameter that reflects our prior belief in false positive, to the number of iterations discarded from the MCMC sample, and to the value of A, the prior upper bound of log odds ratio. In addition, we developed our methods using a conceptual normal model but demonstrated via simulations and applications that this normal model is well connected with widely used real genetic models and is robust to the choice of priors. For example, in application IV when the phenotype is not a case-control status but a quantitative outcome, we kept the same $A = 2$ knowing that the the upper bound for μ, the genetic effect size, in this case can be reasonably assumed to be 0.2. To be more precise, note that μ is a regression coefficient in this setup and is related to the percentage of phenotype variation explained by the SNP via the expression

$$r^2 = \mu^2 \frac{S_X^2}{S_Y^2},$$

where $S_X^2 \approx 0.467$ is the sample variance of the SNP and $S_Y^2 \approx 0.018$ is the sample variance of the phenotype. Since $r^2 \leq 100\%$, thus, $\mu \leq 0.2$. When $A = 0.2$ was assumed, the estimates were largely unchanged compared to results in Table 5: 0.00062 (0, 0.001) for **B.L**, 0.021 (0.000, 0.0474) for **B.H**, 0.0197 (0, 0.0456) for **B.BMA**, 0.0077 (0.000, 0.03996) for **B.Unif** and 0.0084 (0.000, 0.0407) for **B.M**. If a true effect is greater than 2, our Bayesian estimations will be bounded by 2. In practice, if the true OR is greater than $\exp(2) \approx 7.4$, then the putative power of the original association study is very high (unless the sample size is extremely small), resulting in little estimation bias of the naïve estimate. Second, if a Bayesian estimate was close to the upper bound, then one can choose a bigger value such as 6. This modification does not affect the estimation for the cases when the effects are less than 2 (confirmed by our additional simulation studies) but provide better effect estimates when the true effects are indeed greater than 2. The proposed Bayesian methods, however, are not robust to the misspecification of the threshold used. This type of sensitivity was also observed for other existing methods including the likelihood and resampling based methods.

The **NMLE** estimator proposed by Ghosh, Zou and Wright (2008) is the mean of the normalized conditional likelihood, and it can be interpreted as the posterior mean with an improper flat prior on μ which should produce similar results to **B.Unif**. However, unlike **NMLE**, our model allows a point mass on effect being equal to 0 via the spike-and-slab prior, leading to a better performance than **NMLE**. As an average of the conditional **MLE** and the **NMLE** estimators, the **Ghosh** estimator strikes a balance between the two and performs better than both across different settings. Although **Ghosh** and **B.BMA** can have similar performance in some settings, the advantage of the proposed Bayesian estimator is clear and meaningful. For example, the standard error in sample size estimation based

on **B.BMA** is almost twice as small as that based on **Ghosh** when the power of the original association study is low as shown in Figure 5.

Both the likelihood and Bayesian methods correct for threshold effect (i.e., the SNP of interest must pass a significance threshold) by incorporating the threshold value in the models. In practice, another source of bias is the ranking effect. More precisely, suppose that a large number of SNPs are considered but only the effects for top ranked SNPs are estimated. Again, the effect estimate is biased but a likelihood-based correction is cumbersome since all SNPs (with complex correlation structure among them due to linkage disequilibrium) must be considered jointly. The proposed Bayesian method only indirectly models the ranking effect by allowing the SNP of interest to be false positive. So far, the method of choice for this problem remains the bootstrap-based correction method of Sun and Bull (2005). However, the bootstrap method requires the original individual specific data which can be limiting. In contrast, the Bayesian and the likelihood approaches only need the summary statistics such as the reported naïve estimate and the association p-value, and the auxiliary information such as the sample size and the threshold used. In a two-stage setting when both the original discovery scan and a replication study are available, the combined approach proposed by Bowden and Dudbridge (2009) could provide better estimation results.

Although the method proposed here falls within the Bayesian paradigm, it has a clear frequentist component since the sampling distribution is conditional on the significance of the hypothesis test. While a complete Bayesian analysis in which simultaneous testing and estimation is possible for the problems considered here, it must be noted that the current practice among genetic investigators is to perform a large number of individual association tests prior to moving on to the estimation stage, in part due to the computational challenges associated with analyzing 500,000 or more SNPs. It is for this reason and to address the bias incurred by the resulting inference that we chose to use the current model. A full joint Bayesian analysis is the subject of ongoing research.

Acknowledgments. We would like to thank the Editor, an Associate Editor and three reviewers for constructive comments and suggestions that have substantially improved the paper. We would also like to thank Dr. Andrew Paterson for insightful discussions of the association study of complications in type 1 diabetes patients.

SUPPLEMENTARY MATERIAL

Supplement: Additional Derivations and Simulation Plots (DOI: 10.1214/10-AOAS373SUPP; .pdf). The appendix contains derivations related to posterior computation and additional simulation results related to the robustness of the Bayesian model considered to the choice of prior.

REFERENCES

BOWDEN, J. and DUDBRIDGE, F. (2009). Unbiased estimation of odds ratios: Combining genomewide association scans with replication studies. *Genet. Epidem.* **33** 406–418.

BOX, G. E. P. and MEYER, R. D. (1986). An analysis of unreplicated fractional factorials. *Technometrics* **28** 11–18. MR0824728

CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. MR1394738

CLYDE, M. A., DESIMONE, H. and PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91** 1197–1208.

FAYE, L., SUN, L., DIMITROMANOLAKIS, A. and BULL, S. B. (2009). A comprehensive look at the likelihood and bootstrap approaches to overcome the winner's curse in GWAS. *Genetic Epidem.* **33** 782–783.

GARNER, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genet. Epidem.* **31** 288–295.

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEWEKE, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics, 5 (1996)* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 609–620. Oxford Univ. Press, Oxford. MR1425430

GHOSH, A., ZOU, F. and WRIGHT, F. A. (2008). Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *Am. J. Hum. Genet.* **82** 1064–1074.

GÖRING, H., TERWILLIGER, J. D. and BLANGERO, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69** 1357–1369.

HOETING, J., DAVID, M., RAFTERY, A. and VOLINSKY, C. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

IOANNIDIS, J. P., THOMAS, G. and DALY, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10** 318–329.

ISHWARAN, H. and RAO, J. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

JEFFERIES, N. O. (2007). Multiple comparisons distortions of parameter estimates. *Biostatistics* **8** 500–504.

KUO, L. and MALLICK, B. (1998). Variable selection for regression models. *Sankhyā B* **60** 65–81. MR1717076

LIN, P.-I., VANCE, J. M., PERICAK-VANCE, M. A. and MARTIN, E. R. (2007). No gene is an island: The flip–flop phenomenon. *Am. J. Hum. Genet.* **80** 531–538.

MENG, X. and WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83** 1023–1032.

NAIR, R., DUFFIN, K. C. and HELMS, C. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kB pathways. *Nat. Genet.* **41** 199–204.

PATERSON, A. D., WAGGOTT, D., BORIGHT, A. P., HOSSEINI, M., SHEN, E., SYLVESTRE, M.-P. ET AL. (2010). A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both HbA1c and glucose. *Diabetes* **59** 539–549.

ROTHMAN, N., SKIBOLA, C. F., WANG, S. S., MORGAN, G., LAN, Q., SMITH, M. T. ET AL. (2006). Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: A report from the InterLymph Consortium. *Lancet Oncol*. **7** 27–38.

SLAGER, S. L. and SCHAID, D. J. (2001). Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Human Heredity* **52** 149–153.

STALLARD, N., TODD, S. and WHITEHEAD, J. (2008). Estimation following selection of the largest of two normal means. *J. Statist. Plann. Inference* **138** 1629–1638. MR2427293

SUN, L. and BULL, S. B. (2005). Reduction of selection bias in genomewide studies by resampling. *Genet. Epidem.* **28** 352–367.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc*. **82** 528–540. MR0898357

TODD, J. A., WALKER, N. M., COOPER, J. D., SMYTH, D. J., DOWNES, K., PLAGNOL, V. ET AL. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet*. **39** 857–865.

VAN DYK, D. and MENG, X. L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist*. **10** 1–111. MR1936358

VIALLEFONT, V., RAFTERY, A. E. and RICHARDSON, S. (2001). Variable slection and Bayesian model averaging in case-control studies. *Stat. Med*. **20** 3215–3230.

WANG, S. S., CERHAN, J. R., HARTGE, P., DAVIS, S., COZEN, W., SEVERSON, R. K., CHATTERJEE, N. ET AL. (2006). Common genetic variants in proinflammatory and other immunoregulatory genes and risk for non-Hodgkin lymphoma. *Cancer Res*. **66** 9771–9781.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

WU, L. Y., SUN, L. and BULL, S. B. B. (2006). Locus-specific heritability estimation via the bootstrap in linkage scans for quantitative trait loci. *Human Heredity* **62** 84–96.

XIAO, R. and BOEHNKE, M. (2009). Quantifying and corrrecting for the winner's curse in genetic association studies. *Genet. Epidem*. **33** 453–462.

XU, S. (2003). Theoretical basis of the Beavis effect. *Genetics* **165** 2259–2268.

YU, K., CHATTERJEE, N., WHEELER, W., LI, Q., WANG, S., ROTHMAN, N. and WACHOLDER, S. (2007). Flexible design for following up positive findings. *Am. J. Hum. Genet*. **81** 540–551.

ZHONG, H. and PRENTICE, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9** 621–634.

ZÖLLNER, S. and PRITCHARD, J. (2007). Overcoming the winner's curse: Estimating Penetrance parameters from case-control data. *Am. J. Hum. Genet*. **80** 605–615.

L. XU
R. V. CRAIU
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO M5S 3G3
CANADA
E-MAIL: lizhen@utstat.toronto.edu
craiu@utstat.toronto.edu

L. SUN
DALLA LANA SCHOOL OF PUBLIC HEALTH
AND DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
155 COLLEGE STREET
TORONTO, ONTARIO M5T 3M7
CANADA
E-MAIL: sun@utstat.toronto.edu