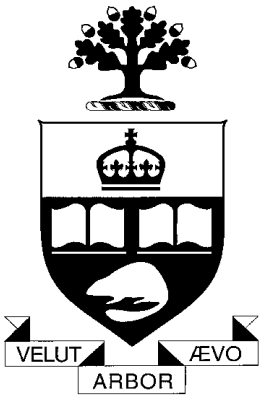

A flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design

by

Laura Faye
Dalla Lana School of Public Health Sciences
University of Toronto

Lei Sun
Department of Statistics &
Dalla Lana School of Public Health Sciences
University of Toronto

Apostolos Dimitromanolakis
Samuel Lunenfeld Research Institute
Mount Sinai Hospital, Toronto

Shelley Bull
Dalla Lana School of Public Health Sciences
University of Toronto

Technical Report No. 1006 Oct 16, 2010

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

A flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design

Laura L. Faye
Dalla Lana School of Public Health, University of Toronto, 6th Floor, Health Sciences Building, 155 College Street, Toronto, ON M5T 3M7, Canada
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 60 Murray Street, Box #18, Toronto, ON M5T 3L9, Canada

Lei Sun
Dalla Lana School of Public Health, University of Toronto, 6th Floor, Health Sciences Building, 155 College Street, Toronto, ON M5T 3M7, Canada
Department of Statistics, University of Toronto, Sidney Smith Hall, 100 St. George St., Toronto, ON M5S 3G3, Canada

Apostolos Dimitromanolakis
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 60 Murray Street, Box #18, Toronto, ON M5T 3L9, Canada

Shelley B. Bull
Dalla Lana School of Public Health, University of Toronto, 6th Floor, Health Sciences Building, 155 College Street, Toronto, ON M5T 3M7, Canada
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 60 Murray Street, Box #18, Toronto, ON M5T 3L9, Canada
bull@lunenfeld.ca

Abstract

The phenomenon known as the Winner's curse is a form of selection bias that affects estimates of genetic association. In genome-wide association studies (GWAS) the bias is exacerbated by the use of stringent selection thresholds and ranking over hundreds of thousands of single nucleotide polymorphisms (SNPs). In this report we develop an improved multi-locus bootstrap method, which accounts for both ranking- and threshold-selection bias in the presence of genome-wide SNP linkage disequilibrium structure. In addition we develop a confidence interval construction method for the bootstrap bias-reduced estimate. The bootstrap method easily adapts to various study designs and alternative test statistics as well as complex SNP selection criteria. The latter is demonstrated by our application to the Wellcome Trust Case Control Consortium

findings, in which the selection criteria was the minimum of the p-values for the additive and genotypic genetic effect models. In contrast, existing likelihood-based bias-reduced estimators account for one SNP at a time, and so are more simple computationally, but do not address ranking across SNPs. Our simulation studies show that the bootstrap bias-reduced estimates are usually closer to the true genetic effect than the likelihood estimates and are less variable. Replication study sample size requirements computed from the bootstrap bias-reduced estimates are adequate 75-90% of the time compared to 53-60% of the time for the likelihood method. The bootstrap methods are implemented in a user-friendly package able to provide point and interval estimation for both binary and quantitative phenotypes in large-scale GWAS in an efficient and flexible manner.

Keywords: Bias-reduction; Case-control studies; Genetic effect estimates; Quantitative traits; Statistical genetics; Winner's curse.

1. Introduction

Bias in genetic effect estimates such as the odds ratio (OR), a phenomenon also known as the Winner's curse or the Beavis effect, can occur in both genome-wide linkage and genome-wide association studies (GWAS) [1,2]. Because the same sample is typically used for both gene discovery and effect estimation, and the genetic effect is estimated only when the test for linkage or association at a genetic marker is significant, the estimate is on average larger in magnitude than the true value [3,4]. Under situations of low power, a dataset that produces a test statistic larger than the critical value for significance is an extreme dataset: the observed test statistic is from the tail of its distribution. This selection effect is exacerbated by the use of stringent significance criteria and the modest power to detect small effects typical in GWAS.

Comparison of estimates from gene discovery studies with those from replication studies demonstrates the upward bias in the magnitude of the original estimates, (hereafter referred to as upward bias) [5,6]. For example, among the 6 associated regions reported in the original Wellcome Trust Case Control Consortium (WTCCC) Type 1 Diabetes (T1D) study, two failed to be replicated by Todd *et al.* [7]; and even among those replicated, the ORs estimated from the replication sample were up to 70% closer to 1 than the corresponding estimate reported in the original discovery sample. This is practically important because if an estimate from the original sample were used to calculate the sample size required for a replication study, upward bias in effect estimation would yield under-estimates of sample size, leaving the investigators with a study underpowered to replicate a true association [5,8,9].

In the GWAS setting, selection of extreme test statistics can arise from application of strict significance criteria as well as from ranking over the whole genome. Threshold selection bias arises when the genetic effect is estimated only for SNPs with p-values below a threshold α. Ranking bias arises when the genetic effect is estimated for SNPs with p-values that are among the K smallest in the experiment with or without the threshold requirement [10]. However, even when selection is by threshold, the ranking effect is present as we demonstrate in the following. Both threshold and ranking selection contribute to the upward bias, and recent work suggests that ranking bias can be more severe in some cases [11]. Modeling ranking or threshold bias requires joint consideration of all SNPs and this can be a difficult task because of the complex correlation structure arising from linkage disequilibrium (LD) among SNPs.

Although the Winner's curse phenomenon has been recognized for some time, practical methods to correct for the selection bias have received attention more recently. The proposed remedies that do not require an additional independent sample fall into two categories: bootstrap

resampling-based methods [9,12-14], and likelihood-based approaches [15-18]. Although proposed for GWAS where both ranking and threshold selection contribute to the bias, all four likelihood approaches model threshold bias independently for each SNP, and thus do not directly account for the effect of competition among SNPs on the estimate for the particular SNP of interest.

The genome-wide bootstrap method developed by Sun and Bull [12] focused on bias-reduced point estimation in the genome-wide linkage setting. Within each bootstrap sample, the entire genome-wide scan was repeated in order to capture the effect of ranking and correlation structure on selection bias in addition to the effect of the significance threshold. Observations not included in a particular bootstrap sample were treated as a corresponding independent out-of-sample estimate. Wu *et al.* [13,19] evaluated the bootstrap method for a quantitative trait locus (QTL) linkage scan and detailed estimation methods for the case of multiple significant markers. Based on comparison studies of alternative bootstrap estimators, these authors recommended a so-called shrinkage estimator in cases of low power, but cautioned that this estimator tends to overcorrect at moderate to high power [12,13]. Yu *et al.* [9] applied the weighted bootstrap estimator proposed by Sun and Bull [12] to association studies but considered only the top-ranked significant SNP. Jeffries [14] proposed the use of quantile-based bootstrap confidence intervals (CI) for bias-reduction; this method considers ranking bias only.

In this report, we extend the bootstrap shrinkage estimator of Sun and Bull [12] to the GWAS setting. We introduce two adjustments crucial to genome wide analysis that also serve to reduce over-correction in the estimates. The first accounts for differences in variance associated with the minor allele frequency (MAF) of a SNP, and the second adjusts for the negative correlation between the within- and out-of-sample estimates used in the shrinkage factor calculation. Simulation studies demonstrate the substantial improvement provided by these corrections. We

also develop a method to construct bootstrap CIs that accounts for both threshold and ranking bias. We begin in Section 2 with the development of the genome-wide bootstrap methods for GWAS data. Application of the methods to the WTCCC T1D GWAS [7] in Section 3 demonstrates feasibility and reveals practical differences between the bootstrap and likelihood methods. To better understand the differences and similarities between the two approaches, we perform simulation studies in Section 4. We close with discussion and practical recommendations in Sections 5 and 6.

2. Methods

2.1 *Genome-wide bootstrap shrinkage estimation*

Let $\hat{\beta}_{N(k)}$ be the naïve estimate of the genetic effect of the k^{th}-ranked SNP reported in a GWAS ($k=1,...,K,\ K \geq 1$), selected by either rank- and/or threshold-based criteria in a set of observations we refer to as the original sample. The genetic effect could be the log odds ratio from a logistic regression for case-control data or the regression coefficient from a linear regression for quantitative outcomes. In each bootstrap sample, we apply the exact same selection criteria as in the original sample. For the k^{th}-selected SNP in the i^{th} bootstrap sample (where i indexes the bootstrap samples with at least k significant SNPs), let $\hat{\beta}_{Di(k)}$ be the within-sample estimate (based on subjects included in bootstrap sample i) and $\hat{\beta}_{Ei(k)}$ the out-of-sample estimate (subjects not included in bootstrap sample i). Let $N_{(k)}$ be the total number of bootstrap samples with at least k SNPs passing the selection criteria, where $N_{(k)} \geq N$ and N is the required minimal number of bootstrap samples, e.g. $N = 100$. (Note that if the original sample produced 2 significant

SNPs, each bootstrap sample could have 0, 1, 2 or more significant SNPs.) A genome-wide (GW) bootstrap estimator for the k^{th} SNP selected in the original sample is

$$\hat{\beta}_{boot(k)} = \hat{\beta}_{N(k)} - \frac{1}{N_{(k)}} \sum_{i=1}^{N_{(k)}} \left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ei(k)} \right) \qquad (2.1)$$

This estimator is an implementation of the method of Sun and Bull [12] proposed for linkage data, called a shrinkage estimator because it shrinks the naïve effect estimate by a bootstrap estimate of the bias in $\hat{\beta}_{N(k)}$.

The out-of-sample estimate $\hat{\beta}_{Ei(\cdot)}$ is meant to mimic an estimate obtained in an independent sample. However, although $\hat{\beta}_{Ei(\cdot)}$ and $\hat{\beta}_{Di(\cdot)}$ are estimated from mutually exclusive sets of observations, they are negatively correlated because the original sample is finite and observations excluded from one sample must be included in the other. This constraint is ignored in the original shrinkage estimator (2.1) developed for linkage data. In addition, unlike linkage analysis, the variance of the parameter estimate for a SNP depends on its allele frequency, which therefore also affects the bias of estimates drawn from the tail region defined by the selection criteria. This needs to be taken into account because the k^{th}-ranked SNP within each of the bootstrap samples may differ from the SNP detected in the original sample and have a different minor allele frequency.

Let $p_{(k)}$ be the MAF of the k^{th}-selected SNP in the original sample, and $p_{i(k)}$ be the MAF for the k^{th}-selected SNP in the i^{th} bootstrap sample (i indexes the bootstrap samples with at least k significant SNPs). Let $\hat{\beta}_{Ni(k)}$ be the estimate from the original sample for the k^{th} ranked SNP *selected in the i^{th} bootstrap sample*. (Note that $\hat{\beta}_{Ni(k)}$ and $\hat{\beta}_{N(k)}$ are different; the latter denotes the estimate from the original sample for the k^{th} ranked SNP *selected in the original sample*.) Let

$\hat{\sigma}^2_{Di(k)}$ be the variance for $\hat{\beta}_{Di(k)}$, $\hat{\sigma}^2_{Ei(k)}$ the variance for $\hat{\beta}_{Ei(k)}$, and $\hat{\sigma}_{DEi(k)}$ their covariance, all estimated empirically. A modified bootstrap shrinkage estimator for the GWAS setting with adjustment for MAF and correlation between $\hat{\beta}_{Di(k)}$ and $\hat{\beta}_{Ei(k)}$ is:

$$\hat{\beta}^*_{boot(k)} = \hat{\beta}_{N(k)} - \frac{\frac{1}{N_{(k)}} \sum_{i=1}^{N_{(k)}} \left(\hat{\beta}_{Di(k)} - \hat{\beta}^*_{Ei(k)}\right)\sqrt{2p_{i(k)}(1-p_{i(k)})}}{\sqrt{2p_{(k)}(1-p_{(k)})}} \qquad (2.2)$$

$$\text{where } \hat{\beta}^*_{Ei(k)} = \hat{\beta}_{Ei(k)} - \frac{\hat{\sigma}_{DEi(k)}}{\hat{\sigma}^2_{Di(k)}}\left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)}\right) \qquad (2.3)$$

We empirically estimate the variance and covariance of $\hat{\beta}_{Di(k)}$ and $\hat{\beta}_{Ei(k)}$ for each SNP individually by taking a separate set of bootstrap samples and computing the sample variance and covariance. In practice, we also truncate the shrinkage estimate at the null, so that the direction of association in the bias-reduced estimate cannot contradict the original test.

2.2 Derivation of the genome-wide bootstrap shrinkage estimator $\hat{\beta}^*_{boot(k)}$

The bootstrap sample and the out-of-sample observations are mutually exclusive and drawn from the finite original sample, which induces negative correlation between the within- and out-of-sample estimates. Their approximate joint distribution is:

$$\begin{pmatrix} \hat{\beta}_{Di(k)} \\ \hat{\beta}_{Ei(k)} \end{pmatrix} \middle| \hat{\beta}_{Ni(k)} = B \sim N\left\{ \begin{pmatrix} B \\ B \end{pmatrix}, \begin{pmatrix} \sigma^2_{Di(k)} & \sigma_{DEi(k)} \\ \sigma_{DEi(k)} & \sigma^2_{Ei(k)} \end{pmatrix} \right\} \qquad \sigma_{DEi(k)} < 0 \qquad (2.4)$$

It follows that conditional on the observed within-sample estimate $\hat{\beta}_{Di(k)}$

$$E\left(\hat{\beta}_{Ei(k)} \middle| \hat{\beta}_{Di(k)} = d, \hat{\beta}_{Ni(k)} = B\right) = B + \frac{\sigma_{DEi(k)}}{\sigma^2_{Di(k)}}(d - B). \qquad (2.5)$$

We correct $\hat{\beta}_{Ei(k)}$ by $\frac{\sigma_{DEi(k)}}{\sigma^2_{Di(k)}}(d - B)$ to remove the correlation (see Supplemental Appendix A for further details of the derivation).

A heuristic explanation of equations (2.3) and (2.5) is as follows. Consider a SNP from the original sample selected in bootstrap sample *i* that was also significant in the original sample. The test statistic $\hat{\beta}_{Di(k)}/\hat{\sigma}_{Di(k)}$ is approximately normally distributed and centered around $\hat{\beta}_{Ni(k)}/\hat{\sigma}_{Ni(k)}$, which was also larger than the critical value. Then, the probability of a significant test statistic for this SNP is high in any bootstrap sample and $\hat{\beta}_{Ei(k)}$ will also be close to $\hat{\beta}_{Ni(k)}$. Now consider a SNP selected in the bootstrap sample that was not significant in the original sample. When $\hat{\beta}_{Di(k)}/\hat{\sigma}_{Di(k)}$ is greater than the critical value, $\hat{\beta}_{Di(k)}$ will tend to be much larger than $\hat{\beta}_{Ni(k)}$. Then the quantity $\left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)}\right) = (d - B)$ in (2.5) will tend to be large and positive and $\sigma_{DEi(k)}$ will be negative, so that the effect on $\hat{\beta}_{Ei(k)}$ will be large and negative. In the GWAS setting with 500,000 or more SNPs, SNPs with null or weak association in the original sample may be selected by chance in many of the bootstrap samples. Therefore it is important to correct the mean of $\hat{\beta}_{Ei(k)}$ for each SNP by subtracting $\frac{\hat{\sigma}_{DEi(k)}}{\hat{\sigma}^2_{Di(k)}}\left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)}\right)$ from each $\hat{\beta}_{Ei(k)}$.

Under the assumption of an additively coded SNP and Hardy-Weinberg equilibrium (HWE), the genotypic variance is $2p(1-p)$, where p is the MAF of the SNP of interest. Within the class of generalized linear regression models, the variance of the regression coefficient $\hat{\beta}$ is inversely proportional to the genotypic variance (see Supplemental Appendix B for details), and selection bias is related to the variance via the power [12]. In order to account for allele frequency differences across SNPs, we rescale each term in the shrinkage factor

by $\sqrt{2p_{i(k)}(1-p_{i(k)})}$.

2.3 *Genome-wide bootstrap confidence interval construction*

We construct a symmetric (1-α)% CI around the point estimate, $\hat{\beta}^{*}_{boot(k)}$ as in equation (2.2), using a bootstrap estimate of the standard deviation of $\hat{\beta}^{*}_{boot(k)}$. This requires a 2-level bootstrap sampling scheme [20]: we draw M 1st level bootstrap samples from the original data, and for each one, compute the bootstrap shrinkage estimate via a 2nd bootstrap level. The CI width is computed with the standard deviation of the M bootstrap estimates.

Stated more formally, let $\hat{\beta}_{Dj(k)}$ be the within-sample estimate for the k^{th}-selected SNP chosen in the j^{th} 1st level bootstrap sample, which serves as the naïve estimate $\hat{\beta}_{Nj(k)}$ in the 2nd level computation. Let $N_{j(k)}\,(\geq N)$ be the total number of 2nd level bootstrap samples, within the j^{th} 1st level bootstrap sample, with at least k SNPs selected. Let $\hat{\beta}_{Dj\_i(k)}$ and $\hat{\beta}_{Ej\_i(k)}$ be the corresponding within- and out-of-sample bootstrap estimates with $p_{j\_i(k)}$ the corresponding MAF (estimated from the original sample) for the i^{th} 2nd level bootstrap sample nested within the j^{th} 1st level bootstrap sample. Let $\hat{\sigma}^{2}_{Dj\_i(k)}$ be the empirically estimated variance of $\hat{\beta}_{Dj\_i(k)}$ and let $\hat{\sigma}_{DEj\_i(k)}$ be the estimated covariance between $\hat{\beta}_{Dj\_i(k)}$ and $\hat{\beta}_{Ej\_i(k)}$. Then for each 1st level bootstrap sample j = 1 to M, we obtain a bootstrap estimate

$$\hat{\beta}^{*}_{boot\_j(k)} = \hat{\beta}_{Nj(k)} - \frac{\frac{1}{N_{j(k)}}\sum_{i=1}^{N_{j(k)}}\left(\hat{\beta}_{Dj\_i(k)} - \hat{\beta}^{*}_{Ej\_i(k)}\right)\sqrt{2p_{j\_i(k)}(1-p_{j\_i(k)})}}{\sqrt{2p_{j(k)}(1-p_{j(k)})}} \quad \text{for } j = 1\ldots M. \qquad (2.6)$$

where $\hat{\beta}^{*}_{Ej\_i(k)} = \hat{\beta}_{Ej\_i(k)} - \frac{\hat{\sigma}_{DEj\_i(k)}}{\hat{\sigma}^{2}_{Dj\_i(k)}}\left(\hat{\beta}_{Dj\_i(k)} - \hat{\beta}_{Nj(k)}\right)$

Defining $s(\cdot)$ as the sample standard deviation, the estimated standard deviation of the M estimates with correction for the MAF is therefore

$$\hat{\sigma}^{*}_{boot(k)} = s\left(\hat{\beta}^{*}_{bootj(k)}\sqrt{2p_{j(k)}(1-p_{j(k)})}\right)\Big/\sqrt{2p_{(k)}(1-p_{(k)})}, \qquad (2.7)$$

which is used to construct an asymptotic (1 - α)% CI for the GW bootstrap estimate $\hat{\beta}^{*}_{boot(k)}$:

$$\hat{\beta}^{*}_{boot(k)} \pm Z_{1-\alpha/2}\hat{\sigma}^{*}_{boot(k)} \qquad (2.8)$$

where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of the standard normal distribution.

2.4 *Conditional likelihood methods*

Several existing bias-reduction methods are based on maximum likelihood, conditional on the test statistic exceeding a critical value. Under a case-control design the likelihood of Zollner and Pritchard [15] is formulated with allele frequency and penetrance parameters but requires external population prevalence data for constrained estimation. A related maximum likelihood estimator, due to Xiao and Boehnke [18], formulates likelihood in terms of risk allele frequency differences between cases and controls. Zhong and Prentice [16,21] apply a standard case-control logistic regression likelihood and construct a weighted average of the naïve log OR estimate obtained from the logistic model and the estimate from a model in which the naïve estimate must be the median of the conditional distribution. We will refer to this estimator as the Adjusted Median Likelihood (AML) estimator. In addition, Ghosh *et al.* [17] consider a more general class of normally distributed estimators with a Wald-like test of significance suitable for association studies of either case-control or quantitative outcomes. Working at the test statistic

level, they take the average of two quantities. The first is the MLE of the mean of the distribution of the test statistic. The second is the mean of a random variable that follows the distribution of the likelihood function for the mean of the distribution of the test statistic, normalized to be a proper density. Multiplying by the standard error of the naïve estimate transforms this quantity from the test statistic level to the genetic effect estimate. We will call this estimator the Normalized Maximum Likelihood Estimator (NMLE). Simulation studies [16-17,22] demonstrated that the AML and NMLE both perform better than the standard MLE under the conditional likelihood. Our simulation studies (see Appendix D in the Supplemental Information) show the NMLE performs better than the AML, based on smaller root mean square error, and therefore we report comparisons of the bootstrap to the NMLE method.

3. Application to the Wellcome Trust Case-Control Consortium Data

We re-analyzed the significant SNPs from the WTCCC T1D [7] sub-study that were also assessed in the Todd *et al.* [6] replication study. The four estimators investigated here as well as in the simulation studies of section 4 are as follows.

Uncorrected

Naïve: original naïve estimator

Bias-reduced, single-SNP likelihood method

NMLE: Normalized Maximum Likelihood Estimator of Ghosh *et al.* [17]

Bias-reduced, genome-wide bootstrap methods

GW Bootstrap without correction: Genome-Wide Bootstrap *without correction* shrinkage estimator, equation (2.1) in Section 2.1

GW Bootstrap: Genome-Wide Bootstrap *with correction* shrinkage estimator, equation (2.2)

in Section 2.2

3.1 *Application methods*

We obtained the individual-level genotypes and phenotypes for the WTCCC T1D sub-study which includes 1963 cases and 2938 controls genotyped at 356,946 SNPs. The WTCCC T1D study reported strong associations (p-value < $5x10^{-7}$) at 5 SNPs and moderate associations (p-value < 10^{-5}) at 7 SNPs. The reported SNPs were the most significant in their regions, in total 601 SNPs met the criteria for moderate association and 472 SNPs met the criteria for strong association. The replication study of Todd *et al.* [6] unequivocally validated 4 WTCCC association findings ($p < 1.35x10^{-9}$), including one SNP meeting the WTCCC criteria for moderate association (rs2542151) and 3 SNPs meeting the WTCCC criteria for strong association (rs12708716, rs17696736, rs2292239). The replication study also provided moderate evidence (p = 0.0231) for rs17388568 from the WTCCC moderate significance table. Although the WTCCC reported rs11171739 as the most significant SNP in the 12q13 region, rs2292239 was more significant in the replication study of Todd *et al.* [6], so we report the latter. The replication study OR estimates are all smaller than the original study naïve estimates, demonstrating the effect of the winner's curse (Table I).

We applied bootstrap and likelihood methods to all SNPs using both the strong and moderate association criteria. In each of the GW bootstrap applications, we used N = 500 samples for the point estimate and M = 100 samples for the variance estimate. We applied the WTCCC selection criteria: minimum p-value of the trend test (1 df) and the genotypic test (2 df) less than the threshold value of $5x10^{-7}$ for inclusion in their strong association table (rs17696736, rs2292239, rs12708716), and less than $1x10^{-5}$ for inclusion in their moderate association table (rs2542151,

rs17388568). Following the WTCCC, we estimated the log odds ratio for the additive genetic effect. For the GW bootstrap method we excluded SNPs with MAF <1%, Hardy-Weinberg equilibrium p-value $<10^{-7}$, and genotyping call rate <99%. As in Ghosh *et al.* [17], we applied the NMLE method using the Wald test (1 df) from the additive logistic regression model. The analysis of Todd *et al.* [6] stratified by geographical subregion, while the WTCCC T1D analysis did not. To facilitate our comparisons, we conducted unstratified analysis of the WTCCC dataset. For this reason the stratified naïve and NMLE estimates reported by Todd *et al.* [6] and Ghosh *et al.* [17] differ slightly from the estimates we present. In Table I, we provide bias-reduced estimates for the five SNPs replicated by Todd *et al.* [6]. In Figure 1, we compare the bootstrap and likelihood bias-reduced estimates using the WTCCC moderate association threshold for a sample of SNPs with p-values between $1x10^{-5}$ and $1x10^{-8}$ that were randomly selected from the 601 significant SNPs.

3.2 *Application Results*

The bootstrap and likelihood methods reduced the naïve estimate for all 5 replicated SNPs (Table I). The amount of shrinkage for the 4 highly significant SNPs was small and similar among the methods. The reduction in the least significant SNP was larger and varied among the methods. This reflects intuition: the small observed p-value implies high power to detect this SNP at the chosen significance threshold and similar studies would also tend to produce significant findings. Therefore, the effect of threshold-selection is minimal. A p-value close to the threshold for selection implies lower power and more bias.

Analysis of all SNPs significant at the moderate association threshold shows the same trend (Figure 1). At highly significant p-values, the bias-reduced estimates are similar and close to the

naïve estimate. The bootstrap corrects for ranking as well as threshold bias, and therefore corrects a little more than the likelihood. Both methods shrink the most when the observed p-value is close to the threshold for selection ($1x10^{-5}$). The likelihood tends to shrink estimates much more than the bootstrap for these marginally significant SNPs, and so we expect the difference between methods to be greatest for lower power SNPs.

Selection bias can be severe for lower power SNPs. We therefore present simulations designed to evaluate the methods in the lower power case, which is most relevant to evaluating differences in performance.

4. Genome-Wide Simulation Studies

4.1 *Design*

We conducted genome-wide simulations to evaluate the performance of the GW bootstrap method and the NMLE when selection is threshold-based. In order to capture the effect of realistic correlation structure among SNPs, we fixed the actual WTCCC genotypes and simulated case-control status under a log additive multiple-SNP logistic model based on varying genetic effect sizes for 5 associated SNPs. In one configuration the associated SNPs had MAF values and corresponding odds ratios (ORs) as described in Table II, which yielded a probability of selection for each SNP of 7%, 11%, 13%, 30% and 49% respectively. In a second configuration the same set of associated SNPs had the same MAF values but larger corresponding ORs so that the probability of selection for each SNP was 60%, 70%, 77%, 90% and 99% respectively (Supplementary Table I). The remaining 356,941 SNPs in the WTCCC genotype dataset had no

genetic effect in our simulation model

We simulated datasets until there were at least 500 replicates with significant estimates for each true positive SNP. We applied a threshold-based significance criterion: a trend test p-value less than the WTCCC strong significance level of 5×10^{-7}. For each dataset we computed the GW bootstrap bias-reduced estimates with MAF and correlation corrections (hereafter referred to as the GW bootstrap), the GW bootstrap without correction, and the NMLE of Ghosh *et al.* [17]. We applied the GW bootstrap using 100 level 1 bootstrap samples. For each simulation we computed the summary statistics and examined the distribution of estimates.

To assess the utility of bias-reduced estimates in planning replication studies, we calculated the sample size required for a replication study with 80% power for the naïve and bias-reduced estimates in turn and compared these values with the actual sample size required (using formulas presented by Slager and Schaid [23]). We assumed the replication sample would be drawn from the same population as the original sample, with an equal number of cases and controls. We examined each associated true positive SNP in our simulation model individually. For each SNP, we estimated the sample size required to replicate the association using the naïve, GW Bootstrap and NMLE estimates (as obtained above) in each of the simulated datasets in which that SNP was significant. For comparison, we computed the power to select the SNP of interest based on the generating OR and the estimated sample size. We report how often the estimated sample size is large enough to achieve 80% power.

At the replication stage, the type I error can be controlled by choice of an appropriate significance criterion, and most false positive SNPs would be eliminated by failure to replicate. However, in order to exclude a SNP, it would be desirable for a 95% CI in the replication sample to be sufficiently precise to exclude an OR that would be of interest, e.g. an OR of 1.15 or larger. Therefore, for each false positive SNP in the simulated datasets we computed the sample size for

replication, using the estimated log OR, and then assuming the computed sample size, calculated the corresponding standard error for the log OR under the null hypothesis of no association. Assuming a 95% CI half-width of log(1.15) = 0.140 would be sufficient to exclude an OR of 1.15, we determined how often this precision was achieved.

In order to evaluate the ability of each method to adapt to the ranking effect, we also compared the mean absolute bias of the naïve estimate (i.e. the amount of shrinkage that is required) to the amount of shrinkage provided by the NMLE and the GW bootstrap stratified by rank.

4.2 *Results*

The naïve estimate for true positive SNPs is upward biased which produces large RMSE, while the NMLE over-corrects slightly with a large RMSE due to high variance (Figure 2, Table II). The GW bootstrap has larger downward bias in comparison to the NMLE, but has smaller RMSE because the majority of bootstrap estimates form a mode just below the true value, and most estimates are within 25% of the true value (Supplemental Table S2). For false positive SNPs, the NMLE method tends to have a higher proportion of estimates close to the true value, however bias and RMSE are larger than the GW bootstrap for MAF <0.3 (Figure 2, Supplemental Table S1 and S2). The GW bootstrap *without correction* shrinks most estimates to a value close to the null. As a result, the method performs well for false positives where the true genetic effect is in fact at the null, but grossly over-corrects for true positive SNPs. The MAF and correlation adjustments reduce this over-correction by 65-100% in the low power case. Similar results were obtained for configuration 2 at high power (Supplemental Table S1, Supplemental Figure S1)

Depending on the effect size and the MAF of the SNP, 75-90% of the time sample size

calculations based on GW bootstrap estimates were at or above the size required for 80% power (Figure 3, Supplemental Table S3 and Figure S2). On average, the sample size based on the GW bootstrap estimate tended to over-estimate the actual sample size required for replication by a factor of 1.25 to 2. In contrast, sample sizes calculated using the NMLE likelihood estimates under-estimated the sample size 38-52% of the time (depending on study design parameters), while the average sample size was larger than that for the GW bootstrap sample size calculations, due to the large variance of the likelihood estimate.

For the false positive SNPs, the sample size for the replication study given by the GW bootstrap was on average 1.8 to 2.6 times larger than that required and was sufficiently large in over 93% of datasets to exclude effect sizes of interest in the replication study (for MAF > 10%). In contrast, the sample sizes given by the NMLE method were slightly less likely to be adequate but were on average 5.2 to 6.9 times larger than required due to many very large estimates of sample size (Supplemental Table S4).

Even when SNP selection is by threshold, there is also a ranking effect as SNPs compete for rank, and so the bias in the naïve estimate depends on the rank achieved by the SNP (Table III). The OR tends to require more correction when a SNP achieves a higher versus a lower rank. The NMLE method, however, overcorrects when the SNP has a lower rank. The GW bootstrap corrects all ranks by about the same amount, which is closer to the correction that is required. As a result, the GW bootstrap has smaller RMSE across ranks.

5. Discussion

In most GWAS, threshold and ranking selection are both present: estimates are reported only for significant SNPs leading to bias away from the null, and for SNPs with the same true effect size, top ranked SNPs are more highly biased than lower-ranked SNPs [10]. This is borne out in our simulations: bias is greater when rank is higher (Table III). The likelihood methods address only the threshold effect, and tend to shrink lower-ranked SNPs the most. The GW bootstrap models both ranking and threshold bias: a higher rank among SNPs of similar significance implies greater ranking bias, even if highly significant p-values imply little threshold bias. On the other hand, a lower rank implies little ranking bias, but the less significant p-values among lower-ranked SNPs implies lower putative power and a larger threshold bias.

The bootstrap method of Sun and Bull [12] was developed for the multi-SNP linkage setting. When simply applied as is to the GWAS setting it seriously over-corrects, as demonstrated by our simulations. The MAF and correlation adjustments we develop in this report optimize the bootstrap for genome-wide association analysis, providing significantly more accurate estimates for associated SNPs.

In appendix D we present results for a bootstrap method that considers only a single SNP at a time, and is analogous to the single-SNP likelihood methods. As the single-SNP bootstrap models only threshold-bias and not ranking bias, it does not reduce bias as well as the GW bootstrap for the genome-wide simulations. The single-SNP bootstrap performed well only in cases of high power, where selection bias is not a concern. In cases where ranking bias is more pronounced than selection bias or is the only bias present, as in the case of selection by rank, single-SNP methods may not reduce the selection bias sufficiently. We recommend the GW bootstrap with correction in preference to both the single-SNP bootstrap and likelihood methods, especially when power is low to moderate.

Confidence intervals for bias-reduced estimates based on the GW bootstrap similarly account

for both threshold and ranking bias, but they need to be interpreted carefully: while the SNP is deemed significant at a genome-wide criterion in the original analysis, the bias-reduced CI could nevertheless cover zero. This does not imply the SNP is no longer significant: the original p-value stands because the confidence interval – hypothesis testing duality does not hold for bias-reduced intervals. However, the bias-reduced CI width does indicate the level of uncertainty in the estimate and may be interpreted as the CI that would be obtained in an independent sample of the same sample size as the original. The variance of the naïve estimate is artificially small due to selection, which is reflected in the poor coverage of the naïve CI. The bias-reduced methods increase the variance by an appropriate amount, which is reflected in their near-nominal coverage. In a simulation study comparing the performance of single-SNP bootstrap and likelihood confidence intervals we report in Appendix D of the Supplemental Information, the 95% CI coverage for both likelihood and bootstrap intervals is near nominal, however the bootstrap intervals are more precise. The computational intensity of the genome-wide bootstrap has precluded a thorough evaluation by simulation, however the underlying principle is the same as for the single-SNP bootstrap. Notably in the WTCCC application, the 95% GW bootstrap bias-reduced CIs covered the replication value for 4 of out 5 SNPs.

In a genome-wide scan, the genetic model is often unknown, so multiple tests may be used in order to determine if a SNP shows significance under one of several genetic models. The WTCCC study, for example, used the minimum p-values from the trend and genotypic tests. To apply the likelihood method in this case would require a power function to be specified and maximized over the parameter of interest, which may not be straight forward. As long as the selection step can be automatically applied to the original dataset, it can be applied in the bootstrap. In addition, the bootstrap is flexible enough to be extended in several respects e.g. multiple-SNP tests, tests for models that include covariates and interactions, or selection criteria

that incorporate external information such as gene ontologies or pathway information.

The genome-wide bootstrap tends to be conservative for true positive SNPs. In planning appropriately powered replication studies, a slightly conservative estimate is desirable. Under-powered studies due to sample size calculations based on optimistic estimates has been cited as a major cause of failure to replicate [3]. As there is generally no way to distinguish between true and false positives, then when true and false positives have similar naïve estimates, they will have similar bias-reduced estimates. Although estimators that perform better for false positives can be specified [12], smaller bias for false positives comes at the expense of larger bias for true positives and vice versa. In planning replication studies, a good estimate for true positive SNPs is more useful than a good estimate for false positive SNPs. Consider a scenario in which the initial genome-wide scan identified a number of significant SNPs with some true positives and some false positives of similar naïve log OR. If the true positive genetic effect estimates are accurate then the replication study using the corresponding sample size estimates will be adequately powered to replicate truly associated SNPs. On the other hand, if true positives are severely under-estimated in order to achieve more accurate false positive estimates, the over-estimated sample size may be too large for the replication study to be feasible. The GW bootstrap performs well for true positives and our simulations show that sample size computed for replication is usually adequate.

The bootstrap method requires a sufficient number of resamples to accurately estimate the effect of selection. If too few bootstrap samples are used, the random variability of bootstrap sampling may produce a poor estimate. As variability depends on the data, it is advisable to run the bootstrap cumulatively with an increasing number of bootstrap samples until stability is achieved. We recommend a minimum starting point of 100 bootstrap samples. Bootstrap methods are computationally intensive, because the same analysis applied to the original dataset

must be applied to each bootstrap sample. If 100 level 1 and 100 level 2 bootstraps (10,000 bootstraps in total) are used to construct a CI, this is computationally equivalent to repeating the original analysis 10,000 times. Computation time scales linearly with the number of bootstraps required. Genome-wide bootstrap analysis of the WTCCC T1D dataset with selection via the trend test took 45 minutes to obtain a point estimate only and 32 hours to obtain confidence intervals using 100 bootstrap samples for both levels of bootstrapping on a moderately powered unix machine (2 x quad-core Intel Xeon E5410 2.33 GHz, 12 GB RAM at 1066MHz). Parallel processing substantially improves computational time.

6. Conclusion

For bias-reduced estimation in GWAS, we recommend the GW bootstrap (with the GWAS-specific adjustments) particularly in cases where complex testing procedures are applicable. The GW bootstrap estimates are slightly conservative, which makes replication study sample sizes computed from the estimates adequate most of the time. Ranking can be a non-trivial source of bias, and among the methods we considered, only the genome-wide method accounts for the effect of all SNPs on the ranking of a SNP of interest. In some studies complex selection criteria make specification of the likelihood difficult, while the bootstrap is adaptable in that any well-defined criteria can be applied in each bootstrap sample. Efficient software implementing the bootstrap method is available [24,25].

Supplementary Appendices

Supplementary material available online: derivation of the MAF and correlation adjustments for the GW bootstrap, additional genome-wide simulation results and single-SNP simulation results.

Acknowledgements

This research was funded by the Canadian Institutes of Heath Research [operating grant number MOP-84287 to SBB and LS, doctoral research award MDR-88001 to LF]. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

References

1. Beavis WD. The power and deceit of QTL experiments: lessons from comparative QTL studies. *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference. American Seed Trade Association,* 1994; 250–266.
2. Voight BF, Cox NJ. Minding your LOD's and q's: how linkage effect size bias can contribute to the winner's curse in replication association studies. In: The American Society of Human Genetics 54th Annual Meeting, October 26-30, 2004. Toronto. Abstract number 277.
3. Göring HH, Terwilliger JD. and Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *The American Journal of Human Genetics* 2001; **69**:1357–1369.
4. Garner C. Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology* 2007; **31**:288-295.
5. Lohmueller K, Pearce CL, Pike M, Lander ES. and Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* 2003; **33**:177-182.
6. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics* 2007; **39**: 857-864.
7. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. 2007; *Nature* **447**:661-678.
8. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genetics in Medicine* 2002; **4**:45-61.

9. Yu K, Chatterjee N, Wheller W, Li Q, Wang S, Rothman N, Wacholder S. Flexible design for following up positive findings. *The American Journal of Human Genetics* 2007; **81**:540–551.
10. Bowden J, Dudbridge F. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genetic Epidemiology* 2009; **33**:406–418.
11. Jeffries NO. Ranking bias in association studies. 2009; *Human Heredity* **67**:267–275.
12. Sun L, Bull SB. Reduction of selection bias in genome wide studies by resampling. *Genetic Epidemiology* 2005; **28**:352-367.
13. Wu LY, Sun L, Bull SB. Locus-specific heritability estimation via the bootstrap in linkage scans for quantitative trait loci. *Human Heredity* 2006; **62**:84-96.
14. Jeffries NO. Multiple comparisons distortions of parameter estimates. *Biostatistics* 2007; **8**:500–504.
15. Zollner S, Pritchard J. Overcoming the winners curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics* 2007; **80**:605–615.
16. Zhong H, Prentice R. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 2008; **9**:621-634.
17. Ghosh A, Zou F, Wright F. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *The American Journal of Human Genetics* 2008; **82**:1064–1074.
18. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology* 2009; **33**:453-62.
19. Wu L, Lee SSF, Sun L, Bull SB. Resampling methods to reduce the selection bias in genetic effect estimation in genome-wide scans. *BMC Genetics* 2005; **6**:S24.
20. Efron B, Tibshirani RJ. *An introduction to the bootstrap* 1993 New York: Chapman and Hall.

21. Zhong H, Prentice R. Correcting the "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology* 2010; **34**:78–91.

22. Faye L, Sun L, Dimitromanolakis A, Bull SB. A comprehensive look at the likelihood and bootstrap approaches to overcome the winner's curse in GWAS. *Genetic Epidemiology* 2009; **33**:788-789.

23. Slager SL. and Schaid D.J. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human Heredity* 2001; **52**:149–153.

24. Sun L, Bull SB, Faye L. Dimitromanolakis, A., Waggott, D., Paterson, A.D. and The DCCT/EDIC Research Group. BR-squared: a practical solution to the winner's curse in genome-wide scans. 2009; In: *The American Society of Human Genetics 59th Annual Meeting*, October 20-24, 2009. Honolulu. Abstract number 186

25. Sun L, Dimitromanolakis L, Faye L, Bull SB. "Documentation for BR2". BR2 home page, University of Toronto. 2010; Accessed June 12 2010. www.utstat.utoronto.ca/sun/Software/BR2/br2-web/br2.html

Table I: Application of bias-reduced methods to WTCCC T1D: point estimates and confidence intervals

			Discovery Samples				Follow-up Samples
				Likelihood	Bootstrap		
SNP	α		Naïve	NMLE	GW w/o correction	GW w correction	Replication
rs17696736	5E-7	2.17E-15	1.39	1.39	1.35	1.37	1.16
rs2292239	5E-7	4.87E-10	1.31	1.27	1.23	1.28	1.28
rs12708716	5E-7	9.24E-08	0.79	0.89	0.89	0.83	0.83
rs17388568	1E-5	5.00E-07	1.26	1.17	1.11	1.19	1.08
rs2542151	1E-5	1.89E-06	1.29	1.13	1.11	1.21	1.29

	95% Confidence Intervals for OR				
		Likelihood	Bootstrap		
SNP	Naïve	NMLE	GW w/o correction	GW w correction	Replication
rs17696736	1.28-1.51	1.28-1.51	0.80-2.27	1.24-1.54	1.09-1.23
rs2292239	1.20-1.42	1.13-1.42	1.00-1.51	1.16-1.41	1.20-1.36
rs12708716	0.72-0.86	0.73-1.01	0.90-1.39	0.75-0.91	0.78-0.89
rs17388568	1.15-1.37	1.07-1.36	0.91-1.34	1.04-1.36	1.01-1.15
rs2542151	1.16-1.43	0.99-1.41	0.91-1.34	1.04-1.41	1.19-1.40

Note: Discovery samples are from the original WTCCC (2007) T1D GWAS; follow-up samples and results are from the replication study of Todd and others (2007). The association p-value is the minimum of the 1 df trend and the 2 df genotypic association tests observed in the WTCCC samples for SNPs meeting criteria for strong significance (p-value < 5E-7) and moderate significance (p-value < 1E-5). Naïve (uncorrected), Adjusted Median Likelihood (AML), Normalized MLE Likelihood (NMLE), genome-wide bootstrap without correction (Bootstrap, GW w/o correction) and with correction (Bootstrap, GW w/ correction) point estimates and confidence intervals were calculated using the discovery WTCCC samples; Replication estimates use follow up samples only.

Table II: Genome-wide simulation: performance of bias reduced point estimators for low power alternative case

Pr(p<α)	MAF	OR	Mean of Estimates (OR)				Percentiles (OR)							
			Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr	Naïve		NMLE		GW Boot w/o corr		GW Boot w/ corr	
							10th	90th	10th	90th	10th	90th	10th	90th
7%	0.33	1.16	1.29	1.16	1.05	1.17	1.25	1.34	1.06	1.32	1.00	1.15	1.12	1.24
11%	0.45	1.16	1.26	1.15	1.03	1.14	1.23	1.29	1.06	1.27	1.00	1.08	1.11	1.19
13%	0.46	1.16	1.26	1.16	1.03	1.15	1.24	1.30	1.06	1.27	1.00	1.10	1.11	1.20
30%	0.29	1.20	1.29	1.18	1.04	1.17	1.26	1.34	1.06	1.31	1.00	1.13	1.12	1.24
49%	0.13	1.34	1.42	1.28	1.06	1.23	1.37	1.50	1.09	1.47	1.00	1.17	1.15	1.33

		Relative Bias (log OR)				RMSE (log OR)			
	log OR	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr
	0.15	0.68	-0.03	-0.68	0.02	0.13	0.11	0.15	0.08
	0.15	0.55	-0.10	-0.82	-0.12	0.10	0.08	0.14	0.04
	0.15	0.57	-0.05	-0.78	-0.09	0.11	0.08	0.14	0.05
	0.18	0.39	-0.11	-0.77	-0.17	0.10	0.10	0.16	0.06
	0.29	0.24	-0.14	-0.83	-0.29	0.10	0.15	0.30	0.14

Note: 500 genome-wide case control datasets where the SNP of interest was significant at p-value threshold α=5E-7 were simulated using the WTCCC T1D genotypes and OR as indicated. Mean of Estimates (average on OR scale), 10th and 90th percentiles (OR scale), Relative bias (average difference between estimated and true log OR divided by true log OR), and root mean squared error (RMSE of log OR estimates) were computed for the uncorrected (naïve), Normalized Maximum Likelihood (NMLE), genome-wide bootstrap without correction (GW Boot w/o correction) and genome-wide bootstrap without correction (GW Boot w/ correction) estimates for true positive SNPs. Pr(p<α) is the probability of obtaining a p-value below threshold α for each SNP. Detailed descriptions of the method are in Section 4.

Table III. Genome-wide simulation: Bias and estimates of bias by rank for naïve, NMLE and GW Bootstrap with correction

Alternative Case - Low Power

	MAF	OR	log OR	Power	Absolute Bias (log OR) in Naïve by Rank			RMSE (log OR) in Naïve by Rank		
					1	2	3	1	2	3
rs12708716	0.33	1.16	0.15	7%	0.12	0.09	0.09	0.13	0.10	0.09
rs11171739	0.45	1.16	0.15	11%	0.09	0.08	0.08	0.09	0.08	0.08
rs17696736	0.46	1.16	0.15	13%	0.09	0.08	0.08	0.09	0.08	0.09
rs9272346	0.29	1.20	0.18	30%	0.08	0.07	0.06	0.08	0.07	0.06
rs6679677	0.13	1.34	0.29	49%	0.07	0.05	0.05	0.08	0.06	0.05

	Reduction Provided by NMLE by Rank			Absolute Bias (log OR) in NMLE by Rank			RMSE (log OR) in NMLE by Rank		
	1	2	3	1	2	3	1	2	3
rs12708716	0.08	0.11	0.13	0.05	-0.01	-0.04	0.11	0.08	0.08
rs11171739	0.08	0.11	0.12	0.00	-0.03	-0.04	0.07	0.06	0.08
rs17696736	0.08	0.10	0.11	0.01	-0.03	-0.02	0.07	0.06	0.08
rs9272346	0.07	0.10	0.13	0.01	-0.04	-0.07	0.08	0.08	0.09
rs6679677	0.09	0.14	0.15	-0.02	-0.09	-0.10	0.11	0.13	0.14

	Reduction Provided by GW Boot w/ corr by Rank			Absolute Bias (log OR) in GW Boot w/ corr by Rank			RMSE (log OR) in GW Boot w/ corr by Rank		
	1	2	3	1	2	3	1	2	3
rs12708716	0.09	0.10	0.10	0.03	-0.01	-0.01	0.08	0.04	0.03
rs11171739	0.10	0.10	0.10	-0.02	-0.02	-0.02	0.04	0.03	0.04
rs17696736	0.10	0.10	0.09	-0.01	-0.02	-0.01	0.04	0.03	0.04
rs9272346	0.10	0.11	0.11	-0.02	-0.04	-0.05	0.05	0.05	0.05
rs6679677	0.15	0.16	0.16	-0.08	-0.10	-0.11	0.10	0.11	0.11

Note: Datasets were simulated with study design parameters described in Table 2 for the low power alternative case, with at least 50 datasets for each SNP at each rank. Reduction provided (average difference between corrected and uncorrected estimate) was computed for the genome-wide bootstrap with correction (GW Boot w/ corr) and Normalized Maximum Likelihood (NMLE). Absolute bias (average difference between the estimate of log OR and the true log OR) and RMSE (root mean square error) was computed for the uncorrected (Naive), NMLE and GW bootstrap with correction methods.

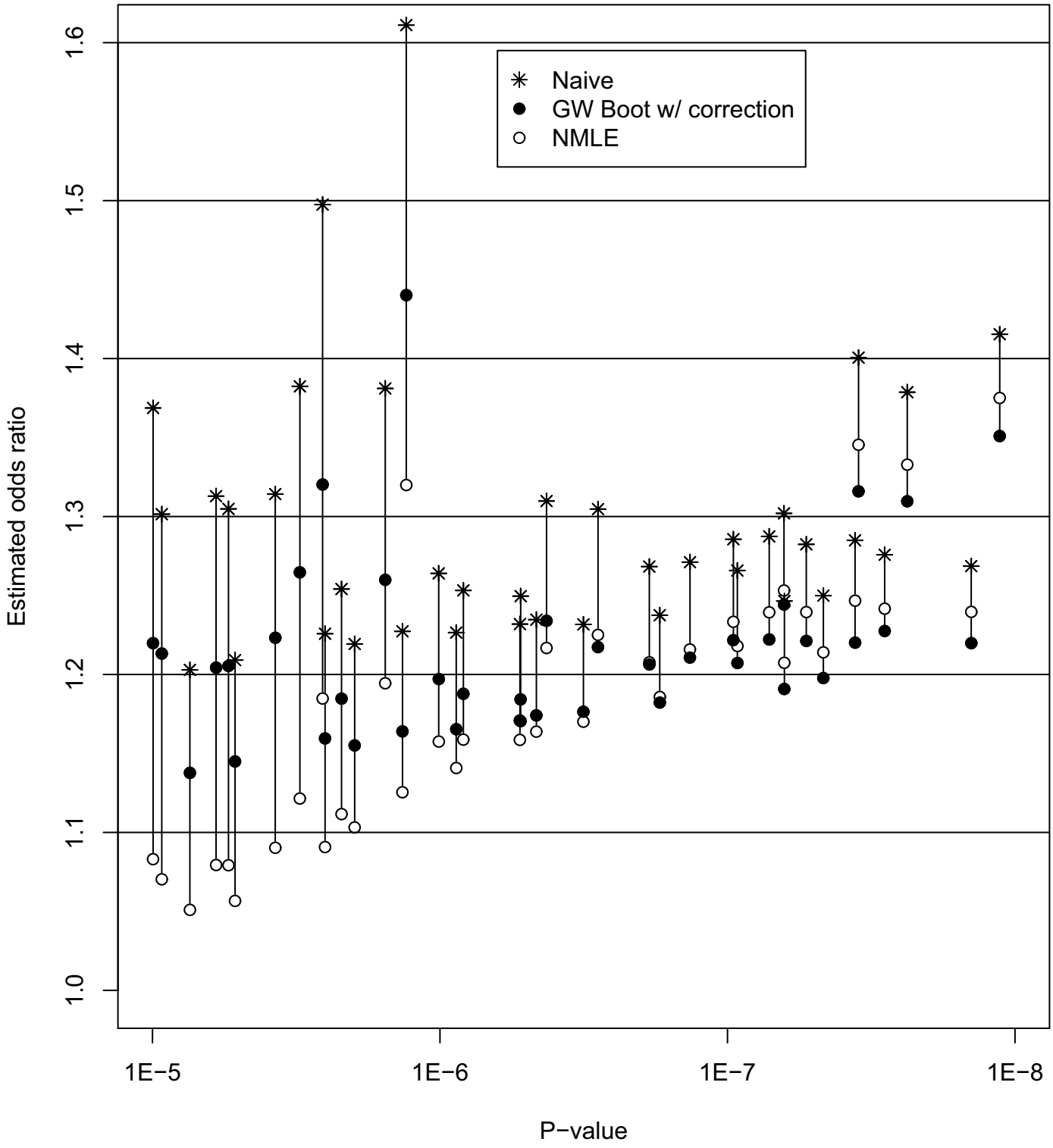

Figure 1. Uncorrected (Naive), normalized maximum likelihood (NMLE) and genome−wide bootstrap with correction (GW Boot w/ correction) bias−reduced estimates for a sample of SNPs meeting the WTCCC moderate association criterion (minimum p−value of the trend test (1 df) and genotypic test (2 df) less than 1E−5). NMLE and GW Boot estimates computed from WTCCC discovery samples using threshold of 1E−5. Vertical lines connect the WTCCC Naive estimate with corresponding bias−reduced estimates.

Top of Figure, Faye L, Figure 2

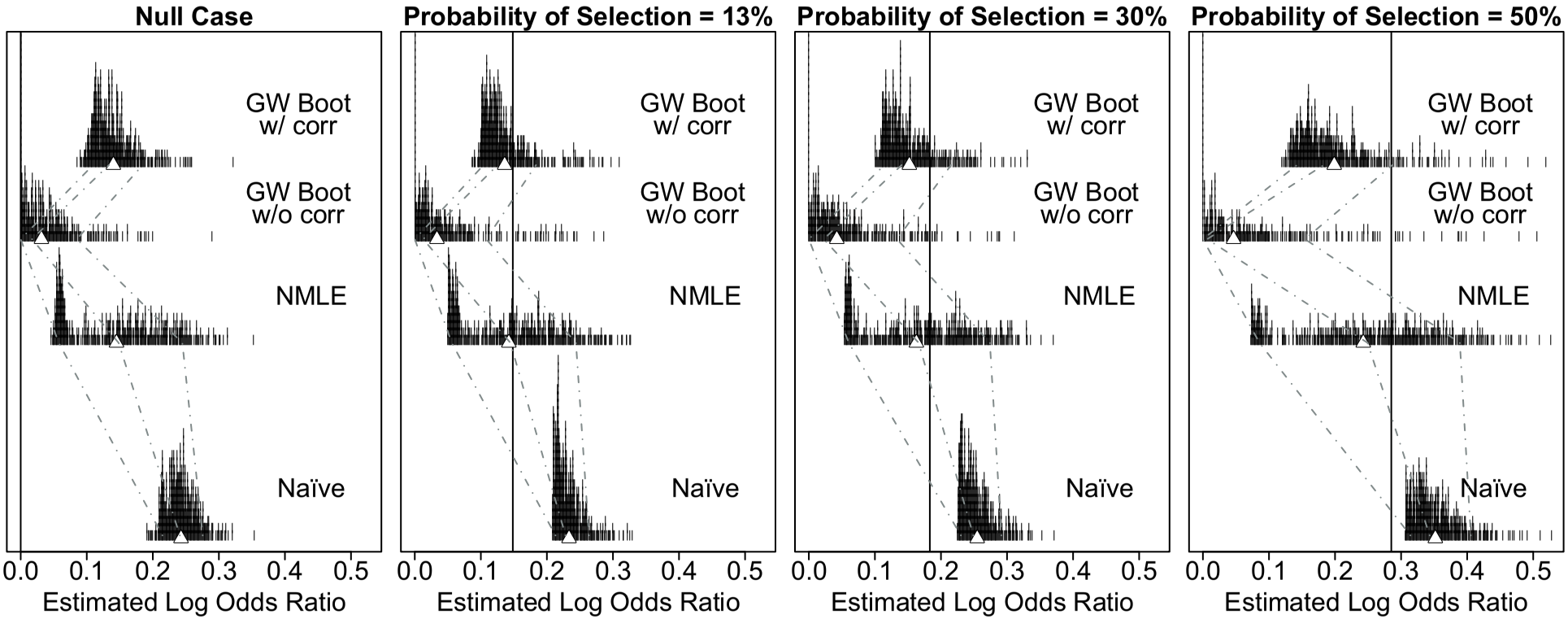

Figure 2. Genome−wide simulation results for effect size estimation in null and low power alternative case. Histograms of genetic effect estimates using the naïve, normalized maximum likelihood (NMLE) and genome−wide bootstrap without correction (GW Boot w/o corr) and with correction (GW Boot w/ corr) from 500 simulated case−control datasets. Open triangle is the mean. The far left plot is the null case for all false positive SNPs with MAF 25−50%. From left to right starting from the second plot, the probability of selection for each SNP (simulated under the alternative) is 13%, 30% and 50%. Dashed lines connect the 10th, 50th and 90th percentiles of the naïve estimates to their corresponding bias−reduced estimates. Simulation study parameters indicated in Table II.

Top of Figure, Faye L, Figure 3

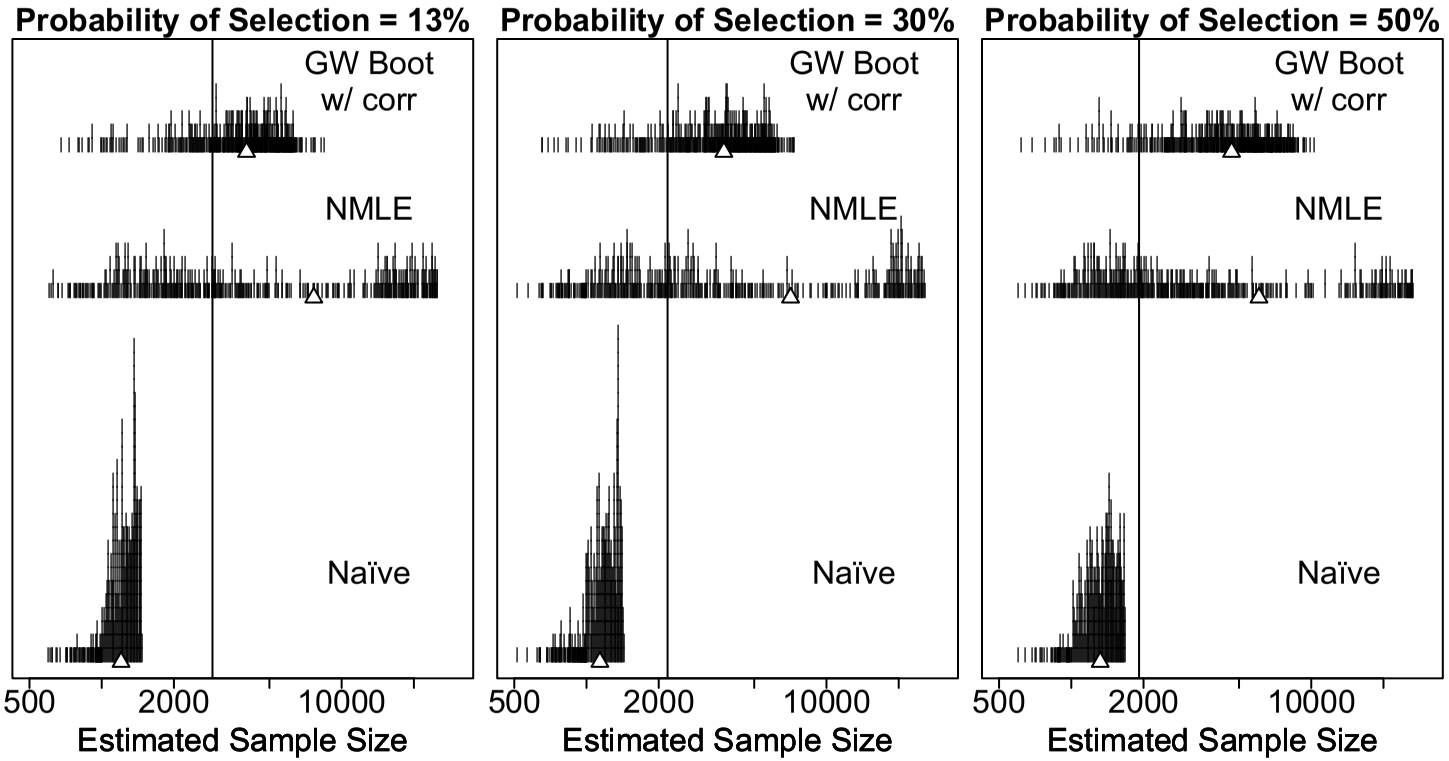

Figure 3. Genome−wide simulation results for replication sample size estimation in low power alternative case. Histograms of estimated sample size required for replication study with 80% power using the uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome−wide bootstrap with correction (GW Boot w/ corr) from 500 simulated case−control datasets. Open triangle is the mean. The probability of selection for each simulated SNP is 13%, 30% and 50% from left to right. Simulation study parameters indicated in Table II.

Table S1. Genome-wide simulation: performance of bias reduced point estimators for log OR

Alternative Case - High Power				Relative Bias (log OR)				RMSE			
	MAF	OR	Pr(p<α)	Naïve	NMLE	GW br2 w/o corr	GW Boot w/ corr	Naïve	NMLE	GW br2 w/o corr	GW Boot w/ corr
rs6679677	0.13	1.35	60%	0.21	-0.09	-0.62	-0.18	0.08	0.12	0.21	0.08
rs9272346	0.29	1.25	70%	0.23	-0.05	-0.50	-0.12	0.06	0.08	0.13	0.06
rs12708716	0.33	1.29	77%	0.08	-0.12	-0.48	-0.21	0.04	0.08	0.14	0.08
rs11171739	0.45	1.29	90%	0.06	-0.08	-0.26	-0.20	0.04	0.08	0.11	0.08
rs17696736	0.46	1.35	99%	0.04	0.01	-0.53	-0.10	0.04	0.06	0.15	0.07

Null Case (OR = 1)	Absolute Bias (log OR)				RMSE			
MAF	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr
0.05-0.1	0.41	0.21	0.03	0.14	0.42	0.22	0.05	0.15
0.1-0.2	0.30	0.16	0.02	0.12	0.31	0.17	0.03	0.14
0.2-0.3	0.25	0.14	0.03	0.10	0.25	0.14	0.04	0.11
0.3-0.4	0.23	0.14	0.04	0.12	0.24	0.14	0.06	0.13
0.4-0.5	0.23	0.13	0.03	0.14	0.23	0.14	0.06	0.15

Note: 500 genome-wide case control datasets where the SNP of interest (in null case, SNP with MAF of interest) was significant at p-value threshold α=5E-7 were simulated using the WTCCC T1D genotypes and OR as indicated. Relative bias (average difference between estimated and true log OR divided by true log OR) and root mean squared error (RMSE) were computed for the uncorrected (Naïve), Normalized Maximum Likelihood (NMLE), genome-wide bootstrap without correction (GW Boot w/o corr) and genome-wide bootstrap without correction (GW Boot w/ corr) estimates for true positive SNPs. Absolute bias (average difference between estimated and true log OR) and RMSE was calculated for false positive SNPs. Pr(p<α) is the probability of obtaining a p-value below threshold α for each SNP. Detailed descriptions of the method are in Section 4.

Table S2. Genome-wide simulation: Proportion of estimates within 10% or 25% of true effect size

Alternative Case (Low Power)					Proportion of estimates within 10% of true logOR				Proportion of estimates within 25% of true logOR			
SNP	MAF	log OR	OR	Power	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr
rs12708716	0.33	0.15	1.16	7%	0	0.13	0.01	0.31	0	0.33	0.04	0.83
rs11171739	0.45	0.15	1.16	11%	0	0.13	0.02	0.19	0	0.30	0.04	0.62
rs17696736	0.46	0.15	1.16	13%	0	0.13	0.01	0.29	0	0.33	0.05	0.77
rs9272346	0.29	0.18	1.20	30%	0	0.18	0.04	0.17	0.09	0.42	0.07	0.50
rs6679677	0.13	0.29	1.34	49%	0.09	0.18	0.02	0.07	0.64	0.44	0.05	0.24

Null Case	Proportion of estimates between 0 and log(1.1)				Proportion of estimates between 0 and log(1.2)			
MAF	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr	Naïve	NMLE	GW Boot w/o corr	GW Boot w/ corr
0.05_to_0.1	0	0.13	0.87	0	0	0.78	1	0.43
0.1_to_0.2	0	0.67	1.00	0	0	0.79	1	0.69
0.2_to_0.3	0	0.66	0.93	0	0	0.91	1	0.93
0.3_to_0.4	0	0.54	0.89	0.01	0	0.81	0.98	0.89
0.4_to_0.5	0	0.39	0.91	0.02	0	0.71	0.97	0.91

Note: The proportion of estimates within 10% (25%) of the true log odds ratio is the proportion of uncorrected (Naïve) or corrected (AML, NMLE, SS Bootstrap) estimates that fall between 0.9β and 1.1β (.75β and 1.25β) where β is the true log odds ratio for the SNP. Simulation study and estimators are described in Table II.

Table S3. Genome-wide simulation: Sample size estimated using bias-reduced methods for replication study with 80% power.

SNP	Actual sample size required for 80% power	Mean estimated sample size for 80% power		
		Naïve	NMLE	GW Boot w/ corr
rs12708716	3096	1169	8283	4005
rs11171739	2944	1246	8485	4332
rs17696736	2895	1198	6975	3937
rs9272346	2177	1141	6817	3814
rs6679677	1916	1320	5982	4652

Note: Actual sample size is sample size required so that the replication study achieves 80% power for the particular SNP. Mean estimated sample size for 80% power is average replication study sample computed from uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/ corr) bias-reduced genetic effect estimates. Simulation study and estimators are described in Table II.

Table S4. Genome-wide simulation: Sample size estimated for false positive SNPs using bias-reduced methods for replication study with 80% power

		Mean estimated sample size for 80% power			Proportion of SNPs for which sample size is adequate		
MAF	Actual sample size required for CI wide enough to exclude OR = 1.15	Naïve	NMLE	GW Bootstrap	Naïve	NMLE	GW Bootstrap
0.05 - 0.1	16165	2027	18187	9119	0	0.65	0.31
0.1 - 0.2	3145	1545	17034	5343	0	0.84	0.97
0.2 - 0.3	2098	1412	14530	4689	0	0.95	0.95
0.3 - 0.4	1759	1355	11620	4193	0	0.86	0.93
0.4 - 0.5	1605	1255	8378	4216	0	0.78	0.94

Note: Actual sample size is sample size large enough so that the log OR CI width is twice log(1.15). Mean estimated sample size for 80% power is average replication study sample computed from naïve and bias-reduced genetic effect estimates for false positive SNPs. Proportion of SNPs for which sample size is adequate is the proportion of SNPs for which the sample size computed from the naive or bias-reduced estimate is larger than the actual sample size required. Simulation study parameters are described in Table II.

Table S5: Single-SNP simulation results: performance of bias-reduced point estimators and confidence intervals

		Relative bias				RMSE			
power	N	Naïve	AML	NMLE	SS Bootstrap	Naïve	AML	NMLE	SS Bootstrap
5%	1650	0.63	-0.04	-0.16	0.28	0.17	0.17	0.13	0.09
10%	2050	0.47	-0.07	-0.18	0.17	0.13	0.16	0.12	0.08
30%	2950	0.26	-0.06	-0.19	0.04	0.08	0.13	0.11	0.05
50%	3600	0.17	-0.05	-0.17	0.00	0.05	0.11	0.10	0.05
90%	5720	0.03	0.00	-0.11	-0.05	0.03	0.04	0.08	0.05
99%	10000	0.01	0.00	-0.02	-0.01	0.03	0.04	0.04	0.04

	95% CI coverage				mean 95% CI width			
power	Naïve	AML	NMLE	SS Bootstrap	Naïve	AML	NMLE	SS Bootstrap
5%	53%	93%	94%	94%	0.32	0.51	0.52	0.36
10%	76%	94%	96%	96%	0.28	0.46	0.46	0.32
30%	92%	93%	95%	98%	0.23	0.37	0.38	0.27
50%	96%	95%	96%	98%	0.21	0.32	0.33	0.25
90%	97%	96%	95%	89%	0.17	0.22	0.23	0.20
99%	95%	95%	95%	93%	0.13	0.13	0.14	0.15

Note: The true odds ratio of the SNP of interest is 1.3 (log(1.3)=0.26) with MAF of 0.25, and the number of cases equals the number of controls (=N/2). 1,000 datasets significant at the p-value threshold of 1E-6 were simulated for each combination of parameters. The relative bias is the average difference between estimated and true log OR divided by true log OR (i.e. 1.3) averaged over 1,000 replicates. RMSE is the square root of the mean squared error. CI coverage is the percentage of the 1,000 CIs that covered the true genetic effect value. CI width is the mean width of the confidence interval. Naïve is the uncorrected estimate; AML, NMLE, and SS Bootstrap are bias-reduced estimates. Detailed descriptions of the methods are in Appendix D of the supplemental information.

Table S6: Single-SNP simulation results: Proportion of estimates within 10% or 25% of true log odds ratio

Alternative Case		Proportion of estimates between 0.262 +/- 0.0262				Proportion of estimates between 0.262 +/- 0.0655			
Power	N	Naïve	AML	NMLE	SS Bootstrap	Naïve	AML	NMLE	SS Bootstrap
5%	1650	0	0.09	0.11	0.21	0	0.20	0.25	0.56
10%	2050	0	0.08	0.14	0.43	0	0.18	0.31	0.67
30%	2950	0.01	0.13	0.19	0.34	0.57	0.38	0.43	0.80
50%	3600	0.34	0.21	0.19	0.31	0.76	0.53	0.49	0.74
99%	10000	0.58	0.58	0.54	0.53	0.96	0.95	0.90	0.91
Null Case		Proportion of estimates between 0 and 0.095				Proportion of estimates between 0 and 0.182			
	N	Naïve	AML	NMLE	SS Bootstrap	Naïve	AML	NMLE	SS Bootstrap
	1000	0	0.99	0.44	0	0	0.99	0.75	0.56

Note: The proportion of naive and bias reduced estimates of log OR for true positive SNPs that fall within 10% (left) or 25% (right) of the true log odds ratio (0.262) and the proportion of estimates for false positive SNPs that fall between 0 and 0.095 (left, corresponds to OR of 1 and 1.1) and 0 and 0.182 (right, corresponds to OR of 1 and 1.2) were computed for datasets simulated as described in Table II.

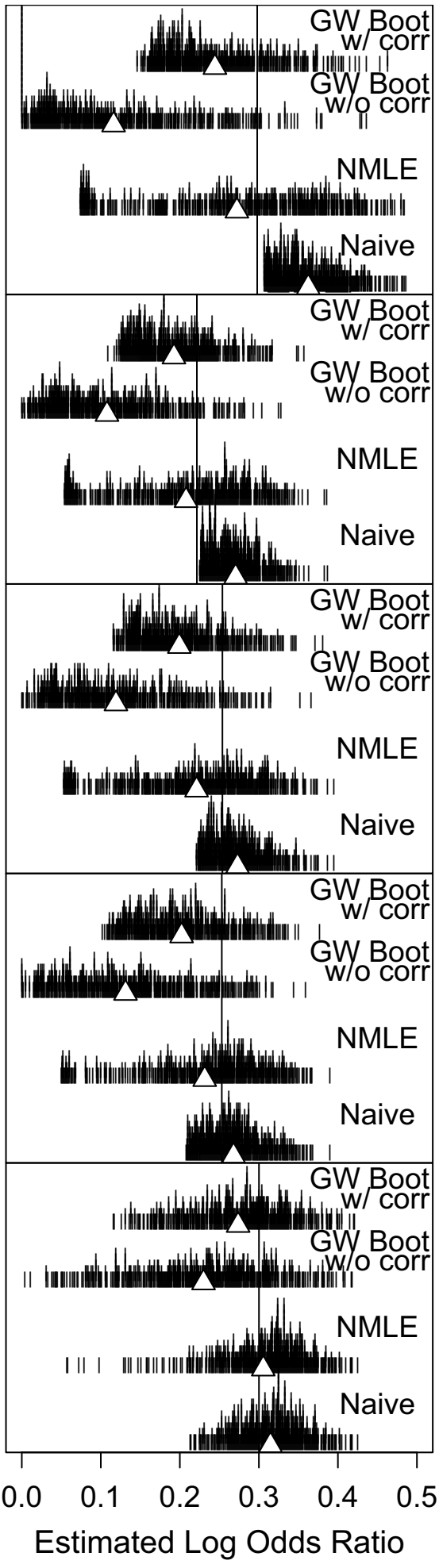

Figure S1. Genome−wide simulation with threshold−based selection (alternative case – high power). Histograms of estimates of genetic effect using the uncorrected (Naïve), normalized maximum likelihood (NMLE), genome−wide bootstrap without correction (GW Boot w/o corr) and with correction (GW Boot w/ corr) methods for 500 simulated significant case−control datasets. Open triangle is the mean. The probability of selection for each simulated SNP is 60%, 70%, 80%, 90%, 99% from top to bottom. Simulation study parameters are the same as for Supplemental Table S1.

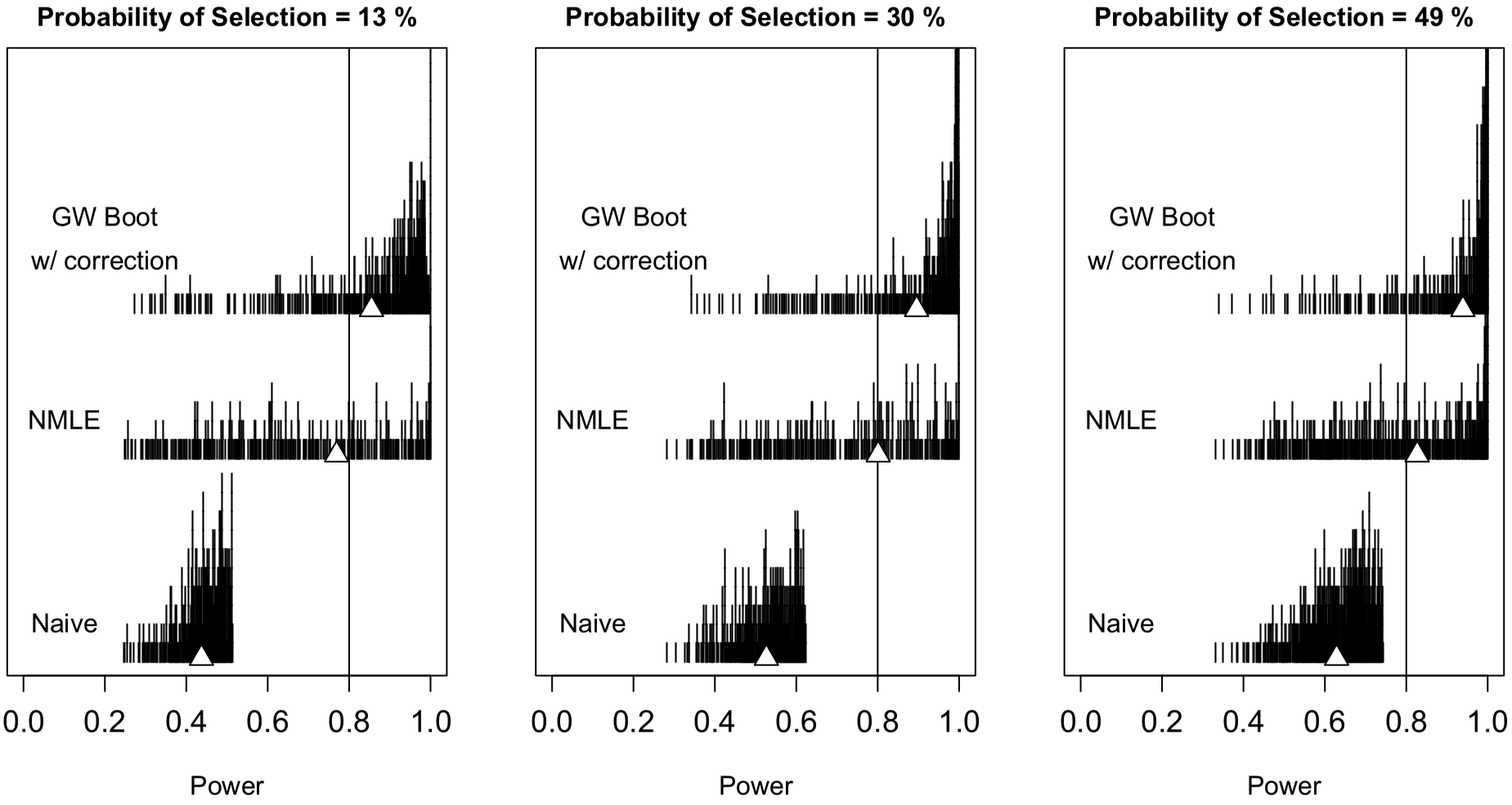

Figure S2. Genome−wide simulation results (alternative case – low power). Histogram of probability of selection in replication study when original genetic effect estimate is used to compute sample size required for replication study with 80% power. Sample size computed using the uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome−wide bootstrap with correction (GW Boot w/ correction) methods for 500 simulated case−control datasets. Open triangle is the mean. The probability of selection in the original study for each simulated SNP is 13%, 30% and 50% from left to right. Simulation study parameters are the same as for Table II.

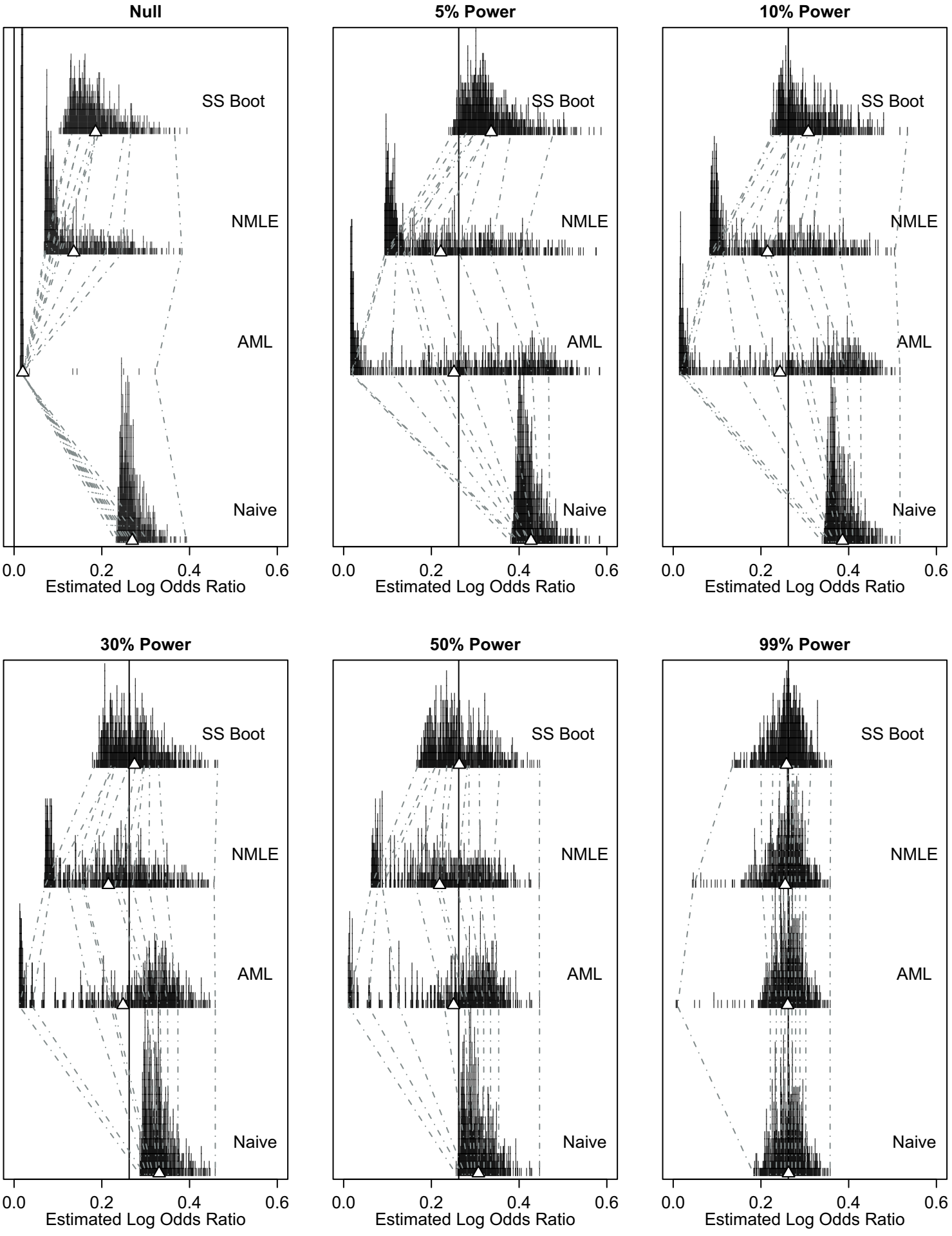
Null
5% Power
10% Power
30% Power
50% Power
99% Power
SS Boot
NMLE
AML
Naive
0.0
0.2
0.4
0.6
Estimated Log Odds Ratio

Figure S3. Single-SNP simulation results. Histograms of uncorrected (Naïve), adjusted median likelihood (AML), normalized maximum likelihood (NMLE) and single-SNP bootstrap (SS Boot) bias reduced estimates from 500 simulated significant case-control datasets. The power for each simulation is null case, 5%, 10%, (top row) 30%, 50%, 99% (bottom row) from left to right. Simulation study parameters are the same as for Supplemental Table S5. Open triangle is the mean, dashed lines connect the deciles, minimum and maximum of the naïve estimates to their corresponding NMLE and SS bootstrap bias-reduced estimates.

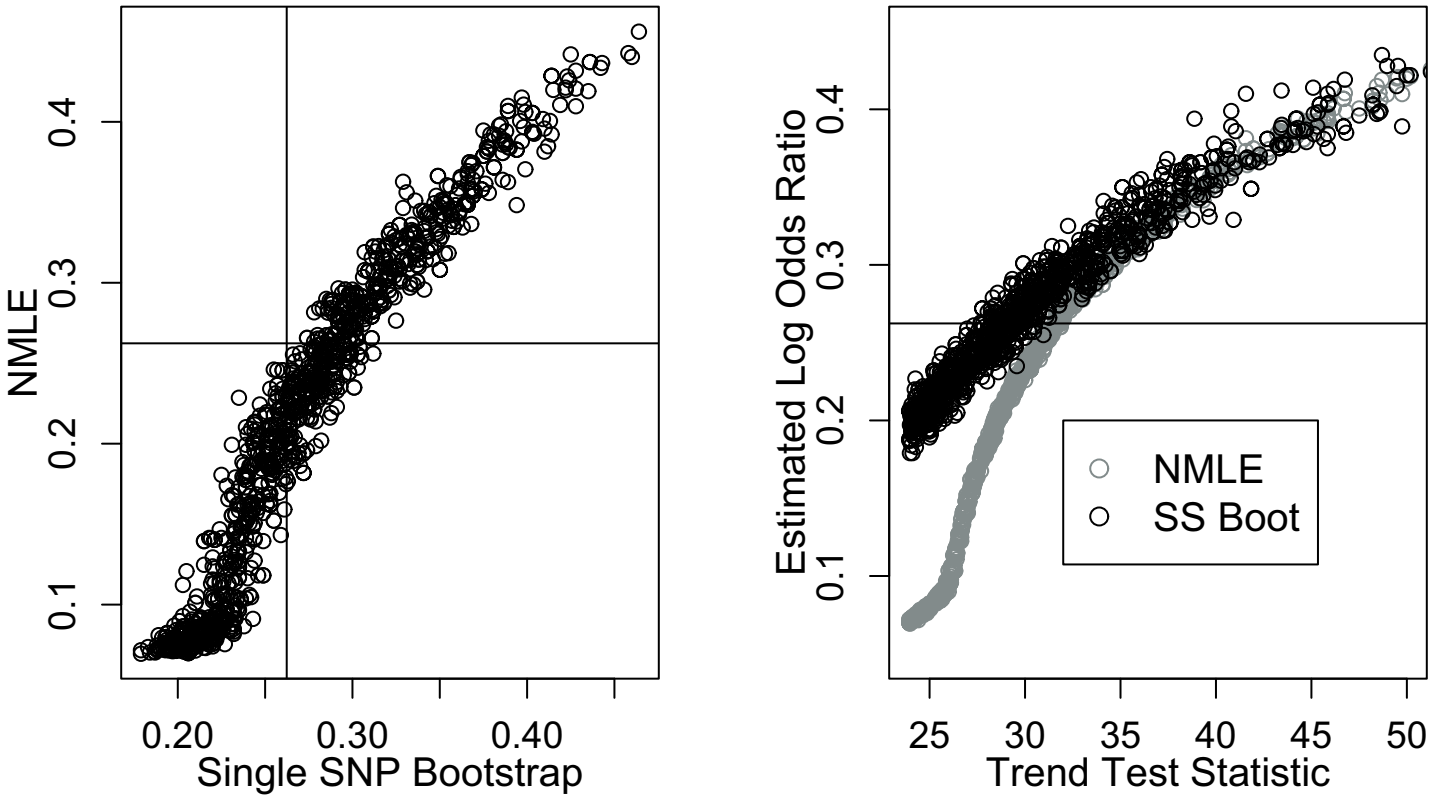

Figure S4. Single−SNP simulation comparison of normalized maximum likelihood (NMLE) and single−SNP bootstrap (SS Boot) estimates of log odds ratio from 1000 simulated significant case−control datasets (left) and comparison of relationship of NMLE and SS bootstrap with trend test statistic used for threshold−based selection. Vertical and horizontal lines are the true log odds ratio. Power of simulation is 30%, odds ratio is 1.3 and minor allele frequency is 0.25, sample size is 2950 and the criterion for significance is trend test p−value less than 1E−6.

The likelihood estimates generally had higher variance than either of the bootstrap methods. When the test statistic is large, the relationship between the likelihood estimate and test statistic is approximately linear, however there is an inflection−point below which estimates are drastically reduced. The consequence is that small changes in the data can result in large changes in the likelihood estimate when the test statistic is close to this threshold. For instance, in the single−SNP simulation where power was 0.05 (Supplemental Figure S3), naïve log odds ratio estimates ranging from 0.43 to 0.44 corresponded to a moderate range of bootstrap estimates (0.29 to 0.38) but to a much larger range of likelihood estimates (0.14 to 0.32).

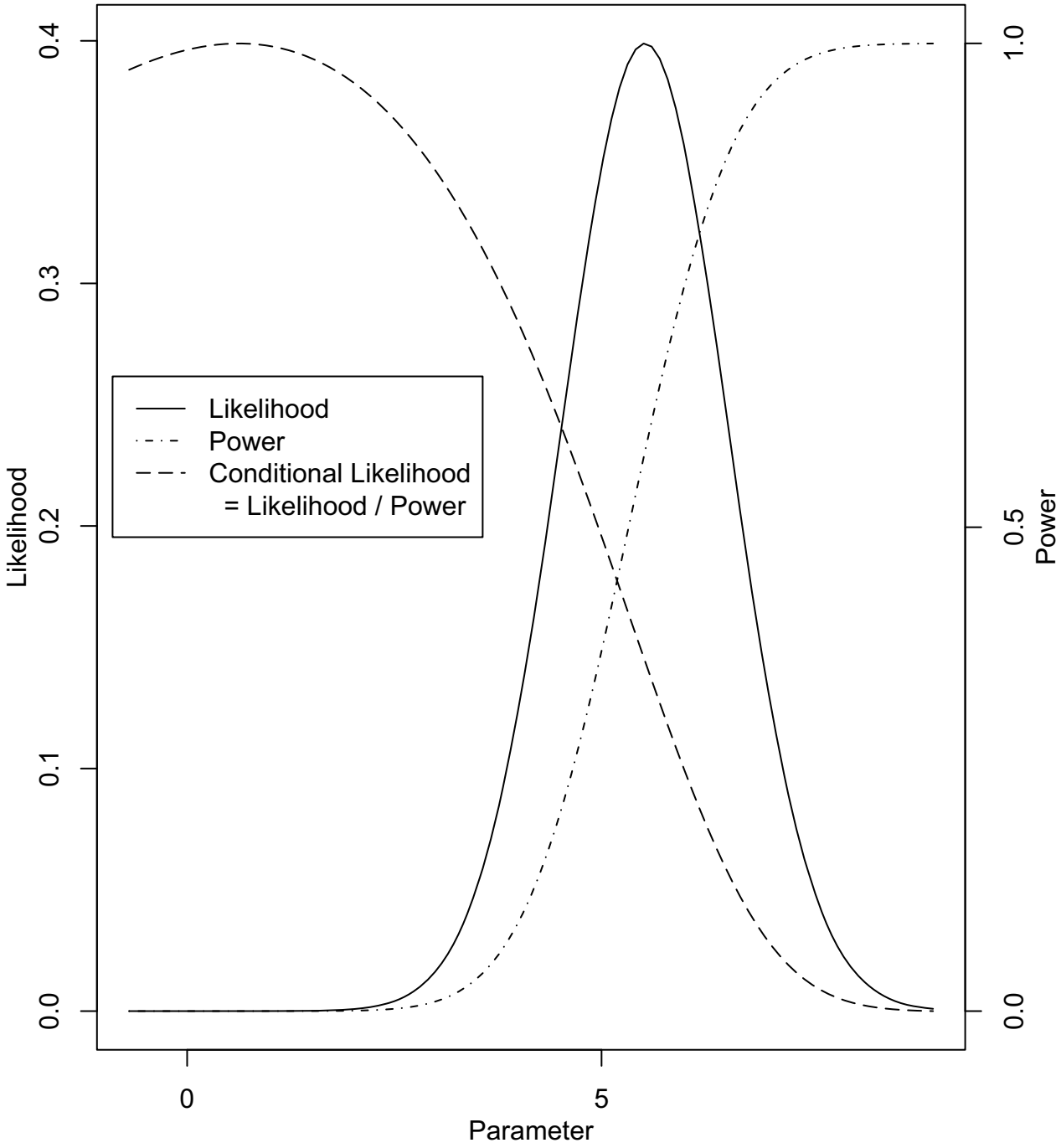

Figure S5. Likelihood, power and conditional likelihood (likelihood conditional on two−sided Wald test significant at p−value threshold of 1E−7) for the mean parameter of the normal distribution where standard deviation is 1. Observed data for the likelihood is Zobs= 5.5. This corresponds to a test statistic for an estimated odds ratio of 1.3 and standard deviation for the estimate of 0.0475.