Biometrics 75, 163–171
March 2019

DOI: 10.1111/biom.12957

Marginal Screening of 2×2 Tables in Large-Scale Case-Control Studies

Ian W. McKeague[*] **and Min Qian**[**]

Department of Biostatistics, Columbia University, 722 W. 168th St, New York, New York 10032, U.S.A.
[*]*email:* im2131@columbia.edu
[**]*email:* mq2158@columbia.edu

Summary. Assessing the statistical significance of risk factors when screening large numbers of 2×2 tables that cross-classify disease status with each type of exposure poses a challenging multiple testing problem. The problem is especially acute in large-scale genomic case-control studies. We develop a potentially more powerful and computationally efficient approach (compared with existing methods, including Bonferroni and permutation testing) by taking into account the presence of complex dependencies between the 2×2 tables. Our approach gains its power by exploiting Monte Carlo simulation from the estimated null distribution of a maximally selected log-odds ratio. We apply the method to case-control data from a study of a large collection of genetic variants related to the risk of early onset stroke.

Key words: Family-wise error rate; Genome-wide association studies; Multiple testing; Post-selection inference.

1. Introduction

Assessing statistical significance when screening large numbers of 2×2 tables arising in case-control studies is a challenging multiple testing problem. The problem becomes especially acute in large-scale genomic studies; see Dudoit and van der Laan (2008) and Sham and Purcell (2014) for comprehensive reviews. One standard approach is to search for significant associations by controlling the family-wise error rate (FWER) using Bonferroni corrections for marginal Fisher's exact or chi-squared tests, but that can be highly conservative, especially when the tables are far from independent. Another standard approach is to use permutation testing: case and control labels are reshuffled at random in order to estimate the "overal" null distribution of observed marginal p-values across the tables, see Churchill and Doerge (1994). However, there is a heavy computational burden with permutation testing.

In this article, we develop a potentially more powerful and computationally efficient approach for detecting the presence of associations by taking into account dependence between the 2×2 tables. Our proposed approach is based on estimating a standardized version of the maximal absolute log-odds ratio, and we develop a method to calibrate such a test statistic using Monte Carlo simulation from its estimated null distribution. We show that this approach leads to substantial computational savings over permutation testing, and superior power to all competing methods. Our approach is especially relevant to genome-wide association studies (GWAS), which attempt to establish associations of genetic variants with disease outcome. As described by Wu et al. (2010), standard GWAS involves genotyping a large number of SNPs, for large numbers of individuals with the disease (cases) as well as healthy controls, in order to identify individual loci that are associated with the outcome. Breast cancer, prostate cancer, and type 2 diabetes have been studied using this design.

There is an extensive literature on methods of testing for the association between a binary exposure variable and a binary outcome from case-control data. For an accessible introduction and references we refer the reader to Chapter 2 of Keogh and Cox (2014). The most thorough study of the asymptotic properties of tests of binary associations based on a single log-odds ratio is due to Kou and Ying (1996), and we will rely on their results extensively in the sequel. We also mention in passing that the problems of testing for homogeneity of odds ratios and estimating a common odds ratio from sequences of dependent 2×2 tables have been studied by Reis et al. (1999) and Kou and Ying (2006).

Typically in high-dimensional genomic screening it is the *marginal* association between a genotype and a phenotype that is of interest. Beyond the permutation testing and Bonferroni-correction methods mentioned earlier, marginal associations can be assessed using, among many other multiple testing procedures, higher criticism (Donoho and Jin, 2015), and sequential goodness-of-fit testing (Carvajal-Rodriguez et al., 2009), both of which are based on test statistics constructed from a set of (almost independent) p-values. For background on the marginal screening approach relevant to the present article, we refer the interested reader to McKeague and Qian (2015a), in which an adaptive resampling test (ART) was developed for screening multiple predictors for an association with a continuous response in the standard linear regression setting. Here, we develop a parallel screening test that is specific to the case-control setting with multiple binary predictors and a binary response. In the sequel, we refer to our proposed new screening procedure as the *binary screening test* (BST). The test statistic used in ART

has non-regular limiting behavior at the null hypothesis of no association, and the calibration of the test was designed to adapt to this non-regularity, albeit at a high computational cost. We show that the BST statistic has the same type of non-regularity, but the simpler form of its limiting distribution allows the direct use of Monte Carlo simulation, making the approach much more tractable and appealing for application in large-scale genomic studies.

The knockoff approach recently developed by Candès et al. (2018), a variation of the knockoff filter of Barber and Candès (2015), provides rigorous control of false discovery rate (FDR) in the setting of high-dimensional logistic regression models, and is applicable to case-control data. However, this approach has not been developed for FWER control. FDR control is most useful in situations where large numbers of predictors are known to be active, as with the Crohn's disease SNP data set analyzed by these authors, rather than testing for the presence of a single active predictor. FWER control is more stringent than FDR control, as noted by Efron (2010) (Section 4.4). Moreover, the knockoff method involves the selection of active predictors *conditionally on all other predictors*, which, as mentioned above, is not a central concern in high-dimensional genomic screening. Another approach that is potentially relevant in our case-control setting is the method of conditional post-selection inference for ℓ_1–penalized likelihood models studied by Taylor and Tibshirani (2018), but a theoretical justification of this approach is only available in the linear model case, and the logistic regression case is only understood at a heuristic level.

The article is organized as follows. Background on multiple 2×2 tables and the sampling design of case-control studies is given in Section 2.1. The proposed binary screening test is developed in Section 2.2, along with an asymptotic result used to validate the approach. Robust confidence intervals for the maximal log-odds ratio are described in Section 2.3. Numerical examples based on simulated data are studied in Section 3.1, and in Section 3.2, we provide an application to SNP data from the Risk Assessment of Cerebrovascular Events Study (RACE, 2017). Various extensions of our approach are discussed in Section 4. Technical assumptions are placed in the Appendix and proofs in Section 2 of the Supplementary Materials.

2. Marginal Screening and Odds Ratios

2.1. *Background*

Let disease status be indicated by $D \in \{0, 1\}$ and consider a p-dimensional vector of $\{0, 1\}$-valued risk factors $\boldsymbol{W} = (W_1, \ldots, W_p)^{\mathsf{T}}$. We are interested in testing whether at least one of the risk factors is marginally associated with D. The sampling design is assumed to be that of a standard unmatched case-control study: independent random samples of M_1 cases and M_2 controls, and total sample size $N = M_1 + M_2$.

Throughout we use the notation N_{1k} and N_{2k} for the numbers of subjects observed to be exposed or unexposed to the kth risk factor, respectively, and X_k for the number of exposed cases. It is convenient to display the resulting data as 2×2 tables cross-classifying the exposure status of the cases and controls:

	Cases	Controls	
Exposed	X_k	$N_{1k} - X_k$	N_{1k}
Unexposed	$M_1 - X_k$	$X_k + N_{2k} - M_1$	N_{2k}
	M_1	M_2	N

The conditional distribution of X_k given fixed margins in the above table is noncentral hypergeometric, with noncentrality parameter given by the odds ratio of disease probability for exposed ($W_k = 1$) versus non-exposed ($W_k = 0$), namely $\theta_k = [p_{11k}/p_{01k}]/[p_{10k}/p_{00k}]$, where $p_{lmk} = P(D = l|W_k = m)$. As is well-known, the conditioning can be reversed in each probability in the definition of θ_k without changing its value, so it is estimable from the case-control data. The maximum likelihood estimator of θ_k when only the margin (M_1, M_2) is fixed (as with the actual data) is given by the empirical odds ratio $\hat{\theta}_k = [X_k(X_k + N_{2k} - M_1)]/[(N_{1k} - X_k)(M_1 - X_k)]$, $k = 1, \ldots, p$.

When both margins are fixed, the MLE of θ_k does not have a closed form expression, so our focus is naturally on the computationally simple estimator $\hat{\theta}_k$. To develop asymptotic theory for $\hat{\theta}_k$, however, it is more tractable initially to study the MLE when the data are generated with both margins fixed, by exploiting the noncentral hypergeometric structure mentioned above. As mentioned in Section 1, a thorough study based on this approach for a single 2×2 table was carried out by Kou and Ying (1996); we will make use of several of their results.

2.2. *Binary Screening Test*

In this section, we introduce the proposed test for detecting whether any of the risk factors are associated with the disease. Formally this means designing a test of the null of whether all $\theta_k = 1$ against the alternative that at least one $\theta_k \neq 1$. The standard approach is to break such a test down into multiple Bonferroni-adjusted tests, but a potentially much more powerful approach is to recast the problem so its implicit variable selection features emerge clearly. That is, we reformulate the problem as testing

$$H_0 : \theta_0 = 1 \text{ versus } H_a : \theta_0 \neq 1, \tag{1}$$

where $\theta_0 = \theta_{k_0}$ with $k_0 \in \arg\max_k |\log \theta_k|/\sigma_k$ and $\sigma_k > 0$ is a prescribed sequence of normalizing constants (defined in (A.1)). The parameter θ_0 represents the odds ratio of the risk factor k_0 having maximal marginal effect (if there is an effect) in terms of standardized units.

The inclusion of the normalizing constants σ_k is technically redundant in the sense that neither hypothesis depends on them, but they provide the flexibility to balance the varying precision in estimating the marginal log-odds of each risk factor. Accordingly, a reasonable strategy is to arrange for σ_k^2 to be proportional to the asymptotic variance of $\log \hat{\theta}_k$, as we do in the sequel. An analogous situation arises in the marginal screening of predictors in linear regression, where pre-standardization of all predictors (to have zero mean and

unit variance) is used to ensure scale-invariance and equal precision of the marginal slope parameter estimates, see McKeague and Qian (2015b) for discussion.

The proposed test statistic is $T_N = \log \hat{\theta}_N$, where $\hat{\theta}_N = \hat{\theta}_{\hat{k}_N}$ is the sample version of θ_0 with $\hat{k}_N \in \arg\max_k |\log \hat{\theta}_k|/\hat{\tau}_k$ estimating k_0, and the normalizer $\hat{\tau}_k$ taken as the standard error of $\log \hat{\theta}_k$ given (in terms of variance) by

$$\hat{\tau}_k^2 = \frac{1}{X_k} + \frac{1}{X_k + N_{2k} - M_1} + \frac{1}{N_{1k} - X_k} + \frac{1}{M_1 - X_k}.$$

The normalization by $\hat{\tau}_k$ has the same asymptotic effect on the test statistic as normalization by σ_k, since $\hat{\sigma}_k^2 = N\hat{\tau}_k^2$ can be shown to be a consistent estimator of σ_k^2.

Calibration of the test is challenging since T_N is non-regular—its asymptotic distribution is non-Gaussian and discontinuous at the null, as we now show by analyzing the local asymptotic behavior of T_N. The local model specifies the vector of marginal odds ratios $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T}$ in the neighborhood of some limiting value $\boldsymbol{\theta}^{(0)}$ by $\log \boldsymbol{\theta} = \log \boldsymbol{\theta}^{(0)} + \boldsymbol{b}/\sqrt{N}$, where $\boldsymbol{b} = (b_1, \ldots, b_p)^\mathsf{T} \in \mathbb{R}^p$ is a vector of local parameters. Here $\boldsymbol{\theta} = \boldsymbol{\theta}^{(N)}$ is indexed by the total sample size $N = M_1 + M_2$, although we will suppress N in the notation. The hypotheses (1) now also depend on sample size: $H_0 : \theta_N = 1$ versus $H_a : \theta_N \neq 1$, where $\theta_N = \theta_{k_N}$, $k_N \in \arg\max_k |\log \theta_k|/\tau_k$, and

$$\tau_k^2 = \frac{1}{E(X_k)} + \frac{1}{E(X_k + N_{2k} - M_1)} + \frac{1}{E(N_{1k} - X_k)} + \frac{1}{E(M_1 - X_k)},$$

with the expectations being conditional on M_1, M_2, N_{1k}, N_{2k}. The following result provides a theoretical basis for the procedures introduced in the sequel.

THEOREM 1. *Suppose Conditions (C1)–(C3) in the Appendix hold, the index* $k_0 = \arg\max_k |\log \theta_k^{(0)}|/\sigma_k$ *is unique when* $\boldsymbol{\theta}^{(0)} \neq \mathbf{1}$, *and* $\tilde{k} = \arg\max_k |b_k|/\sigma_k$ *is unique when* $\boldsymbol{\theta}^{(0)} = \mathbf{1}$ *and* $\boldsymbol{b} \neq \mathbf{0}$. *Then, under the sequence of local models,*

$$\sqrt{N}(\log \hat{\theta}_N - \log \theta_N) \xrightarrow{d} \begin{cases} \sigma_{k_0} Z_{k_0} & \text{if } \boldsymbol{\theta}^{(0)} \neq \mathbf{1}, \\ \sigma_K Z_K + b_K - b_{\tilde{k}} & \text{if } \boldsymbol{\theta}^{(0)} = \mathbf{1}, \end{cases}$$

where $(Z_1, \ldots, Z_p)^T$ *is a normally distributed random vector with mean-zero variance-one marginal distributions and correlation matrix* $\boldsymbol{C_X}$ *defined in Condition (C2),* $K = \arg\max_{k=1,\ldots,p}(Z_k + b_k/\sigma_k)^2$, *and* $(\sigma_1, \ldots, \sigma_p)^T$ *is defined in (A.1).*

The presence of the local parameters in the limiting distribution in Theorem 1 establishes the non-regularity property mentioned earlier. The cause of this non-regularity is post-selection of the empirical odds-ratio: replacing the maximally associated risk factor k_0 by an estimate distorts the limiting behavior of $\log \hat{\theta}_N$.

BST procedure

1. Obtain Monte Carlo draws from the estimated limiting null distribution of $\sqrt{N} \log \hat{\theta}_N$ by plugging-in empirical estimates of $\boldsymbol{C_X}$ and σ_k (see Appendix, Remarks 1 and 2), and setting $\boldsymbol{b} = \mathbf{0}$, $\boldsymbol{\theta}^{(0)} = \mathbf{1}$ in Theorem 1.
2. Calibrate the test at level α by using critical values c_l and c_u given by the lower and upper $\alpha/2$-quantiles of 10,000 draws obtained in step 1.
3. Reject the null hypothesis and conclude that there is at least one significant risk factor if $\sqrt{N} \log \hat{\theta}_N$ falls outside the interval $[c_l, c_u]$.

2.3. *Robust Confidence Intervals*

Another useful application of Theorem 1 is in obtaining Wald-type confidence intervals for $\log \theta_0$ (or for θ_0, by transformation) that are robust to the non-regularity at $\theta_0 = 1$, that is, taking the post-selection into account. To make use of the limiting distribution in Theorem 1, we need to make a sensible choice of the local parameter $\boldsymbol{b}$, which can be viewed as a tuning parameter determining proximity to H_0.

We discuss four types of $100(1-\alpha)\%$ confidence intervals based on the critical values c_l and c_u given in step 2 of the BST procedure, except now the $\hat{k}_N$-th local parameter, $b_{\hat{k}_N}$, is allowed to vary (rather than being set to zero like all the other components of $\boldsymbol{b}$). Each CI takes the form $(\log \hat{\theta}_N - c_u/\sqrt{N}, \log \hat{\theta}_N - c_l/\sqrt{N})$, where c_l and c_u depend on a choice of the value of $b_{\hat{k}_N}$, and are found using Monte Carlo simulation of the corresponding limiting distribution given in Theorem 1 (after plugging-in estimates of $\boldsymbol{C_X}$ and σ_k).

The first method (CI_0) is simply to use the same Monte Carlo calibration as BST (i.e., setting $b_{\hat{k}_N} = 0$); this method should perform well close to the null, but may have poor performance away from the null. The second method ($\mathrm{CI}_{\max}$) is to use the most conservative critical values c_l and c_u. That is, c_u is taken as the maximal upper $\alpha/2$-quantile over all values of $b_{\hat{k}_N}$, and c_l is the minimal lower $\alpha/2$-quantile. The Monte Carlo simulations now need to be carried out over a fine grid of values of $b_{\hat{k}_N}$, but in practice this grid can be confined to an interval that contains the bulk of the mass of $N(0, \hat{\sigma}^2_{\hat{k}_N})$, so when $b_{\hat{k}_N}$ is at the boundary of such an interval there is close agreement with the critical values given by the (regular) Gaussian limit under the non-local alternative. That is, $\mathrm{CI}_{\max}$ adapts to the non-regularity, yet agrees with the usual Wald-type confidence interval sufficiently far away from the null hypothesis of no association.

The third method ($\mathrm{CI}_{\mathrm{boot}}$) aims at a compromise between CI_0 that takes $b_{\hat{k}_N} = 0$ and the highly conservative $\mathrm{CI}_{\max}$ that uses the most extreme critical values c_l and c_u. This is done using bootstrap resampling to select a value of $b_{\hat{k}_N}$ that furnishes the most accurate values of c_l and c_u in terms of providing nominal coverage. Our fourth and recommended method ($\mathrm{CI}_{\mathrm{bag}}$) adapts the well-known technique of *bagging* to stabilize the behavior of $\mathrm{CI}_{\mathrm{boot}}$.

Procedure for $\mathrm{CI}_{\mathrm{boot}}$ and $\mathrm{CI}_{\mathrm{bag}}$

1. Find bootstrap estimates $\log \theta_N^*$ of $\log \theta_0$ from B bootstrap samples that are formed by combining separate bootstrap samples from cases and controls.
2. Select the value of $b_{\hat{k}_N}$ on the grid for which $(\log \hat{\theta}_N - c_u/\sqrt{N}, \log \hat{\theta}_N - c_l/\sqrt{N})$ contains closest to $100(1-\alpha)\%$ of the B bootstrap estimates $\log \theta_N^*$ in step 1.

3. $\mathrm{CI}_{\mathrm{boot}}$ is the interval based on the value of $b_{\hat{k}_N}$ selected in step 2.
4. $\mathrm{CI}_{\mathrm{bag}}$ is the interval formed by averaging the endpoints of a number of independent replications (say 5 or 10) of $\mathrm{CI}_{\mathrm{boot}}$.

3. Numerical Examples

In this section, we study the performance of BST and the corresponding robust confidence intervals using simulated data, and give a real data example.

3.1. *Simulation Study*

We compare the performance of BST to four competing methods:

1) *Bonferroni correction* (Bonf): Marginal chi-squared tests with Bonferroni correction to the smallest p-value (equivalent to Bonferroni–Holm method in this case).
2) *Higher criticism* (HC): Donoho and Jin (2004, 2015)
3) *Sequential goodness-of-fit metatest* (SGoF): Carvajal-Rodriguez et al. (2009)
4) *Permutation test* (Perm): The smallest observed p-value among the p marginal chi-squared tests, denoted $p_{\min}$, is used as a test statistic. To estimate its null distribution, the disease status D is randomly permuted R times, resulting in an empirically adjusted p-value $p^* = (r+1)/(R+1)$, where r is the number of values of $p_{\min}$ based on the reshuffled data that are smaller than the observed value of $p_{\min}$, see Davison and Hinkley (1997), Section 4.3. We take $R = 1000$.

Simulation Models

A) $W_k \sim Ber(0.5)$ for $k = 1, \ldots, p$;
B) $W_k \sim Ber(0.6), k = 1, \ldots, p/2, W_k \sim Ber(0.55), k = p/2+1, \ldots, p$ for cases, and $W_k \sim Ber(0.5), k = 1, \ldots, p/2, W_k \sim Ber(0.6), k = p/2+1, \ldots, p$ for controls;
C) $W_1 \sim Ber(0.65)$, $W_2 \sim Ber(0.6)$, and $W_3 \sim Ber(0.55)$ for cases, $W_k \sim Ber(0.4), k = 1, 2, 3$ for controls, and $W_k \sim Ber(0.5), k = 4, \ldots, p$.

Model A represents the null, whereas B and C represent two types of alternatives. To specify the joint distributions of the risk factors W_k we will consider three scenarios: independent W_k, exchangeable correlation $\mathrm{Corr}(W_j, W_k) = 0.5$ for $j \neq k$, and AR(1) correlation $\mathrm{Corr}(W_j, W_k) = 0.5^{|j-k|}$. The total sample size is taken as $N = 200$ or 400, with a balanced design comprised of $M_1 = N/2$ cases and $M_2 = N/2$ controls. The number of risk factors p varies from 50 to 2000.

Histograms of $\log \hat{\theta}_N$ in each of the simulation models under the independent W_k scenario are displayed in Figure 1, where

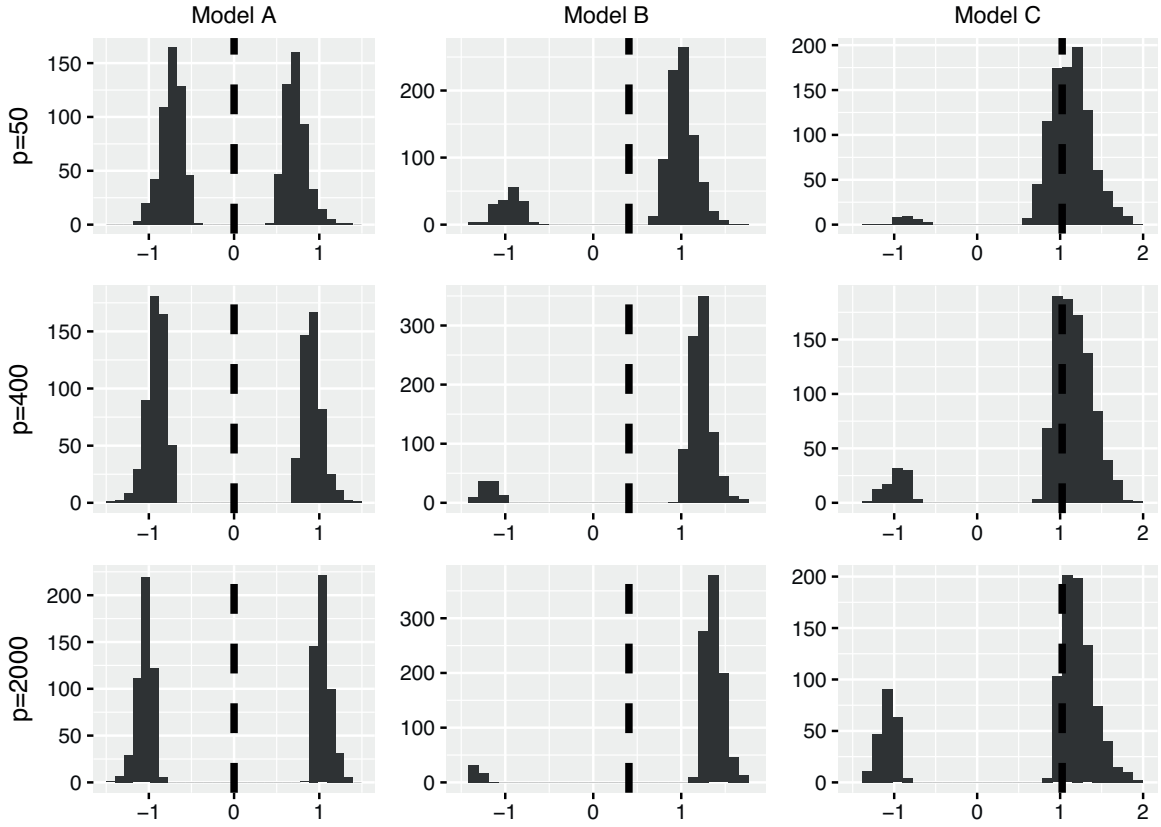

Figure 1. Histograms of $\log \hat{\theta}_N$ in each simulation model for independent risk factors, with $\log \theta_0$ indicated by the vertical dashed line, $N = 200$, $p = 50, 400$, and 2000. The odds ratios of the most active risk factors are $\theta_0 = 1$, 1.5, and 2.79 in models A (first column), B (second column), and C (third column), respectively.

we used 1000 Monte Carlo replications for each histogram. The symmetric bimodal behavior of $\log \hat{\theta}_N$ under the (null) model A is consistent with the result of Theorem 1 (with the local parameter set to 0). For model C, the chance of misidentifying the most active variable (W_1) increases with p, so the histogram is close to normal with mean $\log \theta_0$ for small p (as suggested by Theorem 1) but becomes increasingly bimodal for large p due to more inactive variables having (spuriously) negative estimates of $\log \theta_k$, thus causing a smaller peak to emerge on the left. Model B violates the uniqueness condition of Theorem 1 (that θ_0 correspond to a *unique* risk factor as indexed by k_0), causing the lack of centering of the right peak around $\log \theta_0$; this lack of centering becomes more extreme as p increases. The pronounced left peak in this case is due to the less active variables (W_k, $k = p/2, \ldots, p$) having negative $\log \theta_k$. Note, however, the violation of the uniqueness condition in model B does not affect the calibration of the BST (which is done under the null).

Empirical rejection rates for H_0 at the nominal 5% significance level (again based on 1000 Monte Carlo replications) are reported in Tables 1–3. For Model A, the tables provide type I error rates, or FWER in the case of Bonferroni, which should be compared with the 5% nominal rate; for Models B and C, the tables provide the power of each test. Model B represents a dense signal; note that the number of active predictors increases with p, so we expect the power to increase with p. Model C represents a relatively sparse signal, and the number of active predictors is fixed at 3, so we expect the power to decrease with p. BST has good control of type I error in all cases, although it is slightly anti-conservative for $N = 200$ and $p = 1000$ and 2000; note that this anti-conservativeness disappears for $N = 400$. Bonferroni is highly conservative and has low power, although it is less conservative in the case of independent covariates. HC and SGoF are both highly conservative in the independent and AR(1) cases (Tables 1 and 2), and fail to control type I error in the exchangeable correlation case (Table 3). BST is designed for detecting relatively sparse signals (as in Model C), and does not perform as well in terms of power as HC and SGoF for the dense signal in Model B, in the settings of Tables 1 and 2. The permutation method has good control of type I error throughout, but slightly lower power than BST in both the dense and sparse cases. The advantage of BST over permutation testing and Bonferroni is most evident in the exchangeable correlation case with large p.

Table 3 also presents the runtime (in seconds) of each method. Bonferroni, HC and SGoF have similar speed. BST is 10–50 times slower, but this is to be expected due to the computationally intensive simulation step. The permutation method is 10–80 times slower than BST, and becomes markely slower with larger N; the runtime of BST does vary appreciably with N.

Coverage rates and average widths of the 95% confidence intervals for $\log \theta_0$ based on 1000 Monte Carlo replications for exchangeable correlation structure are reported in Table 4, along with runtime. As expected, CI_0 performs well in model A (the null model), but fails to provide the nominal coverage rate under alternative. On the other hand, confidence intervals produced by $\mathrm{CI}_{\max}$ are much too conservative to provide useful information. $\mathrm{CI}_{\mathrm{boot}}$ (based on $B = 1000$) achieves nominal coverage rates in models A and C, but suffers from undercoverage in the dense signal case (model B) when p is large. This is not surprising since the maximal signal is not unique in model B, which causes instability in the calibration. $\mathrm{CI}_{\mathrm{bag}}$ (averaging the endpoints of 10 copies of $\mathrm{CI}_{\mathrm{boot}}$ with $B = 200$) improved on $\mathrm{CI}_{\mathrm{boot}}$ in all cases by furnishing higher coverage without increasing the average width of the interval.

Table 1

Empirical rejection rates (%) of H_0 *based on 1000 samples generated from models A, B, and C as the number of risk factors* p *ranges from* 50 *to* 2000*, for* $N = 200$*, and independent risk factors.*

Model	p	BST	Bonf	HC	SGoF	Perm
A	50	6.4	2.7	1.3	4.6	5.1
	100	5.5	3.0	1.6	1.7	5.1
	200	5.9	2.6	1.6	2.2	5.2
	400	5.8	1.4	0.8	0.7	4.3
	1000	7.4	2.5	0.8	0.2	3.5
	2000	7.0	1.5	0.4	0.0	3.8
B	50	65.3	48.9	95.8	96.2	63.1
	100	70.4	52.7	99.9	99.6	67.6
	200	81.5	63.3	100	100	77.2
	400	86.4	67.1	100	100	81.2
	1000	94.4	72.4	100	100	90.3
	2000	96.7	80.6	100	100	94.7
C	50	78.7	66.5	10.5	31.5	76.2
	100	68.2	57.9	11.2	11.4	66.4
	200	64.5	54.4	7.6	5.9	61.9
	400	55.3	40.8	4.2	2.1	50.5
	1000	49.7	33.7	3.7	0.1	42.9
	2000	42.7	29.3	3.0	0.0	35.8

Table 2

Empirical rejection rates (%) as in Table 1, $N = 200$*, except for risk factors having AR(1) correlation structure* $\mathrm{Corr}(W_j, W_k) = 0.5^{|j-k|}$.

Model	p	BST	Bonf	HC	SGoF	Perm
A	50	5.8	2.1	2.5	6.4	4.2
	100	6.4	2.9	3.0	3.8	5.6
	200	6.6	3.5	3.1	3.2	5.6
	400	5.7	2.0	1.5	2.0	4.9
	1000	6.6	2.3	1.0	0.0	3.8
	2000	7.9	2.8	1.0	0.2	5.2
B	50	57.2	40.0	87.1	89.3	53.8
	100	65.7	50.7	97.9	96.8	62.8
	200	73.9	57.3	100	99.9	69.0
	400	86.3	64.9	100	100	81.7
	1000	92.8	73.0	100	100	89.0
	2000	96.2	77.6	100	100	92.0
C	50	76.2	61.7	15.1	38.0	73.2
	100	65.2	55.5	17.5	15.0	62.8
	200	58.7	49.4	14.3	10.0	56.3
	400	51.8	37.9	9.5	4.5	48.7
	1000	45.6	31.0	6.5	1.2	39.9
	2000	41.7	27.4	4.2	0.0	33.9

Table 3

Empirical rejection rates (%) as in Table 1 except with exchangeable correlation structure $\mathrm{Corr}(W_j, W_k) = 0.5$ *for* $j \neq k$*, and* $N = 200$ *and* $N = 400$*. The average runtime (seconds) to conduct a single test is given in parentheses.*

N		p	BST	Bonf	HC	SGoF	Perm
200	A	50	5.9 (1.1)	1.7 (0.1)	13.8 (0.1)	11.3 (0.1)	4.9 (54)
		100	6.0 (1.3)	1.2 (0.1)	18.1 (0.1)	12.6 (0.1)	4.9 (104)
		200	5.0 (1.8)	1.1 (0.2)	18.8 (0.2)	12.9 (0.2)	4.5 (201)
		400	6.4 (3.5)	1.4 (0.4)	24.2 (0.5)	15.2 (0.4)	4.8 (394)
		1000	7.0 (16)	1.5 (1.0)	26.7 (1.0)	18.0 (1.0)	5.4 (969)
		2000	8.4 (82)	1.4 (1.9)	29.6 (1.9)	19.9 (1.9)	6.5 (1889)
	B	50	46.9 (0.8)	24.4 (0.1)	8.4 (0.1)	74.9 (0.1)	43.6 (50)
		100	50.4 (1.2)	25.4 (0.1)	96.1 (0.1)	79.5 (0.1)	45.3 (97)
		200	50.6 (1.6)	23.5 (0.2)	99.7 (0.2)	88.2 (0.2)	45.2 (199)
		400	54.6 (3.3)	26.1 (0.4)	100 (0.5)	94.9 (0.4)	48.3 (374)
		1000	58.0 (16)	23.6 (1.0)	100 (1.0)	97.2 (1.0)	49.7 (935)
		2000	64.6 (80)	22.7 (1.9)	100 (1.9)	98.8 (1.9)	56.1 (1848)
	C	50	78.1 (0.8)	59.3 (0.1)	17.7 (0.1)	21.3 (0.1)	75.6 (51)
		100	72.6 (1.2)	52.7 (0.1)	21.2 (0.1)	16.2 (0.1)	68.9 (96)
		200	66.3 (1.6)	45.7 (0.2)	20.1 (0.2)	15.0 (0.2)	62.6 (186)
		400	61.7 (3.3)	36.6 (0.5)	24.3 (0.4)	16.4 (0.4)	57.4 (368)
		1000	56.0 (16)	30.6 (0.9)	26.8 (0.9)	18.4 (0.9)	50.9 (920)
		2000	53.8 (79)	24.9 (1.8)	29.8 (1.8)	20.4 (1.8)	47.4 (1830)
400	A	50	5.9 (0.9)	2.5 (0.1)	15.6 (0.1)	12.7 (0.1)	5.8 (69)
		100	5.9 (1.0)	1.9 (0.1)	20.8 (0.1)	14.4 (0.1)	5.2 (132)
		200	5.8 (1.5)	1.2 (0.3)	22.9 (0.3)	15.6 (0.3)	5.8 (258)
		400	6.5 (3.4)	1.9 (0.5)	24.6 (0.5)	16.6 (0.5)	5.1 (511)
		1000	5.7 (16)	1.8 (1.3)	26.0 (1.3)	17.4 (1.3)	5.6 (1278)
		2000	5.7 (80)	1.2 (2.6)	27.3 (2.6)	17.3 (2.6)	4.8 (2563)
	B	50	76.5 (0.8)	59.5 (0.1)	100 (0.1)	99.5 (0.1)	74.9 (70)
		100	82.6 (1.0)	59.9 (0.1)	100 (0.1)	100 (0.1)	80.2 (137)
		200	85.1 (1.6)	60.3 (0.3)	100 (0.3)	100 (0.3)	83.7 (268)
		400	87.9 (3.5)	58.3 (0.5)	100 (0.5)	100 (0.5)	85.1 (520)
		1000	89.1 (16)	59.6 (1.3)	100 (1.3)	100 (1.3)	86.3 (1277)
		2000	91.2 (85)	56.0 (2.6)	100 (2.6)	100 (2.6)	87.3 (2612)
	C	50	99.0 (0.9)	97.7 (0.1)	26.8 (0.1)	29.4 (0.1)	99.1 (70)
		100	98.5 (1.0)	96.0 (0.1)	24.0 (0.1)	20.4 (0.1)	98.2 (135)
		200	97.7 (1.5)	94.0 (0.3)	26.3 (0.3)	18.6 (0.3)	97.7 (262)
		400	97.1 (3.4)	89.4 (0.5)	25.7 (0.5)	18.5 (0.5)	96.8 (516)
		1000	94.9 (16)	83.8 (1.3)	26.3 (1.3)	17.8 (1.3)	93.7 (1272)
		2000	94.0 (81)	81.5 (2.5)	27.3 (2.5)	17.7 (2.5)	93.4 (2538)

3.2. *Example with Genomic Data*

In this section, we analyze SNP data from the ongoing Risk Assessment of Cerebrovascular Events Study (RACE, 2017). This is a case-control study involving over 5000 imaging-confirmed cases of stroke and 5000 controls, recruited from seven medical centers in Pakistan. The study is aimed at investigating the genetic, biomarker, and lifestyle determinants of stroke and its subtypes. We consider the 1,220 cases of early onset stroke (stroke before age 60) and 1,273 controls for whom both genotypic and phenotypic data are available. We restrict to the first 2000 loci on chromosome 5. Genetic variants by nature have three categories, but we follow the common practice of grouping them into two categories, so the data are represented by 2000 2×2 tables.

The left panel of Figure 2 shows the histogram of all the marginal log-odds ratio estimates $\log \hat{\theta}_k$ along with the rejection region for BST; two SNPs appear to be significant, but the critical region only refers to the most active one. For results on forward-stepwise BST with Mantel–Haenszel adjustment at each step, see the right panel of Figure 2. Five active SNPs are identified in all, each having a highly significant association with the risk of early onset stroke. To obtain a CI for the odds-ratio, we constructed five $\mathrm{CI}_{\mathrm{boot}}$ intervals, each based on 200 bootstrap samples, and averaged their endpoints to form $\mathrm{CI}_{\mathrm{bag}}$. Only two of the five active SNPs identified in forward-stepwise BST are significant in the initial run of BST (left panel of Figure 2). Note, however, that forward-stepwise BST is an ad hoc way of removing the effect of previously detected active SNPs, and should not be compared to a single run of BST, which is designed to detect the most active SNP (in this case rs275430) and can be unduly conservative for detecting additional SNPs. Bonferroni agrees with the initial BST result (left panel of Figure 2) which identified only rs275430 and rs1909745 as

Table 4

Coverage rates (%) and average widths (in brackets) of 95% confidence intervals for $\log \theta_0$*, under the same exchangeable correlation structure as in Table 3. The average runtime (seconds) to construct a single confidence interval is given in parentheses.*

N		p	CI_0	$\mathrm{CI}_{\max}$	$\mathrm{CI}_{\mathrm{boot}}$	$\mathrm{CI}_{\mathrm{bag}}$
200	A	50	94.0 [1.8](0.8)	100 [4.2](37)	97.7 [2.4](38)	99.8 [2.3](38)
		100	94.1 [1.9](0.8)	100 [4.4](38)	95.1 [2.5](40)	99.7 [2.4](42)
		200	94.9 [1.9](1.7)	100 [4.6](40)	96.5 [2.6](45)	99.8 [2.5](49)
		400	93.6 [2.0](3.4)	100 [4.7](46)	95.4 [2.7](60)	99.8 [2.6](74)
		1000	92.8 [2.1](17)	100 [4.9](78)	95.2 [2.7](158)	99.9 [2.7](238)
		2000	91.6 [2.2](79)	100 [5.1](169)	94.4 [2.8](475)	99.9 [2.7](780)
	B	50	65.8 [1.8](0.8)	100 [4.3](37)	93.6 [2.5](38)	96.3 [2.5](40)
		100	64.6 [1.9](1.1)	100 [4.5](38)	93.0 [2.6](40)	96.0 [2.6](41)
		200	65.7 [2.0](1.6)	100 [4.6](40)	92.3 [2.7](44)	95.8 [2.6](48)
		400	65.3 [2.1](3.3)	100 [4.8](46)	91.2 [2.7](60)	94.8 [2.7](74)
		1000	64.1 [2.1](17)	100 [5.0](78)	88.7 [2.7](157)	92.7 [2.7](237)
		2000	64.3 [2.2](78)	100 [5.2](167)	86.7 [2.7](473)	91.2 [2.7](776)
	C	50	91.2 [1.8](0.8)	99.8 [4.3](37)	97.6 [2.3](38)	98.3 [2.3](38)
		100	87.7 [1.9](1.0)	99.9 [4.4](38)	96.9 [2.4](39)	97.5 [2.4](41)
		200	86.3 [1.9](1.5)	100 [4.6](39)	96.6 [2.6](44)	96.8 [2.6](48)
		400	85.4 [2.0](3.3)	100 [4.8](46)	95.8 [2.7](60)	96.3 [2.7](74)
		1000	82.5 [2.1](17)	100 [5.0](79)	93.7 [2.8](158)	94.2 [2.8](238)
		2000	80.1 [2.2](77)	100 [5.1](166)	92.0 [2.9](472)	91.6 [2.9](777)
400	A	50	94.0 [1.3](0.6)	100 [3.0](37)	98.0 [1.7](38)	99.9 [1.6](39)
		100	94.2 [1.3](0.8)	100 [3.1](38)	96.9 [1.7](40)	99.9 [1.7](43)
		200	94.1 [1.4](1.4)	100 [3.2](40)	97.0 [1.8](47)	100 [1.7](55)
		400	93.6 [1.4](3.3)	100 [3.3](47)	96.5 [1.9](72)	100 [1.8](98)
		1000	94.4 [1.5](17)	100 [3.5](79)	98.1 [1.9](223)	99.9 [1.9](369)
		2000	94.3 [1.6](78)	100 [3.6](168)	98.3 [2.0](739)	99.9 [1.9](1292)
	B	50	72.5 [1.3](0.6)	100 [3.0](36)	92.4 [1.6](37)	94.3 [1.6](38)
		100	72.0 [1.3](0.8)	100 [3.1](38)	89.7 [1.7](40)	91.9 [1.7](43)
		200	73.8 [1.4](1.4)	100 [3.3](40)	89.9 [1.6](48)	90.8 [1.6](55)
		400	70.8 [1.5](3.3)	100 [3.4](48)	86.0 [1.7](73)	87.5 [1.7](99)
		1000	74.0 [1.5](16)	100 [3.5](79)	83.8 [1.7](222)	85.6 [1.7](365)
		2000	73.7 [1.6](84)	100 [3.7](185)	81.4 [1.7](766)	82.9 [1.7](1353)
	C	50	99.0 [1.3](0.6)	99.0 [3.0](36)	95.6 [1.2](37)	97.3 [1.2](38)
		100	98.4 [1.3](0.8)	98.5 [3.1](37)	95.9 [1.3](40)	97.0 [1.3](42)
		200	98.0 [1.4](1.4)	98.1 [3.3](39)	95.9 [1.4](47)	97.4 [1.4](54)
		400	96.9 [1.4](3.3)	96.9 [3.4](47)	94.8 [1.5](72)	96.0 [1.5](98)
		1000	96.4 [1.5](17)	97.3 [3.5](80)	94.6 [1.6](223)	96.3 [1.6](368)
		2000	96.0 [1.6](78)	97.8 [3.6](169)	94.4 [1.7](738)	96.1 [1.7](1311)

significant. Higher Criticism rejects the global null, but SGoF fails to reject.

4. Discussion

This article develops a novel approach to post-selection inference for the screening of large numbers of dependent 2×2 tables for the presence of significant associations. BST guarantees rigorous type I error control, has greater power than competing tests, and is computationally much less demanding than permutation testing. Our method is related to the adaptive resampling test (ART) introduced by McKeague and Qian (2015a), but by exploiting the relatively simple statistical structure of 2×2 tables (in contrast to the marginal linear model setting) we are able to avoid the need to use the double bootstrap, and we can base the calibration on direct simulation from the null distribution instead. Further, we are able to provide computationally feasible robust confidence intervals for the odds ratios that take the post-selection into account, whereas robust confidence intervals were not feasible in ART.

In the Supplementary Materials (Section 1), we describe a forward stepwise version of BST based on removing the effects of previously selected risk factors via a Mantel–Haenszel adjustment to the marginal odds-ratio estimates. This approach works well when only a few forward steps are involved, but can become cumbersome beyond that. It is still valid (though more conservative) to use the initial BST critical values to calibrate further steps of BST after previous risk factors have been removed from the dataset; this approach greatly lowers the computational cost in comparison with forward stepwise BST. An important extension of BST is to adjust for pre-specified demographic and environmental effects; this can be done by including all such covariates in

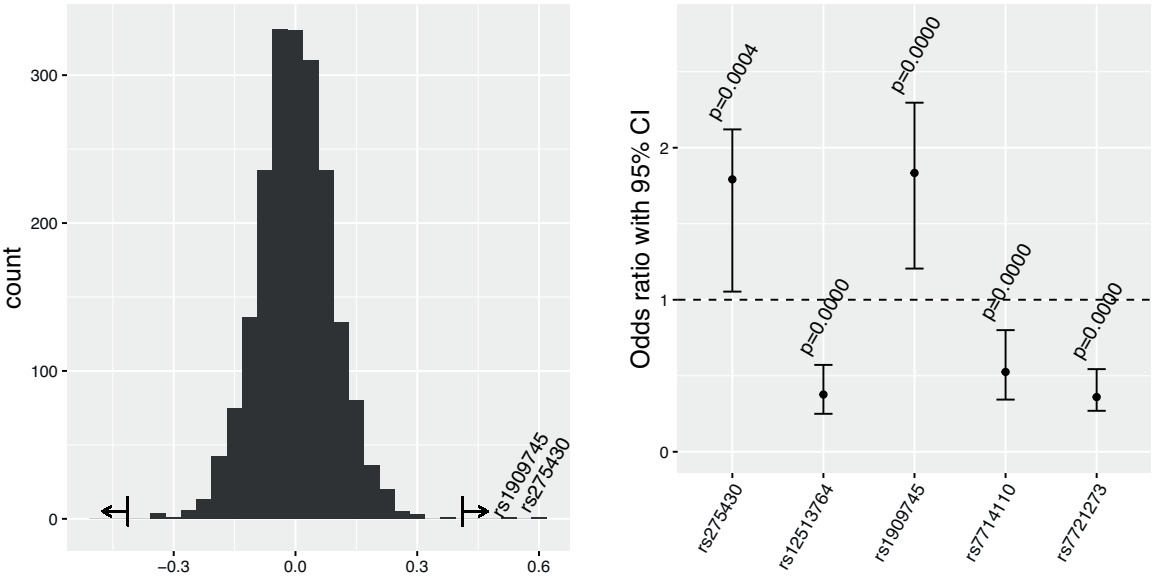

Figure 2. RACE Study. Left panel: histogram of the marginal log-odds ratio estimates $\log \hat{\theta}_k$ for 2000 genetic variants on chromosome 5. The rejection region for BST is indicated by arrows; two SNPs fall in the rejection region as identified by their loci. Right panel: p-values and 95% confidence intervals for the odds ratios of the risk factors detected using forward-stepwise BST at the 5% level of significance.

marginal logistic regression models (along with each risk factor W_k), then using the resulting adjusted estimates of $\log \theta_k$ and its standard error in place of the original $\hat{\theta}_k$ and $\hat{\tau}_k$.

Computational cost can become an issue for BST when p is of the order needed for a genome-wide analysis, as evident from the runtime results in Tables 3 and 4. This occurs when trying to generate realizations from the p-dimensional $N(0, C_X)$ distribution with an unstructured correlation matrix C_X. If a specific structure for C_X can be assumed (e.g., Toeplitz, as for a stationary time series on the lattice $1, \ldots, p$), then methods are available to speed up the simulation, as used in spatial statistics, see for example, Guinness and Fuentes (2017). This also raises the need to estimate C_X taking an assumed covariance structure into account, and there is an extensive literature devoted to this problem. For instance, (Cai et al., 2016) give a thorough account of methods for estimating Toeplitz, banded, spiked, and sparse high-dimensional covariance matrices. For a genome-wide analysis, a simple approach would be to assume a block-diagonal structure for C_X, provided the blocks are justified from biological considerations. If such blocks are still too large to be computationally feasible, splitting them into manageable segments and applying BST separately to each segment would be reasonable.

A further important issue is the suitability of a particular covariance structure for the particular application. The knockoff method of Candès et al. (2018) mentioned in the Introduction faces similar issues. In the application to Crohn's disease SNP data, Candès et al. (2018) approximated the standardized joint distribution of the SNPs as multivariate Gaussian, and estimated the covariance matrix using a method of Wen and Stephens (2010), which shrinks off-diagonal entries of the empirical covariance matrix using genetic distance information estimated from HapMap data. As discussed by Candès et al. (2018), SNP data invariably contain some very high correlations, which presents a challenge to any screening methodology. A way of alleviating this problem, they suggest, is to form clusters of highly correlated SNPs and settle for the detection of such clusters rather than individual SNPs.

A thorough investigation of these issues is beyond the scope of the present article, but we anticipate that it should be feasible to adapt BST along these lines to make it an effective, scalable, and powerful method of post-selection inference for full genome-wide SNP data.

5. Supplementary Materials

See Web Appendices 1 for details on forward-stepwise BST (referenced in Sections 3 and 4), two for the proof of Theorem 1, three for the HC and SGoF procedures (referenced in Sections 3 and 4), along with R code implementing BST, available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Both authors were supported by NIH Grant R01 GM095722-05. We thank Iuliana Ionita-Laza and Xiaoyu Song for help with the RACE study data, and the reviewers for many detailed and perceptive comments.

References

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics* **43**, 2055–2085.

Cai, T. T., Ren, Z., and H. Zhou (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* **10**, 1–59.

Candès, E. J., Fan, Y., Janson, L., and J. Lv (2018). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B* **80**, 551–577.

Carvajal-Rodriguez, A., de Una-Alvarez, J., and E. Rolan-Alvarez (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* **10**, 209.

Churchill, G. A. and R.W. Doerge (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. New York: Cambridge University Press.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* **32**, 962–994.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30**, 1–25.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge New York: University Press.

Guinness, J. and M. Fuentes (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *Journal of Computational and Graphical Statistics* **26**, 88–97.

Keogh, R. H. and Cox, D. R. (2014). *Case-Control Studies*. New York: Cambridge University Press.

Kou, S. G. and Ying, Z. (1996). Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica* **6**, 809–829.

Kou, S. G. and Ying, Z. (2006). Analysis of sequences of dependent 2×2 tables. In *Random Walk, Sequential Analysis and Related Topics*, A. C. Hsiung, Z. Ying, and C.-H. Zhang (eds), 171–198. Singapore: World Scientific.

McKeague, I. W. and M. Qian (2015a). An adaptive resampling test for detecting the presence of significant predictors (with discussion). *Journal of the American Statistical Association* **110**, 1422–1433.

McKeague, I. W. and M. Qian (2015b). Rejoinder to the discussants of "An Adaptive Resampling Test for Detecting the Presence of Significant Predictors." *Journal of the American Statistical Association* **110**, 1459–1462.

Risk Assessment of Cerebrovascular Events (RACE) Study (2017). `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000456.v1.p1` [accession number phs000456.v1.p1].

Reis, I. M., Hirji, K. F. and Afifi, A. A. (1999). Exact and asymptotic tests for homogeneity in several 2×2 tables. *Statistics in Medicine*, **18**, 893–906.

Sham, P. C., and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15**, 335–346.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1–penalized likelihood models. *Canadian Journal of Statistics* **46**, 41–61.

Wen, X. and M. Stephens (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics*, **4**, 1158–1182.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics* **86**, 929–942.

Received October 2017. *Revised July* 2018. *Accepted July* 2018.

Appendix

Conditions for Theorem 1. By Theorem 2.1 of Kou and Ying (1996), the marginal (noncentral hypergeometric) distribution of each X_k conditional on the margins has the following representation as a sum of independent Bernoulli random variables: $X_k \stackrel{d}{=} \sum_{s=1}^{M_1} I(\eta_{sk} \le (1+\theta_k^{-1}\lambda_{sk})^{-1})\delta_{sk}$, where $\{\eta_{sk} : s = 1, \ldots, \min(M_1, N_{1k})\}$ is a sequence of independent uniform-$(0,1)$ random variables, and $\delta_{sk} = I(s \le N_{1k})$. Here, $-\lambda_{sk} \le 0, s = 1, \ldots, \min(M_1, N_{1k})$, are the roots of the Jacobi polynomial

$$\phi_k(z) = \sum_{u=\max(0, M_1 - N_{2k})}^{\min(M_1, N_{1k})} \binom{N_{1k}}{u} \binom{N_{2k}}{M_1 - u} z^u.$$

We will make the mild "structural" assumption that the above representation extends *jointly* over all $k = 1, \ldots, p$. That is, we assume that the M_1 random p-vectors formed by concatenating each summand over $k = 1, \ldots, p$ can be arranged to be independent.

(C1) Model structure: There exist independent random vectors $(\eta_{s1}, \ldots, \eta_{sp}), s = 1, \ldots, M_1$, with each component uniformly distributed on $(0,1)$, such that

$$\boldsymbol{X} \triangleq \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \stackrel{d}{=} \sum_{s=1}^{M_1} \begin{pmatrix} I(\eta_{s1} \le (1+\theta_1^{-1}\lambda_{s1})^{-1})\delta_{s1} \\ \vdots \\ I(\eta_{sp} \le (1+\theta_p^{-1}\lambda_{sp})^{-1})\delta_{sp} \end{pmatrix}$$

conditional on the margins.

(C2) Both the conditional and unconditional correlation matrices of $\boldsymbol{X}$ converge to the same invertible limit $\boldsymbol{C_X}$ as $N \to \infty$.

(C3) Stability conditions: As $N \to \infty$, $M_1/N \to \pi \in (0,1)$, $N_{1k}/N \to q_k \in (0,1)$, $E(X_k)/N_{1k} \to \pi_{1k} \in (0,1)$, and $E(M_1 - X_k)/N_{2k} \to \pi_{2k} \in (0,1)$ for $k = 1, \ldots, p$.

Remark 1. *Under Condition (C2), the matrix* $\boldsymbol{C_X}$ *can be consistently estimated by the sample correlation matrix of the vector of risk factor indicators* $\boldsymbol{W}$ *restricted to the data on the cases. To see this, note that the observations on* $\boldsymbol{W}$ *for the cases are iid and sum to* $\boldsymbol{X}$*, in a parallel fashion to the conditional representation of* $\boldsymbol{X}$ *in Condition (C1). This also makes it clear that Condition (C2) is a mild stability condition, similar to Condition (C3).*

Remark 2. *Under the stability condition (C3),*

$$N\tau_k^2 \to \sigma_k^2 \triangleq \frac{1}{q_k \pi_{1k}} + \frac{1}{(1-q_k)(1-\pi_{2k})} + \frac{1}{q_k(1-\pi_{1k})} + \frac{1}{(1-q_k)\pi_{2k}}, \tag{A.1}$$

so when $\boldsymbol{\theta}^{(0)} = \mathbf{1}$ *we have* $\pi_{1k} = \pi_{2k} = \pi$*, and* $\sigma_k = \{q_k(1-q_k)\pi(1-\pi)\}^{-1/2}$.