Rare variants association testing for a binary outcome when pooling individual level data from heterogeneous studies

Tamar Sofer[*,1,2] and Na Guo[2]

[1]Departments of Medicine and of Biostatistics, Harvard University, Boston, MA

[2]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA

[*]Correspondence: tsofer@bwh.harvard.edu

Abstract

Whole genome and exome sequencing studies are used to test the association of rare genetic variants with health traits. Many existing WGS efforts now aggregate data from heterogeneous groups, e.g. combining sets of individuals of European and African ancestries. We here investigate the statistical implications on rare variant association testing with a binary trait when combining together heterogeneous studies, defined as studies with potentially different disease proportion and different frequency of variant carriers. We study and compare in simulations the type 1 error control and power of the naïve Score test, the saddlepoint approximation to the score test (SPA test), and the BinomiRare test in a range of settings, focusing on low numbers of variant carriers. Taking into account test performance as well as computation considerations, we develop recommendations for association analysis of rare genetic variants. We show that the Score test is preferred when the case proportion in the sample is 50%. Otherwise, for very low number of carriers, BinomiRare is preferred due to computational efficiency and type 1 error control. When there are at least 90 carriers in the

combined sample, the SPA test generally controls the type 1 error and is preferred over BinomiRare due to higher power and wider implementation in software packages. Finally, we recommend to not sample controls in order to generate more balanced case-control ratio, rather, to use appropriate analytic methods. Sampling of controls reduces power.

Introduction

Genetic association studies test the association of genetic variants with a trait. Genome-wide association studies (GWAS) typically test the association of each of single, common, genetic variants across the genome. This is often also done in Whole Genome Sequencing (WGS) studies, that also test rarer genetic variants. In a few examples from the WGS analysis in the Trans-Omics of Precision Medicine (TOPMed) program, investigators used a minor allele frequency threshold (MAF) of 0.001 and allowed for a minimum of 20 minor allele counts for consideration of a variant in association analyses with glycated hemoglobin [1]; a MAF threshold of 0.001 corresponding to at least 32 counts of the rare variant allele was applied in a study of lipids [2]; and variants with 10 counts of the rare allele in the sample were considered in an analysis of brain imaging measures [3]. In other examples, investigators test rare variants associations when studying a specific gene region of interest [e.g. 4, 5].

It is known that tests of the association of a genetic variant with a binary outcome do not control the type 1 error in some settings, and the problem is exacerbated when the genetic variant is rare [6]. Specifically, when the proportion of cases in the study is low, p-values of likelihood-based tests are not well calibrated. A few tests were developed for the association of

single genetic variants, that can also adjust for covariates. The Firth test [7], highlighted by Ma et al. [6], uses a higher order approximation to the likelihood to compute standard errors, and is more well calibrated compared to standard tests. Dey et al. [8] developed the saddlepoint approximation (SPA) to the p-value computation of the Score test based on a cumulant generating function rather than the standard normal distribution approximation, which is better calibrated and has improved control of type 1 error compared to the traditional Score test p-value, and is faster than the Firth test. Lee et al. [8] developed a resampling method for calibrating single-variant tests (as well as variant-set tests), which can also account for covriates. Sofer [9], [10] introduced the BinomiRare test, which is robust to low case proportion and controls the type 1 error for any number of rare allele carriers. In an extensive simulation studies, Ma et al. demonstrated that the count of the rare allele determines the type 1 error and the power of statistical tests. Ma, Blackwell [6] and Sofer [9] considered settings with one or multiple samples with different case proportions, however, they did not consider the scenario in which multiple samples with different frequencies of the genetic variant allele are pooled. This scenario is important, because modern large sequencing studies such as the NHLBI's Trans-Omics in Precision Medicine (TOPMed) and the NHGRI's CCDG aggregate individual level data from WGS studies conducted in diverse populations, where allele frequencies often differ between populations.

We set out to study rare variant association testing when pooling individual level data from various studies with potentially different population characteristics: allele frequency and case proportion. To limit the high number of possible combinations of studies' characteristics, we

focus on two studies of the same sample size and vary the disease prevalence in each of the studies as well as the count of the rare variant allele. We focus on tests that can account for covariates, and that do not use resampling, to limit computation time.

Methods

Logistic association model for two studies

Suppose that individual level data from two studies with n_1 and n_2 individuals respectively are combined. For study $j \in \{1,2\}$ Let the binary outcome $D_{ji} \in \{0,1\}, i_j = 1, \ldots, n_j$ follow a logistic model with

$$logit(p(D_{ji} = 1)) = \beta_{j0} + g_{ji_j}\beta_{jg} ,$$

here assuming no confounders or covariates are adjusted for. When the data are pooled across studies, the model can be written instead as

$$logit\big(p(D_{i=1})\big) = I(individual\ i\ in\ study\ 1)\beta_{01} + I(individual\ i\ in\ study\ 2)\beta_{02} + g_i\beta_g$$
$$= \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where we now add study-specific intercepts in the joint model. Note that this formulation is statistically equivalent to a formulation with the same intercept for all individuals, and a covariate for one of the studies. To simplify exposition, let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, g_i)^T$, $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_g)^T$.

Tests for association of a variant with the outcome

Both the Score and the BinomiRare tests (and the SPA, which is a score test with better calibrated p-value) use estimates of within-sample disease probabilities under the null

hypothesis of no association between the outcome and the genetic variant, i.e. under $H_0: \beta_g = 0$. Clearly, under the null, $\hat{\beta}_{01} = logit(\sum_{i_1=1}^{n_1} D_{i1} \ /n_1) \equiv logit(\pi_1)$, and $\hat{\beta}_{02} = logit(\sum_{i_2=1}^{n_2} D_{i2} \ /n_2) \equiv logit(\pi_2)$, where $expit(x) = exp(x)/[1 + exp(x)]$ is the inverse function of the logit function. The derivative of the $expit(\cdot)$ function is $\frac{\partial}{\partial x} expit(x) = exp(x)/[1 + exp(x)]^2$. For $n = n_1 + n_2$, the score for β_g is derived as:

$$s(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} logL(\boldsymbol{\beta}|\, D) = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} log[expit(\boldsymbol{x}^T{}_i\, \boldsymbol{\beta})^{\boldsymbol{D_i}} (1 - expit(\boldsymbol{x}^T{}_i \boldsymbol{\beta})^{1-D_i}] =$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} D_i log[expit(\boldsymbol{x}^T{}_i\, \boldsymbol{\beta})] + (1 - D_i) log[1 - expit(\boldsymbol{x}^T{}_i \boldsymbol{\beta})] =$$

$$= \sum_{i=1}^{n} \boldsymbol{x}^T{}_i \{\frac{D_i}{1 + exp(\boldsymbol{x}^T{}_i\, \boldsymbol{\beta})} - \frac{(1-D_i) exp(\boldsymbol{x}^T{}_i \boldsymbol{\beta})}{1+exp(\boldsymbol{x}^T{}_i \boldsymbol{\beta})}\} = \sum_{i=1}^{n} \boldsymbol{x}^T{}_i \{D_i - \frac{exp(\boldsymbol{x}^T{}_i \boldsymbol{\beta})}{1+exp(\boldsymbol{x}^T{}_i \boldsymbol{\beta})}\} = \sum_{i=1}^{n} \boldsymbol{x}_i \{D_i - expit(\boldsymbol{x}^T{}_i \boldsymbol{\beta})\},$$

where in the score test for β_g, β_{01} and β_{02} are estimated under the null, and in this setting, the score for β_g simplifies to:

$$U(\beta_g) = \sum_{i=1}^{n} g_i \{D_i - expit[(x_{i1}, x_{i2})^T (\hat{\beta}_{01}, \hat{\beta}_{01})]\}$$

$$= \sum_{i_1=1}^{n_1} g_{1i}(D_{1i} - \pi_1) + \sum_{i_2=1}^{n_2} g_{2i}(D_{2i} - \pi_2)$$

If a genetic variant is rare, then most carriers of the variant are heterozygotes, i.e. most people have $g_i = 0$, a few people have $g_i = 1$, and almost no one has $g_i = 2$, meaning that we can assume a dominant mode of variant association. We then further simplify this expression by introducing additional notation. For study j, let $c_j{}^0$ and $c_j{}^1$ be the number of carriers of a rare

variant allele among people with the outcome $D_{ji} = 0$ and among those with the outcome $D_{ji} = 1$, respectively. Then the score is now:

$$U(\beta_g) = {c_1}^1(1-\pi_1) - {c_1}^0\pi_1 \; + \; {c_2}^1(1-\pi_2) - \; {c_2}^0\pi_2 \; =$$

$$c_1^1(1-\pi_1) - (c_1 - c_1^1)\pi_1 + c_2^1(1-\pi_2) - (c_2 - c_2^1)\pi_2 =$$

$$(c_1^1 - c_1\pi_1) + (c_2^1 - c_2\pi_2)\,.$$

The score for β_g is the sum of scores in each of the two studies, and in each study, the score is a difference between the observed and the expected number of diseased carriers, under the observed disease proportion in the study.

In the standard Score test, the variance of the score for β_g is estimated by deriving the information matrix, and then extracting the appropriate entry from its inverse. For logistic regression, the information matrix is given by:

$$I(\boldsymbol{\beta}) \; = -\frac{\partial}{\partial^T\boldsymbol{\beta}\partial\boldsymbol{\beta}} logL(\boldsymbol{\beta}|\, D) \; = -\frac{\partial}{\partial^T\boldsymbol{\beta}}\sum_{i=1}^{n}\boldsymbol{x}_i\{D_i \; - expit({\boldsymbol{x}^T}_i\boldsymbol{\beta})\} =$$

$$\frac{\partial}{\partial^T\boldsymbol{\beta}}\sum_{i=1}^{n}\boldsymbol{x}_i expit({\boldsymbol{x}^T}_i\boldsymbol{\beta}) = \sum_{i=1}^{n}\boldsymbol{x}_i\frac{exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})}{[1\,+exp({\boldsymbol{x}_i}^T\boldsymbol{\beta})]^2}{\boldsymbol{x}_i}^T.$$

This can be written in a matrix form. Define the following matrices:

$$\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{3}} = \begin{pmatrix} 1 & 0 & g_{11} \\ \vdots & \vdots & \vdots \\ 1 & 0 & g_{1n_1} \\ 0 & 1 & g_{21} \\ \vdots & \vdots & \vdots \\ 0 & 1 & g_{2n_2} \end{pmatrix}, \boldsymbol{W}_{n\times n} = \begin{pmatrix} s_{11} & & & & & \\ & \ddots & & & & \\ & & s_{1n_1} & & & \\ & & & s_{21} & & \\ & & & & \ddots & \\ & & & & & s_{2n_2} \end{pmatrix}$$

where $\boldsymbol{X}$ is the design matrix of the regression of the disease on the variant, accounting for two studies using study-specific intercepts, and $\boldsymbol{W}$ is the diagonal matrix with diagonals, for person ji from study $j, j \in \{1,2\}, i = 1, \dots, n_j$ having $s_{ji} = \frac{exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})}{[1 + exp(\boldsymbol{x}_i{}^T\boldsymbol{\beta})]^2}$. Then

$$I(\boldsymbol{\beta}) = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}.$$

Noting that for the Score test, the information matrix will be evaluated under the null; therefore, the only covariates are the study-specific intercepts, we have that for all individuals in study one $s_{1i} = s_1 = \frac{exp(\widehat{\beta}_{01})}{[1 + exp(\widehat{\beta}_{01})]^2}$, and in study two $s_{2i} = s_2 = \frac{exp(\widehat{\beta}_{02})}{[1 + exp(\widehat{\beta}_{02})]^2}$ Then:

$$I(\boldsymbol{\beta}) = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} = \begin{pmatrix} n_1 s_1 & 0 & c_1 s_1 \\ 0 & n_2 s_2 & c_2 s_2 \\ c_1 s_1 & c_2 s_2 & c_1 s_1 + c_2 s_2 \end{pmatrix}.$$

Using formula for matrix inverse, one can compute the entry of $I(\boldsymbol{\beta})^{-1}$ corresponding to β_g, as

$$[I(\boldsymbol{\beta})^{-1}]_{3,3} = \frac{n_1 n_2}{n_1 s_2 c_2 (n_2 - c_2) + n_2 c_1 s_1 (n_1 - c_1)}.$$

Its inverse is the variance of the score:

$$\widehat{var}\left(U(\beta_g)\right) = \frac{1}{[I(\boldsymbol{\beta})^{-1}]_{3,3}} = \frac{s_2 c_2 (n_2 - c_2)}{n_2} + \frac{s_1 c_1 (n_1 - c_1)}{n_1}$$

which is the sum of the scores for β_g in each of the two studies. The estimator of the variance of the score depends, through s_1, s_2, on the observed outcome proportion in the sample, on the observed variant allele count in the sample, and on the sample size. When the observed variant count is very low compared to the number of individuals in the study, e.g. when c_1 is fixed and $n \to \infty$, we have that $c_j(n_j - c_j)/n_j \approx c_j$ for $j \in \{1,2\}$, so that both the score and its variance do not depend on the sample size, but rather only on the variant count and the

disease proportion in the study. Note that this asymptotic setting is standard for genome sequencing data because as the sample size grows more low-count variants are observed.

The SPA test [11] instead of using the above score variance estimates, computes a p-value based on a obtaining a better approximate distribution of the test statistic using a cumulant generating function, and uses the saddlepoint approximation to solve the resulting optimization problem.

The BinomiRare test [9] only relies on the observed outcome frequency (more generally, the outcome probabilities) in the carriers. It takes the vector of estimated outcome probabilities for the carriers of the rare variants, and uses the Poisson-Binomial distribution [12] to compute a p-value for testing the null hypothesis of no association between the variant and the outcome. Therefore, in our simplified settings that do not use covariates, the BinomiRare tests depends on the numbers of carriers c_1, c_2, diseased carriers c_1^1, c_2^1, and the proportions of diseased individuals in the studies. Because the BinomiRare test does not use a normal approximation to the Poisson-Binomial distribution, it has a discrete probability mass function. Two types of p-values can be computed: the standard p-value, and the mid-p-value. Let $W \sim Poisson - Binomial(\widehat{\boldsymbol{p}})$, with $\widehat{\boldsymbol{p}}$ being the vector of estimated disease probabilities for the $c_1 + c_2$ carriers of the rare-variant of interest in the two studies. The p-value and mid-p-value are defined as:

$$p\text{-value} = \widehat{\Pr}(W = c_1^1 + c_2^1) + \sum_{k=1}^{c_1+c_2} \widehat{\Pr}(W = k) \times 1[\widehat{\Pr}(W = k) < \widehat{\Pr}(W = c_1^1 + c_2^1)] \quad (1)$$

$$\text{mid-}p\text{-value} = \frac{\widehat{\Pr}(W = c_1^1 + c_2^1)}{2} + \sum_{k=1}^{n_c} \widehat{\Pr}(W = k) \times 1[\widehat{\Pr}(W = k) < \widehat{\Pr}(W = c_1^1 + c_2^1)] \quad (2)$$

Clearly, the mid-p-value (2) is less conservative, and therefore will always be smaller than the p-value (1). However, when the number of carriers is small, it may be too liberal.

Simulation studies

In our simulation studies described henceforth, we generated datasets in the simple settings described above, and used the computationally efficient implementation of the Score test $U(\beta_g)$ and $\widehat{var}\left(U(\beta_g)\right)$. For the SPA test, we used the naïve Score test p-value. When it was smaller than 0.05, we re-computed a p-value using the SPAtest R package [13].

Simulation studies: type 1 error control when combining two studies

We considered four settings of rare variant distributions across studies: $(c_1, c_2) = \{(10,10), (100,100), (10,100), (100,10)\}$. We varied the disease proportions in each of the two studies, so that the disease proportion in study 1 was $expit(\beta_{10}) \in \{0.01, 0.05, 0.2, 0.5\}$, and the disease proportion in study 2 was taking the same values across the simulation studies, so that it was always lower or equal to the proportion in study 1. In each of the settings defined by number of carriers and disease proportions we performed 10^8 simulations with $n_1 = n_2 =$

$10{,}000$, and evaluated type 1 error when using the Score, SPA, and BinomiRare tests with a p-value threshold of 10^{-4}. For the BinomiRare test, we used both the p-value (equation (1); pval) and the mid-p-value (equation (2), midp). For comparison, we also studied the following settings:

1. A simulation study with $n_1 = n_2 = 5{,}000$, holding all other parameters the same.
2. A simulation study in which we plugged-in the true, known, per-study disease prevalence $expit(\beta_{10}), expit(\beta_{20})$ rather than estimated them when computing the Score statistics, with provided disease probabilities for computing BinomiRare mid-p-values.
3. Simulation studies in which we generated datasets by sampling controls from each study, with a ratio of up to 3 controls per case and with 1 control per case.

Simulation studies: identifying minimum number of carriers for SPA test

We performed a simulation study with the goal of formulating a recommendation for the minimum number of carriers required for appropriate type 1 error control by the SPA test when combining individual level data from heterogeneous studies. To limit the potential number of simulations, we focused on three possible settings of disease proportion in the two studies: $[expit(\beta_{10}), expit(\beta_{20})] \in \{(0.01, 0.01), (0.01, 0.5), (0.5, 0.5)\}$, and the number of carriers in the two studies was varied so that all possible combinations of $c_1, c_2 \in \{10, 15, 20, 25, 30\}$ were evaluated. We also studied the same settings with the BinomiRare test, with both the pval and midp options. For the SPA test only, we considered additional settings in which $c_1 = c_2 \in \{35, 40, 45, 50, 55\}$.

Simulation studies: power comparisons

To compare power between tests, we used similar settings to those used for type 1 error assessment, with the same follow-up comparisons. To generate datasets for these simulations, we allowed for different probability of disease among carriers of the rare variant, so that $logit(p(D_{ji} = 1)) = \beta_{j0} + g_{ji_j}\beta_{jg}$ with, for simplicity, the same effect size $\beta_{1g} = \beta_{2g} = \beta_g$ in the two studies combined together. Effect sizes varied $\beta_g \in \{\log(2), \log(3), \log(4), \log(5)\}$. We used the same p-value threshold of 10^{-4} as before.

Computing approximate power for BinomiRare test

On the dedicated GitHub repository https://github.com/tamartsi/Binary_combine we provide a function to compute approximate power for the BinomiRare test. The function takes a vector of estimated outcome probabilities in the sample under the null hypothesis of no association between genotype and outcome, an odds ratio parameter, p-value threshold for declaring significance, number of carriers 'n_carrier', and number of simulation iterations. Then, in each simulation iteration it uniformly samples n_carrier outcome probabilities (without replacement). For each sampled carrier, given its outcome probability under the null p_0, the function computes the outcome probability under the alternative hypothesis $p_A = expit[logit(p_0) + \log(OR)]$, and uses the binomial distribution to simulate outcome status using p_A. Then, it uses the BinomiRare test to compute a p-value for the null hypothesis of H_0: $\boldsymbol{p_{car}} = \widehat{\boldsymbol{p}}_{0,car}$, where $\boldsymbol{p_{car}}$ is the true vector of outcome probabilities among the carriers, and $\widehat{\boldsymbol{p}}_{0,car}$, is the vector of estimated outcome probabilities under the null. Finally, the power is

the proportion of p-values computed in the simulations which were lower than the p-value threshold.

Results

Type 1 error when combining two studies

Figure 1 provides type 1 error comparisons when combining two studies with each $n_1 = n_2 = 10{,}000$, with varying disease proportions in the two studies, and four scenarios of number of carriers across the studies. We compare the naïve Score test, the SPA, and the BinomiRare test with the pval (usual p-value, equation (1)) and midp (mid-p-value, equation (2)) options. The figure provides the observed test size divided by the desired type 1 error. Ideally, this number should be 1. Higher numbers indicated inflation (larger number of false detection than expected), and lower numbers indicate deflation, or conservativeness.

Score test: As is already known, we see that the naïve Score test becomes more inflated as the disease prevalence in the total sample becomes low. Overall the Score performance become better with increased number of carriers in the combined sample. However, for a fixed number of carriers, there is a difference in performance depending on which of the two studies have more carriers: when considering the two non-symmetric scenarios, i.e. the scenarios in which $c_1 = 10; c_2 = 100$, and the other way around, we see that the Score test performance depend on the number of carriers in each study. Specifically, in comparison with the settings of $c_1 = c_2 = 10$, if an analyst required at least 100 carriers in the study with higher disease prevalence but allowed the number of carriers in the other study to stay 10, the inflation was reduced

compared to the settings in which 100 carriers were required in the study with lower case proportion. If the analyst further required both studies to have at least 100 carriers, the inflation did not improve much. This suggest that when combining multiple studies, it may be useful to require a minimum number of carriers in the study with the higher disease proportion in order to stabilize the Score test results.

SPA test: Type 1 error control was mostly appropriate when the total number of carriers in the combined two-study sample was 110 or 200, with a few settings with low degree of inflation (see Figure 1). When there were 20 carriers in the combined sample, type 1 error was usually not controlled, other than in the settings in which the disease proportion was equal in the study 1 and study 2. In this case, SPA was often conservative.

BinomiRare: Type 1 error was always controlled when the usual p-value (pval) was used, and usually controlled with the midp option. In a few settings, the BinomiRare with the midp option had low degree of inflation. Due to the discreteness of the Poisson-Binomial distribution (which is not approximated to a normal distribution by this test), the size of the test when using the pval option is often conservative.

Other settings: Comparisons of some of the above simulations to settings where $n_1 = n_2 = 5{,}000$ show that the results are mostly the same, confirming that the properties of the tests mostly depend on the number of carriers (Figure 1 in the Supplementary Information). To address the question of whether and how the results are strongly affected by estimation of disease probabilities, which do depend on sample size, we also compared type 1 error between the main simulation study and a simulation study in which disease probabilities are taken as

known for both the Score and the BinomiRare test (Figure 1 in the Supplementary Information). Type 1 error control often improved for the Score test, but in some cases became worse for the BinomiRare, when using the midp option. To note, BinomiRare was always more conservative than the Score test in this simulation. When sampling controls to reduce case-control ratios, Figure 2 demonstrates that the type 1 error is always controlled by the Score test if the case-control ratio is 1:1 (as expected), but not when the case-control ratio is 1:3. Supplementary Figures 2 and 3 provides all settings under sampling of controls with ratio 1:3 and 1:1, respectively. All tests become very conservative when the total number of carriers in each of the studies is 10 prior to sampling controls because often no carriers are left in the analytic sample after sampling of controls. Further, when sampling controls the SPA test often becomes inflated at times, especially in the 1:1 sampling scenario, likely because the number of carriers remaining in the data after sampling of controls is very low.

Simulations to study the minimum number of carriers for SPA

In the simulations designed to study the minimum number of carriers for SPA, which had up to 60 carriers in the combined sample, type 1 error was not perfectly controlled even when there were 60 carriers and the case proportion was 50% in both studies (Supplementary Figure 4). Because we did not see any pattern related to type 1 error control with respect to the distribution of variant carriers across studies, we also considered scenarios with an equal number of carriers in each study, with up to 55 carriers. Figure 3 provides the type 1 error for the SPA test when the number of carriers was equal in the two combined studies, and ranged from 10 to 55, by increments of 5. When the number of carriers was 45 in each study or 90 in

the combined sample, the type 1 error is controlled. Then, in the setting with 55 carriers in and case prevalence of 0.01 in each study, the type 1 error was $1.14x10^{-4}$, which is larger than expected in the 95% confidence intervals accounting for p-value threshold of $1x10^{-4}$ and $1x10^{8}$ simulations.

Power when combining two studies

Figure 4 compares power between the various tests when the case prevalence was 20% in study 1, 5% in study 2, for a few carriers setting, and comparing the baseline simulations (no sampling of controls), and sampling of controls with case-control ratio of 1:3 and 1:1. The figure provides the estimated power even when tests did not control the type 1 error in the corresponding simulation studies (while highlighting this non-control). For 110 carriers in the combined sample of 20,000 people, the power is higher when there are 100 carriers in the study with 20% cases, compared with 100 carriers in the study with 5% carriers. This is true in other simulations as well: for a fixed number of carriers in the total sample, power is higher when more carriers are in the study with higher case proportion. Power is reduced when controls are sampled, especially when the effect size is small. Among the two settings of 110 carriers in the combined sample, sampling of controls leads to more substantial reduction of power when the number of carriers is 100 in the study with lower case prevalence. This is likely because the sampling is more aggressive (lower total sample size), resulting in a substantially reduced number of carriers after sampling.

We also compared power in the main simulations to the settings when there were only 5,000 individuals in each study (but the same number of carriers), and when the case prevalence was known, providing true outcome probabilities as plug-ins for BinomiRare and Score tests (Supplementary Figure 5). When n=5,000 in each study, the power was about 90-100% of the power in the corresponding setting when there were n=10,000 individuals in each study, suggesting that more precisely estimated outcome probabilities could increase power. However, using the two outcome probabilities did not generally increase the power of BinomiRare, perhaps because in the simple investigated settings the parameter estimates are already quite precise.

Finally, we note that the BinomiRare with the mid-p-value is more powerful than the BinomiRare test with the usual p-value, as is known by definition, with BinomiRare-midp having up to 111% the power of the BinomiRare-pval option. The SPA test was up to 116% more powerful than the BinomiRare-midp test (focusing these comparisons on settings where both tests controlled the type 1 error).

Recommendations

We here address questions that investigators studying rare variants by pooling heterogeneous studies such as TOPMed may have when formulating an analysis plan.

1. If the case proportion is 0.5, use the Score test.
2. When the case proportion is low for testing rare variant associations, it is more powerful to use an appropriate analysis method such as BinomiRare (for very low carrier count) and SPA (for larger carrier counts) than to sample controls if the case proportion is low.

3. To control type 1 error when combining multiple diverse studies together, should a MAC or a number of carriers threshold be applied on each of the contributing studies? When using the BinomiRare test, it is unnecessary, it always controls the type 1 error when using the pval option. When using the SPA test, we saw that type 1 error control is generally good when there are 90 carriers in the combined sample, and there were no patterns suggesting that we need to require a minimum number of carriers in each one of the studies separately.
4. For a fixed number of carriers, power is highly affected by the proportion of cases. Consider restricting the analysis to variants with high number of carriers in the study with higher disease proportion. This will focus the analysis on variants with relatively higher power, while reducing the multiple testing burden.

Discussion

We performed a simulation study to (primarily) assess the type 1 error control of single-variant association tests for a binary outcome when pooling individual level data from heterogeneous studies. The asymptotic framework of our simulations is such that the number of carriers of a variant and the sample sizes were fixed. Variability came from the number of diseased individuals, aka, cases in the sample in general and in the carriers specifically. Despite the fact that testing of rare variant associations suffers from low power, it is still routinely performed as part of genome-wide association studies applied on sequencing or imputed genotype data, or when fine-mapping genomic regions, and investigators need to know when such tests are statistically valid. Our simulations were performed in simplified settings combining two studies,

in the absence of covariates, other than study-specific intercept. This allowed for computational efficiency when running a very large number of simulations, and for reducing the potential number of scenarios to investigate. We found that performance of the Score test largely depends on the number of carriers of the rare variant, number of diseased carriers, and the proportion of diseased individuals in the sample. This is, by design, true for the BinomiRare test as well. Notably, the performance of the tests also depends on the properties of the two combined studies: when the combined sample has a fixed number of carriers, test performance differs according to the number of carriers in each of the combined studies, and the outcome prevalence in each.

As was shown in the past, the Score test controls the type 1 error when the case-control proportion is 1. Like other tests, it is conservative when the number of carriers is very low. In simulations combining two studies of equal sizes, with a total number of 110 carriers, even when there were 50% cases in one of the studies and 100 carriers in that study, type 1 error was controlled regardless of the disease prevalence in the study with 10 carriers (Figure 1, row "Score Test", third column from the left). In these simulations, it seemed like using SPA to re-compute p-values produced better calibration (Figure 1 row "SPA Test", third column from the left). However, in the simulations with very low number of carriers (10 to 30 in each of the studies), when the two studies had 50% cases, the SPA test often did not control the type 1 error, while the usual Score test did (Supplementary Figure 4). When the case proportion is lower, it is clear that as the number of carriers grows the Score test's control of type 1 error

improves. However, we could not point to a single and simple rule, for when it is appropriate to use the Score test.

Given that the Score test controls the type 1 error when there are 50% cases in each of the studies, a natural question is whether it is useful to sample controls to generate a dataset with 50% cases. As we saw in simulations, this indeed led to control of type 1 error, however also to a loss of power. The idea behind sampling of controls is that most of the information is in the cases, and therefore, the loss of information is low. However, as our simulations show, this is not correct. Sampling of controls leads to substantial reduction in the sample size, and therefore to both reduction in the quality of estimation of disease probability model, and to reduction in the number of variant carriers used. The properties of the tests depend on the number of carriers.

BinomiRare could be applied with either the usual p-value, or the mid-p-value. While in our primary simulations where the mid-p-value mostly controlled the type 1 error, and almost always improved upon the SPA when it was inflated, we found that the mid-p-value did not control the type 1 error in some settings, especially when the cases were 50% of the sample in both studies, similarly to the SPA. The usual p-value always controlled the type 1 error. Mid-p-values are preferred (when controlling the type 1 error) because they are less conservative. In all, we recommend using the mid-p-values when the case proportion is lower than 50%, and to compute and report both types of p-values.

We formulated a set of recommendations for investigators in studies such as TOPMed, combining together individual-level data from multiple heterogeneous studies. The recommendations take into account the availability of tests across software packages, type 1 error control across extreme settings combining two heterogeneous studies, and power based on modelling assumption and a small number of simulations in the rare variants settings. We did not assess (a) all possible combinations of two studies in terms of their disease proportion and carrier counts, (b) more than two studies, (c) additive mode of inheritance for slightly higher count variants, (d) power simulation settings with different effect sizes across the combined studies, (e) p-value thresholds lower than 1×10^{-4}, (f) estimation of power while accounting for type 1 error of each test (e.g. by identifying and using the specific p-value threshold yielding the desired type 1 error rate in the power simulations). While doing all these would have been helpful, this is not feasible. Both the number of simulations and the disc memory required for saving a lot of data in order to perform additional computations, would be prohibitive. Therefore, we base our recommendations on simplified and extreme settings of type 1 error control. For power, we primarily show that for rare variants, where dominant mode of inheritance is appropriate as the vast majority of individuals are heterozygotes, the BinomiRare test, which is valid, has often similar performance to the SPA test when it is valid as well, or that SPA test has slightly higher power. For higher frequency variants having homozygotes as well, standard statistical thinking posits that the Score and SPA test will be even more powerful (when valid) because they use additional information, and this is seen to some extent in our simulations as well. Further, while theoretically the BinomiRare test is computationally efficient (in terms of both computer time and memory), current

implementations of the Score test use various approaches to speed up matrix computations, making it very efficient in practice, so that BinomiRare is slower. Therefore, when it is valid, the score test is most desirable. The SPA test is less efficient, and its p-value computation is implemented by software packages as re-computation of the naïve Score p-value when it is <0.05. Similar approach could be taken if using BinomiRare test. Still, SPA test is preferred over BinomiRare when it controls type 1 error, due to better power. An important conclusion is that when using the SPA test, our simulations suggest that we can control the total number of carriers rather than the respective number of carriers in each study, i.e. by requiring at least 90-110 carriers in the combined sample. It is not clear what the appropriate minimum number of carriers is (see Figure 3), and it changes by study characteristics. Therefore, it is critically important to perform replication or other follow-up analysis, as is common in genetic association studies. Our simulations are limited by the use of p-value threshold of 10^{-4}, and there could be some differences when using a lower threshold. Because single variant tests are recommended to use at a gene region level, rather than genome-wide, due to the multiple testing burden, we think that p-value threshold of 10^{-4} suffices, while acknowledging that sometimes rare variant associations are tested in a genome-wide manner.

While this work is focused on performance of statistical tests when pooling together data from heterogeneous studies, it highlights issues that are worth addressing in future work. The level of inflation/deflation of the tests when applied on variants with very low number of carriers, varies in different settings (see for example Supplementary Figure 4). Therefore, to assess overall patterns of inflation/deflation due to population stratification, one may need to rely on

results from testing common variants, in which tests follow their asymptotic properties. To generate QQ-plots comparing the observed versus the expected distributions of test results when testing rare variants, Lee et al. [10], developed a resampling based procedure. It would be useful to extend their approach to other tests.

Author contributions: TS conceptualized and designed the study, performed simulation studies and drafted the manuscript. NG developed data visualization. Both authors critically reviewed the approved the manuscript.

Acknowledgements: The authors thank Seunggeun Lee and Rounak Dey for reviewing the manuscript draft and providing helpful comments. T.S. was supported by grants from National Heart, Lung, and Blood Institute (NHLBI; 1R35HL135818, and 1R21HL145425).

References

1. Sarnowski, C., et al., *Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program.* Am J Hum Genet, 2019. **105**(4): p. 706-718.
2. Natarajan, P., et al., *Deep-coverage whole genome sequences and blood lipids among 16,324 individuals.* Nat Commun, 2018. **9**(1): p. 3391.
3. Sarnowski, C., et al., *Whole genome sequence analyses of brain imaging measures in the Framingham Study.* Neurology, 2018. **90**(3): p. e188-e196.
4. Tuijnenburg, P., et al., *Loss-of-function nuclear factor kappaB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans.* J Allergy Clin Immunol, 2018. **142**(4): p. 1285-1296.
5. Amininejad, L., et al., *Analysis of Genes Associated With Monogenic Primary Immunodeficiency Identifies Rare Variants in XIAP in Patients With Crohn's Disease.* Gastroenterology, 2018. **154**(8): p. 2165-2177.
6. Ma, C., et al., *Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants.* Genet Epidemiol, 2013. **37**(6): p. 539-50.
7. Wang, X., *Firth logistic regression for rare variant association tests.* Front Genet, 2014. **5**: p. 187.

8. Dey, R., et al., *A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS.* The American Journal of Human Genetics, 2017. **101**(1): p. 37--49.
9. Sofer, T., *BinomiRare: A robust test of the association of a rare variant with a disease for pooled analysis and meta-analysis, with application to the HCHS/SOL.* Genet Epidemiol, 2017. **41**(5): p. 388-395.
10. Lee, S., et al., *An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies.* Biostatistics, 2016. **17**(1): p. 1-15.
11. Dey, R., et al., *A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS.* Am J Hum Genet, 2017. **101**(1): p. 37-49.
12. Hong, Y., *On computing the distribution function for the Poisson binomial distribution.* Computational Statistics & Data Analysis, 2013. **59**: p. 41-51.
13. Dey, R. and S. Lee, *SPAtest: Score Test and Meta-Analysis Based on Saddlepoint Approximation*. 2018.

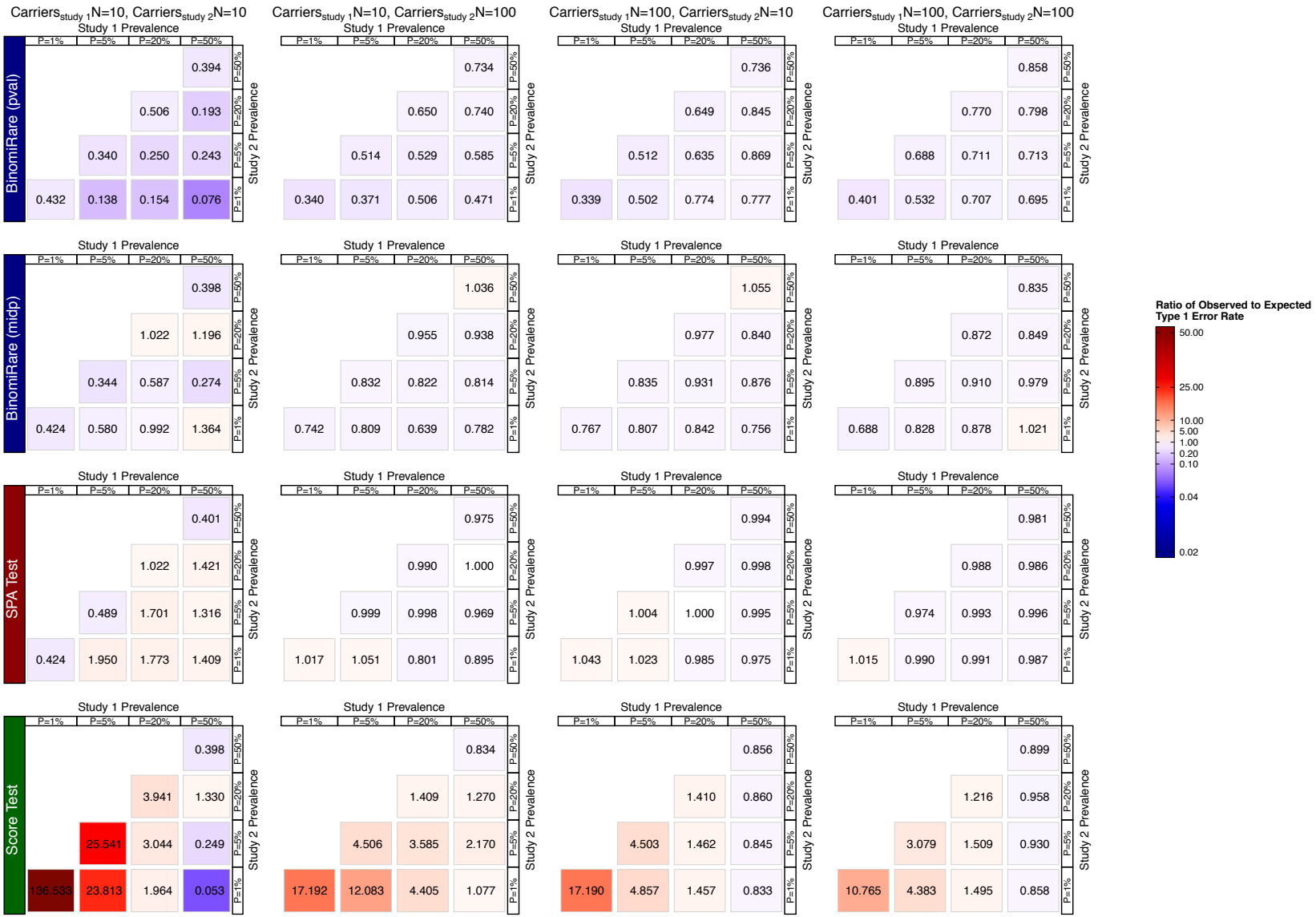


Figure 1: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp), for settings defined by the number of carriers and outcome prevalence in each study. Both studies had n=10,000 individuals. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.

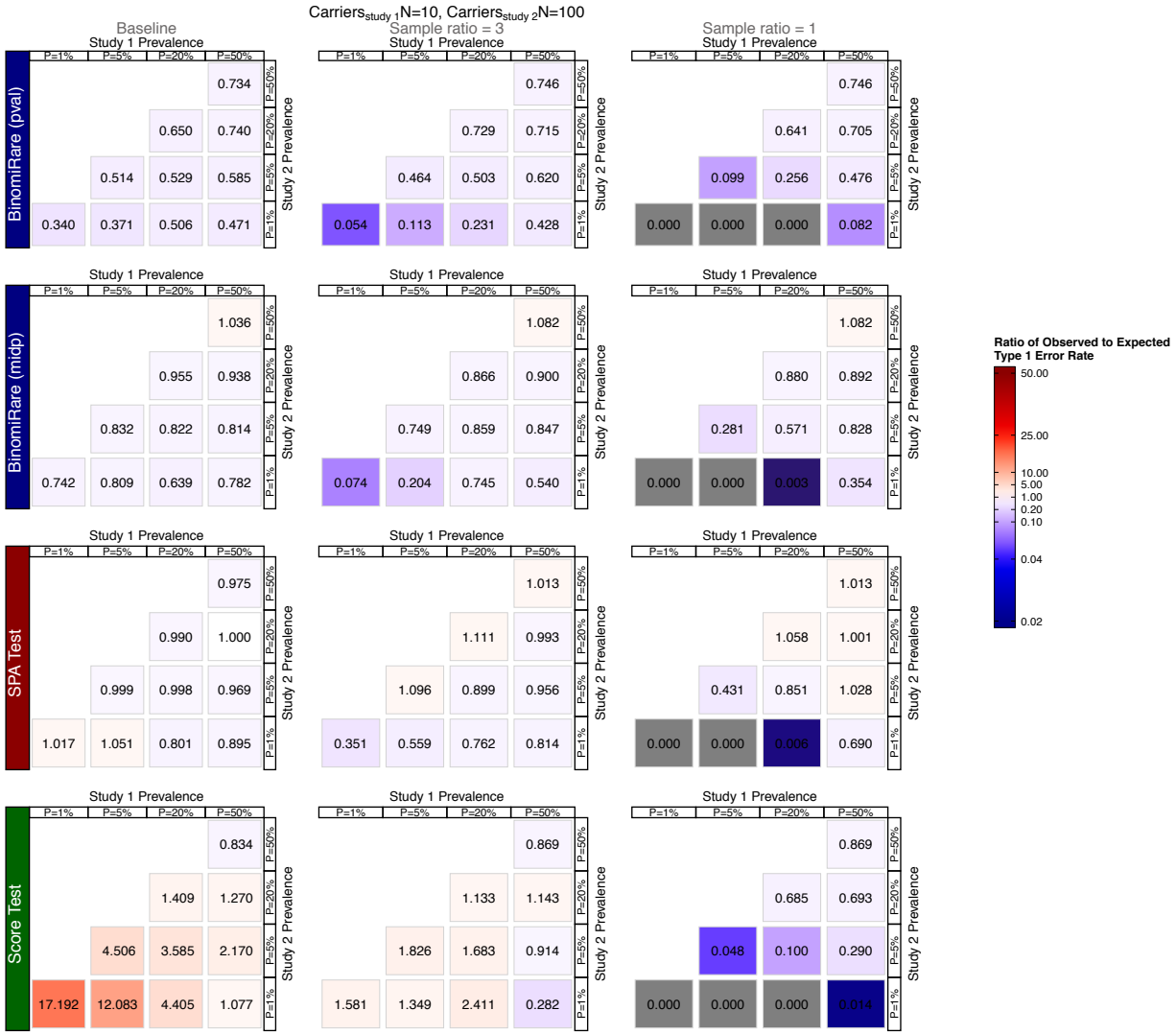


Figure 2: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp). In all settings data from two studies were pooled together, with 10 carriers of the rare variant in study 1, and 100 carriers of the rare variant in study 2. Both studies had n=10,000 individuals. The settings investigated here are defined by the outcome prevalence in each study. The left column ("Baseline") correspond to analysis of the complete data. The middle ("Sample ratio = 3") and right ("Sample ratio = 1") columns provide results for analyses that studies sample sizes by sampling controls to generate samples with the specified control:case ratio. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.

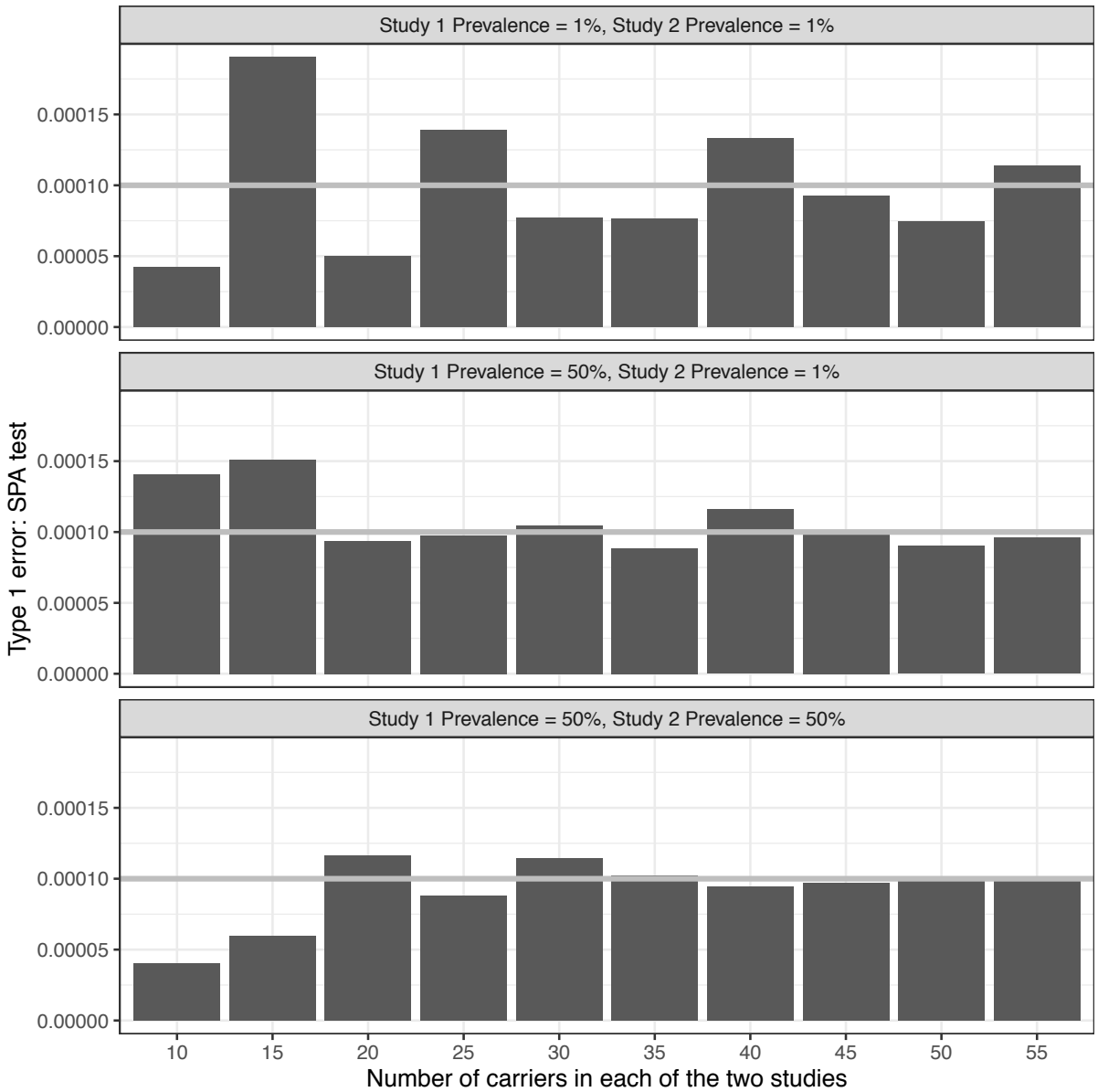


Figure 3: Type 1 error of the SPA test in simulations combining two studies with n=10,000 and equal number of carriers in each study. Simulation settings are defined by the prevalence of the outcome and the number of carriers in each of the studies. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.

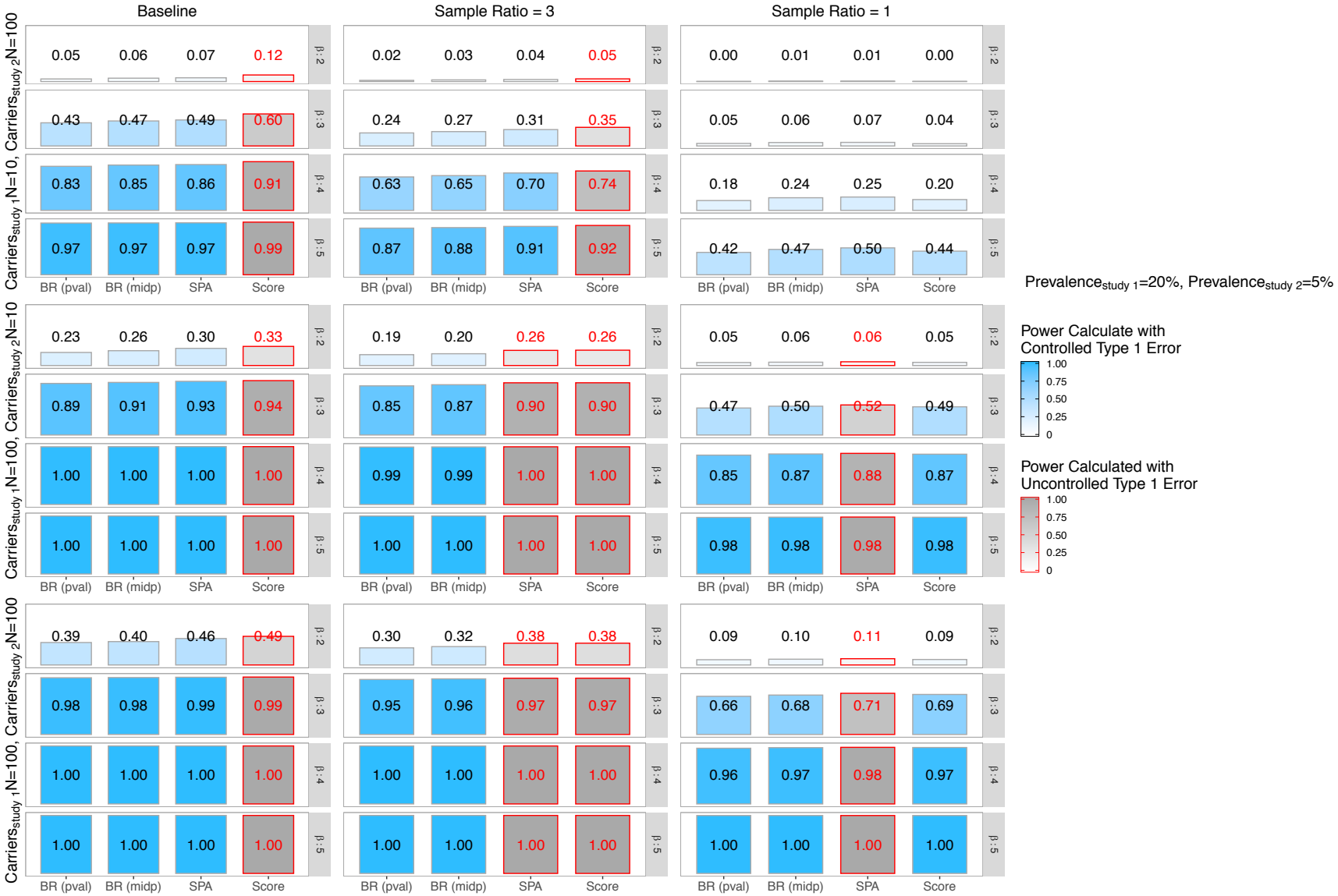
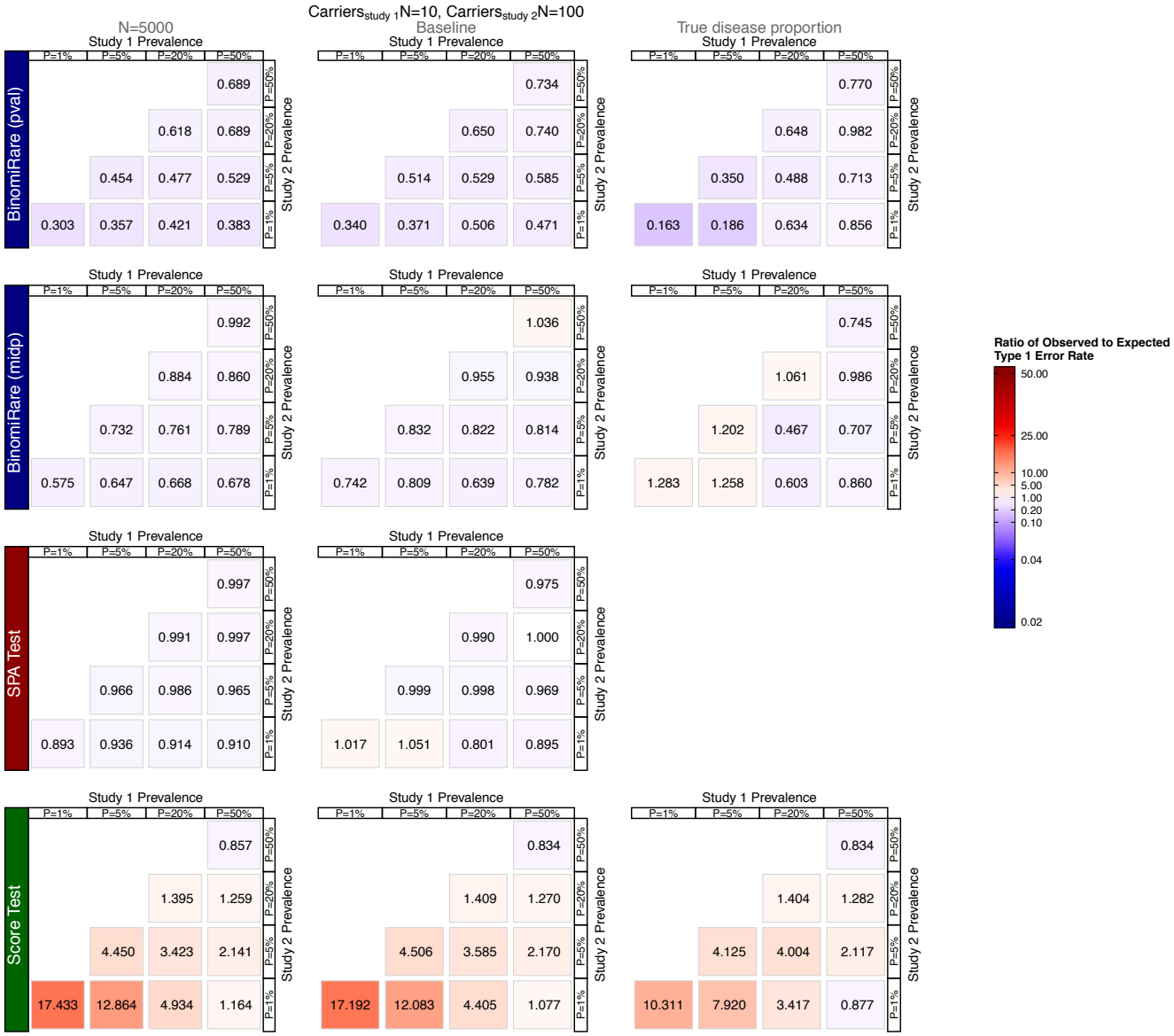
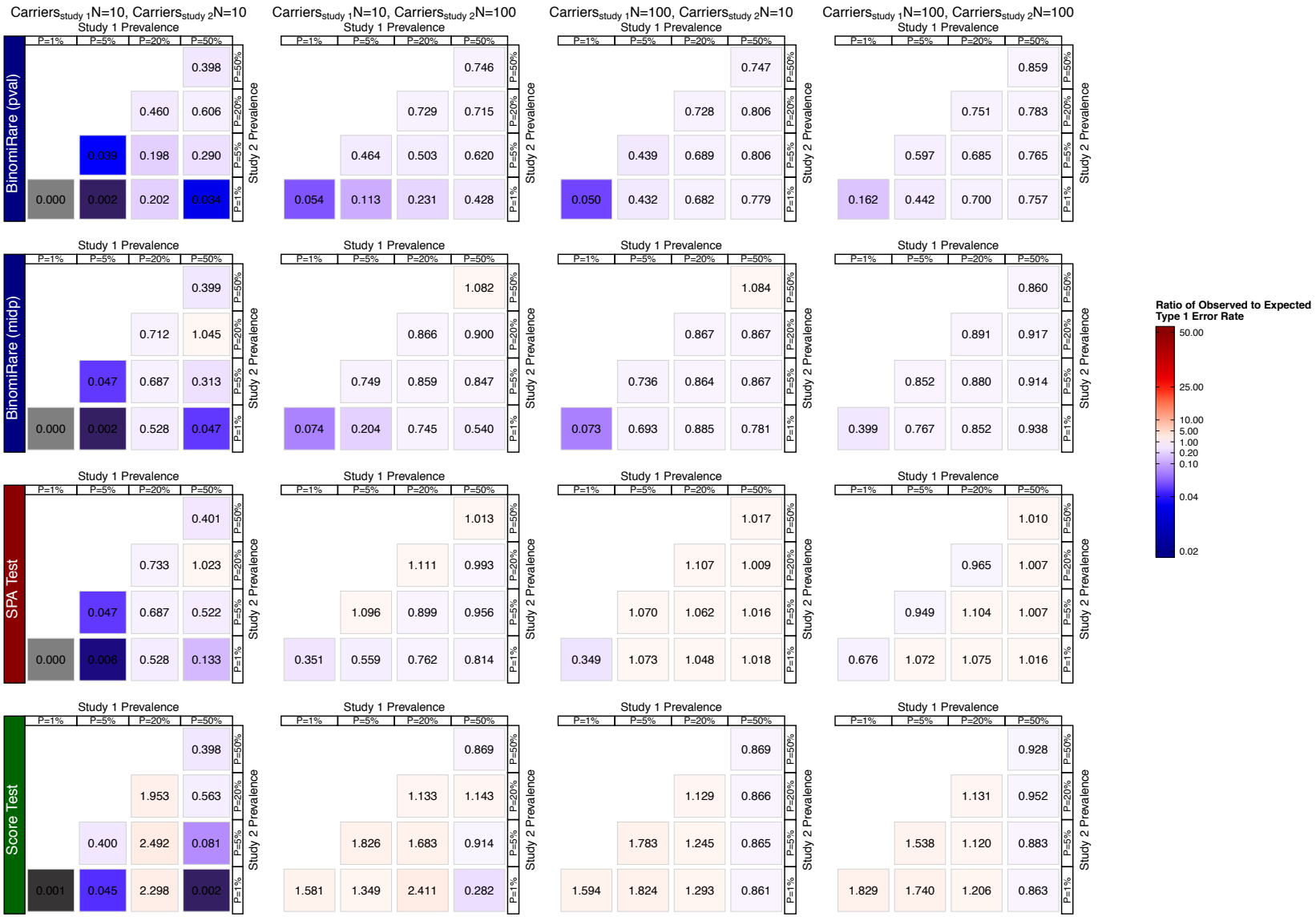


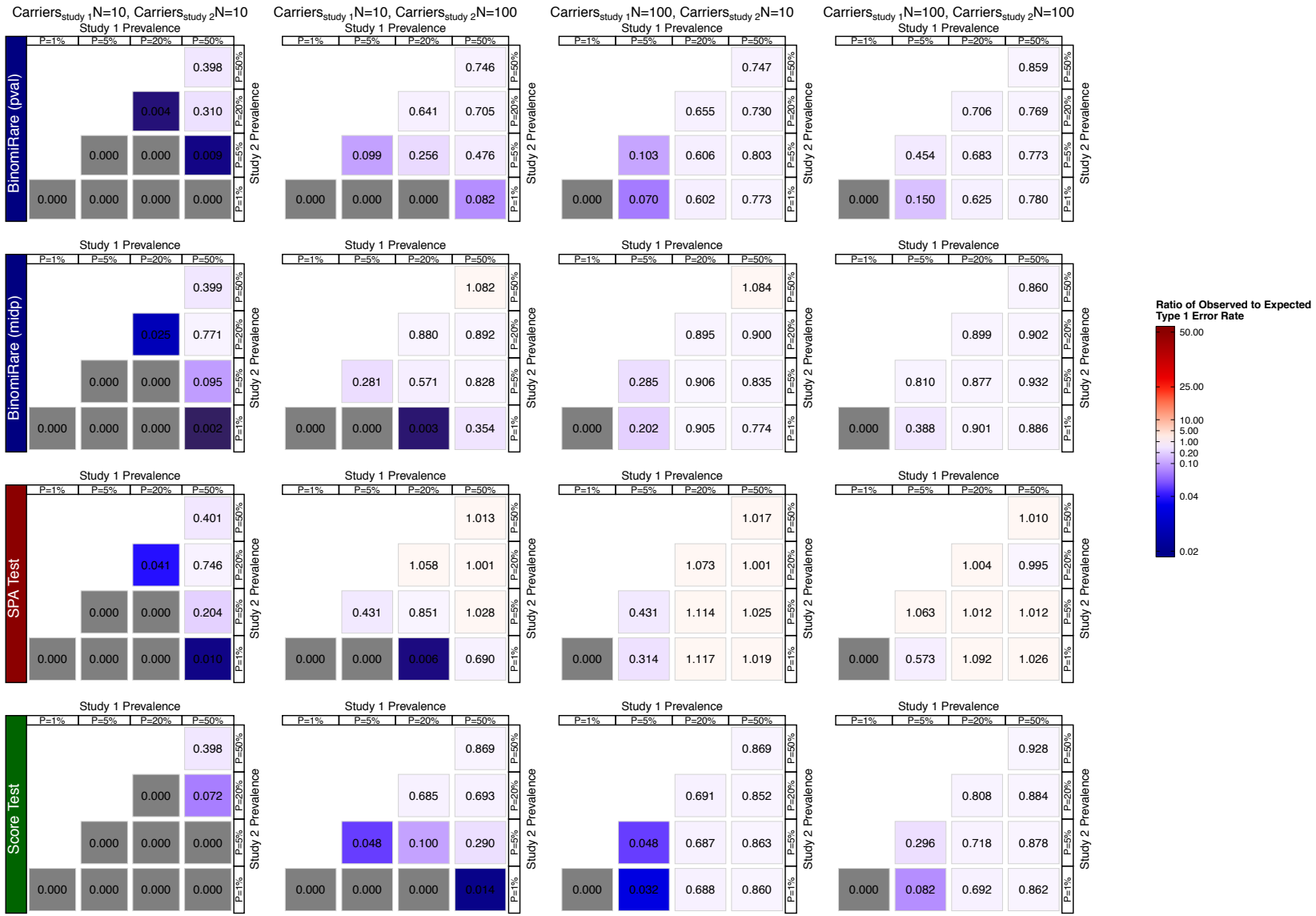
Figure 4: Power estimated in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp). The simulation settings are defined by the number of carriers in each of the studies, and the variant effect size β. The outcome prevalence was fixed at 0.2 in study 1 and 0.05 in study 2. The left column ("Baseline") correspond to analysis of the complete data. The middle ("Sample ratio = 3") and right ("Sample ratio = 1") columns provide results for analyses that studies sample sizes by sampling controls to generate samples with the specified control:case ratio. For each setting we performed 10^4 simulations, and the p-value threshold used for determining significance was 10^{-4}. We color coded the settings according to type 1 error control in the simulations corresponding to the same prevalence, carrier, and sampling settings, but with no variant association.



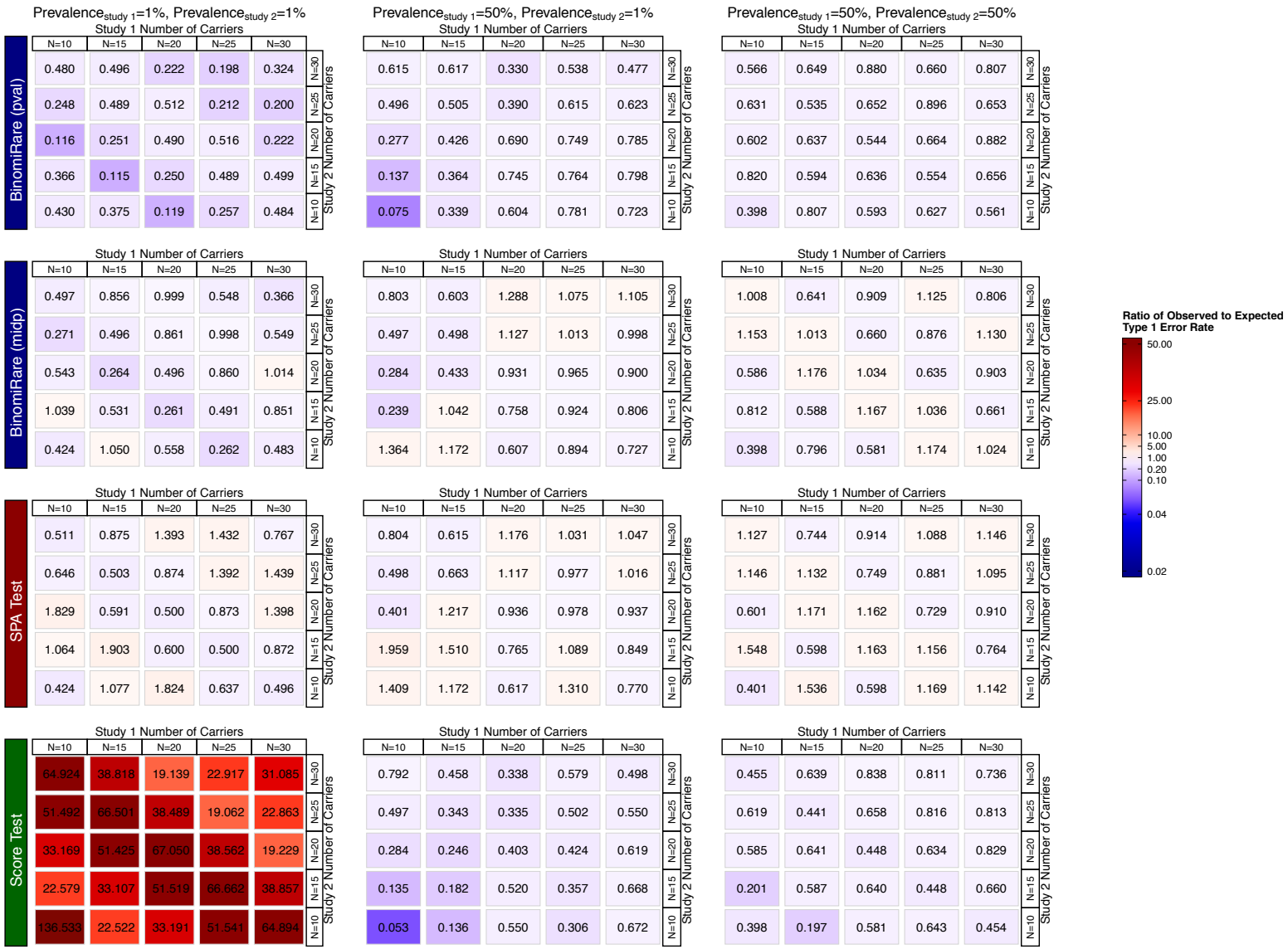
Supplementary Figure 1: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp). In all settings data from two studies were pooled together, with 10 carriers of the rare variant in study 1, and 100 carriers of the rare variant in study 2. The left column ("N=5000") corresponds to settings with 5,000 individuals in each of the studies. The middle column ("Baseline") corresponds to settings with 10,000 individuals in each of the studies, and the right column ("True disease proportion") provides results for analyses that plugged-in the true outcome prevalence in each of the studies in the Score statistic (for the Score test) and provided them the BinomiRare test. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.



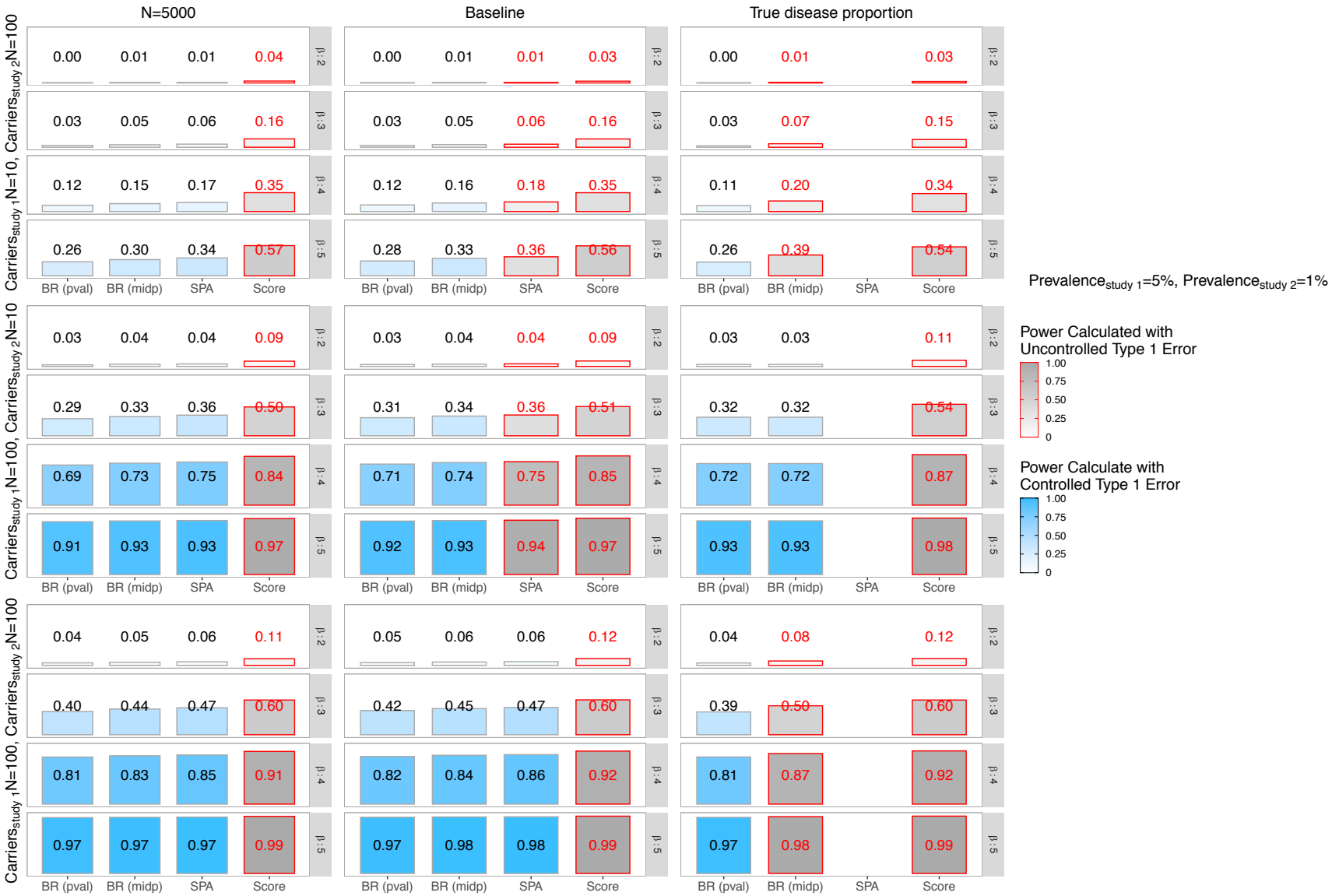
Supplementary Figure 2: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant, and when reducing the sample size by sampling controls to generate samples with up to three controls per case. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp), for settings defined by the number of carriers and outcome prevalence in each study. Both studies had n=10,000 individuals before sampling of controls. Controls were sampled in each study separately. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.



Supplementary Figure 3: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant, and when reducing the sample size by sampling controls to generate samples with one control per case. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp), for settings defined by the number of carriers and outcome prevalence in each study. Both studies had n=10,000 individuals before sampling of controls. Controls were sampled in each study separately. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.



Supplementary Figure 4: Ratios between observed and expected type 1 error rate in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp), for settings defined by the number of carriers and outcome prevalence in each study. Both studies had n=10,000 individuals. For each setting we performed 10^8 simulations, and the p-value threshold used for determining significance was 10^{-4}. Values of 1 correspond to perfect calibration, and values larger (smaller) than 1 correspond to inflation (deflation), or higher (lower) number of detected false associations.



Supplementary Figure 5: Power estimated in simulation studies when testing a binary outcome for association with a rare genetic variant. We compared the naïve score (Score) test, the SPA test, and BinomiRare with the usual p-values (pval) and the mid-p-value (midp). The simulation settings are defined by the number of carriers in each of the studies, and the variant effect size β. The outcome prevalence was fixed at 0.05 in study 1 and 0.01 in study 2. The left column ("N=5000") corresponds to simulations with 5,000 observations in each study. The middle ("Baseline") corresponds to simulations with 10,000 observations in each study. The right column ("True disease proportion") provides results for analyses that plugged-in the true outcome prevalence in each of the studies in the Score statistic (for the Score test) and provided them the BinomiRare test. For each setting we performed 10^4 simulations, and the p-value threshold used for determining significance was 10^{-4}. We color coded the settings according to type 1 error control in the simulations corresponding to the same prevalence, carrier, and sampling settings, but with no variant association.