Ranking Bias in Association Studies

Neal O. Jeffries

Office of Biostatistics Research, National Heart, Lung, and Blood Institute, National Institutes of Health,
Bethesda, Md., USA

Key Words

Estimation bias • Multiple comparisons

Abstract

Background: It is widely appreciated that genomewide association studies often yield overestimates of the association of a marker with disease when attention focuses upon the marker showing the strongest relationship. For example, in a case-control setting the largest (in absolute value) estimated odds ratio has been found to typically overstate the association as measured in a second, independent set of data. The most common reason given for this observation is that the choice of the most extreme test statistic is often conditional upon first observing a significant p value associated with the marker. A second, less appreciated reason is described here. Under common circumstances it is the multiple testing of many markers and subsequent focus upon those with most extreme test statistics (i.e. highly ranked results) that leads to bias in the estimated effect sizes. **Conclusions:** This bias, termed ranking bias, is separate from that arising from conditioning on a significant p value and may often be a more important factor in generating bias. An analytic description of this bias, simulations demonstrating its extent, and identification of some factors leading to its exacerbation are presented.

Copyright © 2009 S. Karger AG, Basel

Introduction

As high dimensional assay technology has become more available for genomic investigations there has been a corresponding increase in methodology to limit false positive findings [1, 2]. Until recently, less attention has been given to the effect sizes of those comparisons that are determined to reach statistical significance. In many instances this may be appropriate as effect sizes may not be of great importance per se; however in other circumstances it may be important to accurately assess the magnitude of the associated effects.

As an example, there has been increased recognition that some form of replication is desirable to support a claim of association found as a single result among a large panel of genetic markers [3]. However, using estimates of effect size derived from an initial study of as many as a million markers can lead to an underpowered evaluation as the initial estimate is likely biased toward more extreme results than are true.

A common rationale given for this bias is that it results from conditioning on the initial estimate meeting p value criteria for declaring significance – often referred to as selection or truncation bias. The term ‘significance bias’ will be used here to stress the conditioning upon a significant p value. Garner [4] describes this in the context of genomewide association scans; Allison et al. [5] describe it in the context of quantitative trait loci of paired siblings. This conditioning on a significant result has the

Table 1. 2×2 table for each biallelic marker

	Case	Control	Total
2 minor alleles	a	b	$a + b$
0 or 1 minor alleles	c	d	$c + d$
	N	N	$2N$

effect of truncating the distribution of the marker's sampling distribution and creating an associated bias. For example, given a normally distributed variable X with mean μ and standard deviation ν , one can show the expected value of X conditional upon $X > c$ (where c may be chosen to generate a significant p value) is given by

$$E[X|X > c] = \mu + \frac{\nu^2 \cdot \phi(c; \mu, \nu^2)}{\text{Prob}[X > c]} \quad (1.1)$$

where $\phi(\cdot)$ denotes the probability density function of normally distributed variable with mean μ and standard deviation ν . The second term on the right hand side of (1.1) captures the bias from conditioning. This description of bias does not recognize the effect multiple testing may exert on creating bias beyond its role in requiring a p value threshold less than 0.05. In this paper multiple testing and focusing on extreme results (i.e. highly ranked results) is shown to generate overestimation in situations where significance bias may have little effect. The terms 'significance bias' and 'ranking bias' are not in common use – the phrases 'truncation bias' and 'selection bias' are used more often but are insufficiently specific and fail to draw the distinction made in this paper.

Though a detailed description is given in the next section the idea of ranking bias is briefly described here. Given a number of test statistics, each one is composed of a trend term (related to true underlying effect size) and some random variation. When looking at the test statistic with the largest positive observed effect size it is likely (under many circumstances) that the random variation component is positive as well – because the large, positive observed effect is the sum of a trend and the random component. This random component captures the difference between the true and observed effect size and positive values correspond to a type of overestimation bias – called ranking bias here.

Allison et al. [5] briefly describe ranking bias but focus the work on significance bias. Göring et al. [6] discuss both significance bias and ranking bias. They make an analogy between the ranking bias discussed here with that corresponding to bias in model selection procedures.

Though they discuss ranking bias their modeling and analytical results are derived solely from models of significance bias. Zöllner and Pritchard [7] use the term 'Winner's Curse' to describe significance bias in a single marker and present an analytical approach toward correcting this bias. Sun and Bull [8] explicitly describe significance bias and more implicitly describe ranking bias in the context of a mathematical model with a number of non-informative markers and 1 informative marker. In the present work we present more general models with a number of informative markers and the relationship between bias (both significance and ranking) and the pattern of the markers' effect sizes is explicitly explored.

Description of Bias

The modeling proposed here is relatively simple both because it is easy to see how to apply this situation to others and because it allows for a transparent partitioning of a statistic into trend and random error terms that facilitates understanding ranking bias. Let N denote the common number of cases and controls drawn in population-based study, G denote the number of biallelic markers under consideration, G_1 the number of markers with different allele frequencies between cases and controls, and G_0 the number with no true difference between cases and controls, i.e. $G = G_1 + G_0$. For simplicity's sake a recessive pattern is assumed to influence the likelihood of disease for the informative markers. For each of the G markers the log odds ratio will be computed from an associated 2×2 table (table 1).

For the j -th marker let θ_j denote the log odds ratio and $\hat{\theta}_j$ its estimate, $\log(ad/bc)$. Then

$$\mathcal{Z}_j = \frac{\hat{\theta}_j - \theta_j}{\widehat{se}_j} \sim N(0,1) \quad (2.1)$$

where

$$\widehat{se}_j = \sqrt{1/a + 1/c + 1/b + 1/d}$$

and $\sim N(0,1)$ indicates an approximate Gaussian distribution with mean 0 and variance 1. Given the observed log odds and associated test statistic

$$T_j = \frac{\hat{\theta}_j}{\widehat{se}_j} \quad (2.2)$$

one can use the relation in (2.1) to obtain p values for the hypothesis that $\theta_j = 0$ as well as to generate confidence intervals

$$\hat{\theta}_j \pm \Phi_{1-\alpha/2}^{-1} \cdot \widehat{se}_j$$

where $1 - \alpha$ is confidence associated with the interval and $\Phi_{1-\alpha/2}^{-1}$ is the associated percentile from a $N(0,1)$ distribution.

To investigate how bias arises it is useful to rewrite the relationship between (2.1) and (2.2). First, write

$$\widehat{sd} = \sqrt{\frac{1}{\hat{p}_{2|Case}} + \frac{1}{1 - \hat{p}_{2|Case}} + \frac{1}{\hat{p}_{2|Control}} + \frac{1}{1 - \hat{p}_{2|Control}}} \quad (2.3)$$

where

$$\hat{p}_{2|Case} = a/N \text{ and } \hat{p}_{2|Control} = b/N \quad (2.4)$$

denote the empirically observed probability of two minor alleles in the case and control populations, respectively. The term $\widehat{sd}$ is an estimate of σ , which is defined by equation (2.3) after replacing the estimated probabilities of having both minor alleles by the true unknown proportions of having both minor alleles. Then we can decompose the test statistic T_j as

$$T_j = \frac{\sqrt{N}\hat{\theta}_j}{\widehat{sd}_j} = \frac{\sqrt{N}(\hat{\theta}_j - \theta_j)}{\widehat{sd}_j} + \frac{\sqrt{N}\theta_j}{\widehat{sd}_j} \quad (2.5)$$

$$= \mathcal{Z}_j + \frac{\sqrt{N}\theta_j}{\widehat{sd}_j} \text{ where } \mathcal{Z}_j = \frac{\sqrt{N}(\hat{\theta}_j - \theta_j)}{\widehat{sd}_j}. \quad (2.6)$$

The decomposition shows the test statistic T_j is composed of a trend term,

$$\frac{\sqrt{N}\theta_j}{\widehat{sd}_j}$$

and a mean 0 random fluctuation term, $\mathcal{Z}_j$. Note that $\mathcal{Z}_j$ and T_j differ in that $\mathcal{Z}_j$ has an approximately normal distribution centered about 0 while for T_j this is only true when $\theta_j = 0$. The $\mathcal{Z}_j$ term captures the difference between estimated log odds and true log odds and its distribution is of primary interest in assessing overestimation bias.

For each j , $\mathcal{Z}_j$ has approximately an unconditional $N(0,1)$ distribution. A critical question concerns the distribution of $\mathcal{Z}_j$ when j corresponds to most extreme observed T_j statistic. We will focus upon the case of largest T_j – the case for smallest, i.e. most negative, statistic is similar and generalizations involving it and other extreme statistics are discussed below.

Let $r_1, r_2, \dots, r_G$ order the T test statistics where

$$T_j = \frac{\sqrt{N}\hat{\theta}_j}{\widehat{sd}_j} \text{ and } T_{r_1} \leq T_{r_2} \leq \dots \leq T_{r_G}.$$

Typically, the distribution of

$$\mathcal{Z}_{r_G} = \frac{\sqrt{N}(\hat{\theta}_{r_G} - \theta_{r_G})}{\widehat{sd}_{r_G}} \quad (2.7)$$

is not that of a normal distribution with mean 0 – this will be demonstrated with simulations below. The reason for the bias is that if a number of markers have similar sized trend components

$$\frac{\sqrt{N}\theta_j}{\widehat{sd}_j}$$

then the random components $\mathcal{Z}_j$ are likely to play a larger role in determining which marker has the maximal observed value, T_{r_G} . Under these conditions of multiple markers with similar trend values the contribution of $\mathcal{Z}_j$ for that with largest T_j value is likely to be positive. From the definition of $\mathcal{Z}_{r_G}$ in (2.7) we see a positive value corresponds to overestimation of the underlying log odds ratio. The expected value of the difference between $\hat{\theta}_{r_G}$ and θ_{r_G} is what is meant by ranking bias, i.e.

$$\text{Ranking Bias} = E[\hat{\theta}_{r_G} - \theta_{r_G}].$$

As discussed further below, it is important to note that r_G , the index of the marker with greatest observed effect size, is random and can change in repeated sampling from the relevant populations.

Simulation Studies Demonstrating Bias

To illustrate the ranking bias a simple set of simulations are presented. In this case-control study there are $G = 500$ markers and there are $N = 500$ cases and 500 controls. In these simulation studies the markers are generated as independent variables. Populations of cases and controls were generated with 480 of the markers having no differences in allele frequencies. For the remaining 20 markers the odds ratios for diseased relative to healthy were 1.05, 1.10, ..., 1.95, and 2.0. The probability of having 2 minor alleles in the control group was randomly chosen for each marker to lie between 0.05 and 0.25 (i.e. the minor allele frequencies varied between 0.224 and 0.5).

Given a simulated sample of cases and controls the following steps were performed:

For each of $G = 500$ markers find

$$T_j = \frac{\sqrt{N}\hat{\theta}_j}{\widehat{sd}_j}, j \in \{1, \dots, G\}.$$

Designate r_G as that index that is associated with largest T_j , i.e.

$$r_G = \text{Arg} \max_{j \in \{1, \dots, G\}} T_j.$$

Fig. 1. Distribution of $Z_{r_G} = N(\hat{\theta}_{r_G} - \theta_{r_G})/\widehat{sd}_{r_G}$ and a standard normal distributions; maximum odds ratio of 2.0.

Because the true log odds ratios (θ_i) are known, one can calculate a realization of

$$Z_{r_G} = \frac{\sqrt{N}(\hat{\theta}_{r_G} - \theta_{r_G})}{\widehat{sd}_{r_G}}.$$

After making many such samples (1000 simulations) the resulting distribution of Z_{r_G} is shown along with the standard normal distribution in figure 1. Of note are the facts that the Z_{r_G} distribution is not centered about 0 (indicating $\hat{\theta}_{r_G}$ typically overestimates θ_{r_G}) and that the distribution appears bimodal. Also, the distribution is relatively narrow compared to the standard normal density – this is perhaps surprising given asymptotic efficiency properties of normal approximations. However, the standard normal distribution is not relevant here as it neglects to take into account that r_G , the index of the maximal T statistic, is random and changes in different samples drawn from the underlying populations. For example, in the 1000 simulations approximately 34% of the time r_G corresponded to that marker with the odds ratio of 2.0, 32% of the time r_G was associated with the 1.95 odds ratio, and 8% of the time with that associated with 1.90, and so forth. Because different markers are chosen as the simulated datasets change a mixture distribution arises as suggested by the non-unimodal character of the distribution.

The observed and true log-odds values associated with the largest test statistic (i.e. $\hat{\theta}_{r_G}$ and θ_{r_G}) for each of the 1000 simulations are shown in figure 2. The discrete ver-

Fig. 2. Comparison of θ_{r_G} (X-axis) and $\hat{\theta}_{r_G}$ (Y-axis) in 1000 simulations.

tical bands correspond to the 12 distinct values of θ_{r_G} that were obtained: $\{0, \log(1.45), \log(1.50), \dots, \log(2.0) = 0.69\}$. The degree to which the points lie above the diagonal 45 degree line indicate the amount of bias in each simulation (fig. 2).

Ranking Bias vs. Significance Bias

Thus far there has been no reliance upon a p value threshold as a means for creating the observed bias. Truncation bias related to conditioning on a significant p value (i.e. significance bias) is a common rationale given for overestimation bias in marker studies and is an independent contributing factor. The bias demonstrated thus far is not conditioned on the largest T statistic first exceeding some threshold, e.g. one that corresponds to a p value = $0.025/G$. Hence it cannot be significance bias. To clarify the difference significance bias is now further discussed.

Recent work involving significance bias focuses upon the effect in a single marker [4, 7]. In our setting we illustrate the idea using the marker with largest true effect size. Let r_G^0 designate this marker, i.e.

$$\frac{\theta_{r_G^0}}{\sigma_{r_G^0}} > \frac{\theta_j}{\sigma_j} \text{ for all } j \neq r_G^0, j \in \{1, \dots, G\}.$$

Because r_G^0 is fixed (though unknown) it is the case that for large sample size N

$$Z_{r_G^0} = \frac{\sqrt{N}(\hat{\theta}_{r_G^0} - \theta_{r_G^0})}{\widehat{sd}_{r_G^0}} \sim \text{Normal}(0,1) \quad (4.1)$$

Table 2. Ranking and significance bias on a log odds ratio scale

	$\hat{\theta}_{r_G^0}$ significance bias: $E[\hat{\theta}_{r_G^0} T_{r_G^0} > 3.90] - \theta_{r_G^0}$	$\hat{\theta}_{r_G}$ ranking bias: $E[\hat{\theta}_{r_G} - \theta_{r_G}]$	$\hat{\theta}_{r_G}$ combined significance and ranking bias: $E[\hat{\theta}_{r_G} - \theta_{r_G} T_{r_G} > 3.90]$
Scenario 1: n = 500, $G_1 = 20$ informative ORs {1.05, ..., 2.0}	0.072	0.191	0.192
Scenario 2: n = 500, $G_1 = 40$ {1.025, ..., 2.0}	0.072	0.230	0.230
Scenario 3: n = 500, $G_1 = 5$ {1.20, ..., 2.0}	0.072	0.113	0.129
Scenario 4: n = 500, $G_1 = 1$ {2.0}	0.072	0.069	0.085
Scenario 5: n = 500, $G_1 = 1$ {4.0}	0.000	0.000	0.000
Scenario 6: n = 1000, $G_1 = 20$ {1.05, ..., 2.0}	0.002	0.112	0.112
Scenario 7: n = 500, $G_1 = 10$ {1.04, ..., 1.40}	0.330	0.448	0.471
Scenario 8: n = 500, $G_1 = 10$ {1.30, ..., 1.30}	0.395	0.476	0.543

and hence $E[\hat{\theta}_{r_G^0}] = \theta_{r_G^0}$. However, a bias is introduced if one conditions on $T_{r_G^0} = \sqrt{N}\hat{\theta}_{r_G^0}/\widehat{sd}_{r_G^0}$ exceeding a threshold value so that the associated p value is sufficiently small. For instance, if $T_{r_G^0} > 3.90$ then the associated p value will be less than 0.025/500 and in our simulations the marker would be judged showing significant association using a Bonferroni criterion. However, using standard derivations concerning conditional expectations of normally distributed random variables one can show that

$$E[\mathcal{Z}_{r_G^0} | T_{r_G^0} > 3.90] \approx E[\mathcal{Z}_{r_G^0} | \mathcal{Z}_{r_G^0} > 3.90 - \sqrt{N}\theta_{r_G^0}/\sigma_{r_G^0}] \quad (4.2)$$

$$= \frac{\phi(3.90 - \sqrt{N}\theta_{r_G^0}/\sigma_{r_G^0})}{1 - \Phi(3.90 - \sqrt{N}\theta_{r_G^0}/\sigma_{r_G^0})} \quad (4.3)$$

where ϕ and Φ denote the density and cumulative distribution function of a standard normal distribution. Consequently it follows from (4.1) that

$$E[\hat{\theta}_{r_G^0} | T_{r_G^0} > 3.90] = \theta_{r_G^0} + \text{significance bias} \quad (4.4)$$

where significance bias

$$\approx \frac{\sigma_{r_G^0}}{\sqrt{N}} \cdot \left(\frac{\phi(3.90 - \sqrt{N}\theta_{r_G^0}/\sigma_{r_G^0})}{1 - \Phi(3.90 - \sqrt{N}\theta_{r_G^0}/\sigma_{r_G^0})} \right). \quad (4.5)$$

The approximation arises because $\widehat{sd}_{r_G^0}$ has been replaced by $\sigma_{r_G^0}$ and the asymptotic normal distribution was used. From the result in (4.5) one can show how changes in sample size, odds ratio, allele frequency (affecting the σ term), and p value threshold change the significance bias. Such an assessment was performed by Garner [4].

However, in most circumstances r_G^0 is not known, i.e. in a study with a large number of markers and little a priori knowledge of which markers have the strongest ef-

fects, only r_G is observed and the preceding development based upon r_G^0 is not really applicable in practice. What is of practical interest is $E[\hat{\theta}_{r_G} | T_{r_G} > 3.90]$. However analytical evaluation of this expression is more complicated than assessing bias for r_G^0 because r_G is random. Consequently, simulations may be used to explore the following types of overestimation bias

significance bias related to $\hat{\theta}_{r_G^0} : E[\hat{\theta}_{r_G^0} | T_{r_G^0} > 3.90] - \theta_{r_G^0}$,

ranking bias related to $\hat{\theta}_{r_G} : E[\hat{\theta}_{r_G} - \theta_{r_G}]$,

and the combination of ranking bias and significance bias related to $\hat{\theta}_{r_G} : E[\hat{\theta}_{r_G} - \theta_{r_G} | T_{r_G} > 3.90]$.

Sets of simulations were performed to examine the relative size of biases across varying conditions. Each design has 500 markers with a smaller number of informative markers (G_1) ranging between 20 and 1; the results are shown in table 2. Scenario 1 corresponds to the simulations generating figure 1, the maximum true odds ratio is 2 (log odds ratio of 0.69) and the maximum effect size is given by $\theta_{r_G^0}/\sigma_{r_G^0} = 0.69/3.465$ where 3.465 is the associated σ value derived from the underlying allele probabilities (the probability of 2 minor alleles is 0.1685 in the control group) and odds ratio. An estimate of σ is given by $\widehat{sd}$ in equation (2.3). From equation (4.5) an analytical estimate of the significance bias applied to this particular marker is 0.072 as shown in the first column of table 2 (all biases in this section are reported on a log odds ratio scale). In the 1000 simulations the observed bias for this marker arising from conditioning on its T statistic exceeding 3.90 was 0.075 – in good agreement with the analytic result. The simulations' estimate of ranking bias related to $\hat{\theta}_{r_G}$ is about 0.191 in this case – approximately 2.5 times larger than the significance bias of $\hat{\theta}_{r_G^0}$. As an aside it is worth noting that other markers with smaller

true effect sizes will have larger significance bias, but smaller likelihood of reaching the threshold. Finally we can get an estimate of the combined effects of ranking bias and significance bias for $\hat{\theta}_{r_G}$ by examining just those simulations in which $T_{r_G} > 3.90$. In these 1000 simulations $T_{r_G} > 3.90$ in 99% of the occasions so the additional significance bias induced by conditioning on this common event is essentially negligible as it raises the bias only from 0.191 to 0.192 (table 2).

Next we investigate how results change when the number of informative markers change from the design in Scenario 1. Scenario 2 has a similar structure but has 40 informative markers taking values 1.025, 1.05, ..., 1.975, and 2.0. Because the sample size, standard deviations, and maximum odds ratio of 2 stay the same the significance bias associated with $\hat{\theta}_{r_G}^0$ remains the same at 0.072. However, the changes in informative effect sizes induce changes in the ranking bias and combined ranking and significance bias associated with $\hat{\theta}_{r_G}$. These figures have increased to 0.230 for both bias measures. Conversely, when the of informative markers are reduced to only 5 as in Scenario 3, taking values 1.20, ..., 1.80, 2.0, then the two bias measures associated with $\hat{\theta}_{r_G}$ are reduced to 0.113 and 0.129 though the significance bias associated with $\hat{\theta}_{r_G}^0$ remains unchanged. These three scenarios show how bias in $\hat{\theta}_{r_G}$ depends upon the distribution of all markers, not just the conditions associated with the single marker having the largest true effect size.

In Scenario 4 the more extreme case of a single informative marker with an odds ratio of 2 is considered. Here the ranking bias of $\hat{\theta}_{r_G} = 0.069$ is considerably less because there is a relatively low probability that any other marker will be associated with r_G – in 92.7% of the simulations the marker with an odds ratio of 2 generated the largest T statistic. Therefore, the significance bias of $\hat{\theta}_{r_G}^0$ and $\hat{\theta}_{r_G}$ are quite comparable as shown in the first and third column. In Scenario 5 the effect size of the single informative marker is increased so that this marker produces the smallest p value in all the simulations so there is essentially no ranking bias. Further, the sample size is sufficiently large so that the associated T statistic will exceed the 3.90 cutoff with an extremely high probability so there is no detected significance bias. Scenario 6 changes Scenario 1 by increasing the sample size. As in Scenario 5 this essentially eliminates the significance bias associated with the marker having an odds ratio of 2. However, there is still considerable bias associated with $\hat{\theta}_{r_G}$ indicating ranking bias is a more important factor than significance bias in this case.

In Scenario 7 more modest effect sizes are considered. Here there are 10 informative ORs between 1.04 and 1.40.

In this case 10,000 simulations were considered as there was low power (7.1%) for any of the 10 markers with informative ORs to exceed the Bonferroni threshold. Because the power is low, one might expect higher significance bias and this is in fact the case. The marker with OR = 1.4 had significance bias of 0.330 – indicating the observed log-odds ratio was typically twice as high as the true effect ($\log(1.4) = 0.336$) when the criterion was met. However, in this case the ranking bias was extensive as well with values of 0.448 (unconditional) and 0.471 (conditional upon a significant statistic). Ranking bias remains high because those conditions that create large significance bias (e.g. a large Bonferroni threshold relative to true effect size) also tend to create large ranking bias (many different markers could potentially be highest ranked, with large bias for any one marker being highest ranked). In general it is difficult to find conditions in which significance bias will appreciably exceed the rank bias – either unconditional or conditional upon the highest ranked test statistic exceeding a threshold. Indeed, as it is usually the case that $\hat{\theta}_{r_G} \geq \hat{\theta}_{r_G}^0$ and $\theta_{r_G} \leq \theta_{r_G}^0$ one would not expect significance bias, $E[\hat{\theta}_{r_G}^0 - \theta_{r_G}^0 | T_{r_G}^0 > 3.90]$, to exceed the conditional ranking bias, $E[\hat{\theta}_{r_G} - \theta_{r_G} | T_{r_G} > 3.90]$.

Finally, Scenario 8 shows the case when all the informative markers have the same effect size. As described in the Discussion and Appendix sections one should expect high ranking bias and this is in fact the case.

Among the conclusions to be drawn from these comparisons of different types of bias are (1) significance bias associated with a single, fixed marker may not be particularly relevant; (2) ranking bias is distinct from significance bias; (3) ranking bias is likely the greater of the two sources when many variables are considered. Finally, as shown by comparing the first 4 scenarios in table 2 that have many common design elements, the effect sizes of all markers play a role in determining ranking bias – efforts to understand bias associated with highly ranked markers must take this into account.

Discussion

Appendix A shows that ranking bias (scaled by the marker's standard deviation and sample size) for independent markers is given approximately by

$$E\left[\frac{\sqrt{N}(\hat{\theta}_{r_G} - \theta_{r_G})}{\widehat{sd}_{r_G}}\right] = E[Z_{r_G}] \approx \frac{1}{\sqrt{2\pi}} \sum_{j=1}^G E\left[e^{\frac{-M_j^2}{2}}\right] \quad (5.1)$$

where

$$M_{-j} = \max_{k \neq j} T_k - \sqrt{N} \frac{\theta_j}{\sigma_j} = \max_{k \neq j} \left(Z_k + \sqrt{N} \left(\frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right) \quad (5.2)$$

for j in $1, \dots, G$. In words, M_{-j} represents the largest of all T_k (besides T_j) minus the trend term corresponding to the j -th marker. As discussed further in the appendix, some understanding of the bias can be deduced from relation (5.1). First, and perhaps most importantly, it is clear the bias depends in principle on the effect sizes, θ_j/σ_j , for all G of the markers. In practice, perhaps a few or even just one marker is relevant, but in principle one needs to take into account the true effect sizes of all.

As the sample size increases then each of the M_{-j} increase toward infinity in absolute value (with probability 1 under common conditions) so that the bias tends to zero as we expect assuming there are some informative markers and one is more informative than the others. On the other hand, bias is considerably worse when the effect sizes, θ_j/σ_j , are the same for all $j = 1, \dots, G$ as would be the case is there were no informative markers. If the markers are independent then in this case

$$E[Z_{r_G}] \approx \frac{G}{\sqrt{2\pi}} E \left[e^{-\frac{1}{2}(\max\{X_1, \dots, X_{G-1}\})^2} \right]$$

where the X_j have independent standard normal distributions. Intermediate situations (some informative markers of different sizes with a fixed sample size) are associated with varying degrees of bias.

An important related question that underlies equation (5.1) is: given repeated samples of size N from the underlying populations, how many different markers could reasonably be selected as r_G ? If one marker has an effect size that is much larger than all the rest then in repeated samples we would expect r_G to almost always correspond to that marker and no appreciable bias should result – empirically this is seen in Scenario 5 of table 2. Analytically this is demonstrated in the appendix when $\gamma \rightarrow \infty$. If, on the other hand, there are a number of markers that have non-trivial probabilities of producing the largest test statistic in a given sample then in these cases the corresponding Z_j terms play a larger role in determining which marker is chosen and we would expect a correspondingly greater degree of bias given the relationship between bias and Z_j . One way of thinking about these factors is to see that the distribution of Z_{r_G} is a mixture of the conditional distributions Z_j given $r_G = j$ with the weights of the mixture given by the probability that $r_G = j$. These conditional distributions are typically not centered about 0 – in

Table 3. Bias of markers exceeding threshold under scenario 1

Marker rank	Prob [$T_{r_j} > 3.90$]	$E[\hat{\theta}_{r_j} - \theta_{r_j} T_{r_j} > 3.90]$
1st highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{500}}$	0.990	0.192
2nd highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{499}}$	0.925	0.143
3rd highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{498}}$	0.741	0.124
4th highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{497}}$	0.438	0.128
5th highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{496}}$	0.193	0.132
6th highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{495}}$	0.054	0.131
7th highest ranked: $\hat{\theta}_{r_j} = \hat{\theta}_{r_{494}}$	0.007	0.129

fact they are subject to a kind of truncation bias similar in form to that arising from conditioning on a significant p value, but instead arising from conditioning on $r_G = j$. When there are a number of markers with non-negligible $\text{Prob}[r_G = j]$ this mixture leads to non-negligible bias for Z_{r_G} . Appendix A provides technical details.

The results thus far presented only focus upon the marker with largest test statistic. A natural question is how much bias may be present if attention is paid to all markers that exceed a significance threshold. This corresponds to the common strategy of following-up on all markers that meet a stringent significance threshold. For each simulation under Scenario 1 in table 2 the number of markers with T scores > 3.90 was recorded and the difference between the observed $\hat{\theta}$ and the true underlying θ was recorded. Table 3 provides information about all highly ranked markers that exceed the threshold. In 99% of the simulations at least one marker met the threshold requirement. The bias associated with the highest ranking marker exceeding the threshold is 0.192 when measured on a log odds scale – this is essentially the same information as reported in the first row of table 2. Again it is worth noting that this is not reporting the results for the marker with true log-odds ratio of 2.0, but rather the average bias for the marker with the highest observed T statistic – in 34% of the simulations this did correspond to the marker with odds ratio of 2.0, in 32% of occasions the marker with odds ratio of 1.95, etc.... The other rows of table 3 provide new information: in 92.5% of the simu-

lations the second highest ranked marker met the Bonferroni criteria and the bias for these instances was estimated as 0.143 log-odds units. Table 3 gives information for the 3rd, 4th, 5th, 6th, and 7th mostly highly ranked markers (no more than 7 markers ever met the criteria) and we see the bias conditional upon exceeding the threshold remains considerable. This table indicates that the ranking bias is not just a problem for the most highly ranked marker – it can be present and substantial for all those markers meeting significance criteria.

Conclusions

In this report the focus has been on measuring bias associated with the largest effect size. Ranking bias is potentially present in any study with multiple tests where attention is drawn to outcomes associated with the most extreme observed effect sizes. Here we have seen that focus upon highly ranked results, rather than conditioning on significant p values, can be responsible for much of overestimation bias. Some authors have put forth methods to correct for both ranking and significance bias. The bootstrap/cross-validation approach put forth by Sun and Bull [8] can accommodate both types of bias. A similar bootstrap approach by Jeffries [9, 10] has been employed to correct for ranking bias in microarray and diagnostic modeling contexts. Analytic approaches have been put forth by Zöllner and Pritchard [7], Ghosh et al. [11], and Zhong and Prentice [12] however these appear to address significance bias but not ranking bias. The more extensive dependence of rank bias on potentially all effect sizes (as opposed to significance bias that depends only upon one marker's effect size) complicates analytical solutions. It is likely that analytic approaches to biased overestimation from genomic studies that ignore the need to consider all markers' effect sizes may be missing an important aspect of the problem, i.e. the distribution of the most extreme test statistic is a mixture distribution depending, in principle, upon all markers' parameters.

The modeling in this paper was very simple: a single stage study (i.e. no two-stage or higher stage designs) with a recessive genetic model and Wald test approach to evaluating significance. The simple approach allowed for a simple analytic description of the problem – more complicated designs are likely prone to the same types of biases and work remains for examining the role of ranking bias in these circumstances.

Appendix A: Calculation of the Bias

Here an effort is made to sketch the degree of bias that may be expected and link this magnitude to some factors such as sample size, distribution of true effect sizes, and the number of markers. Simplifying assumptions will be employed as necessary. Recall that the j -th test statistic T_j may be decomposed into a trend and random fluctuation component as

$$T_j = Z_j + \sqrt{N} \frac{\theta_j}{\widehat{sd}_j}$$

where $Z_j = \sqrt{N}(\hat{\theta}_j - \theta_j)/\widehat{sd}_j$ and Z_j is approximated by a standard normal distribution. Then we may calculate the expected value of Z_{r_G} (and hence obtain a measure of the bias of $\hat{\theta}_{r_G} - \theta_{r_G}$) as

$$E[Z_{r_G}] = \sum_{j=1}^G E[Z_{r_G} | r_G = j] P[r_G = j] \quad (\text{A.1})$$

and

$$E[Z_{r_G} | r_G = j] P[r_G = j] = \left(\int z_j f_{z_j | r_G=j}(z_j) dz_j \right) P[r_G = j] \quad (\text{A.2})$$

$$= \frac{\int z_j f_{z_j}(z_j, r_G = j) dz_j}{P[r_G = j]} P[r_G = j] \quad (\text{A.3})$$

$$= \int z_j f_{z_j}(z_j, r_G = j) dz_j \quad (\text{A.4})$$

where $f_{z_j | r_G=j}$ is the conditional distribution of Z_j given $r_G = j$ and $f_{z_j}(z_j, r_G = j)$ describes the joint distribution of Z_j and the event $r_G = j$.

$$\text{Now } r_G = j \text{ if and only if } T_j > \max_{k \neq j} T_k \quad (\text{A.5})$$

$$\text{if and only if } Z_j > \max_{k \neq j} \left(Z_k + \sqrt{N} \left(\frac{\theta_k}{\widehat{sd}_k} - \frac{\theta_j}{\widehat{sd}_j} \right) \right) \quad (\text{A.6})$$

Here we have used the decomposition in (2.5) to show how the bias depends on the random fluctuation terms, Z_j and Z_k . To simplify we will approximate the $\widehat{sd}_j$ and $\widehat{sd}_k$ terms by σ_j and σ_k (the associated true population values). Then we obtain

$$E[Z_{r_G}] = \sum_{j=1}^G \int z_j f_{z_j}(z_j, r_G = j) dz_j \quad (\text{A.7})$$

$$\approx \sum_{j=1}^G E \left[Z_j \cdot I_{[Z_j > M_{-j}]} \right] \quad (\text{A.8})$$

where

$$M_{-j} = \max_{k \neq j} \left(Z_k + \sqrt{N} \left(\frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right) \quad (\text{A.9})$$

and

$$= I_{[A]} \text{ is a function } = 1 \text{ if event } A \text{ is true, } 0 \text{ otherwise.} \quad (\text{A.10})$$

To proceed further we condition the expectation on M_{-j} and recall that $\mathcal{Z}_j$ is marginally distributed as a standard Gaussian random variable and thus, for independent markers,

$$E[\mathcal{Z}_{r_G}] \approx \sum_{j=1}^G E\left[\mathcal{Z}_j \cdot I_{[\mathcal{Z}_j > M_{-j}]} \right] \quad (\text{A.11})$$

$$= \sum_{j=1}^G E\left[E\left[\mathcal{Z}_j \cdot I_{[\mathcal{Z}_j > M_{-j}]} \middle| M_{-j}\right]\right] \quad (\text{A.12})$$

$$= E\left[\sum_{j=1}^G \int_{M_{-j}}^{\infty} \mathcal{Z}_j f_{\mathcal{Z}}(z) dz\right] \quad (\text{A.13})$$

$f_{\mathcal{Z}}$ denotes a Gaussian distribution and the expectation in (A.13) is necessary because M_{-j} contains random elements $\mathcal{Z}_k$. The integral may be explicitly rewritten as

$$E[\mathcal{Z}_{r_G}] \approx \frac{1}{\sqrt{2\pi}} \sum_{j=1}^G E\left[e^{-\frac{M_{-j}^2}{2}}\right]. \quad (\text{A.14})$$

From (A.14) one sees that bias is inversely related to the absolute value of the M_{-j} terms. Some consequences of this derivation are as follows.

Consider the effect of increasing the sample size holding all else constant. We focus upon the case in which there exists a single true maximum effect size and designate that marker's index by r_G^0 , i.e.

$$\frac{\theta_{r_G^0}}{\sigma_{r_G^0}} > \frac{\theta_j}{\sigma_j} \text{ for all } j \neq r_G^0.$$

It is worthwhile to examine M_{-j} for the case when $j = r_G^0$ and $j \neq r_G^0$ separately where we assume only one variable (with index r_G^0) has the most positive effect size, i.e. there are no ties. Then

$$\lim_{N \rightarrow \infty} M_{-j} = \lim_N \max_{k \neq j} \left(\mathcal{Z}_k + \sqrt{N} \left(\frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right) \quad (\text{A.15})$$

$$= -\infty \text{ with probability 1 if } j = r_G^0 \quad (\text{A.16})$$

$$= \infty \text{ if } j \neq r_G^0. \quad (\text{A.17})$$

In either case we have that $M_{-j}^2 \rightarrow \infty$ so from (A.14) one sees $\lim_{N \rightarrow \infty} E[\mathcal{Z}_{r_G}] = 0$.

The case for expanding the differences among effect sizes is similar – at least for the simplified example below. For a given pattern of effect sizes among the G variables (again with no ties for the largest effect size), consider a new pattern of effect sizes given by multiplying each original effect by a constant $\gamma > 0$. Then if r_G^0 designates the marker with the most positive true effect size

$$\lim_{\gamma \rightarrow \infty} M_{-j} = \lim_{\gamma \rightarrow \infty} \max_{k \neq j} \left(\mathcal{Z}_k + \gamma \sqrt{N} \left(\frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right) \quad (\text{A.18})$$

$$= -\infty \text{ if } j = r_G^0 \quad (\text{A.19})$$

$$= \infty \text{ if } j \neq r_G^0. \quad (\text{A.20})$$

Consequently the same conclusion of no bias follows. If one reverses the limiting action of γ so that $\gamma \rightarrow 0$ from above then

$$\lim_{\gamma \downarrow 0} M_{-j} = \max_{k \neq j} \mathcal{Z}_k \quad (\text{A.21})$$

where the $\mathcal{Z}_k$ are standard normal statistics and the bias is then positive. This situation corresponds to the situation of no variables showing differential expression.

The case for increasing G , the number of variables is less clear cut as it depends upon the combination of effect sizes. Empirically it seems that adding variables with effect sizes at or near the size of the largest preexisting effect sizes exaggerates the bias effects for θ_{r_G} . In terms of figuring the change of M_{-j} terms as above there is more ambiguity as some terms M_{-j}^2 terms will likely increase, others decrease, and some new terms will be introduced.

References

- Benjamini Y, Hochberg Y: Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;57:289–300.
- Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003;100:9440–9445.
- Editorial: Freely associating. *Nat Genet* 1999; 22:1–2.
- Garner C: Upward bias in odds ratio estimates from genome-wide association studies. *Genet Epidemiol* 2007;31:288–295.
- Allison D, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, Amos CI: Bias in estimates of quantitative-trait locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am J Hum Genet* 2002;70:575–585.
- Göring H, Terwilliger JD, Blanger J: Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001;69:1357–1363.
- Zöllner S, Pritchard JK: Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 2007;80:605–615.
- Sun L, Bull SB: Reduction of selection bias in genomewide studies by resampling. *Genet Epidemiol* 2005;28:352–367.
- Jeffries N: Multiple comparisons distortions of parameter estimates. *Biostatistics* 2007;8: 500–504.
- Jeffries N: Performance of a genetic algorithm for mass spectrometry proteomics. *BMC Bioinformatics* 2004;5:180.
- Ghosh A, Zou F, Wright FA: Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am J Hum Genet* 2008;82:1064–1074.
- Zhong H, Prentice RL: Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 2008, in press.