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

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 \cite{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.

A common rationale given for this bias is that it results from conditioning on the initial estimate meeting \textit{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 \textit{p} value. Garner \cite{4} describes this in the context of genomewide association scans; Allison et al. \cite{5} describe it in the context of quantitative trait loci of paired siblings. This conditioning on a significant result has the

Key Words
Estimation bias • Multiple comparisons

Abstract

\textbf{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 \textit{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 the most extreme test statistics (i.e. highly ranked results) that leads to bias in the estimated effect sizes. \textbf{Conclusions:} This bias, termed ranking bias, is separate from that arising from conditioning on a significant \textit{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.
to generate a significant value of that corresponding to bias in model selection procedures. They make an analogy between the ranking bias discussed here with both significance bias and ranking bias. Göring et al. called ranking bias here. The 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 – the terms ‘significance bias’ and ‘ranking bias’ are not in common usage. Allison et al. [5] briefly describe ranking bias but focus on 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.

### 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 $2 \times 2$ table (Table 1).

<table>
<thead>
<tr>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
<tr>
<td>$N$</td>
<td>$N$</td>
<td>$2N$</td>
</tr>
</tbody>
</table>

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.

### Table 1. $2 \times 2$ table for each biallelic marker

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Control</td>
<td>Total</td>
</tr>
<tr>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
<tr>
<td>$N$</td>
<td>$N$</td>
<td>$2N$</td>
</tr>
</tbody>
</table>
where \(1 - \alpha\) is confidence associated with the interval and \(\Phi_1^{1-\alpha/2}\) 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

\[
\hat{s}_d = \sqrt{\frac{1}{\hat{p}_{2|\text{Case}}} + \frac{1}{1 - \hat{p}_{2|\text{Case}}} + \frac{1}{\hat{p}_{2|\text{Control}}} + \frac{1}{1 - \hat{p}_{2|\text{Control}}}}
\]

then the random components \(Z_j\) are likely to play a larger role in determining which marker has the maximal observed value, \(T_{c^*}\). Under these conditions of multiple markers with similar trend values the contribution of \(Z_j\) for that with largest \(T_j\) value is likely to be positive. From the definition of \(Z_{r_G}\) 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 \(\theta_{r_G}\) 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 0.224 and 0.5).

For each \(j\), \(Z_j\) has an approximately normal distribution centered about 0 while for \(T_j\) this is only true when \(\theta_j = 0\). The \(Z_j\) 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 \(j\), \(Z_j\) has approximately an unconditional \(N(0,1)\) distribution. A critical question concerns the distribution of \(Z_j\) 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 \(r_1, r_2, ..., r_G\) order the \(T\) test statistics where

\[
T_j = \frac{\sqrt{N} \hat{\theta}}{sd_j} \quad \text{and} \quad T_1 \leq T_2 \leq \cdots \leq T_G.
\]

Typically, the distribution of

\[
Z_G = \frac{\sqrt{N} (\hat{\theta}_{r_G} - \theta_{r_G})}{sd_{r_G}}
\]

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

\[
\frac{\sqrt{N} \theta_j}{sd_j}
\]

Downloaded by: 54.70.40.11 - 11/18/2017 2:44:38 PM
Because the true log odds ratios (θ) are known, one can calculate a realization of
\[ Z = \frac{\sqrt{N} (\hat{\theta}_r - \theta_r)}{\hat{sd}}. \]

After making many such samples (1000 simulations) the resulting distribution of \( Z_{rg} \) is shown along with the standard normal distribution in figure 1. Of note are the facts that the \( Z_{rg} \) distribution is not centered about 0 (indicating \( \hat{\theta}_r \) typically overestimates \( \theta_r \)) 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. \( \hat{\theta}_{rg} \) and \( \theta_{rg} \)) for each of the 1000 simulations are shown in figure 2. The discrete vertical bands correspond to the 12 distinct values of \( \theta_{rg} \) that were obtained: \{0, log(1.45), log(1.50), ..., 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 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.
\[ \frac{\theta_{r_G^0}}{\sigma_{r}^0} > \frac{\theta_{r_j}}{\sigma_j} \text{ for all } j \neq r_G^0, j \in \{1, ..., G\}. \]

Because \( r_G^0 \) is fixed (though unknown) it is the case that for large sample size \( N \)
\[ Z = \frac{\sqrt{N} (\hat{\theta}_r - \theta_r)}{\hat{sd}} \sim \text{Normal}(0,1) \quad (4.1) \]
such an assessment was performed by Garner.

For instance, if \( T_{rc} > 3.90 \), then the associated \( p \) value will be less than \( 0.025/500 \) and in our simulations the marker would be judged significant according to a Bonferroni criterion. However, using standard derivations concerning conditional expectations of normally distributed random variables one can show that

\[
E[Z_c | T_{rc} > 3.90] < E[Z_c | Z_c < 3.90 - \sqrt{N} \hat{\theta}_G / \sigma_c] \tag{4.2}
\]

\[
= \phi \left( 3.90 - \sqrt{N} \hat{\theta}_G / \sigma_c \right) \left( \frac{1 - \Phi(3.90 - \sqrt{N} \hat{\theta}_G / \sigma_c)}{\sqrt{N}} \right) \tag{4.3}
\]

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

\[
E[\hat{\theta}_G | T_{rc} > 3.90] = \hat{\theta}_G + \text{significance bias} \tag{4.4}
\]

where significance bias

\[
\approx \frac{\sigma_c}{\sqrt{N}} \left( \phi \left( 3.90 - \sqrt{N} \hat{\theta}_G / \sigma_c \right) - \phi \left( 3.90 - \sqrt{N} \hat{\theta}_G / \sigma_c \right) \right) \tag{4.5}
\]

The approximation arises because \( \hat{\theta}_G \) has been replaced by \( \sigma_c \) 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 \( \sigma \) term), and \( p \) value threshold change the significance bias. Such an assessment was performed by Garner.

However, in most circumstances \( r_c^2 \) 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 effects, only \( r_G \) is observed and the preceding development based upon \( r_c^2 \) is not really applicable in practice. What is of practical interest is \( E[\hat{\theta}_G | T_{rc} > 3.90] \). However analytical evaluation of this expression is more complicated than assessing bias for \( r_c^2 \) because \( r_G \) is random. Consequently, simulations may be used to explore the following types of overestimation bias

significance bias related to \( \hat{\theta}_G : E[\hat{\theta}_G | T_{rc} > 3.90] \),

ranking bias related to \( \hat{\theta}_G : E[\hat{\theta}_G - \hat{\theta}_G] \),

and the combination of ranking bias and significance bias related to \( \hat{\theta}_G : E[\hat{\theta}_G - \hat{\theta}_G | T_{rc} > 3.90] \).

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 (\( G_i \)) 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 \( \theta_G / \sigma_G = 0.69/3.465 \) where 3.465 is the associated \( \sigma \) 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 \( \sigma \) is given by \( s/\hat{d} \) 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 analytical result. The simulations’ estimate of ranking bias related to \( \hat{\theta}_G \) is about 0.191 in this case – approximately 2.5 times larger than the significance bias of \( \hat{\theta}_G \). As an aside it is worth noting that other markers with smaller

| Scenario | \( n = 500 \), \( G_i = 20 \) | \( \hat{\theta}_G \) significance bias: \( E[\hat{\theta}_G | T_{rc} > 3.90] - \hat{\theta}_G \) | \( \hat{\theta}_G \) ranking bias: \( E[\hat{\theta}_G - \hat{\theta}_G] \) | \( \hat{\theta}_G \) combined significance and ranking bias: \( E[\hat{\theta}_G - \hat{\theta}_G | T_{rc} > 3.90] \) |
|----------|-------------------|--------------------------|--------------------------|--------------------------|
| Scenario 1 | \( n = 500 \), \( G_i = 40 \) \{1.05, ... , 2.0\} | 0.072 | 0.191 | 0.192 |
| Scenario 2 | \( n = 500 \), \( G_i = 5 \) \{1.20, ... , 2.0\} | 0.072 | 0.230 | 0.230 |
| Scenario 3 | \( n = 500 \), \( G_i = 1 \) \{2.0\} | 0.072 | 0.113 | 0.129 |
| Scenario 4 | \( n = 500 \), \( G_i = 1 \) \{4.0\} | 0.000 | 0.000 | 0.000 |
| Scenario 5 | \( n = 500 \), \( G_i = 1 \) \{4.0\} | 0.002 | 0.112 | 0.112 |
| Scenario 6 | \( n = 1000 \), \( G_i = 20 \) \{1.05, ... , 2.0\} | 0.330 | 0.448 | 0.471 |
| Scenario 7 | \( n = 500 \), \( G_i = 10 \) \{1.04, ... , 1.40\} | 0.395 | 0.476 | 0.543 |
| Scenario 8 | \( n = 500 \), \( G_i = 10 \) \{1.30, ... , 1.30\} | 0.130 | 0.192 | 0.230 |
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 \( \hat{\theta}_{rG} \) by examining just those simulations in which \( T_{rG} > 3.90 \). In these 1000 simulations \( T_{rG} > 3.90 \) 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 \( \hat{\theta}_{rG} \) 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 \( \hat{\theta}_{rG} \). These figures have increased to 0.230 for both bias measures. Conversely, when the number 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 \( \hat{\theta}_{rG} \) are reduced to 0.113 and 0.129 though the significance bias associated with \( \hat{\theta}_{rG} \) remains unchanged. These three scenarios show how bias in \( \hat{\theta}_{rG} \) 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 \( \hat{\theta}_{rG} = 0.069 \) 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 \( \hat{\theta}_{rG} \) and \( \hat{\theta}_{rG} \) are quite comparable as shown in the first and second 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 \( \hat{\theta}_{rG} \) 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 rank test statistic exceeding a threshold. Indeed, as it is usually the case that \( \theta_{rG} \geq \theta_{rG} \) and \( \theta_{rG} \leq \theta_{rG} \), one would not expect significance bias, \( E[\theta_{rG} - \theta_{rG} | T_{rG} > 3.90] \), to exceed the conditional ranking bias, \( E[\theta_{rG} - \theta_{rG} | T_{rG} > 3.90] \).

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

\[
E \left[ \frac{\sqrt{N} (\hat{\theta}_{rG} - \theta_{rG})}{sd_{rG}} \right] = E[Z_{rG}] \approx \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{G} E \left[ e^{\frac{-M_{j}^{2}}{2}} \right]
\]  

(5.1)
where
\[ M_j = \max_{k \neq j} T_k - \sqrt{N} \frac{\theta_j}{\sigma_j} = \max_{k \neq j} \left( Z_k + \sqrt{N} \left( \frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right) \] (5.2)

for \( j \) in 1, ..., \( G \). 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, \( \theta_j/\sigma_j \), 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, \( \theta_j/\sigma_j \), are the same for all \( j = 1, \ldots, G \) as would be the case if there were no informative markers. If the markers are independent then in this case
\[ E[Z_j] \approx \frac{G}{\sqrt{2\pi}} e^{-\frac{1}{2} \left( \max_{i=1}^{G} \left( x_i - x_i \right) \right)^2} \]

where the \( X_i \) 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 \( \gamma \to \infty \). 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 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 \( \theta \) 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-

<table>
<thead>
<tr>
<th>Marker rank</th>
<th>( \text{Prob} \left( T_{G} &gt; 3.90 \right) )</th>
<th>( E[\hat{\theta}<em>{G} - \theta</em>{G} \mid T_{G} &gt; 3.90] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1998} )</td>
<td>0.990</td>
<td>0.192</td>
</tr>
<tr>
<td>2nd highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1999} )</td>
<td>0.925</td>
<td>0.143</td>
</tr>
<tr>
<td>3rd highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{498} )</td>
<td>0.741</td>
<td>0.124</td>
</tr>
<tr>
<td>4th highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1497} )</td>
<td>0.438</td>
<td>0.128</td>
</tr>
<tr>
<td>5th highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1496} )</td>
<td>0.193</td>
<td>0.132</td>
</tr>
<tr>
<td>6th highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1495} )</td>
<td>0.054</td>
<td>0.131</td>
</tr>
<tr>
<td>7th highest ranked: ( \hat{\theta}<em>{G} = \hat{\theta}</em>{1494} )</td>
<td>0.007</td>
<td>0.129</td>
</tr>
</tbody>
</table>
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

\[
T_j = Z_j + \sqrt{N} \frac{\theta_j}{sd_j}
\]

where \( Z_j = \sqrt{N} (\hat{\theta}_j - \theta) / sd_j \) and \( Z_j \) is approximated by a standard normal distribution. Then we may calculate the expected value of \( Z_{r_0} \) (and hence obtain a measure of the bias of \( \hat{\theta}_r - \theta_{r_0} \)) as

\[
E[Z_{r_0}] = \sum_{j=1}^{6} E[Z_{r_0} | r_0 = j] P[r_0 = j]
\]

and

\[
E[Z_{r_0} | r_0 = j] P[r_0 = j] = \left( \int z_j f_{z_j | r_0 = j} (z_j) dz_j \right) P[r_0 = j] \]

\[
= \int z_j f_{z_j | r_0 = j} (z_j) dz_j P[r_0 = j] \]

\[
= \int z_j f_{z_j} (z_j, r_0 = j) dz_j
\]

where \( f_{z_j | r_0 = j} \) is the conditional distribution of \( Z_j \) given \( r_0 = j \) and \( f_{z_j} (z_j, r_0 = j) \) describes the joint distribution of \( Z_j \) and the event \( r_0 = j \).

Now \( r_0 = j \) if and only if \( T_j > \max_{k \neq j} T_k \)

\[
\text{if and only if } Z_j > \max_{k \neq j} \left( Z_{r_0} + \sqrt{N} \left( \frac{\theta_k}{sd_k} - \frac{\theta_j}{sd_j} \right) \right)
\]

Here we have used the decomposition in (2.5) to show how the bias depends on the random fluctuation terms, \( Z_j \) and \( Z_{r_0} \). To simplify we will approximate the \( sd_j \) and \( sd_{r_0} \) terms by \( \sigma_j \) and \( \sigma_{r_0} \) (the associated true population values). Then we obtain

\[
E[Z_{r_0}] = \sum_{j=1}^{6} \int z_j f_{z_j} (z_j, r_0 = j) dz_j
\]

\[
\approx \sum_{j=1}^{6} E[Z_j \cdot I_{Z_j > M_j}]
\]

where

\[
M_j = \max_{k \neq j} \left( Z_k + \sqrt{N} \left( \frac{\theta_k}{\sigma_k} - \frac{\theta_j}{\sigma_j} \right) \right)
\]

and

\[
I_{\{A\}} \text{ is a function } = 1 \text{ if event } A \text{ is true, } 0 \text{ otherwise.}
\]
To proceed further we condition the expectation on $M_{j}$ and recall that $Z_{k}$ is marginally distributed as a standard Gaussian random variable and thus, for independent markers,
\begin{align}
E[Z_{k}] & \approx \sum_{j=1}^{G} E[Z_{j} \cdot I[Z_{j} > M_{j}]] \\
& = \sum_{j=1}^{G} E[E[Z_{j} \cdot I[Z_{j} > M_{j}]] | M_{j}]
\end{align}
(A.11)
\begin{align}
& = \sum_{j=1}^{G} \int_{M_{j}}^{\infty} z f_{Z}(z) d z \\
& = E[\sum_{j=1}^{G} \int_{M_{j}}^{\infty} z f_{Z}(z) d z]
\end{align}
(A.12)
\begin{align}
& = E[\sum_{j=1}^{G} \int_{M_{j}}^{\infty} e^{-\frac{z^{2}}{2}} d z]
\end{align}
(A.13)
where $f_{Z}$ denotes a Gaussian distribution and the expectation in (A.13) is necessary because $M_{j}$ contains random elements $Z_{k}$. The integral may be explicitly rewritten as
\begin{align}
E[Z_{k}] & \approx \frac{1}{2\pi} \sum_{j=1}^{G} e^{-\frac{M_{j}^{2}}{2}}.
\end{align}
(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 $r_{0}$, i.e.
\begin{align}
\frac{\theta_{r_{0}}}{\sigma_{}\theta_{}} > \frac{\theta_{j}}{\sigma_{}\theta_{}} \quad \text{for all } j \neq r_{0}.
\end{align}

It is worthwhile to examine $M_{j}$ for the case when $j = r_{0}$ and $j \neq r_{0}$ separately where we assume only one variable (with index $r_{0}$) has the most positive effect size, i.e. there are no ties. Then
\begin{align}
\lim_{N \to \infty} M_{j} &= \lim_{N \to \infty} \max_{k \neq j} \left[ Z_{k} + \sqrt{N} \left( \frac{\theta_{r_{0}}}{\sigma_{r_{0}}} - \frac{\theta_{j}}{\sigma_{j}} \right) \right] \\
& = -\infty \text{ with probability 1 if } j = r_{0}
\end{align}
(A.15)
\begin{align}
& = -\infty \text{ if } j \neq r_{0},
\end{align}
(A.16)
In either case we have that $M_{j}^{2} \to \infty$ so from (A.14) one sees
\begin{align}
\lim_{N \to \infty} E[Z_{r_{0}}] = 0.
\end{align}
(A.17)

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 $r_{0}$ designates the marker with the most positive true effect size
\begin{align}
\lim_{\gamma \to \infty} M_{j} &= \lim_{\gamma \to \infty} \max_{k \neq j} \left[ Z_{k} + \gamma \sqrt{N} \left( \frac{\theta_{r_{0}}}{\gamma \sigma_{r_{0}}} - \frac{\theta_{j}}{\gamma \sigma_{j}} \right) \right] \\
& = -\infty \text{ if } j = r_{0}
\end{align}
(A.18)
\begin{align}
& = -\infty \text{ if } j \neq r_{0}.
\end{align}
(A.19)
Consequently the same conclusion of no bias follows. If one reverses the limiting action of $\gamma$ so that $\gamma \to 0$ from above then
\begin{align}
\lim_{\gamma \to 0} M_{j} &= \max_{k \neq j} Z_{k}
\end{align}
(A.20)
where the $Z_{k}$’s 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 $\theta_{r_{0}}$. 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


Ranking Bias in Association Studies

Hum Hered 2009;67:267–275

Downloaded by: 54.70.40.11 - 11/18/2017 2:44:38 PM