Testing multiple biological mediators simultaneously

Simina M. Boca, Rashmi Sinha, Amanda J. Cross, Steven C. Moore, Joshua N. Sampson*
Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA

Associate Editor: Dr. Igor Jurisica

1 INTRODUCTION

Mediation analysis or causal inference offers numerous methods for testing if a single variable mediates the relationship between a known exposure and an outcome (Baron and Kenny, 1986; MacKinnon et al., 2002; MacKinnon, 2008; Biesanz et al., 2010; Taylor and MacKinnon, 2012). These methods assume that a known exposure, E, affects an outcome, Y, and aim to test whether this effect is at least partially transmitted through a mediator, M. For example, in biology, these methods have suggested that the negative impact of lead exposure on cognition is mediated by a decrease in the volumes of specific brain regions (Caffo et al., 2008), and that the association between certain variants in the FTO gene and increased body-weight is mediated by a lowered response to satiety cues (Wardle et al., 2008).

With new technologies, such as microarrays (Brown, 1995), next-generation sequencing (Shendure and Ji, 2008), and high-throughput metabolomics (Dettmer et al., 2006), it is possible to simultaneously test whether 100s or 1000s of biomarkers mediate a known relationship. In our motivating study, investigators aim to identify metabolites that mediate the association between increased fish consumption and a reduced risk of colorectal adenoma (Sinha et al., 1999). Our current objective is to define a testing procedure that accounts for “multiple comparisons” and maintains a desired Family Wise Error Rate (FWER). Guided by the methods developed for testing direct associations, we develop a permutation approach for testing multiple mediators. Specifically, we design a permutation method that tests whether any biomarker meets a common definition of a mediator (MacKinnon et al., 2002). We say that M is a mediator if it is both associated with the exposure, E and, conditional on E, associated with the outcome, Y.

Our first step is to define a permutation method for testing a single mediator. Our defined method uses the Freedman and Lane (1983) approach for testing the conditional association between M and Y, in contrast to previous methods (Taylor and MacKinnon, 2012) that use the Manly (1997) approach, and will therefore be more robust to outliers (Anderson and Robinson, 2001). Unfortunately, as has already been noted (Anderson and Robinson, 2001), there can be no exact permutation method for testing a conditional association. In addition to having its own value, our presentation of the single mediator test aims to illustrate the inherent difficulty in designing a permutation test for a composite null hypothesis. When testing mediation, the null hypothesis allows for either an association between E and M or a conditional association between M and Y. Approaches designed for testing direct associations, which permute only the exposure or outcome, cannot simulate such a composite null hypothesis.

This paper is structured so that methods for testing mediators can be described by comparing and contrasting them with methods for testing associations. We begin by describing permutation tests for a single association and the extensions needed for testing multiple associations. Then, we introduce permutation methods for testing a single mediator and the extensions for testing multiple mediators. The following simulations demonstrate that the potential increase in power from replacing the Bonferroni correction with a joint correction is greater when testing for mediation than when testing for direct association. Finally, we apply our method to our motivating example and offer a brief discussion.

*To whom correspondence should be addressed
Table 1. Tests discussed in the paper. The rows indicate whether single or multiple testing is performed and the type of adjustment for multiple testing, Bonferroni (Bonf) or joint. The columns indicate whether the test evaluates associations or mediations, and among mediation methods, whether \( \alpha \) and \( \beta \) are each evaluated separately (\( \alpha, \beta \)) or as a product (\( \alpha \times \beta \)).

<table>
<thead>
<tr>
<th>Association</th>
<th>Mediation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>Test 0</td>
</tr>
<tr>
<td>Multiple</td>
<td>Test 0B</td>
</tr>
<tr>
<td>joint</td>
<td>Test 0J</td>
</tr>
</tbody>
</table>

2 METHODS

2.1 Testing a Single Association

For testing a single association, we assume that there are two normally distributed variables, \( X \) and \( Y \), related by equation (1)

\[
Y = \kappa_0 + bX + \epsilon,
\]

where \( \epsilon \sim N(0, \sigma^2) \). The null hypothesis can be stated as either \( H_0 : b = 0 \) or \( H_0 : \rho_{X,Y} = 0 \), where \( \rho_{X,Y} \) is the correlation between \( X \) and \( Y \). Although less common, the latter framework facilitates comparisons with testing mediators. The null hypothesis can be tested by:

Test 0: We declare \( X \) to be significantly associated with \( Y \) if \( p_b \leq 0.05 \) (or |\( \hat{\rho}(X, Y) | \geq t_1 \)),

where \( \hat{\rho}(X, Y) \) is the sample correlation coefficient, \( p_b \) is the p-value, and \( t_1 \) is the 95th percentile for the null distribution of |\( \hat{\rho}(X, Y) | \). Both \( p_b \) and \( t_1 \) can be calculated by asymptotic theory or permutation. For the latter, we can permute values of \( Y \) to obtain \( \hat{\rho}(X, \pi(Y)) \), where \( \pi(\cdot) \) indicates a permutation of an original variable, and calculate |\( \hat{\rho}(X, \pi(Y)) | \) for each permuted dataset. Then, \( p_b \) is the proportion of permuted values greater than or equal to the observed value and \( t_1 \) is the 95th percentile of the permuted values. The tests discussed in this paper are summarized in Table 1.

2.2 Testing Multiple Associations

When we test the association between \( X \) and \( Y \) multiple variables, \( X_1, \ldots, X_K \), we calculate |\( \hat{\rho}(X_i, Y) | \) and its corresponding p-values, \( p_{bi} \), for \( i \in \{1, \ldots, K\} \). To maintain a FWER of 0.05, we have two options. First, we can estimate the threshold by Bonferroni correction, where \( t^B \) is the \( 100 \times (1 - 0.05/K)^{1/K} \) percentile of the null distribution estimated for |\( \hat{\rho}(X_i, Y) | \) and consider the test:

Test 0B: We declare \( X_i \) to be significantly associated with \( Y \) if \( p_{bi} \leq 0.05/K \) (or |\( \hat{\rho}(X_i, Y) | \geq t^B_i \)).

For the permutation estimates of \( t^B_i \) to converge to their true quantiles, we depend on the inherent assumption in equation 1, that the full joint distribution of \( X \) is determined by the marginal distributions, or that the joint distribution of \( X \) is constant for all \( Y \) under the null hypothesis (Huang et al., 2006). Similar assumptions are required for tests described here and elsewhere.

The second option, which uses the max correction (Westfall and Young, 1993), is to estimate the 95th percentile of the distribution of \( \hat{\rho}_{\text{max}} \equiv \max_i(\{\hat{\rho}(X_i, \pi(Y))\}) \) by permuting \( Y \). Here, we use the term “joint correction” because the distribution of the maximal test statistic depends on the joint distribution of the multiple variables. We let \( t_j \) be the 95th percentile of this permutation distribution and consider the test:

Test 0J: We declare \( X_i \) to be significantly associated with \( Y \) if |\( \hat{\rho}(X_i, Y) | \geq t_j \).

Note that we only permute \( Y \) once for each dataset, as opposed to once for each \( X_i \) for Test 0J.

2.3 Testing a Single Mediator

For testing a single mediator, we start by assuming that the exposure, putative mediator, and outcome are normally distributed and related by equations 2 and 3. The directed acyclic graph (DAG) in Figure 1 illustrates this relationship.

\[
M = \kappa_M + \alpha E + \epsilon_M, \quad Y = \kappa_Y + \gamma E + \beta M + \epsilon_Y, \tag{2}
\]

where \( E \sim N(0, \sigma^2_E) \), \( \epsilon_M \sim N(0, \sigma^2_M) \), and \( \epsilon_Y \sim N(0, \sigma^2_Y) \). Therefore, \( Y \) can also be described by

\[
Y = \kappa_Y + \gamma^* E + \epsilon_Y^*, \tag{4}
\]

where \( \epsilon_Y^* \sim N(0, \sigma^2_Y + \beta^2 \sigma^2_M + (\gamma + \alpha \beta)^2 \sigma^2_E) \). We can describe two general approaches for testing if \( M \) is a mediator (Sobel, 1982; MacKinnon et al., 2002). We can either evaluate whether both \( \alpha \neq 0 \) and \( \beta \neq 0 \) or whether their product \( \alpha \beta \neq 0 \) for interpretation, we note that \( \alpha \beta \equiv \gamma^* - \gamma \), which is the difference between the total and direct effect of \( E \). To facilitate comparison across metabolites, with concentrations that vary by orders by magnitude, we prefer normalized versions of \( \alpha \) and \( \beta \). Therefore, we replace \( \alpha \) and \( \beta \) by \( \rho_{EM} \) and \( \rho_{MY|E} \), respectively, where \( \rho_{EM} \) is the correlation between \( E \) and \( M \) and \( \rho_{MY|E} \) is the conditional correlation between \( M \) and \( Y \). The sample correlation coefficients, \( \hat{\rho}(E, M) \) and \( \hat{\rho}(r_{MY|E}, r_{Y|E|E}) \), offer estimates of these parameters. Here, \( r_{Y|E} \) and \( r_{MY|E} \) are the residuals from regressing \( Y \) on \( E \) and \( M \) on \( E \) respectively.

The first test we consider is:

Test 1: We declare \( M \) to be significant if \( p_{\rho} \leq 0.05 \) and \( p_{\rho} \leq 0.05 \) (or |\( \hat{\rho}(E, M) | \geq t_1(\alpha) \) and |\( \hat{\rho}(r_{MY|E}, r_{Y|E|E}) | \geq t_1(\beta) \)).

where \( p_{\rho} \) is the corresponding p-value and \( t_1(\alpha) \) is the 95th percentile for the null distribution of |\( \hat{\rho}(E, M) | \). \( p_{\rho} \) and \( t_1(\beta) \) are defined similarly for |\( \hat{\rho}(r_{MY|E}, r_{Y|E|E}) | \). All values can be calculated
by asymptotic theory or permutation. An overall p-value for Test 1 can be defined as \( \max(p_{B1}, p_{B2}) \). The exact steps are presented in Algorithms 1 (for \( \alpha \)) and 2 (for \( \beta \)). Freedman and Lane (Freedman and Lane, 1983) offer the rationale for Algorithm 2, even though, like all permutation tests for a conditional association, it is not exact (Anderson and Robinson, 2001).

Algorithm 1: Permutation Algorithm for \( p_B \) and \( t_1(\alpha) \).

1. Permute \( E \) to obtain \( \pi(E) \).
2. Calculate \( \hat{\rho}(\pi(E), M) \).
3. Repeat steps 1 and 2 to obtain a distribution of \( |\hat{\rho}(\pi(E), M)| \). \( p_B \) is the proportion of \( |\hat{\rho}(\pi(E), M)| \) exceeding \( |\hat{\rho}(E, M)| \) and \( t_1(\alpha) \) is 95\(^{th}\) percentile of the distribution.

Algorithm 2: Permutation Algorithm for \( p_B \) and \( t_1(\beta) \).

1. Permute \( r_{Y|E} \) to obtain \( \pi(r_{Y|E}) \).
2. Regress \( \pi(r_{Y|E}) \) on \( E \) to obtain a new set of residuals, \( r_{Y|E}^* \).
3. Calculate \( \hat{\rho}(r_{M|E}, r_{Y|E}^*) \).
4. Repeat steps 1, 2, and 3 to obtain a distribution of \( \hat{\rho}(r_{M|E}, r_{Y|E}^*) \). Calculate p-value and threshold as usual.

The second test uses the statistic \( S = |\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r_{Y|E})| \):

**Test 2:** We declare \( M \) to be significant if \( p_S \leq 0.05 \) (or \( S \geq t_1(S) \)), where \( p_S \) and \( t_1(S) \) have their usual meaning. The exact steps are presented in Algorithm 3, which is described so that it can be adapted for testing multiple mediators. Otherwise, we would have performed two sets of permutations and let \( t_1(S) \) be the maximum of two thresholds, the 95\(^{th}\) percentile for the distribution of \( |\hat{\rho}(\pi(E), M)\hat{\rho}(r_{Y|E}, r_{M|E})| \) and the 95\(^{th}\) percentile for \( |\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r_{Y|E})| \). In this format, tests 1 and 2 would be equivalent. To obtain \( t_1(S) \), we could perform Algorithms 1 and 2 and then let \( t_1(S) = t_1(\beta)\hat{\rho}(E, M) \) if \( |\hat{\rho}(E, M)| \geq |\hat{\rho}(r_{M|E}, r_{Y|E})| \) and \( t_1(S) = t_1(\alpha)\hat{\rho}(r_{M|E}, r_{Y|E}) \) if \( |\hat{\rho}(E, M)| < |\hat{\rho}(r_{M|E}, r_{Y|E})| \). However, we choose to formally describe the steps for test 2 by Algorithm 3. In this form, the algorithm can be easily adapted for multiple mediators. When considering non-normally distributed variables, test 2 could, at least in theory, produce inflated type I errors. The test and algorithm described in supplementary section S5 offers a more conservative alternative.

### 2.4 Testing Multiple Mediators

For testing multiple mediators, we start by assuming that the exposure, putative mediators, and outcome are normally distributed and related by equations 5 and 6. The DAG in Figure 2 illustrates this relationship:

\[
M_i = \kappa_{M_i} + \alpha_i E + \epsilon_{M_i}, \quad Y = \kappa_Y + \gamma E + \sum_{i=1}^{K} \beta_i M_i + \epsilon_Y, \quad \gamma > 0, \quad \kappa_Y > 0, \quad \epsilon_{M_i} \sim N(0, \sigma_{M_i}^2), \quad \epsilon_Y \sim N(0, \sigma_Y^2).
\]

When we test multiple mediators, \( M_1, \ldots, M_K \), we calculate \( \hat{\rho}(E, M_i), \hat{\rho}(r_{M|E}, r_{Y|E}), \) and \( S_i \) for each metabolite individually. We can maintain a FWER across all tests by using Bonferroni corrected thresholds:

**Test 1B:** We declare \( M_i \) to be significant if \( p_{B_i} \leq 0.05/K \) and \( p_{B_i} \leq 0.05/K \) (or \( |\hat{\rho}(E, M_i)| \geq t_{B_i}(\alpha) \) and \( |\hat{\rho}(r_{M|E}, r_{Y|E})| \geq t_{B_i}(\beta) \)).

**Test 2B:** We declare \( M_i \) to be significant if \( p_{S_i} \leq 0.05/K \) (or \( S_i \geq t_{S_i}(S) \)).

Here, \( t_{B_i}(\alpha) \), \( t_{B_i}(\beta) \), and \( t_{S_i}(S) \) are the 100 \times (1 - 0.05/K)\(^{th}\) percentiles of the appropriate null distribution. Again, the quality

---

**Algorithm 3:** Permutation Algorithm for \( p_S \) and \( t_1(S) \).

1. If \( \hat{\rho}(E, M) < |\hat{\rho}(r_{M|E}, r_{Y|E})| \), then calculate \( \hat{\rho}(E, M) \) as in Algorithm 1 and \( S^* = |\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r_{Y|E})| \).
2. If \( \hat{\rho}(E, M) \geq |\hat{\rho}(r_{M|E}, r_{Y|E})| \), calculate \( \hat{\rho}(r_{M|E}, r_{Y|E}) \) as in Algorithm 2 and \( S^* = |\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r_{Y|E})| \).
3. Repeat steps 1 and 2 to obtain a distribution of \( S^* \).

Calculate p-value and threshold as usual.
of the permuted estimates of $t_J^p(\alpha)$, $t_J^p(\beta)$, and $t_J^p(S)$, depend strongly on our assumption that the joint distribution of $Y$ and $M$, conditioned on $E$, is uniquely determined by the marginal distributions defined in equations 5 and 6 (Xu and Hsu, 2007).

We can also maintain a FWER by using a joint correction through taking the maximal test statistics. First, letting $\hat{\rho}_{M|E}^{\text{max}} = \max(|\hat{\rho}(E, M_i)|)$ and $\hat{\rho}_{M|E}^{\text{max}} = \max(|\hat{\rho}(r_{M_i|E}, r_{Y|E})|)$, we define test 1J:

Test 1J: We declare $M_i$ to be significant if $p_{J}^{\text{max}} \leq 0.05$ and $p_{J}^{\text{max}} \leq 0.05$ (or $|\hat{\rho}(E, M_i)| \geq t_J(\alpha)$ and $|\hat{\rho}(r_{M_i|E}, r_{Y|E})| \geq t_J(\beta)$),

where $t_J(\alpha)$ and $t_J(\beta)$ are the 95th percentiles of the appropriate null distributions. Tests 1J and 1B will be nearly identical for independent mediators. The more interesting option is the extension of Test 2, where we define a single statistic $S_{\text{joint}} = \max(S_i)$:

Test 2J: We declare $M_i$ to be significant if $p_J \leq 0.05$ (or $\hat{S}_i \geq t_J(S)$).

To define the permutation algorithm for $p_J$ and $t_J(S)$, we consider the following partition of the possible mediators:

$$A = \{i : |\hat{\rho}(E, M_i)| < |\hat{\rho}(r_{M_i|E}, r_{Y|E})|\},$$

$$B = \{i : |\hat{\rho}(E, M_i)| \geq |\hat{\rho}(r_{M_i|E}, r_{Y|E})|\}.$$

We then go through the steps in Algorithm 4.

Algorithm 4: Permutation Algorithm for $p_J$ and $t_J(S)$.

1. For $i \in A$, calculate $\hat{\rho}(\pi(E), M_i)$ and $S_i^\pi = |\hat{\rho}(\pi(E), M_i)||\hat{\rho}(r_{M_i|E}, r_{Y|E})|.$
2. For $i \in B$, calculate $\hat{\rho}(r_{M_i|E}, r_{Y|E})$ and $S_i^\pi = |\hat{\rho}(E, M_i)||\hat{\rho}(r_{M_i|E}, r_{Y|E})|.$
3. Calculate $S_{\text{joint}}^\pi = \max(S_i^\pi).$
4. Repeat steps 1, 2, and 3 to obtain a distribution of $S_{\text{joint}}^\pi.$

Calculate p-value and threshold as usual.

2.5 Beyond Normality

Our framework only needs to be slightly modified for the scenario where the outcome is not normally distributed. For example, mediation for binary outcomes can be tested using the same algorithms. We continue to let $r_{Y|E}$ be the residuals after fitting a linear regression with $Y$ as the dependent variable and let $\hat{\rho}(r_{M_i|E}, r_{Y|E})$ be the sample correlation coefficients. Then, we will still assume that the DAG in Figure 2 is true, but we will replace equation 6 with:

$$Pr(Y = 1|E, M_1, \ldots, M_K) = e^{K_H\gamma + \sum_{i=1}^K \beta_i M_i} + e^{K_H\gamma + \sum_{i=1}^K \beta_i M_i}.$$

However, we will need to slightly modify the algorithms in case/control studies, where the data is collected retrospectively. Thus, in order for $\hat{\rho}(E, M)$ to consistently estimate the correlation between $E$ and $M$ in the overall population, it is necessary to weight each case by $\upsilon$ and each control by $(1 - \upsilon)(1 - \rho)$, where $\upsilon$ represents the prevalence of the outcome in the overall population and $\rho$ represents the proportion of cases in the sample, as in VanderWeele and Vansteelandt (2010). We note that in order for this analysis to be performed, one must have prior knowledge of $\upsilon$, as this cannot be estimated from the data.

The distribution of $E$ is taken to be normal only out of convenience, and the same analysis can be performed for other distributions. In the case where $E$ is discrete and each possible value is represented multiple times in the dataset, an exact permutation test exists for testing the conditional associations between the mediators and the outcome, instead of the Freedman-Lane approximation. (Brown and Maritz, 1982).

2.6 Simulations

We simulate studies where all the variables are normally distributed and follow equations 5 and 6 (Figure 2) for the single mediator case and equations 5 and 6 (Figure 2) for the multiple mediator case. The marginal variance of all variables was fixed to be 1. We also simulate case-control studies with a binary outcome $Y$ following equation 8 and a population prevalence $\upsilon = 0.2$. The sample size, $n$, for each study was either 100 or 1000 individuals. For studies with a binary outcome, $n$ was equally divided among cases and controls. For each simulated study, 20,000 permutations were performed to obtain the permutation distributions. All methods were implemented in the R programming language (R Core Team, 2012).

2.6.1 Single Mediator

We first consider the scenario where $E$, $M$, and $Y$ are normally distributed. To estimate the type I error rates, we set $(\alpha, \beta)$ to be either $(0, 0)$, $(0, \upsilon)$, or $(\upsilon, 0)$, where $\upsilon$ is the sample size dependent effect size. As with all simulated studies, we take $\gamma = 0$. Supplementary material shows that letting $\gamma \neq 0$ does not qualitatively affect the results. We then simulate 10,000 studies, and define the observed type I error rate to be the fraction of studies where $M$ is declared to be statistically significant. To estimate power, we set $(\alpha, \beta)$ to $(0, \upsilon)$, $(0.5\upsilon, 1.5\upsilon)$, or $(1.5\upsilon, 0.5\upsilon)$. We then simulate 1000 studies, and define power to be the fraction of studies where $M$ is significant. So that all marginal variances equal 1, we let $\sigma^2 = 1$, $\sigma^2_M = 1 - \alpha^2$, and $\sigma^2_Y = 1 - \beta^2\sigma^2_{M} - (\alpha\beta + \gamma)^2$. We consider $\upsilon = 0.2$ for $n = 100$, and $\upsilon = 0.08$ for $n = 1000$.

Next, we consider the scenario where $Y$ is binary. The general design is similar, but for type I error, we set $(\alpha, \beta)$ to either $(0, 0)$, $(0, \upsilon)$, or $(\upsilon, 0)$, where $\beta$ now refers to $\beta_1$ in equation 8, and for power we set $(\alpha, \beta)$ to $(0.5\upsilon, 1.5\upsilon)$, $(0.5\upsilon, 1.5\upsilon)$, and $(1.5\upsilon, 0.5\upsilon)$. We consider $\upsilon = 0.2$ for $\upsilon = 0.5\upsilon$, $\upsilon = 0.08$, $\upsilon = 0.2$ when $n = 1000$. The combinations of $\upsilon$, $\alpha$, and $\beta$ we consider are listed in Tables 2 (for normally-distributed outcome) and S1 (for case-control study, in Supplementary materials).

2.6.2 Multiple Mediators

The simulations are similar to the single mediator scenario. However, here we must not only define effects sizes, but also the proportion of mediators with each effect size. Again, we first consider the scenario with normally distributed variables. To obtain the FWER, we consider either $K = 10$ or $K = 100$ null mediators. We set $K_1$, $K_2$, and $K_3$ mediators to have $(\alpha, \beta) = (0, 0)$, $(0, \upsilon)$, and $(\upsilon, 0)$, respectively. Here, we consider
The power, we also add a single mediator with observed FWER to be the proportion of the studies where at least one mediator is declared to be statistically significant. To obtain the power, we also add a single mediator with $(\alpha, \beta) = (es, es)$. We discuss the scenario of multiple mediators in section S3 of the supplementary material. We then simulate 10,000 studies, and define power to be the proportion of studies where this true mediator is declared significant. When $K = 10$ or 11 mediators, we consider 100 subjects and $es = 0.3$. When $K = 100$ or 101, we consider 1000 subjects and $es = 0.1$. So that all marginal variances equal 1, we let $\sigma_0^2 = 1$, $\sigma_M^2 = 1 - \alpha_i^2$, and $\sigma_Y^2 = 1 - \sum_{i=1}^K \beta_i^2 \sigma_M^2 - (\sum_{i=1}^K \alpha_i \beta_i + \gamma)^2$. In the Supplementary material, we also consider simulations with more than one true mediator, including scenarios where one is “screening” mediators.

The DAG in Figure 2 implies that the putative mediators are independent conditional on exposure. We also want to explore the power to be the proportion of studies where this true mediator is declared significant. When $K = 10$ or 11, we consider 100 subjects and $es = 0.3$. When $K = 100$ or 101, we consider 1000 subjects and $es = 0.1$. So that all marginal variances equal 1, we let $\sigma_0^2 = 1$, $\sigma_M^2 = 1 - \alpha_i^2$, and $\sigma_Y^2 = 1 - \sum_{i=1}^K \beta_i^2 \sigma_M^2 - (\sum_{i=1}^K \alpha_i \beta_i + \gamma)^2$. In the Supplementary material, we also consider simulations with more than one true mediator, including scenarios where one is “screening” mediators.

The prevalence of colorectal adenoma within this age group at the time of the original study was assumed to be 0.228 (Dr. Brooks Cash, personal communication). Of the 416 known metabolites measured, we considered only the 149 metabolites present in all study participants. These metabolite values were batch-normalized and log-transformed. We normalized the exposure, mediators, and outcome by using the residuals from regression analyses that included gender, age, current smoking status, and body mass index.

<table>
<thead>
<tr>
<th>Table 2. Single mediator, normally-distributed outcome: Type I error and power are estimated by simulation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>1000</td>
</tr>
<tr>
<td>1000</td>
</tr>
<tr>
<td>1000</td>
</tr>
</tbody>
</table>

3 RESULTS

3.1 Simulations: Single Mediator

The type I error rates for Test 1 and Test 2 depend on the values of $\alpha$ and $\beta$. When both $\alpha = 0$ and $\beta = 0$, both tests are extremely conservative, with type I error rates close to 0.0025 = 0.05. When either $(\alpha > 0$ and $\beta = 0$) or $(\alpha = 0$ and $\beta > 0$), type I error rates are closer to 0.05 (Tables 2 and S1). Note three features. First, similar trends are observed for normal and binomial outcomes. Second, although permutation methods for testing conditional associations are not exact, the type I error rates are below or close to 0.05 (data not shown). Third, Test 1 and Test 2 produce essentially identical results for both type I error and power.

When running simulations under the alternative hypothesis, we found that the threshold for significance for Test 2, $t_1(S)$, increased with effect size. Therefore, when $\alpha = 0$ and $\beta = 0$, we considered the permutation statistic $\hat{\alpha}$, which we used to compute the observed $\hat{\alpha}$ or $\hat{\beta}$ and therefore increase with the observed values.

3.2 Simulations: Multiple Mediators

The FWER depend on the values of $\alpha_i$ and $\beta_i$. When all the mediators have $\alpha_i = \beta_i = 0$, all four tests are extremely conservative (Tables 3 and 4). In these simulations, where all mediators were conditionally uncorrelated, we found that power for Test 2 was larger than the power for Tests 1B, 1J or 2B. In general, all three tests (1B, 1J, 2B) performed nearly identically, so we only report Test 2B here and in tables 3 and 4. The results from tests 1B and 1J are provided as part of supplementary tables S2 and S3. For example, with 1000 subjects, 100 null mediators, and $\alpha_i = \beta_i = 0$ for the true mediator, we found the power for detecting an association by Test 2 to be 0.657, 0.626, 0.277, and 0.243 when $(K_1, K_2, K_3) = (100, 0, 0), (70, 30, 0), (70, 0, 30)$, and $(60, 20, 20)$. In contrast, we found the power from Tests 1B, 1J, or 2B, which again are essentially equivalent, to be between 0.13 and

2.7 Navy Colorectal Adenoma Study

The original Navy Colorectal Adenoma case-control study (Sinha et al., 1999) was a study of colorectal adenoma risk factors. A follow-up study was conducted to investigate circulating metabolites in relation to self-reported diet and colorectal adenoma, in 129 cases and 129 controls. Serum metabolites were measured by Metabolon Inc., whose methods have been previously described. (Sreekumar et al., 2009; Suhre et al., 2011). For this analysis, the exposures of interest were the daily intake of red meat and the daily intake of fish (g/day), inferred from dietary questionnaires and the outcome of interest was the presence of colorectal adenoma. The prevalence of colorectal adenoma within this age group at the time of the original study was assumed to be 0.228 (Dr. Brooks Cash, personal communication). Of the 416 known metabolites measured, we considered only the 149 metabolites present in all study participants. These metabolite values were batch-normalized and log-transformed. We normalized the exposure, mediators, and outcome by using the residuals from regression analyses that included gender, age, current smoking status, and body mass index.
0.14. Importantly, note that the power is greatly improved by the joint test here, while if we generated independent variables when testing associations, the joint test could offer no improvement (data not shown). Similar improvements were observed when considering the scenario with multiple true mediators (supplementary table S4).

Tables 3 and 4 show that as \( K \) increases, the power for test 2J decreases, whereas its FWER increases. Both trends can be explained by the fact that the expected value of \( S^2 \) is higher when the null variable has one association, with either \( E \) or \( Y \). For example, assume that \( M \) is associated with \( E \). The result is that variable \( i \) will likely be included in group \( B \) and the expected value of \( (S^2)^i \) will be approximately \[ (\rho(E, M_i))^2 \times \text{var}(\hat{\rho}(r_{Y|E}, r_{M_i|E})). \]

Now, consider \( M^*_i \), associated with neither \( E \) nor \( Y \). \( (S^2)^i \) should only be slightly larger than \( \text{var}(\hat{\rho}(r_{Y|E}, r_{M^*_i|E})) \times \text{var}(\hat{E}, M_i) \), a comparatively small number. Because \( S^2 \) increases with \( K \), the threshold for significance increases which, in turn, leads to a decrease in statistical power. Furthermore, note that when \( \rho(E, M_i) \neq 0 \), \( \rho(E, M_i) \) is more likely to be comparatively large, and \( S_i \) only requires a single chance event, namely for \( |\hat{\rho}(r_{Y|E}, r_{M_i|E})| \) to be large, to be statistically significant. Hence, lowering \( K \) increases the FWER toward 0.05.

The improvement by the joint test 2J, as compared to 2B, is a result of the fundamental difference between the Bonferroni and joint corrections. The Bonferroni correction fixes a percentile, \( 100 \times (1 - 0.05/K) \), and therefore lets \( t_{2B} \) vary across mediators. The joint correction fixes \( t_{2J} \) and lets the corresponding percentile, \( 100 \times P(S_i \geq t_{2J}) \) vary across all mediators. When testing associations, \( t_{2B} \) is effectively independent of the underlying truth and \( t_{2J} \) when all variables are independent. In stark contrast, as we observed in our previous simulations, when testing mediators, \( t_{2B} \) is strongly dependent on the true values of \( \alpha \) and \( \beta \). Moreover, \( t_{2B} \) will tend to be larger for true mediators, when both \( \alpha \neq 0 \) and \( \beta \neq 0 \), resulting in the Bonferroni correction being especially tough on the true mediators. The improvement by Test 2J, as compared to 1B or 1J, results from the fact that the thresholds for significance for each association are likely driven by two different metabolites.

As with testing associations, the Bonferroni correction is increasingly conservative as the correlation between variables increases. Figure 3 shows that, as the correlation within the 5 sets of null mediators increases, the power for the joint correction increases, from 0.670 for \( \rho = 0 \) to 0.768 for \( \rho = 1 \), an increase of 15.9%. However, the power for Test 2B stays approximately constant. In particular, as \( \rho \) increases, the power of the joint correction gets closer to the power of the “limiting scenario” (0.760), where there are 6 independent mediators, one of which is true; this does not hold for the Bonferroni correction.

### 3.3 Navy Colorectal Adenoma Study

In the Navy Colorectal Adenoma Study, red meat consumption was associated with an increased risk of colorectal adenoma (\( \rho = 0.010 \), while fish consumption was associated with decreased risk (\( \rho = 0.075 \)), adjusting for gender, age, current smoking status, and body mass index. Although no metabolite could be identified as a potential mediator for the association with red meat, Test 2J suggested that increased docosahexaenoate (DHA, fish oil) may link fish consumption with a decreased risk of colorectal adenoma (\( \rho_S = 0.062 \), Table 5). DHA was positively associated with

<table>
<thead>
<tr>
<th>Number of mediators of each type</th>
<th>FWER</th>
<th>Test 2B</th>
<th>Test 2J</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K ) ( K ) ( K ) ( K ) ( K )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>100</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of mediators of each type</th>
<th>Power</th>
<th>Test 2B</th>
<th>Test 2J</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K ) ( K ) ( K ) ( K ) ( K )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>101</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>70</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>101</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of mediators of each type</th>
<th>FWER</th>
<th>Test 2B</th>
<th>Test 2J</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K ) ( K ) ( K ) ( K ) ( K )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>70</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>100</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of mediators of each type</th>
<th>Power</th>
<th>Test 2B</th>
<th>Test 2J</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K ) ( K ) ( K ) ( K ) ( K )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>101</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>70</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>101</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
fish consumption (p < 0.001) and negatively associated with adenoma (p = 0.013). This result agrees with previous research (Cheng et al., 2003; Chapkin et al., 2007) and supports ongoing studies like the seaAOod Polyp Prevention Trial (http://www.seafood-trial.co.uk/). DHA also had the smallest p-value by Test 2B (0.007), but this evidence would not have appeared convincing after adjusting for 149 tests.

Table 5. Results for the Navy Colorectal Adenoma study. For each dietary intake of interest, we list the most likely mediator, as suggested by Tests 2B and 2J. For the top mediator, we report p_B for Test 2B (i.e. compare to 0.05 divided by 149 tests) and p_J for Test 2J (i.e. compare directly to 0.05). 1-SGPA abbreviates 1-stearoylglycerophosphoethanolamine and DHA abbreviates docosahexaenoate.

<table>
<thead>
<tr>
<th></th>
<th>Test 2B</th>
<th>Test 2J</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metab p-value</td>
<td>Metab FWER</td>
</tr>
<tr>
<td>Red meat</td>
<td>1-SGPA 0.069</td>
<td>glycerol 0.577</td>
</tr>
<tr>
<td>Fish</td>
<td>DHA 0.006</td>
<td>DHA 0.062</td>
</tr>
</tbody>
</table>

4 DISCUSSION

We developed permutation methods for testing multiple putative mediators. Although such methods can be applied in a variety of settings, we considered the specific application of a modern epidemiological study that measures 100s or 1000s of similar biomarkers (e.g. gene expression, protein, or metabolite levels). We show, via simulations, that testing putative mediators by using the joint correction has substantially higher power over Bonferroni correction even when all biomarkers are conditionally independent. We apply our approach to the Navy Colorectal Adenoma Study and find evidence suggesting that DHA may mediate the protective effect of fish consumption on adenoma risk, which would not have been found using Bonferroni.

We first defined a permutation method for testing a single mediator. This method was used to lay the groundwork for describing our tests of multiple mediators, but is also novel in the single mediator literature. Whereas prior permutation methods (Taylor and MacKinnon, 2012) used the Manly (1997) approach for testing the conditional association between M and Y, we used the Freedman and Lane (1983) approach which, in general, appears to be a more robust approach. (Anderson and Robinson, 2001). We then extended this method to testing multiple mediators, using either the Bonferroni correction or a joint correction. The key component of our methods, required to handle the composite null hypothesis in both the single and multiple mediator scenarios, is to use two sets of permutations.

A simpler, but incorrect, alternative to our approach would be to repeatedly permute Y to obtain one null distribution of S_{max}. Then, repeatedly permute E to obtain another null distribution, and let p_{max} be the maximum of the two p-values. However, if a subset of metabolites was associated with the exposure and a different subset was associated with the outcome, then the observed S_{max} would appear extreme as measured by either simple null distribution.

As with any approach for testing mediation, our method presumes that the causal paths considered in Figures 1 and 2 are correct. Thus, we assumed that the only causal paths that may exist are from the exposure to the biomarkers, from the biomarkers to the outcome, and from the exposure directly to the outcome. Unmeasured confounders might also lead to incorrect inferences. Such a confounder might, for example, induce a relationship between M and Y, causing β to appear to be non-zero. If this same metabolite were associated with E, the result would be a false positive. Despite this limitation, causal analysis still offers one of the best methods for testing putative mediators. Given the popularity of high-throughput technologies, investigators will soon require methods for testing multiple putative mediators. This manuscript is one of the first to introduce possible options.

5 ACKNOWLEDGMENTS

This study was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH) and utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (http://biowulf.nih.gov).

REFERENCES


