BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 214–220
doi:10.1093/bioinformatics/btt633

Systems biology

Advance Access publication November 6, 2013

Testing multiple biological mediators simultaneously

Simina M. Boca, Rashmi Sinha, Amanda J. Cross, Steven C. Moore and Joshua N. Sampson*

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA

Associate Editor: Dr. Igor Jurisica

ABSTRACT

Motivation: Modern biomedical and epidemiological studies often measure hundreds or thousands of biomarkers, such as gene expression or metabolite levels. Although there is an extensive statistical literature on adjusting for 'multiple comparisons' when testing whether these biomarkers are directly associated with a disease, testing whether they are biological mediators between a known risk factor and a disease requires a more complex null hypothesis, thus offering additional methodological challenges.

Results: We propose a permutation approach that tests multiple putative mediators and controls the family wise error rate. We demonstrate that, unlike when testing direct associations, replacing the Bonferroni correction with a permutation approach that focuses on the maximum of the test statistics can significantly improve the power to detect mediators even when all biomarkers are independent. Through simulations, we show the power of our method is 2–5× larger than the power achieved by Bonferroni correction. Finally, we apply our permutation test to a case-control study of dietary risk factors and colorectal adenoma to show that, of 149 test metabolites, docosahexaenoate is a possible mediator between fish consumption and decreased colorectal adenoma risk.

Availability and implementation: R-package included in online Supplementary Material.

Contact: joshua.sampson@nih.gov

Supplementary information: Supplementary materials are available at *Bioinformatics* online.

Received on April 8, 2013; revised on August 10, 2013; accepted on October 29, 2013

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; Biesanz *et al.*, 2010; MacKinnon, 2008; MacKinnon *et al.*, 2002; 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 100 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 article 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.

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)

$$Y = \kappa_0 + bX + \epsilon \tag{1}$$

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 95*th* 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 $\pi(Y)$, 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 95*th* percentile of the permuted values. The tests discussed in this article are summarized in Table 1.

2.2 Testing multiple associations

When we test the association between Y and multiple variables, $X_1, ..., X_K$, we calculate $|\hat{\rho}(X_i, Y)|$ and its corresponding P-values, p_{b_i} for $i \in \{1, ..., K\}$. To maintain a FWER of 0.05, we have two options. First, we can estimate the threshold by Bonferroni correction, where t^i_B is the $100 \times (1 - 0.05/K)th$ 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 $p_{b_i} \leq 0.05/K$ (or $|\hat{\rho}(X_i, Y)| \geq t^i_B$).

For the permutation estimates of t^i_B 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 95*th* percentile of the distribution of $\hat{\rho}_{\max} \equiv \max_i(|\hat{\rho}(X_i, \pi(Y))|)$ 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 95*th* 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.

Table 1. Tests discussed in the article

Type	Association	Mediation	
		$\alpha,\ \beta$	$\alpha \times \beta$
Single	Test 0	Test 1	Test 2
Multiple			
Bonf	Test 0B	Test 1B	Test 2B
joint	Test 0J	Test 1J	Test 2J

Note: 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 α and β are each evaluated separately ($\alpha,\ \beta$) or as a product ($\alpha \times \beta$).

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.

$$M = \kappa_M + \alpha E + \epsilon_M, \tag{2}$$

$$Y = \kappa_Y + \gamma E + \beta M + \epsilon_Y, \tag{3}$$

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 (MacKinnon *et al.*, 2002; Sobel, 1982). 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 α and β. Therefore, we replace α and β by ρ_{EM} and $\rho_{MY|E}$, respectively, where ρ_{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 $\hat{\rho}(r_{M|E}, r_{Y|E})$, offer estimates of these parameters. Here, $r_{Y|E}$ and $r_{M|E}$, are the residuals from regressing Y on E and M on E, respectively.

The first test we consider is as follows:

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

where p_α is the corresponding P-value and $t_1(\alpha)$ is the 95*th* percentile for the null distribution of $|\hat{\rho}(E, M)|$. p_β and $t_1(\beta)$ are defined similarly for $|\hat{\rho}(r_{M|E}, r_{Y|E})|$. All values can be calculated by asymptotic theory or permutation. An overall P-value for Test 1 can be defined as $\max(p_\alpha, p_\beta)$. The exact steps are presented in Algorithms 1 (for α) and 2 (for β). 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_α 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_α 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.

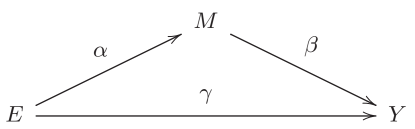

Fig. 1. A scenario with a single possible mediator between exposure and outcome

Algorithm 2: Permutation Algorithm for p_β 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, $r^{\pi}_{Y|E}$.
(3) Calculate $\hat{\rho}(r_{M|E}, r^{\pi}_{Y|E})$.
(4) Repeat steps 1, 2 and 3 to obtain a distribution of $\hat{\rho}(r_{M|E}, r^{\pi}_{Y|E})$.
    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 $|\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r^{\pi}_{Y|E})|$. 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.

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 $\hat{\rho}(\pi(E), M)$ as in
    Algorithm 1 and $S^{\pi} = |\hat{\rho}(\pi(E), M)\hat{\rho}(r_{Y|E}, r_{M|E})|$.
(2) If $|\hat{\rho}(E, M)| \geq |\hat{\rho}(r_{M|E}, r_{Y|E})|$, calculate $\hat{\rho}(r_{M|E}, r^{\pi}_{Y|E})$ as in
    Algorithm 2 and $S^{\pi} = |\hat{\rho}(E, M)\hat{\rho}(r_{M|E}, r^{\pi}_{Y|E})|$.
(3) Repeat steps 1 and 2 to obtain a distribution of $S^{\pi}$. Calculate
    $P$-value and threshold as usual.
```

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:

$$M_i = \kappa_{M_i} + \alpha_i E + \epsilon_{M_i}, \tag{5}$$

$$Y = \kappa_Y + \gamma E + \sum_{i=1}^{K} \beta_i M_i + \epsilon_Y, \tag{6}$$

where $E \sim N(0, \sigma^2_E)$, $\epsilon_{M_i} \sim N(0, \sigma^2_{M_i})$ and $\epsilon_Y \sim N(0, \sigma^2_Y)$. Therefore, Y can also be described by

$$Y = \kappa^*_Y + \gamma^* E + \epsilon^*_Y, \tag{7}$$

where $\epsilon^*_Y \sim N(0, \sigma^2_Y + \sum_{i=1}^{K} \beta^2_i \sigma^2_{M_i} + (\gamma + \sum_{i=1}^{K} \alpha_i \beta_i)^2 \sigma^2_E)$

When we test multiple mediators, $M_1, \ldots, M_K$, we calculate $\hat{\rho}(E, M_i)$, $\hat{\rho}(r_{M_i|E}, r_{Y|E})$ 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 $p_{\alpha_i} \leq 0.05/K$ and $p_{\beta_i} \leq 0.05/K$ (or $|\hat{\rho}(E, M_i)| \geq t^i_B(\alpha)$ and $|\hat{\rho}(r_{M_i|E}, r_{Y|E})| \geq t^i_B(\beta)$).

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

Here, $t^i_B(\alpha)$, $t^i_B(\beta)$ and $t^i_B(S)$ are the $100 \times (1 - 0.05/K)th$ percentiles of the appropriate null distribution. Again, the quality of the permuted estimates of $t^i_B(\alpha)$, $t^i_B(\beta)$ and $t^i_B(S)$ depends 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).

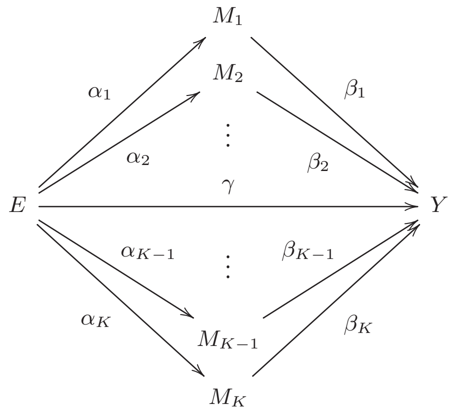

Fig. 2. A scenario with K possible mediators between exposure and outcome

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

Test 1J:. We declare M_i to be significant if $p^{\max}_{\alpha_i} \leq 0.05$ and $p^{\max}_{\beta_i} \leq 0.05$ [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 95*th* percentiles of the null distributions of $\hat{\rho}^{\max}_{EM}$ and $\hat{\rho}^{\max}_{MY}$, respectively, and $p^{\max}_{\alpha_i}$ and $p^{\max}_{\beta_i}$ compare the values of the observed statistics for mediator i to these 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 $S_{\text{joint}} = \max_i(S_i)$:

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

To define the permutation algorithm for p_{J_i} 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_i} and $t_J(S)$.

```
(1) For $i \in A$, calculate $\hat{\rho}(\pi(E), M_i)$ and $S^{\pi}_i = |\hat{\rho}(\pi(E),
    M_i)\hat{\rho}(r_{Y|E}, r_{M_i|E})|$.
(2) For $i \in B$, calculate $\hat{\rho}(r_{M_i|E}, r^{\pi}_{Y|E})$ and $S^{\pi}_i = |\hat{\rho}(E, M_i)\hat{\rho}(r_{M_i|E}, r^{\pi}_{Y|E})|$.
(3) Calculate $S^{\pi}_{\text{joint}} = \max_i(S^{\pi}_i)$.
(4) Repeat steps 1, 2 and 3 to obtain a distribution of $S^{\pi}_{\text{joint}}$. Calculate
    the $P$-value, $p_{J_i}$, for each mediator as the proportion of $S^{\pi}_{\text{joint}}$
    exceeding $S_i$ and the threshold $t_J(S)$ as the 95th percentile of this
    distribution.
```

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:

$$Pr(Y=1|E, M_1, \ldots, M_K) = \frac{e^{K_Y+\gamma E+\sum_{i=1}^{K}\beta_i M_i}}{1+e^{K_Y+\gamma E+\sum_{i=1}^{K}\beta_i M_i}} \tag{8}$$

However, we will need to slightly modify the algorithms in case/control studies, where the data are collected retrospectively. Thus, 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 υ/q and each control by $(1-\upsilon)/(1-q)$, where υ represents the prevalence of the outcome in the overall population and q represents the proportion of cases in the sample, as in VanderWeele and Vansteelandt (2010). We note that for this analysis to be performed, one must have prior knowledge of υ, 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 (3) and (4) (Fig. 1) for the single mediator case and Equations (5) and (6) (Fig. 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 (α, β) to be either (0, 0), (es, 0) or (0, es), where es 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 (α, β) to (es,es), (0.5es,1.5es) or (1.5es, 0.5es). 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 $\sigma_E^2 = 1$, $\sigma_M^2 = 1-\alpha^2$ and $\sigma_Y^2 = 1-\beta^2\sigma_M^2-(\alpha\beta+\gamma)^2$. We consider es = 0.2 for $n = 100$, and es = 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 (α, β) to either (0, 0), (es$_1$, 0) or (0, es$_2$), where β now refers to β_1 in Equation (8), and for power we set (α, β) to (es$_1$, es$_2$), (0.5es$_1$, 1.5es$_2$) and (1.5es$_1$, 0.5es$_2$). We consider es$_1$ = 0.2, es$_2$ = 0.5 when $n = 100$ and es$_1$ = 0.08, es$_2$ = 0.2 when $n = 1000$. The combinations of n, α and β we consider are listed in Table 2 (for normally distributed outcome) and Supplementary Table 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 $(\alpha, \beta) = (0, 0), (0, \text{es})$, and (es, 0), respectively. Here, we consider (K_1, K_2, K_3) = (10, 0, 0), (7, 3, 0), (7, 0, 3) and (6, 2, 2) for K = 10 and (K_1, K_2, K_3) = (100, 0, 0), (70, 30, 0), (70, 0, 30) and (60, 20, 20) for K = 100. We then simulate 10 000 studies and define the 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 (α, β) = (es, es). We discuss the scenario of multiple mediators in Supplementary Section S3 of the Supplementary Material. We then simulate 1000 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 $\sigma_E^2 = 1$, $\sigma_{M_i}^2 = 1-\alpha_i^2$ and $\sigma_Y^2 = 1-\sum_{i=1}^{K}\beta_i^2\sigma_{M_i}^2-(\sum_{i=1}^{K}\alpha_i\beta_i+\gamma)^2$. In the Supplementary Material, we also consider simulations with more than one true mediator, including scenarios where one is 'screening' mediators.

Table 2. Single mediator, normally distributed outcome: Type I error and power are estimated by simulation

n	α	β	Type I error: Test 1	Type I error: Test 2
100	0.00	0.00	0.004	0.004
100	0.20	0.00	0.023	0.023
100	0.00	0.20	0.026	0.026
1000	0.00	0.00	0.001	0.001
1000	0.08	0.00	0.034	0.034
1000	0.00	0.08	0.035	0.035

n	α	β	Power: Test 1	Power: Test 2
100	0.20	0.20	0.242	0.242
100	0.10	0.30	0.131	0.131
100	0.30	0.10	0.122	0.122
1000	0.08	0.08	0.510	0.510
1000	0.04	0.12	0.217	0.217
1000	0.12	0.04	0.229	0.229

The DAG in Figure 2 implies that the putative mediators are independent conditional on exposure. We also want to explore the benefit of Test 2J when the mediators are conditionally dependent. Specifically, we consider the scenario with one true mediator (es = 0.1) and 100 null mediators (n = 1000), each with $\alpha_i = \beta_i = 0$ and divided into five correlated blocks of equal size. Two mediators in the same block have a correlation of ρ, which varies from 0 to 1, whereas any mediators in two different blocks are independent. We calculate the power based on 10 000 simulated studies.

Next, we consider the scenario where Y is binary. The design is similar, but we now set (α_i, β_i) to either (0, 0), (es$_1$, 0), (0, es$_2$) or (es$_1$, es$_2$), where β_i now refers to Equation (8). When $K = 10$ or 11, we consider n = 100, es$_1$ = 0.3 and es$_2$ = 0.6, whereas for $K = 100$ or 101, we take n = 1000, es$_1$ = 0.1 and es$_2$ = 0.2. The combinations of K_1, K_2, K_3 and K_4 we consider are listed in Tables 3 and 4, with K_4 being the number of true mediators (and therefore, just 0 or 1.).

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

Table 3. Multiple mediators, normally distributed outcome: FWER and power are estimated by simulation

Number of mediators of each type					FWER	
K	K_1	K_2	K_3	K_4	Test 2B	Test 2J
10	10	0	0	0	<0.001	0.013
10	7	3	0	0	0.009	0.048
10	7	0	3	0	0.007	0.050
10	6	2	2	0	0.012	0.048
100	100	0	0	0	<0.001	0.022
100	70	30	0	0	0.005	0.051
100	70	0	30	0	0.006	0.051
100	60	20	20	0	0.007	0.052
Number of mediators of each type					Power	
K	K_1	K_2	K_3	K_4	Test 2B	Test 2J
11	10	0	0	1	0.291	0.666
11	7	3	0	1	0.291	0.447
11	7	0	3	1	0.287	0.452
11	6	2	2	1	0.298	0.403
101	100	0	0	1	0.133	0.657
101	70	30	0	1	0.133	0.268
101	70	0	30	1	0.140	0.277
101	60	20	20	1	0.139	0.243

Note: Different effect sizes are considered: K_1 mediators have $(\alpha, \beta) = (0, 0)$, K_2 have $(\alpha, \beta) = (\text{es}, 0)$, K_3 have $(\alpha, \beta) = (0, \text{es})$ and K_4 have $(\alpha, \beta) = (\text{es}, \text{es})$.

Table 4. Multiple mediators, case-control study: FWER and power are estimated by simulation

Number of mediators of each type					FWER	
K	K_1	K_2	K_3	K_4	Test 2B	Test 2J
10	10	0	0	0	0.001	0.016
10	7	3	0	0	0.009	0.052
10	7	0	3	0	0.007	0.046
10	6	2	2	0	0.013	0.048
100	100	0	0	0	<0.001	0.021
100	70	30	0	0	0.004	0.051
100	70	0	30	0	0.007	0.048
100	60	20	20	0	0.009	0.051
Number of mediators of each type					Power	
K	K_1	K_2	K_3	K_4	Test 2B	Test 2J
11	10	0	0	1	0.197	0.546
11	7	3	0	1	0.197	0.380
11	7	0	3	1	0.163	0.340
11	6	2	2	1	0.173	0.287
101	100	0	0	1	0.102	0.534
101	70	30	0	1	0.102	0.239
101	70	0	30	1	0.053	0.161
101	60	20	20	1	0.074	0.137

Note: Different effect sizes are considered: K_1 mediators have $(\alpha, \beta) = (0, 0)$, K_2 have $(\alpha, \beta) = (\text{es}_1, 0)$, K_3 have $(\alpha, \beta) = (0, \text{es}_2)$ and K_4 have $(\alpha, \beta) = (\text{es}_1, \text{es}_2)$.

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 446 known metabolites measured, we considered only the 149 metabolites present in all study participants. The 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.

3 RESULTS

3.1 Simulations: single mediator

The type I error rates for Test 1 and Test 2 depend on the values of α and β. When both $\alpha = 0$ and $\beta = 0$, both tests are extremely conservative, with type I error rates close to $0.0025 = 0.05^2$. When either ($\alpha \gg 0$ and $\beta = 0$) or ($\alpha = 0$ and $\beta \gg 0$), type I error rates are closer to 0.05 (Table 2 and Supplementary Table 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 $\alpha = \beta$ was set to be 0.1, 0.2, 0.3, 0.4 and 0.5, we found the corresponding thresholds to be 0.032, 0.050, 0.068, 0.086 and 0.103. Because we are testing a composite null hypothesis, each permuted statistic still uses the observed $\hat{\alpha}$ or $\hat{\beta}$ and therefore increases with the observed values.

3.2 Simulations: multiple mediators

The FWER depends on the values of α_i and β_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 2J 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 $\alpha = \beta = 0.1$ for the true mediator, we found the power for detecting an association by Test 2J to be 0.657, 0.268, 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 0.14. Importantly, note that the power is greatly improved by the joint test here, whereas 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_1 decreases, the power for Test 2J decreases, whereas its FWER increases. Both trends can be explained by the fact that the expected value of S_i^π is higher when the null variable has one association, with either E or Y. For example, assume that M_i is associated with E. The result is that variable i will likely be included in group B and the expected value of $(S_i^\pi)^2$ will be approximately $[\rho(E, M_i)]^2 \times \mathrm{var}(\hat{\rho}(r_{Y|E}, r_{M_i|E}))$. Now, consider M_i^*, associated with neither E nor Y. $(S_{i^*}^\pi)^2$ should only be slightly larger than $\mathrm{var}(\hat{\rho}(r_{Y|E}, r_{M_i^*|E})) \times \mathrm{var}(\hat{\rho}(E, M_i^*))$, a comparatively small number. Because S_i^π increases with K_2 and K_3, 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 than 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_1 increases the FWER toward 0.05.

The improvement by the joint Test 2J, as compared with 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_B^i vary across mediators. The joint correction fixes t_J and lets the corresponding percentile, $100 \times P(S_i \geq t_J)$ vary across all mediators. When testing associations, t_B^i is effectively independent of the underlying truth and $t_B^i \approx t_J$ when all variables are independent. In stark contrast, as we observed in our previous simulations, when testing mediators, t_B^i is strongly dependent on the true values of α and β. Moreover, t_B^i 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 with 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 five 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 ρ increases, the power of the joint correction gets closer to the power of the 'limiting scenario' (0.760), where there are six 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 ($P = 0.010$), whereas fish consumption was associated with a decreased risk ($P = 0.075$), 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 ($p_J = 0.062$, Table 5). DHA was positively associated with fish consumption ($P < 0.001$) and negatively associated with adenoma ($P = 0.013$). This result agrees with previous research (Chapkin *et al.*, 2007; Cheng *et al.*, 2003) and supports ongoing studies like the seaAFOod Polyp Prevention Trial (http://www.seafood-trial.co.uk/). DHA also had the smallest P-value by Test 2B (0.006), but this evidence would not have appeared convincing after adjusting for 149 tests.

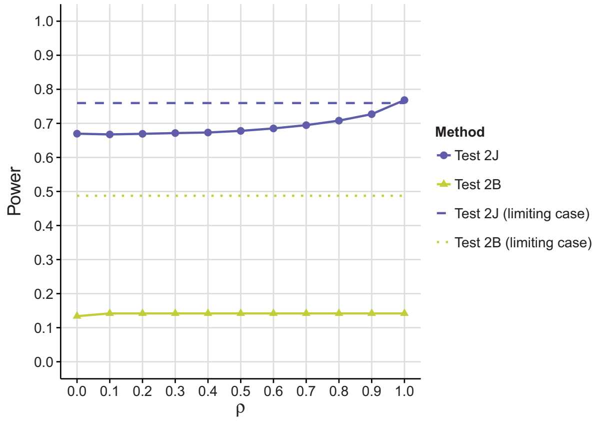

Fig. 3. The solid lines illustrate the power for Test 2B (light) and Test 2J (dark) as a function of ρ (within block correlation) when there are 100 null mediators and 1 true mediator. Mediators in different blocks are uncorrelated. The dashed lines show the limiting power, when the 100 null mediators are replaced with 5 independent null mediators

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 100 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_{joint}. Then, repeatedly permute E to obtain another null distribution, and let p_S be the maximum of the two P-values. However, if a

Table 5. Results for the Navy Colorectal Adenoma study

Dietary item	Test 2B		Test 2J	
	Metabolite	*P*-value	Metabolite	FWER
Red meat	1-SGPA	0.069	glycerol	0.577
Fish	DHA	0.006	DHA	0.062

Note: 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_S for Test 2B (i.e. compare with 0.05 divided by 149 tests) and p_J for Test 2J (i.e. compare directly with 0.05). 1-SGPA (1-stearoylyglycerophosphoethanolamine) and DHA (docosahexaenoate).

subset of metabolites was associated with the exposure and a different subset was associated with the outcome, then the observed S_{joint} 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.

Funding: Support was from the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH). Computation was performed using the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (http://biowulf.nih.gov).

Conflict of Interest: None declared.

REFERENCES

Anderson,M.J. and Robinson,J. (2001) Permutation tests for linear models. *Aust. N. Z. J. Stat.*, **43**, 75–88.

Baron,R.M. and Kenny,D.A. (1986) The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, **51**, 1173–1182.

Biesanz,J.C. *et al.* (2010) Assessing mediational models: testing and interval estimation for indirect effects. *Multivariate Behav. Res.*, **45**, 661–701.

Brown,O.P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Brown,B.M. and Maritz,J.S. (1982) Distribution-free methods in regression. *Aust. J. Stat.*, **24**, 318–331.

Caffo,B. *et al.* (2008) Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function? *Am. J. Epidemiol.*, **167**, 429–437.

Chapkin,R.S. *et al.* (2007) Colon cancer, fatty acids and anti-inflammatory compounds. *Curr. Opin. Gastroenterol.*, **23**, 48–54.

Cheng,J. *et al.* (2003) Increased intake of n-3 polyunsaturated fatty acids elevates the level of apoptosis in the normal sigmoid colon of patients polypectomized for adenomas/tumors. *Cancer Lett.*, **193**, 17–24.

Dettmer,K. *et al.* (2006) Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.*, **26**, 51–78.

Freedman,D. and Lane,D. (1983) A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.*, **1**, 292–298.

Huang,Y. *et al.* (2006) To permute or not to permute. *Bioinformatics*, **22**, 2244–2248.

MacKinnon,D.P. (2008) *Introduction to Statistical Mediation Analysis*. Erlbaum Psych Press, New York.

MacKinnon,D.P. *et al.* (2002) A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods*, **7**, 83.

Manly,B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd edn. Chapman & Hall, London.

R Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Sinha,R. *et al.* (1999) Well-done, grilled red meat increases the risk of colorectal adenomas. *Cancer Res.*, **59**, 4320–4324.

Sobel,M.E. (1982) Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhart,S. (ed.) *Sociological Methodology*. Vol. 13, American Sociological Association, Washington, DC, pp. 290–312.

Sreekumar,A. *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, **457**, 910–914.

Suhre,K. *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54–60.

Taylor,A.B. and MacKinnon,D.P. (2012) Four applications of permutation methods to testing a single-mediator model. *Behav. Res. Methods*, **44**, 1–39.

VanderWeele,T.J. and Vansteelandt,S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.*, **172**, 1339–1348.

Wardle,J. *et al.* (2008) Obesity associated genetic variation in FTO is associated with diminished satiety. *J. Clin. Endocrinol. Metab.*, **93**, 3640–3643.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. Vol. 279, Wiley-Interscience, New York.

Xu,H. and Hsu,J.C. (2007) Applying the generalized partitioning principle to control the generalized familywise error rate. *Biom. J.*, **49**, 52–67.