Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Global Test for High-dimensional Mediation: Testing Groups of Potential Mediators

Vera Djordjilović[1] | Christian M. Page[2,3] | Jon Michael Gran[1,2] | Therese H. Nøst[4] | Torkjel M. Sandanger[4] | Marit B. Veierød[1] | Magne Thoresen[1]

[1]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Norway
[2]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Norway
[3]Center for Fertility and Health, Division of Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway
[4]Department of Community Medicine, The Arctic University of Norway, Tromsø, Norway

Correspondence
Vera Djordjilović, Department of Biostatistics, University of Oslo, Norway,
Email: vera.djordjilovic@medisin.uio.no

Abstract

We address the problem of testing whether a possibly high-dimensional vector may act as a mediator between some exposure variable and the outcome of interest. We propose a global test for mediation, which combines a global test with the intersection-union principle. We discuss theoretical properties of our approach and conduct simulation studies which demonstrate that it performs equally well or better than its competitor. We also propose a multiple testing procedure, ScreenMin, that provides asymptotic control of either familywise error rate or false discovery rate when multiple groups of potential mediators are tested simultaneously. We apply our approach to data from a large Norwegian cohort study, where we look at the hypothesis that smoking increases the risk of lung cancer by modifying the level of DNA methylation.

KEYWORDS:
Multiple mediators, High-dimensional data, Familywise error rate, False discovery rate

1 | INTRODUCTION

Over the last years, we have witnessed an increased interest for causal mediation analysis in genetic epidemiology, genomics, epigenomics, and neuroscience. For example, researchers in epidemiology hypothesise that epigenetic alterations might in certain cases mediate the effect of environmental exposure on the outcome of interest.[1,2,3] In neuroscience, functional magnetic resonance imaging was used to investigate whether the effect of temperature on reported pain is mediated by brain response measured at thousands of voxels.[4] What these problems have in common is that instead of a single variable M on the path between an exposure X and an outcome Y, there is a high-dimensional vector $\boldsymbol{M}$ (Figure 1).

With the recent developments in causal mediation analysis, various methods for settings with multiple mediators have been suggested.[5,6,7,8] However, most of these are not well suited for high-dimensional settings. The methods usually aim at decomposing the total effect into different pathways involving components of $\boldsymbol{M}$, which seems to be an overly ambitious task in most biomedical applications involving high-dimensional mediators. While it can be safely assumed that the components of $\boldsymbol{M}$ are highly correlated due to unmeasured common causes, no information is usually available about their causal ordering. They are thus typically considered *en bloc*, without trying to detangle the effects within $\boldsymbol{M}$. Although this simplifies the problem, the dimensionality of $\boldsymbol{M}$ still poses difficulties and standard inferential methods cannot be readily applied in this context.

A common approach in high-dimensional inference in general is to perform some initial screening of variables to identify candidate variables worthy of further investigation. When only a small fraction of variables is expected to play a role in the problem

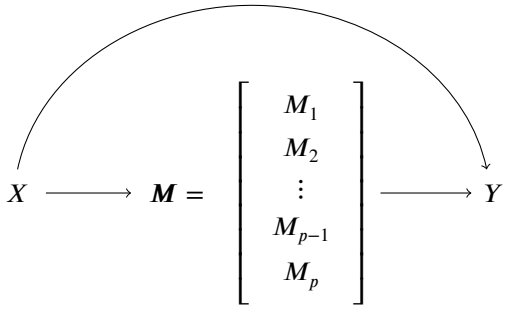

FIGURE 1 Causal diagram of the mediation model with an exposure X, an outcome Y, and a vector $\boldsymbol{M} = (M_1, \dots, M_p)^\top$ of potential mediators.

at hand, this preliminary screening step can significantly reduce the problem by discarding a large number of unpromising variables. In this work, we propose a test to identify promising mediator candidates in such a screening step. More precisely, we test whether $\boldsymbol{M}$ is a) associated to X; and b) associated to Y conditionally on X. We propose a global test for mediation, which adapts the global test[9] to the mediation setting. The global test was first proposed for testing an association between a clinical outcome and a group of functionally related genes. It offers a valuable alternative in situations when standard methods fail – such as when the number of genes greatly exceeds the sample size. It has been proved to enjoy certain optimality properties[10] and is especially suitable for detecting departures from the null hypothesis characterised by a large number of small effects. Our proposed procedure inherits these favourable properties.

Note that for the above-mentioned associations to represent actual mediation, certain causal assumptions need to hold. Generally, there should be no unobserved confounding between exposure and mediator, exposure and outcome, and mediator and outcome, and no exposure induced mediator-outcome confounding.[11] To accommodate the fact that these assumptions will never hold in practice without any covariate adjustment, our proposed procedure easily extends to include covariates.

In high-dimensional applications, many groups of potential mediators are usually considered simultaneously. Finding an efficient way of addressing the problem of multiplicity, while controlling the inclusion of false positives, is an important task. To address this issue, we propose a multiple testing procedure, ScreenMin, that provides asymptotic control of the familywise error rate and can be easily adapted to control the false discovery rate. The behavior of the procedure is investigated in a small simulation study.

The problem of high-dimensional mediation is quite recent and has so far received limited attention in the literature. Zhang et al.[12] addressed the problem of selecting variables that act as mediators among a very high-dimensional set of potential mediators. Huang and Pan[13] and Chén et al.,[4] similar to us, considered a mediation effect of a group of potential mediators and proposed a testing and an estimating procedure, respectively. As opposed to the global test for mediation, both of these approaches dealt with the issue of high-dimensionality by employing orthogonal transformations of $\boldsymbol{M}$. We compare our proposal with the testing procedure put forward in Huang and Pan[13] in two simulation studies.

The manuscript is organized as follows. In Section 2, we describe the model for the triple $(X, \boldsymbol{M}, Y)$. The global test for mediation is presented in Section 3. The ScreenMin procedure for testing multiple groups of potential mediators is proposed in Section 4. The results of the simulation studies are reported in Section 5. In Section 6, we analyze data from a large Norwegian cohort study, where we look at the hypothesis that smoking increases the risk of lung cancer by modifying the level of DNA methylation at specific CpG sites. Some concluding remarks and open questions are in the Discussion; proofs and technical details are in the Appendix.

2 | MODEL SPECIFICATION

Let Y denote the outcome of interest, X the exposure variable, and $\boldsymbol{M} = \left(M_1, \dots, M_p\right)^\top$ a p-dimensional vector of potential mediators. Let us assume that subject matter considerations suggest that the causal structure of the triple $(X, \boldsymbol{M}, Y)$ takes on the form depicted in Figure 1. To describe it, we specify two models: one for the vector of mediators and one for the outcome. The linear model is assumed for the mediator

$$\boldsymbol{M} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha} X + \boldsymbol{\epsilon}_M, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^\top$ is a vector of regression coefficients associated to X, $\boldsymbol{\epsilon}_M \sim \mathsf{N}_p(0, \Sigma)$ is a random disturbance independent of X, and Σ is an arbitrary positive definite matrix.

The generalized linear model is assumed for the outcome

$$g\left[\mathrm{E}(Y \mid X, \boldsymbol{M})\right] = \beta_0 + \boldsymbol{M}^\top \boldsymbol{\beta} + \gamma X, \tag{2}$$

where g is a suitable link function, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^\top$ is a vector of regression coefficients pertaining to $\boldsymbol{M}$, and $\gamma \in \mathbb{R}$. Common choices for the link function include identity, logit and log function, leading to the linear, logistic, and Poisson regression models, respectively. We will pay special attention to the first two choices for continuous and binary outcomes.

3 | GLOBAL TEST FOR MEDIATION

Based on the structure depicted in Figure 1, we are interested in determining whether $\boldsymbol{M}$ has the properties of a mediator. In a setting without confounders, this corresponds to determining whether $\boldsymbol{M}$ is a) associated to X; b) associated to Y conditionally on X. In the context specified by (1) and (2) this is formulated as a testing problem

$$H : \boldsymbol{\alpha} = 0 \text{ or } \boldsymbol{\beta} = 0, \quad \text{against} \quad H_A : \boldsymbol{\alpha} \neq 0 \text{ and } \boldsymbol{\beta} \neq 0. \tag{3}$$

The hypothesis H can alternatively be written as

$$H = H_1 \cup H_2,$$

where $H_1 : \boldsymbol{\alpha} = 0$ and $H_2 : \boldsymbol{\beta} = 0$. Both H_1 and H_2 define nested submodels and could be tested with standard methods, such as likelihood ratio tests. However, the performance of the standard methods deteriorates when the number of potential mediators p approaches the sample size of the experiment n, and they are not defined when p exceeds n. To circumvent this issue, it is possible to consider orthogonal transformations of $\boldsymbol{M}$,[4,13] which allows for reformulating the problem within a setting of p smaller models to be estimated separately. In this paper, we follow a different strategy, and propose a procedure based on a combination of the global test[10] with the intersection-union principle[14].

3.1 | The proposed procedure

We propose the following procedure for testing H:

Procedure 1. Global test for mediation. Test H_1 and H_2 at level α with a global test,[10] and reject H at level α if both H_1 and H_2 are rejected.

The validity of the global test for mediation follows trivially from the validity of the global test and the intersection-union principle described by Berger,[15] who gave general properties of tests constructed on the basis of this principle. In particular, if p_1 and p_2 are p-values of tests of H_1 and H_2, then $p = \max\left\{p_1, p_2\right\}$ is a p-value of H.

The global test was first proposed in the context of gene set analysis,[9] as a means of testing the association between a group of genes and a phenotype of interest. Given that the test does not degenerate when the number of model parameters exceeds the sample size, it provided a conceptually simple, yet effective answer to the problem of low sample size typical for genomic applications. The approach was further developed as a general answer to the problem of testing a point null hypothesis against a high-dimensional alternative in generalized linear models,[10] where it was shown that it is locally most powerful on average in the neighbourhood of the null hypothesis. What makes the global test particularly attractive is that it is a score test, which means that it does not require estimation of the parameters under the alternative hypothesis. Furthermore, the test statistic, which is a quadratic form in residuals under the null hypothesis, features an $n \times n$ matrix, instead of a potentially much larger $p \times p$ matrix, see below for details.

Before we have a look at the particular form that the global test takes on in our setting, we make one observation. The hypothesis H_1 concerns a p-dimensional parameter in the model (1) with a single explanatory variable X and a multivariate response $\boldsymbol{M}$. In order to avoid a challenging task of estimating a large covariance matrix Σ, as well as making any assumptions about its structure, we propose to test the hypothesis H_1^* instead. Namely, if the exposure X is continuous and normally distributed in the population of interest, under the assumptions stated in Section 2, the hypothesis H_1 is equivalent to

$$H_1^* : \tilde{\boldsymbol{\alpha}} = 0,$$

where $\tilde{\boldsymbol{\alpha}}$ is a vector of regression coefficients of $\boldsymbol{M}$ in the model in which the roles of the explanatory and response variables are reversed

$$X = \tilde{\alpha}_0 + \boldsymbol{M}^\top \tilde{\boldsymbol{\alpha}} + \epsilon_X, \tag{4}$$

where $\epsilon_X \sim \mathsf{N}(0, \sigma_X^2)$ is independent of $\boldsymbol{M}$. We are now in the standard context with a scalar response and a high-dimensional vector of explanatory variables. It should be stressed that, as opposed to model (1), model (4) does not represent the assumed data generating process and is used only as a means for testing the hypothesis of interest. In particular, we do not interpret the parameter $\tilde{\alpha}$. The assumption of normality of X is not overly restrictive since an analogous procedure can be applied whenever the conditional mean of X given $\boldsymbol{M}$ belongs to the family of generalized linear models.

Let us now assume that n independent triplets are observed: $(X_1, \boldsymbol{M}_1, Y_1), \ldots, (X_n, \boldsymbol{M}_n, Y_n)$. Let $\boldsymbol{X}_{n\times 1} = (X_1, \ldots, X_n)^\top$, $\boldsymbol{M}_{n\times p} = (\boldsymbol{M}_1 \ldots, \boldsymbol{M}_n)^\top$, and $\boldsymbol{Y}_{n\times 1} = (Y_1, \ldots, Y_n)^\top$. Note that now, $\boldsymbol{M}$ stands for the $n \times p$ matrix, instead of a p-variate vector, with rows corresponding to statistical units, and columns corresponding to potential mediators. The statistic for the global test of H_1^* is

$$Q_X = n^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X)^\top \boldsymbol{M}\boldsymbol{M}^\top (\boldsymbol{X} - \boldsymbol{\mu}_X),$$

where $\boldsymbol{\mu}_X = \tilde{\alpha}_0 \mathbf{1}_{n\times 1}$ is the mean of $\boldsymbol{X}$ under the null hypothesis H_1^*.

Analogously, the test statistic for H_2 is

$$Q_Y = n^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_Y)^\top \boldsymbol{M}\boldsymbol{M}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_Y),$$

where $\boldsymbol{\mu}_Y = g(\beta_0 \mathbf{1}_{n\times 1} + \gamma \boldsymbol{X})$ is the expected value of $\boldsymbol{Y}$ under the null hypothesis, and $g(\boldsymbol{a})$ for $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ denotes $[g(a_1), \ldots, g(a_n)]^\top$.

In order to be able to use the statistics Q_X and Q_Y, a few adjustments pertaining to nuisance parameters are called for. First, the expressions for Q_X and Q_Y involve parameters $\tilde{\alpha}_0$, β_0, and γ, which are usually unknown and need to be estimated from the data. Secondly, depending on the type of the distribution of X and Y, distributions of the associated test statistics might further depend on unknown dispersion parameters. Let us for illustrational purposes assume that the model for X is linear, while the model for Y is logistic. To obtain the test statistic whose null distribution does not depend on σ_X^2, Goeman et al.[16] scaled the original test statistic by plugging in the maximum likelihood estimate of σ_X^2, so that

$$\hat{Q}_X = \frac{(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_X)^\top \boldsymbol{M}\boldsymbol{M}^\top (\boldsymbol{X} - \hat{\boldsymbol{\mu}}_X)}{n\hat{\sigma}_X^2},$$

is independent of σ_X^2, where $\hat{\sigma}_X^2$ is the maximum likelihood estimate of σ_X^2 under H_1^*. The test statistic $\hat{Q}_X$ is a ratio of dependent quadratic forms in normal random variables[16], and although its exact null distribution is not available in closed form, fast algorithms for computing associated probabilities are available, and already implemented in the `R` package `globaltest`.

The expression for Q_Y contains $\boldsymbol{\mu}_Y$, the expected value of $\boldsymbol{Y}$ under the null hypothesis. Furthermore, the distribution of Q_Y depends on the diagonal covariance matrix of $\boldsymbol{Y}$ under the null hypothesis given by $\operatorname{diag}\{\boldsymbol{\mu}_Y * (1 - \boldsymbol{\mu}_Y)\}$, where $*$ denotes element by element multiplication. If we plug in the estimate of $\boldsymbol{\mu}_Y$ and normalize the test statistic so that it does not depend on the unknown covariance matrix, we obtain the approximate pivot $\hat{Q}_Y$. More precisely, the test statistic $\hat{Q}_Y$ is defined as

$$\hat{Q}_Y = \frac{(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_Y)^\top \boldsymbol{M}\boldsymbol{M}^\top (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_Y)}{(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_Y)^\top \boldsymbol{D} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_Y)},$$

where $\boldsymbol{D}$ is an $n \times n$ diagonal matrix equal to $\boldsymbol{M}\boldsymbol{M}^\top$ at the diagonal, and zero otherwise. In this case, the distribution of the $\hat{Q}_Y$ is asymptotically that of a ratio of two dependent quadratic forms in normal random variables. The quality of the approximation offered by this asymptotic result hinges on the strong eigenvalue structure of $\boldsymbol{M}\boldsymbol{M}^\top$: when it is weak and $\boldsymbol{M}\boldsymbol{M}^\top$ is close to a diagonal matrix, the approximation will be poor. In such cases, it is recommended to rely on permutation tests, if possible. We refer to Goeman et al[16] for further details.

3.2 | Remarks

Covariates. We have so far assumed the absence of confounders between exposure and mediator, exposure and outcome, and mediator and outcome. The described procedure can however be easily adjusted for covariates; it is sufficient to include them in models (2) and (4), and adjust the expected means $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ under H_1^* and H_2 accordingly. Note that while the dimension of the vector of potential mediators p is allowed to exceed n, the number of the so-called *nuisance* covariates is assumed such that the estimation of the associated null models is feasible.

Power. The power of the global test for mediation depends on the power of the global test and the power of the intersection union procedure.

The global test is locally most powerful on average in the neighbourhood of the null hypothesis. Since the alternative hypothesis is high-dimensional, there is no hope in finding a test which would have optimal power against *all* alternatives. Therefore, a test that is optimal on average will still have many alternatives against which it has zero power. The global test thus shifts power from the alternatives deemed less interesting to alternatives in more interesting directions. These directions are given by the large variance principal components of $\boldsymbol{M}$. Therefore, the global test will usually have high power in detecting alternatives for which large variance principal components explain most of the variability of the response. On the contrary, it will have very low power in detecting alternatives in which low variance principal components dominate the relation with the response. Fortunately, in biological applications the first scenario is more plausible, since low variance components usually reflect noisy measurements rather than a biological signal. For a more detailed discussion of the power of the global test, its comparison to the F-test and the approach based on the principal components, see Goeman et al.[10]

The intersection union principle offers simple and intuitive solutions for testing union hypotheses. The union hypothesis is composite, and the size of the intersection union test will be different for different points of the null hypothesis. The test is exact when either component hypothesis, H_1 or H_2, is sufficiently far from the null, but is conservative otherwise. Berger[17] proposed a method for constructing a more powerful test of a specified size. One could invert this test to obtain the associated p-value. Motivated by applications where many "no mediation" hypotheses are considered at the same time, we follow a different strategy. In Section 4, we propose a multiple testing procedure that attenuates the power issue of the intersection union test when many union hypotheses are tested simultaneously.

Issues related to high-dimensional mediation. As stated at the beginning of this section, we test whether $\boldsymbol{M}$ is a potential mediator candidate. This is different from testing whether any component of $\boldsymbol{M}$ mediates the effect of X on Y. Consider the example in Figure 2 showing three instances of a disjunctive effect, i.e. some of the mediators are associated with either the exposure or the outcome, but not both. In all three cases, M_1 is associated with X, M_2 is associated with Y and thus hypothesis (3) is false. Nevertheless, only in the third instance (right panel) $\boldsymbol{M}$ mediates the effect of X on Y. This illustrates that a) rejecting (3) does not imply that any M_i, $i = 1, \dots, p$, is a true mediator, and b) the causal structure within $\boldsymbol{M}$ plays a critical role. In our approach, we make no assumptions regarding this structure since such information is typically unavailable in high-dimensional applications, which means that we are unable to distinguish between these cases. Given this limitation, the global test for mediation is better suited for the exploratory, as opposed to confirmatory, analysis. Rejecting hypothesis (3) should be interpreted as an indication that $\boldsymbol{M}$ contains potential mediators that deserve further investigation.

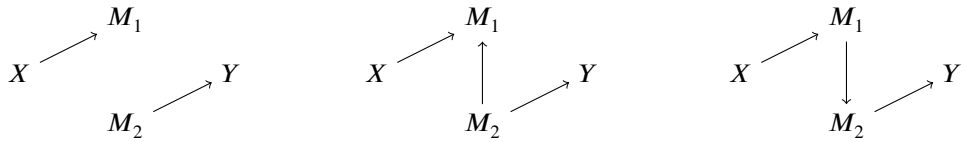

FIGURE 2 Three instances of a disjunctive effect.

Implementation. Performing the global test for mediation in `R` is straightforward with the package `globaltest`. It requires specifying the models for X and Y and performing two separate global tests, whose results are then combined according to Procedure 1.

4 | TESTING MULTIPLE GROUPS OF POTENTIAL MEDIATORS

Consider m disjoint groups of potential mediators and let $\mathcal{H} = \{H_1, \dots, H_m\}$ be a collection of "no mediation" hypotheses (3), where for each $i = 1, \dots, m$, the hypothesis H_i is given by the union of two component hypotheses: $H_i = H_{i1} \cup H_{i2}$. We are interested in controlling the familywise error rate (FWER) for $\mathcal{H}$. We first describe a simple procedure building on Procedure 1, and then propose a modified two step procedure, that exploits independence of the statistics used for testing H_{i1} and H_{i2} (see

TABLE 1 The two step approach to testing m union hypotheses.

	p-value matrix	$\min p$	$\max p$
H_1	$p_{11}\ p_{12}$	$\min\left\{p_{11}, p_{12}\right\}$	$\max\left\{p_{11}, p_{12}\right\}$
$\vdots$	$\vdots\ \vdots$	$\vdots$	$\vdots$
H_m	$p_{m1}\ p_{m2}$	$\min\left\{p_{m1}, p_{m2}\right\}$	$\max\left\{p_{m1}, p_{m2}\right\}$

the Appendix).

Single step procedure. Let p_{ij} be a p-value of H_{ij}, $i = 1, \dots, m$ and $j = 1, 2$. According to the intersection-union principle, the p-value of H_i is $\max p_i = \max\left\{p_{i1}, p_{i2}\right\}$. A straightforward solution to the problem of multiple testing is to adjust these p-values. The simplest solution is offered by the Bonferroni or Holm[18] correction; however, this can result in a conservative procedure when the number of false hypotheses is small[19], especially in combination with the intersection union test.

Two step procedure: ScreenMin. We propose a modification of the above procedure that attenuates its conservativeness by exploiting the independence of p_{i1} and p_{i2} under the model described in Section 2 (see the Appendix). The key is to introduce an extra step and consider the minimum p-value for each row (Table 1). These per-row minimal p-values are used to screen hypotheses and discard those that we know we will be unable to reject even before seeing the relevant p-value. In this way, we lessen the burden of multiple testing. We call the modified procedure ScreenMin.

Procedure 2 (ScreenMin)**.**

Step 1. Consider a set of per-row minimal p-values: $\left\{\min p_1, \dots, \min p_m\right\}$, where $\min p_i = \min\left\{p_{i1}, p_{i2}\right\}$. For a given $c \in (0, 1)$, let $\mathcal{S} = \left\{i : \min p_i \leq c\right\}$ be an index set of selected hypotheses.

Step 2. Consider a subset of per row maximum p-values $\left\{\max p_i : i \in \mathcal{S}\right\}$, and correct them for multiplicity by the Bonferroni or Holm correction. Denote these adjusted p-values $\max p_i^*$. Set the adjusted p-value of H_i as

$$p_i^* = \begin{cases} \max p_i^* & i \in \mathcal{S} \\ 1 & i \notin \mathcal{S}. \end{cases}$$

Theorem 1. When p-values $\left\{p_{ij}, i = 1, \dots, m, j = 1, 2\right\}$ are jointly independent, the ScreenMin procedure provides asymptotic control of FWER for the family $\mathcal{H} = \left\{H_1, \dots, H_m\right\}$.

The proof is given in the Appendix.

Remarks.

- Theorem 1 assumes joint independence of the $2m$ component p-values. In the context of mediation analysis, independence of p_{i1} and p_{i2}, i.e. independence within rows of the $m \times 2$ p-value matrix (see Table 1) follows from the specification of the outcome (2) and mediator model (4). On the other hand, independence between rows, i.e. within sets $\left\{p_{11}, \dots, p_{m1}\right\}$ and $\left\{p_{12}, \dots, p_{m2}\right\}$, concerns test statistics associated to different groups of potential mediators, and in practice might or might not hold. Nevertheless, independence between rows is only a sufficient condition for the validity of the proposed procedure, and preliminary empirical results obtained in Simulation study 3 show that type I error control is maintained even under strong positive dependence (supplementary material). Further investigation of the error control in the settings in which row independence is not satisfied is left for future work.

- In practice one needs to decide on an appropriate threshold c to be used in Step 1. To control the FWER at level α, one can set $c = \alpha/m$ in Step 1, and use the Bonferroni or Holm correction in Step 2, which leads to a uniformly more powerful procedure with respect to the single step procedure.

- In many applications involving a large number of hypotheses FWER is too stringent a criterion, and one is interested in controlling the false discovery rate (FDR) instead. ScreenMin can be easily adapted: it is sufficient to set $c = \alpha/m$ in Step 1, and use an FDR procedure, such as Benjamini and Hochberg,[20] in Step 2. This will result in asymptotic control of the FDR.

- The ScreenMin procedure is not restricted to a global test or any specific model; it is appropriate whenever one is considering a collection of union hypotheses such that the two p-values associated to each of them are independent.

5 | SIMULATION STUDIES

We conducted three simulation studies to evaluate the performance of the global test for mediation (Section 3) and the ScreenMin procedure (Section 4). The first two studies are concerned with the global test, while the third one looks at the proposed multiple testing procedure.

5.1 | Simulation study 1

The first simulation study reproduces the settings considered in Huang and Pan,[13] so the number of potential mediators was set to $p = 50$, and two sample sizes were considered: $n = 50$ and $n = 500$. The exposure variable X was drawn randomly from the set $\{1, 2, 3\}$. The parameter $\boldsymbol{\alpha}$ in model (1) was set to $a\mathbf{1}_{p\times 1}$, where $a \in [0, 0.15]$ for $n = 500$, and $a \in [0, 0.4]$ for $n = 50$. The matrix Σ has a compound symmetry structure: its diagonal is set to 1, and all off-diagonal elements equal $\rho = 0.3$. The model considered by Huang and Pan[13] allows for the interaction between the exposure and mediators, so that the outcome model becomes

$$Y = \beta_0 + \boldsymbol{M}^\top \boldsymbol{\beta} + \gamma X + X\boldsymbol{M}^\top \boldsymbol{\beta}_C + \epsilon_Y.$$

With the variance of the error term set to 1, this model was used to simulate 1000 observations of Y for each considered setting: a range of values for a, and 6 different configurations of the parameter $(\boldsymbol{\beta}, \boldsymbol{\beta}_C)$ as shown in Figure 3. These parameter configurations were chosen so that different situations are explored: nonsparse mediation effects in the first and third row; sparse mediation effects in the second row; all positive effects in the first column, and mixed positive and negative effects in the second column (the so-called cancellation effect).

We compared our approach with the τ_p test,[13] referred to as the HP test in the following, in terms of power and type I error control. Results are shown in Figure 3. We observe that the global test controls Type I error in all settings considered in this study quite accurately, while the HP test tends to be conservative in some settings (the second column, rows two and three). The power is similar for the two approaches, especially for the settings with all mediation effects of the same sign (the first column), while the global test for mediation performs slightly better in settings with mixed positive and negative effects (the second column). Note that the good performance of the global test is particularly encouraging since data are simulated from an outcome model different from the one used to derive the global test.

We also investigated a special case, the so-called disjunctive effect,[13] in which each of the $p = 50$ mediators is associated with either the exposure or the outcome, but not both. We set the first 25 elements of $\boldsymbol{\alpha}$ and the last 25 elements of $\boldsymbol{\beta}$ to 0.15, and $\boldsymbol{\beta_C}$ to a null vector. The exposure, Σ, and the variance of the error term in the outcome model were set as above. For $n = 50$, the power of the global test for $\alpha = 0.05$ is 0.85, and the power of the HP test is 0.29. In terms of mediation analysis, disjunctive effect may be a false positive or a true mediation effect, depending on the casual structure within $\boldsymbol{M}$ (see the discussion in Section 3.2). In particular, when Σ is a diagonal matrix, and components of $\boldsymbol{M}$ are independent, we know that a disjunctive effect is not a mediation effect. In that case, HP is under the null hypothesis and controls Type I error (obtained level for $n = 50$ is < 0.01) correctly concluding that $\boldsymbol{M}$ is not a mediator. On the other hand, hypothesis (3) is false, and the global test for mediation rejects it with probability 0.45.

5.2 | Simulation study 2

In the second study, instead of employing a structured covariance matrix, we assumed a more realistic structure of dependence between potential mediators by using the gene expression dataset `chimera` available in the R package `simPATHy`. We set $p = 50$ as before and used all the 78 observations of the first 50 genes in this dataset to estimate the $p \times p$ covariance matrix of genes, which was subsequently used as Σ in the model (1). The vector $\boldsymbol{\beta}$ was chosen so that large principal components of $\boldsymbol{M}$ have large regression coefficients. We wrote Σ in a singular value decomposition as $\Sigma = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix, and $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix with diagonal elements $\lambda = \left(\lambda_1, \ldots, \lambda_p\right)^\top$, and then set the vector of regression coefficients as

$$\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1/2}\lambda.$$

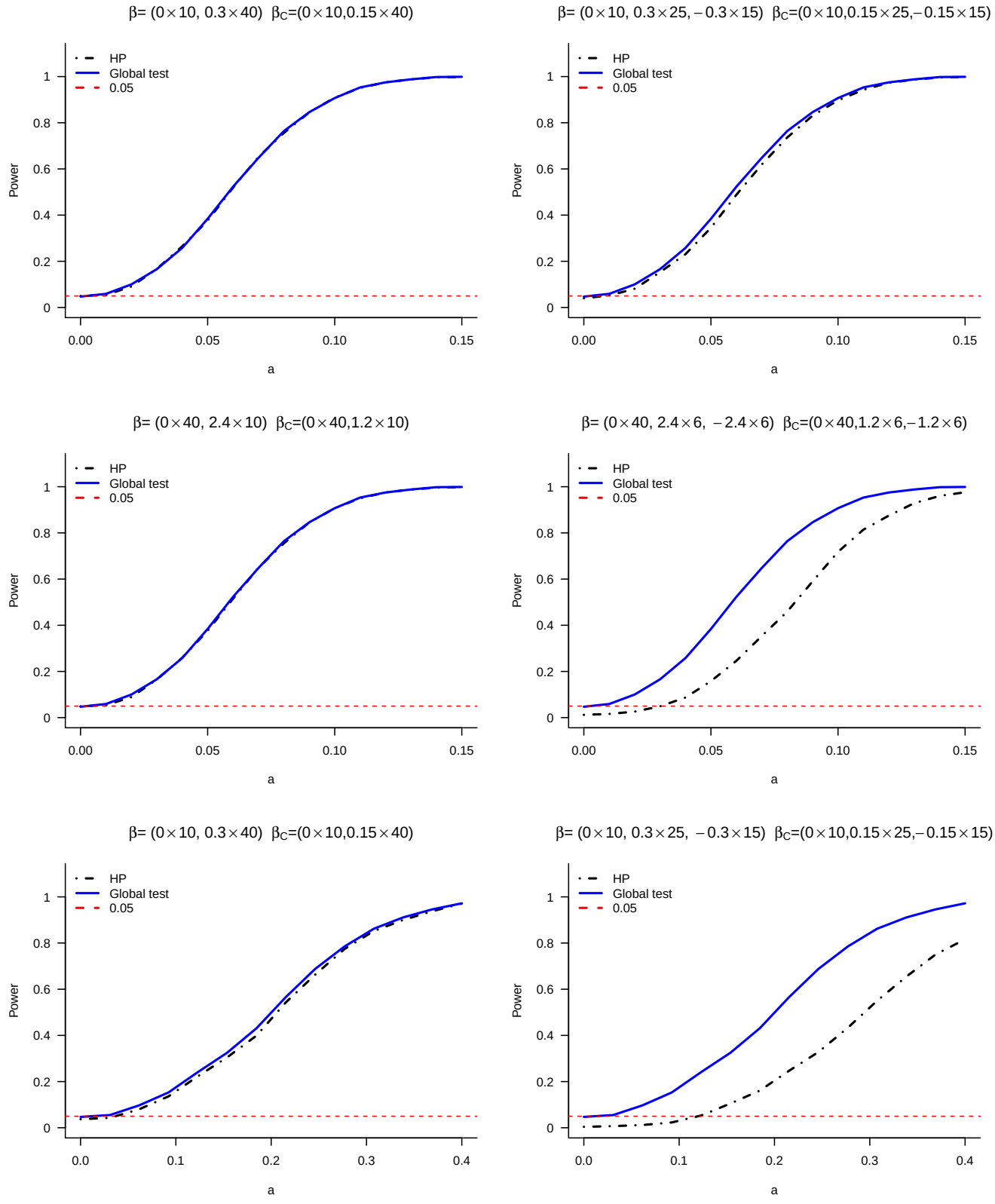

FIGURE 3 Power of the global test for mediation and the HP test in six different scenarios given by parameter configurations. $(0 \times 10, 0.3 \times 40)$ is to be read as: the first 10 elements equal 0, and the remaining 40 elements equal 0.3. The top 2 rows show power for $n = 500$, the bottom row for $n = 50$.

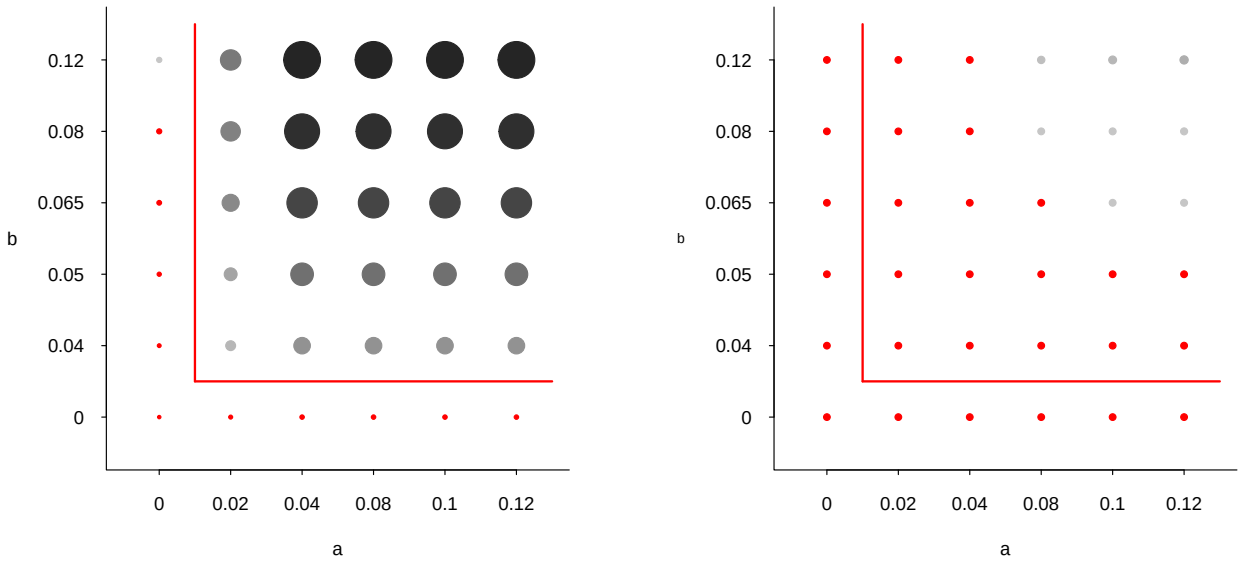

FIGURE 4 Power of the global test for mediation (left) and the HP test (right) for different values of parameters a and b whose meaning is explained in the text. Parameter pairs (a, b) below and to the left from the red lines belong to the null hypothesis. The size of the circle for each parameter pair is proportional to the estimated power based on 1000 Monte Carlo runs, i.e., the number of times the null hypothesis was rejected at 5% level. Red circles denote power, or Type 1 error, lower than 5%.

In this way the i-th principal component of $\boldsymbol{M}$ has a regression coefficient $\lambda_i^{1/2}$. We varied the signal strength via a scalar $b \in \{0, 0.04, 0.05, 0.065, 0.08, 0.12\}$ that multiplied the parameter $\boldsymbol{\beta}$. The exposure variable X was drawn randomly from a normal distribution with zero mean and variance 4. The first 40 elements of $\boldsymbol{\alpha}$ in model (1) were set to a, with six different values of a: $a \in \{0, 0.02, 0.04, 0.08, 0.1, 0.12\}$. The remaining 10 elements of $\boldsymbol{\alpha}$ were set to zero. Note that the choices $a = 0$ and $b = 0$ allow us to study behavior under the null hypothesis. The parameter γ in model (2) was set to 0.02. For each combination of a and b we simulated $B = 1000$ samples of size $n = 200$ of $(\boldsymbol{M}, Y)$. We computed power, or Type I error when under the null hypothesis, as the number of draws in which the null hypothesis was rejected at 5% level. The results are shown in Figure 4. Both approaches control type I error as before, but the global test for mediation clearly outperforms the HP test in terms of power under the considered alternatives.

5.3 | Simulation study 3

The ScreenMin procedure can be applied whenever the two components of each union hypothesis are independent (or better, the corresponding test statistics are independent). The method is thus not restricted to the specific problem given in (3). For illustration purposes, we consider a simpler setting giving rise to a collection of union hypotheses. We consider two random vectors $X^{(1)} \sim \mathsf{N}_m(\boldsymbol{\mu}^{(1)}, \Sigma)$ and $X^{(2)} \sim \mathsf{N}_m(\boldsymbol{\mu}^{(2)}, \Sigma)$, with $m = 50$, and a collection $\mathcal{H} = \{H_i;\, i = 1, \ldots, m\}$, where $H_i = H_{i1} \cup H_{i2}$, and $H_{ij} : \mu_i^{(j)} = 0$, $i = 1, \ldots, m$; $j = 1, 2$. Each H_{ij} is tested against a one-sided alternative: $\mu_i^{(j)} > 0$. For H_{ij}, a p-value is obtained as $1 - \Phi(X_i^{(j)})$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

We consider different configurations for the parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$. Three different levels of effect sparsity are emulated by setting the number of non-zero elements of $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ to 5, 25, and 50. This corresponds to sparse, medium, and rich effects, respectively. Similarly, we consider three different effect sizes. When H_{ij} is true, then $\mu_i^{(j)} = 0$. Otherwise, the mean is shifted so that the probability of rejecting H_{ij} at level 0.05 is either 0.8, 0.9 or 0.99, corresponding to the weak, medium and strong effect, respectively. The associated probability of rejecting the false union hypothesis is then 0.64, 0.81 and 0.98, respectively. The matrix Σ is either an identity matrix (results reported here) or a compound symmetry matrix with all off-diagonal elements equal to 0.5 (supplementary material).

For every combination of sparsity and effect size, we simulated $N = 10000$ pairs $(X^{(1)}, X^{(2)})$ and used them to estimate the underlying FWER and power of the ScreenMin procedure. The FWER is computed as the proportion of simulated pairs in which at least one true union hypothesis is rejected. The power, computed for each simulated pair as the proportion of rejected false hypotheses among all false hypotheses, is averaged over 10000 simulations.

TABLE 2 Estimated FWER (multiplied by 100) at $\alpha = 0.05$ for the four methods based on $N = 10000$ simulated datasets. The number of hypotheses is $m = 50$. For the false hypotheses, three levels of effect size – weak, medium, and strong – correspond to the power 0.64, 0.81, and 0.98, respectively, to reject a false union hypothesis. Sparsity pattern (a, b) is to be read as: hypotheses $H_{11}, \dots, H_{1a}$ and $H_{21}, \dots, H_{2b}$ are false; $H_{1(a+1)}, \dots, H_{1m}$ and $H_{2(b+1)}, \dots, H_{2m}$ are true.

Effect	Sparsity pattern	Procedures		
		single step	ScreenMin	oracle
weak	(50,50)	< 0.1	< 0.1	< 0.1
	(25,25)	< 0.1	< 0.1	< 0.1
	(5,5)	< 0.1	0.1	< 0.1
	(50,25)	0.6	2.0	1.6
	(50,5)	1.3	4.6	1.5
	(25,5)	0.5	4.0	0.6
medium	(50,50)	< 0.1	< 0.1	< 0.1
	(25,25)	< 0.1	< 0.1	< 0.1
	(5,5)	< 0.1	0.1	< 0.1
	(50,25)	1.2	2.2	2.6
	(50,5)	2.1	4.5	2.4
	(25,5)	0.9	3.7	1.0
strong	(50,50)	< 0.1	< 0.1	< 0.1
	(25,25)	< 0.1	< 0.1	< 0.1
	(5,5)	< 0.1	0.1	< 0.1
	(50,25)	2.3	2.6	4.4
	(50,5)	3.5	4.3	4.0
	(25,5)	1.5	3.8	1.7

We compared our procedure with a) the single step procedure that considers $\max p_i, i = 1, \dots, m$ and corrects these with a Bonferroni correction; b) an oracle procedure that uses m_0, the number of true hypotheses in $\mathcal{H}$, instead of m, as a Bonferroni correction factor.

The results are reported in Tables 2 and 3. Results regarding the estimated FWER in Table 2 show that all considered methods are conservative, and only ScreenMin approaches the nominal level ($\alpha = 0.05$) for a specific sparsity pattern (5,5). This behavior can be explained by the inherent conservativeness of the intersection union tests.

When it comes to power, Table 3 offers a few insights. First, while the sparsity pattern has no impact on the single step procedure, it plays an important role in the performance of the remaining methods. ScreenMin achieves higher power in sparse settings, such as $(5, 5)$ or $(25, 5)$, as opposed to $(50, 50)$. This is due to the fact that in less sparse settings with many false null hypotheses, the beneficial impact of screening is limited. On the other hand, the performance of oracle improves in less sparse settings (compare $(25, 25)$ with $(5, 5)$), but the improvement diminishes with increasing effect size.

Second, if we compare the three methods for a fixed sparsity pattern and effect size, we see that ScreenMin exhibits the best performance. It is not difficult to show that ScreenMin is uniformly more powerful than the single step procedure, so that comparison is not surprising. The comparison between ScreenMin and oracle shows that the information on $\min p_i$ used in the

screening is more helpful than the information provided by m_0 – the number of true union hypotheses – especially when the effect is weak. The difference between the two diminishes with increasing effect size.

TABLE 3 Power estimates (multiplied by 100) at $\alpha = 0.05$ for the four methods based on $N = 10000$ simulated datasets. The number of hypotheses is $m = 50$. For the false hypotheses, three effect sizes – weak, medium, and strong – correspond to the power 0.64, 0.81, and 0.98, respectively, to reject a false union hypothesis. Sparsity pattern (a, b) is to be read as: hypotheses $H_{11}, \dots, H_{1a}$ and $H_{21}, \dots, H_{2b}$ are false; $H_{1(a+1)}, \dots, H_{1m}$ and $H_{2(b+1)}, \dots, H_{2m}$ are true.

Effect	Sparsity pattern	Procedures		
		single step	ScreenMin	oracle
	(50,50)	7.5	11.9	–
	(25,25)	7.5	16.6	12.2
	(5,5)	7.1	27.8	7.7
weak	(50,25)	7.5	13.4	12.1
	(50,5)	7.6	15.0	8.2
	(25,5)	7.5	19.6	8.1
	(50,50)	18.9	22.9	–
	(25,25)	18.8	30.3	26.8
	(5,5)	18.9	47.4	20.0
medium	(50,25)	18.8	25.0	26.9
	(50,5)	18.6	26.7	19.7
	(25,5)	19.1	34.4	20.2
	(50,50)	65.6	66.1	–
	(25,25)	65.8	74.6	74.5
	(5,5)	65.9	88.5	67.3
strong	(50,25)	65.8	67.3	74.6
	(50,5)	65.8	68.2	67.2
	(25,5)	65.6	76.1	66.9

6 | DATA APPLICATION: SMOKING, DNA METHYLATION, AND LUNG CANCER RISK

Smoking is the major risk factor of lung cancer.[21,22] Currently, the interest is in understanding the biological mechanisms underlying this relationship. One of the working hypothesis is that the tobacco exposure alters DNA methylation patterns over time, which in turn affect individual's lung cancer risk. We test this hypothesis with data from the Norwegian Woman and Cancer (NOWAC) prospective cohort study (`https://site.uit.no/nowac/`).[23] Data are not publicly available due to third party restrictions.

Our data consist of 125 case-control pairs matched by time since blood sampling and year of birth, identified in the NOWAC cohort. Smoking was coded as "Never", "Former", and "Current" smoker. Methylation levels were measured in peripheral blood samples, on average 3.88 years prior to diagnosis (range: 0.29 to 7.92 years).[24] Potential mediators are groups of CpG sites whose differential methylation is hypothesized to mediate the effect of smoking on lung cancer. We considered a subset of measured CpG sites associated to smoking related genes. According to a recent systematic review,[25] we selected unique gene identifiers from the collection of 151 CpG sites reported in two or more studies. These CpG sites map to 72 genes, and for each of these genes we grouped all the associated CpG sites. The average number of CpGs per gene/group was 45. A multinomial model was considered for smoking in (4), while a conditional logistic model was used for lung cancer, and both were adjusted for potential

TABLE 4 Lung cancer study: p-values for the seven genes selected by the ScreenMin procedure. In the first column, p-values for the hypotheses of no association between a given gene and the smoking status; in the second column, p-values for the hypotheses of no association between a given gene and the lung cancer status.

Gene	p_1	p_2
F2RL3	5.48×10^{-5}	5.35×10^{-1}
AHRR	1.76×10^{-4}	5.68×10^{-1}
GFI1	5.72×10^{-6}	4.24×10^{-1}
MYO1G	6.61×10^{-6}	4.84×10^{-1}
ITGAL	1.72×10^{-6}	3.41×10^{-1}
VARS	1.61×10^{-5}	8.97×10^{-1}
CLDND1	2.37×10^{-4}	9.89×10^{-1}

confounders: age, time since blood sampling, and white blood cell composition.[26] Since data come from a case-control study and the prevalence of lung cancer is very low, we used only controls to test the association between smoking and methylation in the multinomial model.[11, p.28] We applied the ScreenMin procedure to this collection of 72 hypotheses of no mediation. Setting a threshold $c = 0.05/72$ in Step 1 resulted in selecting seven hypotheses to be tested in Step 2. However, the results, reported in Table 4, show that while the associations between smoking and methylation seem strong, there is no evidence of any association between methylation and lung cancer status in the outcome model. Our results hence do not support the hypothesis that smoking-induced methylation changes mediate the effect of smoking on lung cancer risk. Similar conclusions were reached in a recent two step Mendelian randomization analysis.[27]

The obtained results are considered exploratory and are to be interpreted with caution. Variables such as occupational exposure or family history of lung cancer might confound the relations between smoking, DNA methylation and lung cancer.[21] Another issue of concern in observational studies is reverse causation. Our study is prospective, but samples close to diagnosis might be influenced by disease onset. Finally, in this illustrative example we do not address the problem of measurement error. Information about the extent of environmental exposure (smoking) comes from self-administered questionnaires and is thus likely to suffer from well-known sources of bias. On the other hand, complex processes involved in measuring molecular markers introduce the problem of technical variability. Some work addressing the problem of measurement error is already available[28,29]; however, this important issue certainly deserves further study.

7 | DISCUSSION

In this work, we have addressed the problem of testing whether a potentially high-dimensional vector acts as a mediator between some exposure variable and an outcome of interest. The global test for mediation was motivated by a growing need to assess the presence of mediation in high-dimensional settings, such as genomics, epigenomics, genetic epidemiology and neuroscience. The proposed method can, for instance, be applied to test whether an established effect of the exposure on the outcome is mediated by a group of functionally related genes, a group of neighbouring CpG sites, or a group of neighbouring voxels in the brain.

We have considered a joint significance test. An alternative approach for testing mediation is the so-called *product significance* test based on $\hat{\boldsymbol{\alpha}}^\top \hat{\boldsymbol{\beta}}$. The downside of the product method is the complexity of the null distribution which needs to be approximated either analytically,[30,31] or by bootstrapping.[32] Analytical approximations can be unsatisfactory when the sample size is small, while bootstrap can be computationally expensive in high-dimensional settings. Furthermore, a large empirical study comparing different methods for testing mediation[31] indicated that the joint test has the best overall performance. This finding has been confirmed within the context of multiple mediators[33] and Huang[34] provided a first theoretical insight into this result.

We have assumed the association between M and X (and M and Y) is linear (on a logit scale if the outcome is binary). In some situations, it might be appropriate to consider non-linear effects and replace (1) and (2) with semi-parametric models. The hypotheses of no-association H_1 and H_2 can then be modified accordingly and tested by tests proposed by Liu et al[35] and Wu et al[36] in continuous and binary models, respectively. Following that, the intersection union test is performed without any change.

When many union hypotheses of no mediation are tested by a joint test, controlling type I error by applying standard multiple comparison procedures to $\{\max p_i, i = 1, \ldots, m\}$ can be conservative. This is especially the case if both component null hypotheses are true for many union hypotheses,[33,34] which seems to be a plausible scenario in genomics applications. To attenuate this conservativeness, one can try to exploit a particular feature of the mediation setting, i.e. independence of the p-values for the hypotheses in the mediator and outcome model.[37] Following this reasoning, the proposed ScreenMin procedure reduces the burden of multiple testing by filtering out unpromising hypotheses. ScreenMin is shown to maintain type I error control under independence of p-values pertaining to different union hypotheses (independence between rows of the p-value matrix), while empirical results indicate error is also controlled under positive dependence. On the other hand, the proposed method relies heavily on the independence between the columns of the p-value matrix, which may be violated in applications with unobserved confounding factors. The threshold for selection c in Step 1 plays an important role, and we have proposed a value that leads to a uniformly more powerful procedure with respect to a single step Bonferroni procedure. An interesting question is whether it is possible to further improve upon the proposed threshold.

When the hypothesis (3) is rejected, we cannot claim that any component of $\boldsymbol{M}$ is mediating the effect of X on Y. Whether this is the case depends also on the relations among the components of $\boldsymbol{M}$. Therefore, the main value of the proposed approach is in identifying promising groups of potential mediators. This is especially useful in situations with a very high number of potential mediators that allow for meaningful grouping (i.e. according to function or location). To go beyond the step of screening for potential mediators and actually estimate mediation effects in a high-dimensional settings, more work is still needed, e.g. on causal assumptions, the impact of exposure-mediator interactions, and sensitivity analysis.[11]

Once an interesting group of potential mediators is identified, the question of interest is which components of $\boldsymbol{M}$ are responsible for the rejection. Zhao and Luo[38] have recently proposed a lasso type approach for structural equations model selection that aims to identify significant pathways between X and Y. Related to this, also recently, Song et al.[39] proposed a global test of mediation where they utilize Bayesian variable selection to identify active mediators. An alternative solution we are currently pursuing is to estimate parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in models (1) and (2), and test significance of all components of $\boldsymbol{M}$ simultaneously.

We believe that the proposed global test for mediation and the ScreenMin procedure can find their place in the exploratory analysis of high-dimensional data. By downsizing the pool of potential mediators and identifying those groups that deserve to be studied further, they can be seen as tools for generating new hypotheses to be investigated with more refined causal inference methods.

References

1. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015; 6: 10192. doi: 10.1038/ncomms10192

2. Chen R, Meng X, Zhao A, et al. DNA hypomethylation and its mediation in the effects of fine particulate air pollution on cardiovascular biomarkers: A randomized crossover trial. *Environ Int* 2016; 94: 614–619.

3. Houtepen LC, Vinkers CH, Carrillo-Roa T, et al. Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nat Commun* 2016; 7: 10967. doi: 10.1038/ncomms10967

4. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* 2017; 19(2): 121–136.

5. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 2008; 40(3): 879–891.

6. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods* 2014; 2(1): 95–115.

7. Daniel R, De Stavola B, Cousens S, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics* 2015; 71(1): 1–14.

8. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology* 2017; 28(2): 258–265.

9. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; 20(1): 93–99.

10. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc Series B Stat Methodol* 2006; 68(3): 477–493.

11. VanderWeele T. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press . 2015.

12. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* 2016; 32(20): 3150–3154.

13. Huang YT, Pan WC. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* 2016; 72(2): 402–413.

14. Gleser L. On a Theory of Intersection Union Tests. *Institute of Mathematical Statistics Bulletin* 1973; 2(233): 9.

15. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 1982; 24(4): 295–300.

16. Goeman JJ, Van Houwelingen HC, Finos L. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika* 2011; 98(2): 381–390.

17. Berger RL. Likelihood ratio tests and intersection-union tests. In: Birkhauser. 1997 (pp. 225–237).

18. Holm S. A simple sequentially rejective multiple test procedure. *Scand Stat Theory Appl* 1979; 6(2): 65–70.

19. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med* 2014; 33(11): 1946–1978.

20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 1995; 57(1): 289–300.

21. Spitz MR, Wu X, Wilkinson A, Wei Q. Cancer of the lung. In: Schottenfeld D, Fraumeni Jr JF., eds. *Cancer epidemiology and prevention.*Oxford University Press. 2006 (pp. 638-658).

22. Vineis P, Alavanja M, Buffler P, et al. Tobacco and cancer: recent epidemiological evidence. *J Natl Cancer Inst* 2004; 96(2): 99–106.

23. Lund E, Dumeaux V, Braaten T, et al. Cohort Profile: The Norwegian Women and Cancer Study – NOWAC – Kvinner og kreft. *Int J Epidemiol* 2008; 37(1): 36-41.

24. Sandanger TM, Nøst TH, Guida F, et al. DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Sci Rep* 2018; 8(1): 16714. doi: 10.1038/s41598-018-34334-6

25. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics* 2015; 7(1): 113. doi: 10.1186/s13148-015-0148-3

26. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 2012; 13(1): 86. doi: 10.1186/1471-2105-13-86

27. Battram T, Richmond R, Baglietto L, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. *bioRxiv* 2018: 287888. doi: 10.1101/287888

28. Valeri L, Lin X, VanderWeele TJ. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Stat Med* 2014; 33(28): 4875–4890.

29. Valeri L, Reese SL, Zhao S, et al. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight?. *Epigenomics* 2017; 9(3): 253-265.

30. Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol* 1982; 13: 290–312.

31. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods* 2002; 7(1): 83–104.

32. Bollen KA, Stine R. Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociol Methodol* 1990; 2: 115–140.

33. Barfield R, Shen J, Just AC, et al. Testing for the indirect effect under the null for genome-wide mediation analyses. *Genet Epidemiol* 2017; 41(8): 824–833.

34. Huang YT. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *Ann Appl Stat* 2018; 12(3): 1535–1557.

35. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least lquares lernel lachines and linear mixed models. *Biometrics* 2007; 63(4): 1079-1088.

36. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010; 86(6): 929–942.

37. Sampson JN, Moore SC, Boca SM, Heller R. FWER and FDR control when testing multiple mediators. *Bioinformatics* 2018; 34(14): 2418-2424.

38. Zhao Y, Luo X. Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749* 2016.

39. Song Y, Zhou X, Zhang M, et al. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *bioRxiv* 2018: 467399. doi: https://doi.org/10.1101/467399

□

APPENDIX

PROOFS AND TECHNICAL DETAILS

Independence of p_{i1} and p_{i2}.

Consider a joint distribution of X, $\boldsymbol{M}$ and Y and the factorization of its probability density function

$$f(x, \boldsymbol{m}, y) = f(x) f(\boldsymbol{m} \mid x) f(y \mid x, \boldsymbol{m}),$$

where $f(\cdot)$ denotes a generic density function, and $f(x \mid y)$ denotes the density of the conditional distribution of X given Y. With model (4), we make use of an alternative factorization

$$f(x, \boldsymbol{m}, y) = f(\boldsymbol{m}) f(x \mid \boldsymbol{m}) f(y \mid x, \boldsymbol{m}). \qquad (\star)$$

Although this factorization is not representative of the assumed data generating process, it shares certain favorable statistical properties. Consider three vectors of parameters $(\boldsymbol{\mu}_M, \psi_M)$, $(\tilde{\boldsymbol{\alpha}}, \psi_X)$, and $(\boldsymbol{\beta}, \gamma, \psi_Y)$ pertaining to the marginal distribution of $\boldsymbol{M}$, the conditional distribution of X given $\boldsymbol{M}$, and conditional distribution of Y given X and $\boldsymbol{M}$, respectively, where ψ_A denotes potential nuisance parameters for $A = X, Y, \boldsymbol{M}$. From $(\star)$, the associated likelihood function decomposes as

$$l(\boldsymbol{\mu}_M, \psi_M, \tilde{\boldsymbol{\alpha}}, \psi_X, \boldsymbol{\beta}, \gamma, \psi_Y) = l(\boldsymbol{\mu}_M, \psi_M) + l(\tilde{\boldsymbol{\alpha}}, \psi_X) + l(\boldsymbol{\beta}, \gamma, \psi_Y).$$

The three vectors of parameters are thus L-independent in the terminology of Barndorff-Nilesen (Information and Exponential Families in Statistical Theory, 2014, p.26). As a consequence, inference is decomposed into three independent problems that can be solved separately. In particular, all mixed second order derivatives of the log-likelihood function are zero, implying, at most asymptotic, independence of the corresponding maximum likelihood estimators.

In the manuscript, we allow the dimension of $\boldsymbol{M}$ to be higher than the sample size making the maximum likelihood estimation inadequate. To address this issue, we resort to the global test which assumes the components of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\alpha}}$ are randomly drawn

from zero-centered normal distributions. The parameters of interest are then not the vectors $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\alpha}}$, but rather the variances of the associated distributions σ^2_{β} and $\sigma^2_{\tilde{\alpha}}$. The above arguments apply analogously to these parameters, and ensure, at most asymptotic, independence of the statistics $\hat{Q}_X$ and $\hat{Q}_Y$, resulting in the independence of the associated p-values.

Proof of Theorem 1.

In order to prove that the testing procedure controls FWER for $\mathcal{H}$, it is sufficient to show that it controls FWER for $\left\{H_i, i \in \mathcal{S}\right\}$. We thus need to prove that $\max p_i$ remains a valid p-value for H_i after conditioning on the selection event $\mathcal{S}$, which given the independence of p-values, amounts to conditioning on the event $i \in \mathcal{S}$. Recall that a random variable U is a valid p-value if under the null hypothesis its distribution is either uniform or stochastically greater than the uniform, that is if $\mathsf{P}(U \leq u) \leq u$ for each $0 \leq u \leq 1$. We thus in the following show that under H_i, the conditional distribution of $\max p_i$ given the event $i \in \mathcal{S}$ has this property.

The hypothesis H_i is true when at least one hypothesis among H_{i1} and H_{i2} is true. Since our test statistics are continuous, we can assume that the distribution of the p-value associated to the true hypothesis is uniform. Let F denote the cumulative distribution function of the p-value of the remaining hypothesis. Using the independence of p_{i1} and p_{i2}, and the definition of the selection event, the conditional distribution of interest is given by

$$\mathsf{P}(\max p_i \leq u \mid i \in \mathcal{S}) = \begin{cases} \frac{uF(u)}{F(c)+c-cF(c)}, & 0 < u \leq c; \\ \frac{uF(c)+cF(u)-cF(c)}{F(c)+c-cF(c)}, & c < u \leq 1. \end{cases} \tag{1}$$

We distinguish two scenarios in which H_i is true:

- **Both H_{i1} and H_{i2} are true.** In this case $F(u) = u$, for $u \in [0, 1]$, and

$$\mathsf{P}(\max p_i \leq u \mid i \in \mathcal{S}) = \begin{cases} \frac{u^2}{c(2-c)} \leq u & 0 < u \leq c; \\ \frac{2u-c}{2-c} \leq u, & c < u \leq 1; \end{cases}$$

  and thus $\max p_i$ remains a valid p-value after conditioning on the selection event.

- **Exactly one hypothesis among H_{i1} and H_{i2} is true.** In this case $F(u) \geq u$ for $u \in [0, 1]$. The distribution F depends on the sample size and the alternative, but with increasing sample size it converges to F^*, where $F^*(0) = 0$, and $F^*(u) = 1$, $u \in (0, 1]$. Therefore, the conditional distribution of interest (1) also depends on the sample size and is not necessarily greater than the uniform; however, it is straightforward to show that with increasing sample size it converges to the uniform distribution. □