A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP6240.

Challenges Raised by Mediation Analysis in a High-Dimension Setting

Michaël G.B. Blum,[1,2] Linda Valeri,[3] Olivier François,[1] Solène Cadiou,[4] Valérie Siroux,[4] Johanna Lepeule,[4] and Rémy Slama[4]

[1]Laboratoire Techniques de l'Imagerie Médicale et de la Complexité (TIMC-IMAG; UMR 5525), French National Centre for Scientific Research (CNRS), University Grenoble Alpes, La Tronche, France
[2]OWKIN, Paris, France
[3]Department of Biostatistics, Columbia University Mailman School of Public Health, New York, New York, USA
[4]Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Institute for Advanced Biosciences (IAB) joint research center, Institut national de la santé et de la recherché médicale (Inserm), CNRS, University Grenoble-Alpes, Grenoble, France

Background: Mediation analysis is used in epidemiology to identify pathways through which exposures influence health. The advent of high-throughput (omics) technologies gives opportunities to perform mediation analysis with a high-dimension pool of covariates.

Objective: We aimed to highlight some biostatistical issues of this expanding field of high-dimension mediation.

Discussion: The mediation techniques used for a single mediator cannot be generalized in a straightforward manner to high-dimension mediation. Causal knowledge on the relation between covariates is required for mediation analysis, and it is expected to be more limited as dimension and system complexity increase. The methods developed in high dimension can be distinguished according to whether mediators are considered separately or as a whole. Methods considering each potential mediator separately do not allow efficient identification of the indirect effects when mutual influences exist among the mediators, which is expected for many biological (e.g., epigenetic) parameters. In this context, methods considering all potential mediators simultaneously, based, for example, on data reduction techniques, are more adapted to the causal inference framework. Their cost is a possible lack of ability to single out the causal mediators. Moreover, the ability of the mediators to predict the outcome can be overestimated, in particular because many machine-learning algorithms are optimized to increase predictive ability rather than their aptitude to make causal inference. Given the lack of overarching validated framework and the generally complex causal structure of high-dimension data, analysis of high-dimension mediation currently requires great caution and effort to incorporate *a priori* biological knowledge. https://doi.org/10.1289/EHP6240

Introduction

Mediation analysis is used to help in deciphering mechanisms that relate causes to their consequences. It has been used in many areas of research, including, for example, social psychology, to understand which factors can bridge the gap between intentions and behaviors; cognitive psychology, to analyze how information is transformed into a response; in intervention research, to assess whether an intervention on a specific factor can trigger a positive outcome; or in epidemiology, to quantify to what extent the total effect of a given (environmental or genetic) factor on a health or biological outcome is explained by a so-called indirect effect, through intermediate (e.g., biological) variables on the pathway between exposure and outcome (MacKinnon et al. 2007; VanderWeele 2015, 2016). In environmental epidemiology, for example, it could help in quantifying to what extent air pollution affects respiratory health through oxidative pathways (Romieu et al. 2004).

With the advent of high-throughput screening technologies, there are settings in which one aims to perform mediation analysis with high-dimension data, that is, a data set in which the number of potential mediators p is larger than the number of observations n. For example, in environmental epigenetics, there is increasing evidence that specific environmental exposures such as atmospheric pollutants exposure could influence DNA methylation (Abraham et al. 2018; Gruzieva et al. 2017), which is currently typically assessed using chips measuring methylation in 10^5–10^6 cytosine-phosphate-guanine (CpG) dinucleotide sites on the genome. Given the role of DNA methylation in gene expression and, consequently, health, methylation at multiple CpG sites could contribute to (high dimension) mediation of the effects of atmospheric pollutants on health. High-dimension mediation also may be relevant to analyses of other types of omics data, such as genomic, transcriptomic, metabolomic, and microbiota data.

From a statistical viewpoint, such analyses considering a high-dimension set of potential mediators raise challenges; in particular, approaches used in the case of a single mediator cannot be extended to higher dimensions in a straightforward way. After briefly reviewing the classical case of mediation in a low-dimension setting, we will discuss the issues of mediation analysis with a high-dimension set of covariates and of quantification of the mediated effects, mentioning some of the existing statistical tools, in particular, with regard to applications in environmental epidemiology.

Discussion

Mediation analysis was developed as path analysis in the genetic field, and later in the area of social sciences, and then further formalized in biomedical research in connection with regression modeling in the counterfactual outcome framework of causal inference. For a detailed presentation in the context of causal inference, the reader can refer to Chapters 2 and 5 of VanderWeele (2015), or other sources (Imai et al. 2010; VanderWeele 2016).

If one assumes that a part of the effect of exposure E on outcome Y is mediated by mediator M (Figure 1), then the proportion of the association between E and Y that occurs through M is termed the indirect (or mediated) effect of E. The fraction of the effect of E that occurs independently of M (represented by the direct arrow from E to Y in Figure 1) is called the direct effect. The addition of the direct and indirect effects is termed the total effect.

Mediation with a Single Mediator

Assuming that both Y and M are continuous variables, with further assumptions on the distribution of the error term, one can write several linear regression models:

Address correspondence to R. Slama, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Inserm-CNRS-University Grenoble-Alpes, IAB joint research center, Allée des Alpes, Site Santé, 38706 La Tronche, France. Telephone: 33 476549402. Email: remy.slama@univ-grenoble-alpes.fr

The authors declare they have no actual or potential competing financial interests.

Received 17 September 2019; Revised 14 April 2020; Accepted 15 April 2020; Published 6 May 2020.

One model relating the exposure E to the outcome Y taking M into account is as follows:

$$\mathbb{E}(Y) = \theta_0 + \theta_1 E + \theta_2 M + \theta_3 EM + \theta_4 \boldsymbol{C} \quad \text{(Exposure–outcome model)} \quad (1)$$

Where $\mathbb{E}$ is the mathematical expectation, E is the exposure variable, $\boldsymbol{C}$ is a matrix including all potential confounders of the exposure–outcome, exposure–mediator, and mediator–outcome associations (i.e., $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$ in Figure 1), EM is the exposure–mediator interaction term, and $\theta_0, \ldots, \theta_4$ are the estimated parameters.

One regression model relating the exposure to the mediator is as follows:

$$\mathbb{E}(M) = \beta_0 + \beta_1 E + \beta_2 \boldsymbol{C}' \quad \text{(Exposure–mediator model)} \quad (2)$$

where $\boldsymbol{C}'$ represents the confounders for the exposure–mediator association.

In this linear model setting, one can define the controlled direct effect, corresponding to the average change in the outcome for an increase by one in exposure, assuming that the mediator remains fixed at a given value identical in all subjects. In contrast, the natural direct effect corresponds to the average change in the outcome for an increase by one in exposure, assuming that the mediator level does not vary and is set in each subject at the value it would have in the absence of exposure. In the simpler case of lack of interaction between the exposure and the mediator ($\theta_3 = 0$), the controlled direct effect and the natural direct effect are identical. The natural indirect effect is defined as the average change in outcome, under a given fixed exposure, when the mediator level changes to the level it would have attained if the exposure had increased by one.

The estimation of direct and indirect effects requires several assumptions: *a*) a lack of exposure–outcome confounding (i.e., in the example of Figure 1, efficient adjustment for $\boldsymbol{C}_1$) and *b*) a lack of mediator–outcome confounding (efficient control for $\boldsymbol{C}_3$). These are the only conditions required for the estimation of the controlled direct effect.

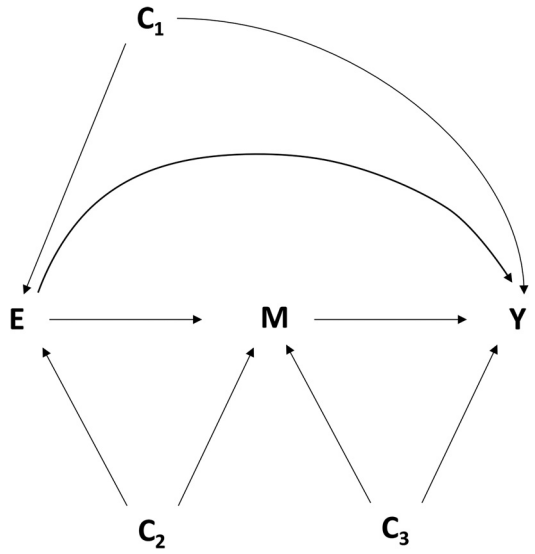

Figure 1. Example of the effect of a single exposure E whose effect on outcome Y is mediated by a single mediator M. Exposure–outcome ($\boldsymbol{C}_1$), exposure–mediator ($\boldsymbol{C}_2$) and mediator–outcome confounders ($\boldsymbol{C}_3$) need to be controlled for. Adapted from VanderWeele (2015).

Two additional assumptions, required for the identification of natural direct and indirect effects, are *c*) a lack of uncontrolled exposure–mediator confounding and *d*) the lack of mediator–outcome confounder affected by the exposure. Under these hypotheses, and that of the absence of exposure–mediator interaction ($\theta_3 = 0$), θ_1 provides an estimate of the direct effect of E on Y, whereas the product $\theta_2\beta_1$ is an estimate of the indirect effect through M. This product is used in the Sobel mediation test, whose null hypothesis is $H_0{:}\theta_2\beta_1 = 0$.

This corresponds to a composite null hypothesis (Baron and Kenny 1986; MacKinnon et al. 2002), implying both θ_2 and β_1. It can be handled either by testing for the product $\theta_2\beta_1$ being null, corresponding to a product significance test, or by testing separately for both θ_2 and β_1 being null, which is termed a joint significance test.

Mediation with Multiple Mediators

VanderWeele and Vansteelandt (2014) and VanderWeele (2015) have discussed the expansion of mediation analysis to the case of several mediators $M_1, \ldots, M_p$, where p is much lower than n. One option is to consider each mediator independently and to fit one exposure–outcome model (Equation 1) and one exposure–mediator model (Equation 2) for each mediator $M_1, \ldots, M_p$. This approach works fine as long as there is no mediator M_i influencing another mediator M_j and no mediator–mediator interaction. If an influence of one mediator over another exists, the abovementioned assumption of the lack of confounding of the mediator–outcome association by a factor influenced by exposure no longer holds. Indeed, in the situation of Figure 2 in which mediator M_1 influences M_2, if one considers solely mediator M_2, then M_1 acts as a confounder of the relation between mediator M_2 and outcome Y influenced by E. Several approaches have been proposed in a low-dimension setting to estimate direct effects in this case, such as marginal structural modeling and structural mean models (VanderWeele 2015).

An alternative to the mediator-by-mediator approach is to treat all mediators as a whole (VanderWeele and Vansteelandt 2014). In this case, any possible influence between mediators (such as that of M_1 on M_2), including interactions, can be

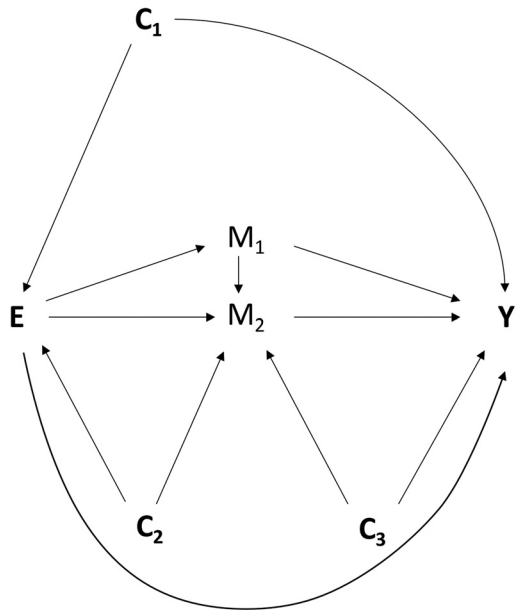

Figure 2. Example of mediation with two mediators M_1 and M_2 influencing each other.

ignored, as long as one is not interested in estimating path-specific effects of sequential mediators. In practice, for a continuous outcome, one can fit an outcome model without mediators:

$$\mathbb{E}(Y) = \theta_0' + \theta_1' E + \theta_3' \boldsymbol{C} \quad (3)$$

and the following model for p mediators $M_1, \ldots, M_p$:

$$\mathbb{E}(Y) = \theta_0 + \theta_1 E + \theta_2^1 M_1 + \theta_2^2 M_2 + \cdots + \theta_2^p M_p + \theta_3 \boldsymbol{C}, \quad (4)$$

thus allowing us to estimate the indirect effect mediated by all mediators $M_1, \ldots, M_p$ as a whole by the difference $\theta_1' - \theta_1$. Causal interpretation of this quantity depends on correct model specification and lack of unmeasured confounding and of interaction, as in the single-mediator case, so that $\boldsymbol{C}$ should include all exposure–outcome, mediator–outcome, and exposure–mediator confounders.

A third option relying on inverse probability weighting has been proposed that assumes exposure to be categorical, with few categories (VanderWeele and Vansteelandt 2014). This approach does not require models for the mediators, but a model predicting the outcome as a function of exposure and the mediators is necessary (VanderWeele and Vansteelandt 2014). Finally, estimating so-called interventional (in)direct effects, which requires weaker assumptions than identification of natural (in)direct effects, is also an option (Vansteelandt and Daniel 2017). Bellavia et al. (2019) have provided other examples of methods applied when considering the mediating effects of (low-to-intermediate) chemical mixtures on health. We will now assume that the number of mediators p is of the same or of a larger order of magnitude than the number of observations (Figure 3).

Identification of Mediators in High Dimension

Generally, several issues need to be considered when trying to translate the mediation analysis framework to the case of high-dimension mediation. These issues relate to the knowledge about the causal model underlying the data and, relatedly, the identification of the mediators, to the correction for multiple testing, to the consideration of composite tests, and to the estimation of the share of the effect of the considered exposure explained by mediators.

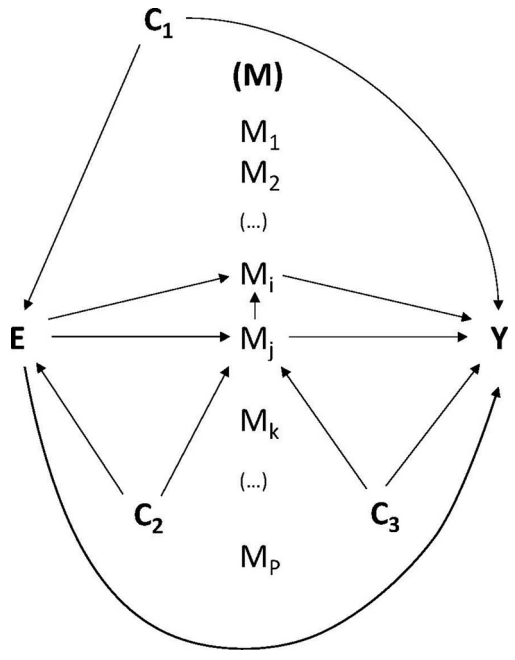

Figure 3. High-dimension mediation. Hypothesized relation between an exposure E; a health outcome Y; an exposure–outcome confounder $\boldsymbol{C}_1$; a high-dimension mediator $M = (M_i)_{i \leq p}$, where p is typically larger than the number of observations in the data set, an exposure–mediator confounder $\boldsymbol{C}_2$; and a mediator–outcome confounder $\boldsymbol{C}_3$. Causal influences also exist among the candidate mediators (here, M_j influences M_i). p is typically much larger than the number of observations n in the data set.

The approach to mediation analysis as developed in the framework of causal inference (VanderWeele 2015) makes the assumption that the causal structure underlying the data is known *a priori*, as opposed to inferred from the data. This means that the directions of the relations between E, M, Y and all potential confounders are known. However, *a priori* knowledge of the causal structure is less likely to be available as the dimension of the set of (potential) mediators increases, that is, as the biological system considered becomes more complex. Moreover, as in the low-dimension case, exposure–outcome, exposure–mediators, and mediators–outcome confounders need to be identified and controlled. In Table 1, we list different approaches proposed to perform mediation analysis in a high-dimension setting.

Independent consideration of each potential mediator. One technically relatively simple option has been to consider all potential mediators independently, in separate models. Küpers et al. (2015) for example used a two-step approach consisting of detecting associations between maternal smoking and genome-wide methylation levels using an epigenome-wide association study (corrected for multiple testing), and then performing a series of mediation analyses of the maternal smoking effects on birth weight mediated by each of the methylation hits (i.e., the potential mediators) identified in the first step. Such an approach makes among other the strong assumptions of lack of correlation or interactions between mediators, as discussed above for multiple mediators, which is unlikely to hold in many settings. It is tempting to try controlling for the potential mediators that may influence the mediator considered in the second step, but unfortunately the tools classically used in low-dimension settings to control for confounding (*a priori* identification of potential confounders and adjustment for these factors based on a multiple regression model), cannot be applied in a straightforward way when confounders need to be identified from a high-dimension vector, unless knowledge of the causal relations within the biological layer corresponding to mediators is known *a priori*. Trying to identify the confounders in a data-driven approach is challenging; indeed, in this setting, if some correlation exists among the mediators, then univariate approaches are expected to suffer from a high false detection rate, even when multiple correction techniques are used. A simulation study aiming to relate a large number of (weakly) correlated factors to a health outcome showed that false discovery rate (FDR)-correction techniques yielded false detection rates far higher than the expected value of 5%, as a result of this correlation between disease predictors (Agier et al. 2016), a situation that may also happen when modeling associations of candidate mediators with disease risk.

An approach related to that from Küpers et al. (2015), used to identify if epigenetic marks mediate the relationship between genotypes and disease status, considered the causal inference test (CIT) (Liu et al. 2013). CIT is related to the Baron and Kenny (1986) procedure in that it corresponds to a chain of mathematical conditions that must be satisfied to conclude that each potential mediator causally influences the outcome (Millstein et al. 2009). This approach can be seen as attempting to reconstruct the whole causal structure underlying the data, but it does so considering each potential mediator separately, which, again, may be challenging if there is correlation between mediators (Wang and Michoel 2017).

Permutation tests for mediators. In the two abovementioned approaches (Küpers et al. 2015; Liu et al. 2013), candidate mediators are tested separately, paralleling genome-wide association

Table 1. Overview of the approaches and models for high-dimension analysis reviewed.

Name of approach	Reference	Assumptions, method, comment
Separate consideration of the potential mediators		
Successive tests of association of the potential mediators with the exposure followed by the Sobel mediation test	Küpers et al. 2015	Approaches can be used to overcome the limited power of the Sobel test. Assumes lack of uncontrolled confounders and mutual influences between mediators.
Causal inference test	Liu et al. 2013	Assumes lack of uncontrolled confounders.
Permutation test	Boca et al. 2014; Sampson et al. 2018	Tests multiple putative mediators while controlling the family-wise error rate. Replacing Bonferroni correction with a permutation approach improves statistical power (MultiMed R package).
Joint significance test	Huang 2018	Separate tests of exposure–mediator and mediator–outcome associations.
Test for a composite null hypothesis	Huang 2019	Test statistic is derived by accounting for the composite nature of the null hypothesis. It is less conservative than the Sobel test.
Simultaneous consideration of the potential mediators		
Inverse probability weighting approach	VanderWeele and Vansteelandt 2014	More efficient if exposure is categorical with a small number of categories. Can accommodate exposure–mediator and mediator–mediator interactions.
R package HIMA dimension reduction approach	Zhang et al. 2016	Uses variable selection to reduce the number of mediators (HIMA R package).
Joint test of a group of mediators	Huang and Pan 2016	Component-wise testing to evaluate several mediators *en bloc* rather than testing the marginal contribution of each individual mediator. Spectral decomposition of the mediators.
Directions of mediations	Chén et al. 2018	Builds linear combinations among the potential mediators to construct polymediators.
Sparse principal component–based high-dimension mediation analysis	Zhao et al. 2020	Dimension reduction of the potential mediators via sparse principal component analysis.
Mediation analysis for composition data	Sohn and Li 2019	Tests several mediators *en bloc*; well-suited for compositional data (i.e., proportions of a whole, as can be the case for microbiome data).
Distance-based test for mediation analysis (applied to microbiome data)	Zhang et al. 2018	Reduces multiple testing burden by using a distance-based approach in which all mediators are tested simultaneously. Implies the existence of a relevant distance that can be used between mediators.
Global test for high-dimension mediation	Djordjilović et al. 2019	Global approach for mediation to test simultaneously a group of mediators.

studies (GWAS). Boca et al. (2014) developed a permutation test that allows controlling the family-wise error rate while testing a large number of mediators, again under the assumption of a lack of unmeasured confounding. Permutation tests account for the underlying correlation between mediators and do not suffer from the problem of the Bonferroni correction, which is increasingly conservative as the correlation between mediators increases (Boca et al. 2014).

Composite tests. Under specific hypotheses, several approaches commonly used to test mediation, such as the product test, are overly conservative (Barfield et al. 2017; MacKinnon et al. 2002). This issue can be addressed by computing empirical *p*-values based on bootstrapping, which provides an increased power to detect mediation for a given sample size (Barfield et al. 2017). Boca et al. (2014) and Sampson et al. (2018) developed several statistical procedures, which are implemented in the R package MultiMed, to increase the power of such composite analyses when testing multiple mediators. Huang (2018) developed a joint significance test in the context of multiple mediators. Huang (2019) also leveraged the composite nature of the null hypothesis to construct a new test statistic that is less conservative than the Sobel test.

Empirical estimation of the null distribution. The limited power of the Sobel test can also be viewed as related to wrong assumptions regarding the theoretical null distribution. When a test statistic assumes a given distribution under the null hypothesis (e.g., a chi-squared distribution) while the real distribution under the null hypothesis is modified (e.g., by unmeasured confounders), hypothesis testing and FDR-control procedures may be invalidated (Devlin and Roeder 1999; Efron 2004; François et al. 2016; Strimmer 2008). In a high-dimension setting, this issue can be identified by displaying the distribution of *p*-values or of *z*-scores (Efron 2004). Solutions have been reviewed in the GWAS context and include empirical null distribution and genomic inflation factors to calibrate *p*-values, that is, the correction of *p*-values in a way ensuring that their distribution is flat under the null hypothesis (Efron 2004; François et al. 2016; Strimmer 2008). We illustrate this in Figure 4. Simulations were performed using the mediation model of Equation 4, assuming that, of 5,000 putative mediators, 500 variables were involved in the indirect path relating the exposure to the outcome. Raw *p*-values of Sobel test testing for mediation (Figure 4, red histogram) were shifted toward values closer to 1 compared with the expected distribution, which is, a mixture of a uniform distribution and a distribution with an excess of small *p*-values. After application of empirical null hypothesis testing techniques, the (adjusted) distribution of *p*-values of the Sobel test (blue distribution) became closer to the expected one.

Mediators considered as a whole. In the context of high-dimension mediation, as outlined above with a few mediators, there can be mutual influences between mediators. For example, in the epigenetic field, the methylation level on one CpG site can influence the methylation level of other CpG sites. This is the case for

alterations in the methylation of DNA-methyltransferase (DNMT) genes, which may alter the level of methyltransferase enzymes, which in turn impact the methylation of several other genes (Zhang and Xu 2017). These relations among mediators can create confounding and hamper identification of mediators causally linked with the health outcome when considering each mediator separately (VanderWeele and Vansteelandt 2014). For example, in Figure 3, if one tries to consider mediators separately, the proportion mediated by M_j must be identified for the proportion mediated by M_i to be properly quantified (VanderWeele 2015). Identifying the mediator(s) responsible for this confounding bias is, as discussed above, challenging in a high-dimension setting.

An alternative to the separate consideration of each candidate mediator is to follow the logic of the approach highlighted above for multiple mediators considering all (potential) mediators as a whole. In this situation, a subtle understanding of the causal relations within the high-dimension mediators is not necessary and only identification and control for confounders outside the set of mediators is required. In particular, mutual influences or interactions among mediators can be ignored in this setting, provided one does not aim to identify specific causal mediator(s), or, in the case of sequential mediators, the effect of a specific causal path.

When using classical (least squares or maximum likelihood for logistic regression) estimators, the exposure–outcome regression model including all potential mediators (Equation 4) provides estimation of regression parameters with a prohibitively large variance as p increases and reaches a fraction of n (Sur and Candès 2019; Vittinghoff and McCulloch 2007). Once p is larger than n, the model cannot be estimated by ordinary least squares or maximum likelihood anymore. Instead, one of the multivariate variable selection or dimension reduction techniques proposed to relate high-dimension variables to one or a few unidimensional variables can be used (Chadeau-Hyam et al. 2013). Several approaches have been developed in this spirit, which we describe below, starting with approaches relying on variable selection and then presenting those related to dimension reduction.

High-dimension mediation based on variable selection. Zhang et al. (2016) implemented in the R package HIMA a three-step approach that, first, excludes candidate mediators that are not strongly associated with the health outcome in an univariate approach then, second, uses a regularized multivariate mediation model allowing further restriction to a smaller group of mediators whose mediation effect is tested in a last, third, step. A limitation of this approach is that it assumes a lack of confounding or residual confounding, similarly to the abovementioned approaches considering each mediator independently (to which it is actually related). One way to overcome this would be to rely on iterative sure independence screening (or ISIS) in the first step, which was designed to cope with situations such as that of a covariate not marginally associated with the outcome but related with it conditionally on another covariate (Fan et al. 2009).

High-dimension mediation based on dimension reduction. Another approach, called the directions of mediation, was used to determine which brain locations mediate the relationship between the application of a thermal stimulus and self-reported pain (Chén et al. 2018). It does not attempt to identify true mediators but, rather, seeks linear combinations of mediators that capture the mediators' effect according to a criterion related to the ability to predict the outcome and to be explained by exposure E. This approach is therefore related to dimension reduction techniques such as (supervised) principal component analysis (PCA). Relatedly, Huang and Pan (2016) developed a test for mediation for high-dimension mediators relying on spectral decomposition of the set of mediators, followed by a series of univariate regression models with the independent components. As an extension, Zhao et al. (2020) introduced sparsity into the PCA-type analysis used by Huang and Pan (2016).

Advantage can be taken of the nature of the layer of mediators. If, for example, a distance can be defined among the potential mediators, then it may be used to decrease the dimension of the mediation layer. This approach has been applied to characterize to what extent the effect of diet on body mass index is mediated by changes in the composition of the gut microbiota (Zhang et al. 2018). Data on microbiota composition obtained from 16S rRNA gene sequencing can be classified according to operational taxonomic units, from which a distance based on DNA sequence divergence (or distance of species in the phylogenetic tree) can be calculated. Of course, here a central assumption relates to the relevance of the proposed distance for the health outcome considered in the microbiota example that microbiota diversity, as quantified from DNA sequence, influences the health outcome considered. As discussed by Zhang et al. (2018), approaches that accommodate different types of distances, without having to *a priori* choose one of them, have been proposed.

All of these techniques are limited when it comes to the causal interpretation: Indeed, these approaches will generally allow identifying sets or combination of covariates with predictive power, which does not imply that they are the causal agents.

Issues related to overfitting. Options to *a priori* reduce the dimension of the mediators' layer are all the more relevant because of issues related to overfitting. Indeed, in the context of high-dimension mediators, the ability of the mediators to predict the outcome can be overestimated; this is all the more a concern because many machine-learning algorithms tend to be optimized to increase their predictive ability rather than their aptitude to make causal inference (Hernán et al. 2019). Mistaking the predictive ability of a model with its value for causal inference may lead to overestimating the share of the mediated effect. For example, using an approach such as the least absolute shrinkage and selection operator, or LASSO, to relate the candidate mediators to the health outcome may lead to a model with a very high predictive value owing to the fact that the ability to predict a unidimensional variable increases with the number of potential predictors, but which may include noncausal predictors of the outcome among the candidate mediators (Leng et al. 2006; Meinshausen and Bühlmann 2010). In order to limit such overestimation of the estimated effect of the candidate mediators on the health outcome, algorithms and strategies targeted for counterfactual prediction and causal inference should be preferred over those favoring the model's predictive ability. This will be best achieved through some knowledge on the confounders external to the set of mediators (which may be more easily *a priori* available than information on the relation among the potential mediators) and on factors that are certainly not confounders of associations between the exposure and mediators or outcome, such as those influenced by exposure (Hernán et al. 2019). Purely statistical considerations will also weight on the efficiency of a strategy in terms of counterfactual inference; for example, maximizing the predictive ability of a model may not be the most relevant choice, whereas trying to maximize stability (Meinshausen and Bühlmann 2010) or specificity is expected to be more favorable. Finally, approaches relying on a targeted minimum loss estimator (TMLE), which provides efficient prediction while remaining in a causal inference framework, are certainly worth considering here (Lendle et al. 2013; Zheng and van der Laan 2018).

Estimation of the Proportion of Effect Mediated

Measures of mediated effects are an expected output of mediation analysis (MacKinnon et al. 1995). Again, translating in high dimension the approach used with a single mediator is not

straightforward. Some analyses with high-dimension mediators may report mediated effects for each of the candidate marker, which is not informative about the overall mediated effect because single components of a set of high-dimension mediators can have opposite effects (Küpers et al. 2015; Zhang et al. 2016). In addition, because of correlation and interactions among mediators, the sum of the proportion mediated can be more than 100% (VanderWeele and Vansteelandt 2014).

Alternatively, one might wish to estimate the global proportion of effect mediated by all potential mediators considered simultaneously. Insight may, again, be gained from the GWAS field. In GWAS, methodological efforts have been devoted to estimate how much of the population variability in a phenotypic trait is explained by genetic variation among individuals. Some statistical models to estimate the influence of genetic polymorphisms do not rely on identification of true causal markers but, rather, assume a polygenic model where each marker has a (possibly) infinitesimal effect (Yang et al. 2011). Similarly, a polymediator approach transposing this logic and considering mediators as a whole could be applied: If one relied on one of the above-mentioned approaches allowing one to build a linear combination of the candidate mediators (Chén et al. 2018; Huang and Pan 2016), then the proportion mediated by this new (unidimensional) variable could be estimated as in the single-mediator case. This approach has been applied by Zhao et al. (2020), who relied on a sparse PCA of the candidate mediators to identify a linear combination of parameters quantifying the activity of various brain regions likely to mediate the effect of a learning task on reaction time; they provided an estimate of the indirect effect.

Conclusion

The increasing interest for omics data and the role of epigenetic marks, RNA, and protein levels or of the microbiota on disease phenotypes make it very appealing to try to quantify to what extent the effect of environmental (Bind et al. 2014), infectious, behavioral, social (Huang 2018), or genetic (Liu et al. 2013) factors on health is mediated by these biological layers. As we discussed, generalizing mediation analysis techniques developed for one or a few mediators to high-dimension mediation is not straightforward. A key conceptual issue relates to the fact that mediation analysis assumes *a priori* knowledge of the causal relations between the exposure, the mediator, the outcome, and confounders. In a high dimension, the causal relations between all considered factors (e.g., among CpGs) is unlikely to be accurately known *a priori*. It is optimistic to expect the causal structure to be unraveled by machine-learning techniques because many models developed in data science tend to excel in predictive ability but may be of more limited use in making causal inference (Hernán et al. 2019), in particular in a context of a rather

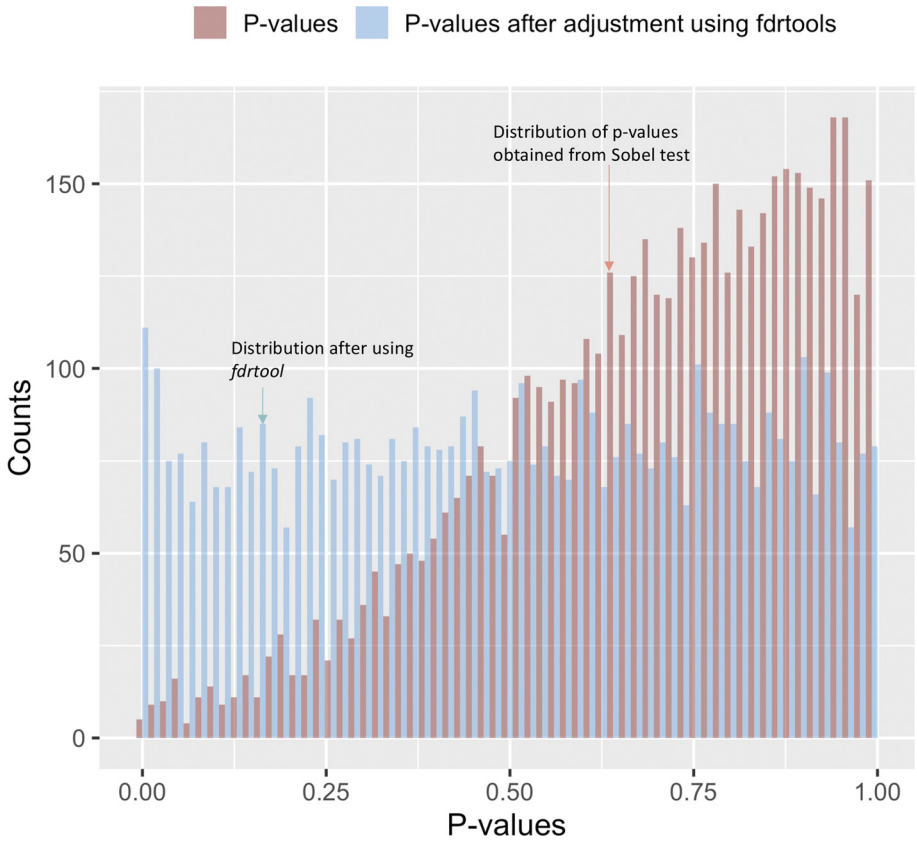

Figure 4. Raw distribution of the *p*-values of the Sobel mediation test for 5,000 simulated variables that are putative mediators (in red, not uniform) and corrected distribution (blue) after using the fdrtool package (R version 3.6.1; R Development Core Team). After correction, the distribution is closer to that expected under the simulated causal model, which assumes the presence of mediators, so that one observes a mixture of a uniform distribution and a distribution with an excess of small *p*-values. The distribution of the raw *p*-values should be uniform except for an excess of small *p*-values corresponding to true mediators. The fact that the (red) distribution is not uniform may indicate several deviations from the null model such as confounding factors or poor standardization of the test statistic. The red histogram indicates that the Sobel test is too conservative (MacKinnon et al. 1995). Here we use the R package fdrtool that implements an empirical null distribution approach to transform initial *p*-values to uniformly distributed *p*-values and that provides control of the false discovery rate (Strimmer 2008). To perform simulations, we consider the mediation model of Equation 4, where there are 500 random mediators influenced by the environment that affect the simulated outcome according to Equation 4. We considered 4,500 additional putative mediators distributed according to a multivariate distribution that did not depend on environment and outcome (see code on GitHub https://github.com/mblumuga/opinion_mediation/blob/master/Simus_Sobel_FDR.R).

limited number of training samples. TMLE represents a way to try to accommodate both aims that is certainly worth further considering in a high-dimension context (Lendle et al. 2013). Although we took the example of high-dimension mediators, most of the issues discussed here also apply to problems in which the mediators have an intermediate dimension (with p lower than n but still relatively high), as would, for example, happen if someone wished to quantify to what extent the association of socioeconomic status and a health outcome is mediated by a set of several hundred exposures assessed in a population of a few thousand subjects.

The fact that we mostly discussed data-driven approaches should not let the reader believe that this is the only way forward. On the contrary, any effort to incorporate in models biological knowledge should be undertaken. This may, for example, be done by restricting analyses to a subset of genes with high *a priori* plausibility for an effect on the outcome or by reducing the dimension of the layer of potential mediators using a biologically relevant distance, as is done for microbiote data, on the basis, for example, of the phylogenetic distance between the microbiote species (Zhang et al. 2018). Once such options have been considered and, if relevant, implemented, one may then turn to more data-driven approaches.

High-dimension omics layers may have a complex and hard to identify causal structure, for example, in the case of mutual influences among the mediators (CpG sites or protein levels). Such situations in which a mediator is also a confounder for the relation between another mediator and the outcome influenced by the exposure of interest are hard to handle rigorously considering each mediator separately, and for such high-dimension omics layers, it is unlikely that molecular biology will soon unravel all causal relations among all variables. Currently, approaches considering each potential mediator separately or treating the potential mediators as a whole coexist. Issues identified in the (low dimension) case of multiple mediators tend to indicate that approaches considering mediators as a whole (rather than individually) should be preferred with a high-dimension mediator (VanderWeele and Vansteelandt 2014).

Issues less specific to high-dimension mediation analysis add to the abovementioned issues; these include reverse causation (Liu et al. 2013), measurement error in the exposure or in the mediator(s) (Valeri et al. 2017), and reliance on observational studies to test mediation (Richmond et al. 2014).

Data collection should generally be guided by power calculations, which are challenging in high-dimension mediation given that the question of sample size requirements for mediation analysis is not even completely solved in a single-mediator setting (VanderWeele 2015). Simulation studies prior to the design of a new study should be considered but are challenging with complex data structures; a few examples exist (Barfield et al. 2017; Boca et al. 2014; Huang 2018).

Further extensions of the case that we discussed (Figure 3) could be worth considering. First, we did not single out the multimediator case with ordering, where there may be several successive ordered layers of potential mediators:

$$E \rightarrow M_1 \rightarrow M_2 \rightarrow \ldots \rightarrow M_k \rightarrow Y, \quad (5)$$

with each $M_1, \ldots M_k$ possibly being highly dimensional.

In addition, not only the mediator, but also the exposure, could be multidimensional. This situation has been considered in a case–control study considering the mediating role of DNA methylation in the association between genetic polymorphisms (assessed from a genome-wide screening leading to about 300,000 genetic polymorphisms) and arthritis risk (Liu et al. 2013). Exposome studies in which methylome or metabolome data are available can lead to a similar data structure (Maitre et al. 2018). Further, the outcome could be multidimensional, corresponding to outcome- or diseasome-wide studies (VanderWeele 2017). This would imply a move from the rather simple three-variable system corresponding to the typical original mediation framework (MacKinnon et al. 2007) to a much more complex three-layer system. Finally, all these situations could be combined, for example, in a study assessing in the same population multiple layers of interconnected omics layers, from a large exposome to, for example, microbiota, methylome, transcriptome, proteome, or diseasome data. There are descriptive tools for exploring the relations within and between such layers, for example, in the literature referring to multimodal data, with approaches such as sparse generalized canonical correlation analysis (Garali et al. 2018). However, a rigorous causal analysis of such data, whose collection on hundreds or thousands of subjects is now feasible (Maitre et al. 2018), would require knowledge on the causal relations between (and possibly within) each data layer, which may, in many situations, be very difficult to attain. Inferring causal structure from data without strong *a priori* is an expanding field of research (Scanagatta et al. 2019; Uusitalo 2007). The approaches initially used have tended to have a complexity increasing at least exponentially with the number of possible nodes in the causal diagram to infer, but alternatives have been recently suggested that may make the problem tractable also in high dimension (Zheng et al. 2018).

Omics platforms generate a huge amount of data, opening the way for joint analysis of environmental exposures, intermediate biological layers of data, and health outcomes. High-dimension mediation analysis constitutes one promising framework to handle such multiple layers of data in a causal inference framework; however, it is still in its infancy and raises numerous challenges. These challenges should be tackled by biostatisticians, biologists, and epidemiologists in order to better understand the determinants of health, to make efficient use of the data generated beyond their use for prediction, and avoid making a data cemetery out of the promised knowledge Eldorado (Hunter 2006).

Acknowledgments

This article arose from discussions in a Data Challenge on High Dimension Mediation Analysis held in Aussois, France, on 7–9 June 2017. We warmly thank all the participants of the challenge. This work was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the Investissements d'Avenir program (ANR-15-IDEX-02) and has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the National Research Agency funded projects ETAPE (ANR-18-CE36-0005) and GUMME (ANR-18-CE34-0013).

References

Abraham E, Rousseaux S, Agier L, Giorgis-Allemand L, Tost J, Galineau J, et al. 2018. Pregnancy exposure to atmospheric pollution and meteorological conditions and placental DNA methylation. Environ Int 118:334–347, PMID: 29935799, https://doi.org/10.1016/j.envint.2018.05.007.

Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. 2016. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environ Health Perspect 124(12):1848–1856, PMID: 27219331, https://doi.org/10.1289/EHP172.

Barfield R, Shen J, Just AC, Vokonas PS, Schwartz J, Baccarelli AA, et al. 2017. Testing for the indirect effect under the null for genome-wide mediation analyses. Genet Epidemiol 41(8):824–833, PMID: 29082545, https://doi.org/10.1002/gepi.22084.

Baron RM, Kenny DA. 1986. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173–1182, PMID: 3806354, https://doi.org/10.1037//0022-3514.51.6.1173.

Bellavia A, James-Todd T, Williams PL. 2019. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environ Int 123:368–374, PMID: 30572168, https://doi.org/10.1016/j.envint.2018.12.024.

Bind MA, Lepeule J, Zanobetti A, Gasparrini A, Baccarelli A, Coull BA, et al. 2014. Air pollution and gene-specific methylation in the Normative Aging Study: association, effect modification, and mediation analysis. Epigenetics 9(3):448–458, PMID: 24385016, https://doi.org/10.4161/epi.27584.

Boca SM, Sinha R, Cross AJ, Moore SC, Sampson JN. 2014. Testing multiple biological mediators simultaneously. Bioinformatics 30(2):214–220, PMID: 24202540, https://doi.org/10.1093/bioinformatics/btt633.

Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al. 2013. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen 54(7):542–557, PMID: 23918146, https://doi.org/10.1002/em.21797.

Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. 2018. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics 19(2):121–136, PMID: 28637279, https://doi.org/10.1093/biostatistics/kxx027.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55(4):997–1004, PMID: 11315092, https://doi.org/10.1111/j.0006-341x.1999.00997.x.

Djordjilović V, Page CM, Gran JM, Nøst TH, Sandanger TM, Veierød MB, et al. 2019. Global test for high-dimensional mediation: testing groups of potential mediators. Stat Med 38(18):3346–3360, PMID: 31074092, https://doi.org/10.1002/sim.8199.

Efron B. 2004. Large-scale simultaneous hypothesis testing. J Am Stat Assoc 99(465):96–104, https://doi.org/10.1198/016214504000000089.

Fan J, Samworth R, Wu Y. 2009. Ultrahigh dimensional feature selection: beyond the linear model. J Mach Learn Res 10:2013–2038, PMID: 21603590.

François O, Martins H, Caye K, Schoville SD. 2016. Controlling false discoveries in genome scans for selection. Mol Ecol 25(2):454–469, PMID: 26671840, https://doi.org/10.1111/mec.13513.

Garali I, Adanyeguh IM, Ichou F, Perlbarg V, Seyer A, Colsch B, et al. 2018. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. Brief Bioinform 19(6):1356–1369, PMID: 29106465, https://doi.org/10.1093/bib/bbx060.

Gruzieva O, Xu CJ, Breton CV, Annesi-Maesano I, Antó JM, Auffray C, et al. 2017. Epigenome-wide meta-analysis of methylation in children related to prenatal NO_2 air pollution exposure. Environ Health Perspect 125(1):104–110, PMID: 27448387, https://doi.org/10.1289/EHP36.

Hernán MA, Hsu J, Healy B. 2019. A second chance to get causal inference right: a classification of data science tasks. Chance 32(1):42–49, https://doi.org/10.1080/09332480.2019.1579578.

Huang YT. 2018. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. Ann Appl Stat 12(3):1535–1557, https://doi.org/10.1214/17-AOAS1120.

Huang YT. 2019. Genome-wide analyses of sparse mediation effects under composite null hypotheses. Ann Appl Stat 13(1):60–84, https://doi.org/10.1214/18-AOAS1181.

Huang YT, Pan WC. 2016. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 72(2):402–413, PMID: 26414245, https://doi.org/10.1111/biom.12421.

Hunter DJ. 2006. Genomics and proteomics in epidemiology: treasure trove or "high-tech stamp collecting"? Epidemiology 17(5):487–489, PMID: 16906050, https://doi.org/10.1097/01.ede.0000229955.07579.f0.

Imai K, Keele L, Tingley D. 2010. A general approach to causal mediation analysis. Psychol Methods 15(4):309–334, PMID: 20954780, https://doi.org/10.1037/a0020761.

Küpers LK, Xu X, Jankipersadsing SA, Vaez A, la Bastide-van Gemert S, Scholtens S, et al. 2015. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. Int J Epidemiol 44(4):1224–1237, PMID: 25862628, https://doi.org/10.1093/ije/dyv048.

Lendle SD, Subbaraman MS, van der Laan MJ. 2013. Identification and efficient estimation of the natural direct effect among the untreated. Biometrics 69(2):310–317, PMID: 23607645, https://doi.org/10.1111/biom.12022.

Leng C, Lin Y, Wahba G. 2006. A note on the Lasso and related procedures in model selection. Stat Sin 16(4):1273–1284, https://doi.org/10.5705/ss.2011.029a.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31(2):142–147, PMID: 23334450, https://doi.org/10.1038/nbt.2487.

MacKinnon DP, Fairchild AJ, Fritz MS. 2007. Mediation analysis. Annu Rev Psychol 58:593–614, PMID: 16968208, https://doi.org/10.1146/annurev.psych.58.110405.085542.

MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. 2002. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods 7(1):83–104, PMID: 11928892, https://doi.org/10.1037/1082-989x.7.1.83.

MacKinnon DP, Warsi G, Dwyer JH. 1995. A simulation study of mediated effect measures. Multivariate Behav Res 30(1):41, PMID: 20157641, https://doi.org/10.1207/s15327906mbr3001_3.

Maitre L, de Bont J, Casas M, Robinson O, Aasvang GM, Agier L, et al. 2018. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. BMJ Open 8(9):e021311, PMID: 30206078, https://doi.org/10.1136/bmjopen-2017-021311.

Meinshausen N, Bühlmann P. 2010. Stability selection. J R Stat Soc Series B Stat Methodol 72(4):417–473, https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Millstein J, Zhang B, Zhu J, Schadt EE. 2009. Disentangling molecular relationships with a causal inference test. BMC Genet 10:23, PMID: 19473544, https://doi.org/10.1186/1471-2156-10-23.

Richmond RC, Al-Amin A, Smith GD, Relton CL. 2014. Approaches for drawing causal inferences from epidemiological birth cohorts: a review. Early Hum Dev 90(11):769–780, PMID: 25260961, https://doi.org/10.1016/j.earlhumdev.2014.08.023.

Romieu I, Sienra-Monge JJ, Ramirez-Aguilar M, Moreno-Macías H, Reyes-Ruiz NI, Estela del Rio-Navarro B, et al. 2004. Genetic polymorphism of GSTM1 and antioxidant supplementation influence lung function in relation to ozone exposure in asthmatic children in Mexico City. Thorax 59(1):8–10, PMID: 14694237.

Sampson JN, Boca SM, Moore SC, Heller R. 2018. FWER and FDR control when testing multiple mediators. Bioinformatics 34(14):2418–2424, PMID: 29420693, https://doi.org/10.1093/bioinformatics/bty064.

Scanagatta M, Salmerón A, Stella F. 2019. A survey on Bayesian network structure learning from data. Prog Artif Intell 8(4):425–439, https://doi.org/10.1007/s13748-019-00194-y.

Sohn MB, Li H. 2019. Compositional mediation analysis for microbiome studies. Ann Appl Stat 13(1):661–681, https://doi.org/10.1214/18-AOAS1210.

Strimmer K. 2008. A unified approach to false discovery rate estimation. BMC Bioinformatics 9:303, PMID: 18613966, https://doi.org/10.1186/1471-2105-9-303.

Sur P, Candès EJ. 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. Proc Natl Acad Sci USA 116(29):14516–14525, PMID: 31262828, https://doi.org/10.1073/pnas.1810420116.

Uusitalo L. 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol Modell 203(3–4):312–318, https://doi.org/10.1016/j.ecolmodel.2006.11.033.

Valeri L, Reese SL, Zhao S, Page CM, Nystad W, Coull BA, et al. 2017. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? Epigenomics 9(3):253–265, PMID: 28234025, https://doi.org/10.2217/epi-2016-0145.

VanderWeele TJ. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press.

VanderWeele TJ. 2016. Mediation analysis: a practitioner's guide. Annu Rev Public Health 37:17–32, PMID: 26653405, https://doi.org/10.1146/annurev-publhealth-032315-021402.

VanderWeele TJ. 2017. Outcome-wide epidemiology. Epidemiology 28(3):399–402, PMID: 28166102, https://doi.org/10.1097/EDE.0000000000000641.

VanderWeele TJ, Vansteelandt S. 2014. Mediation analysis with multiple mediators. Epidemiol Methods 2(1):95–115, PMID: 25580377, https://doi.org/10.1515/em-2012-0010.

Vansteelandt S, Daniel RM. 2017. Interventional effects for mediation analysis with multiple mediators. Epidemiology 28(2):258–265, PMID: 27922534, https://doi.org/10.1097/EDE.0000000000000596.

Vittinghoff E, McCulloch CE. 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 165(6):710–718, PMID: 17182981, https://doi.org/10.1093/aje/kwk052.

Wang L, Michoel T. 2017. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. PLoS Comput Biol 13(8):e1005703, PMID: 28821014, https://doi.org/10.1371/journal.pcbi.1005703.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82, PMID: 21167468, https://doi.org/10.1016/j.ajhg.2010.11.011.

Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. 2016. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics 32(20):3150–3154, PMID: 27357171, https://doi.org/10.1093/bioinformatics/btw351.

Zhang J, Wei Z, Chen J. 2018. A distance-based approach for testing the mediation effect of the human microbiome. Bioinformatics 34(11):1875–1883, PMID: 29346509, https://doi.org/10.1093/bioinformatics/bty014.

Zhang W, Xu J. 2017. DNA methyltransferases and their roles in tumorigenesis. Biomark Res 5:1, PMID: 28127428, https://doi.org/10.1186/s40364-017-0081-z.

Zhao Y, Lindquist MA, Caffo BS. 2020. Sparse principal component based high-dimensional mediation analysis. Comput Stat Data Anal 142:106835, https://doi.org/10.1016/j.csda.2019.106835.

Zheng X, Aragam B, Ravikumar P, Xing EP. 2018. DAGs with NO TEARS: continuous optimization for structure learning. [Poster.] In: *Proceedings of the 32nd Conference on Advances in Neural Information Processing Systems*. 2–8 December 2018. San Diego, CA: NIPS, 5740.

Zheng W, van der Laan MJ. 2018. Mediation analysis with time-varying mediators and exposures. In: *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. van der Laan MJ, Rose S, eds. Cham, Switzerland: Springer International publishing, 277–299.