Human Molecular Genetics, 2016, Vol. 25, No. R2 R149–R156

doi: 10.1093/hmg/ddw197
Advance Access Publication Date: 20 July 2016
Invited Review

INVITED REVIEW

Challenges and novel approaches for investigating molecular mediation

R.C. Richmond[1,2], G. Hemani[1,2], K. Tilling[1,2], G. Davey Smith[1,2,†] and C.L. Relton[1,2,†,*]

[1]MRC Integrative Epidemiology Unit, University of Bristol, UK and [2]School of Social and Community Medicine, University of Bristol, UK

*To whom correspondence should be addressed at: Caroline Relton, MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, Bristol, UK BS8 2BN, UK. Tel: +44 (0) 117 33 14028; Fax: +44 (0) 117 33 14052, Email: caroline.relton@bristol.ac.uk

Abstract

Understanding mediation is useful for identifying intermediates lying between an exposure and an outcome which, when intervened upon, will block (some or all of) the causal pathway between the exposure and outcome. Mediation approaches used in conventional epidemiology have been adapted to understanding the role of molecular intermediates in situations of high-dimensional omics data with varying degrees of success. In particular, the limitations of observational epidemiological study including confounding, reverse causation and measurement error can afflict conventional mediation approaches and may lead to incorrect conclusions regarding causal effects. Solutions to analysing mediation which overcome these problems include the use of instrumental variable methods such as Mendelian randomization, which may be applied to evaluate causality in increasingly complex networks of omics data.

Introduction

New technologies permit the genotyping and profiling of gene expression, epigenetics and metabolites, allowing the collection of high-dimensional molecular phenotype data on a large number of individuals. This “omics revolution” (1,2) offers the potential to vastly improve the granularity of measurements related to the processes of normal development and disease pathogenesis.

Recent applications of omics technologies within large-scale population-based studies present new opportunities for identifying novel biomarkers for both risk factors and disease. Furthermore, different forms of omic data can be combined with increasingly complex models (3) and may help to interrogate otherwise opaque networks in confirming observed risk factor and disease associations from observational epidemiology and identifying new ones (4) (Fig. 1).

However, as these molecular intermediates are influenced by both endogenous and exogenous factors and by disease processes, they are prone to the many limitations of observational epidemiological study including confounding, bias and reverse causation (Box 1) (5). We are therefore presented with the challenge of understanding the causal nature of correlations between measures of interest. Statistical methods are required to dissect causal relationships and to construct a causal framework involving molecular intermediates (6,7).

What Is Mediation Analysis and Why Is Understanding Mediation Useful?

A mediator (M) is a variable that is on the causal path from an exposure (E) to an outcome variable (Y). It causes the outcome and is itself caused by the exposure. There are a variety of statistical

[†]The authors wish it to be known that, in their opinion, the last 2 authors should be regarded as joint Senior Authors.

Received: June 6, 2016. **Revised:** June 6, 2016. **Accepted:** June 20, 2016

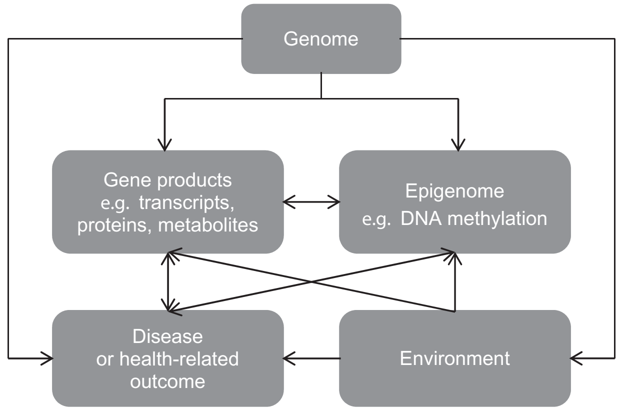

Figure 1. The interplay between genomics, other "omics" and environmental factors in relation to disease or health-related outcomes.

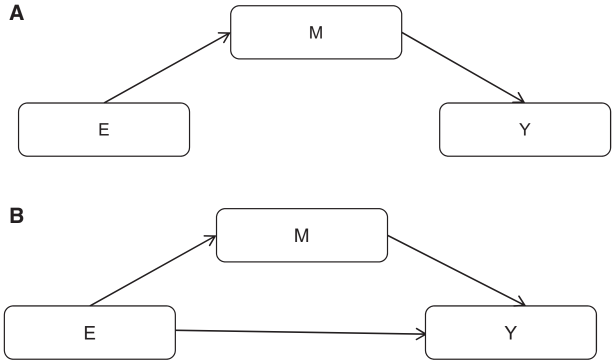

Figure 2. A simplistic representation of mediation. (A) Complete mediation - M is the only mechanism by which E can change Y. (B) Partial mediation - In practice, it is more likely that E has an effect on Y other than those operating by changing M. Mediation aims to partition the total (causal) effect of E on Y into mediated effects (effects that operate by changing the mediator, M) and non-mediated effects.

methods that have been introduced for analysing mediation, from simple regression-based systems and structural equation models to more novel parametric and semi-parametric methods (8), and these have been widely implemented (Fig. 2).

Understanding mediation is useful for identifying potential modifiable risk factors lying between an exposure and an outcome which, when intervened upon, will block (some or all of) the causal pathway between the exposure and outcome. For example, elevated levels of non-fasting remnant and LDL cholesterol levels are modifiable intermediates of cardiovascular disease. These may be intervened upon to alter the downstream risk of cardiovascular disease, when underlying risk factors are either difficult, as in the case of adiposity (9), or indeed impossible to alter, as in the cases of the underlying genetic factors related to cholesterol levels (10).

Mediation approaches have been adapted to understanding the role of molecular intermediates in causal pathways, using high-dimensional omics data (4,11–18). However, these approaches have been applied with varying degrees of success as each approach has different strengths and challenges due to their underlying assumptions.

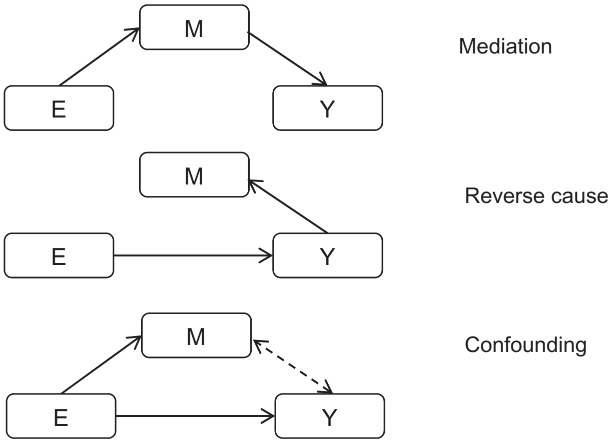

Figure 3. Distinguishing mediation from reverse causation and confounding In a situation of mediation, the effect of the exposure (E) on an outcome (Y) is mediated through an intermediate (M). In a situation of reverse cause, E influences Y which then has an effect on M. In a situation of common cause (confounding), E has an independent effect on both M and Y, so inducing a spurious association between M and Y.

Exposure – Outcome Mediation

One of the most widely cited approaches for evaluating mediation in an epidemiological setting is that originally outlined by Baron and Kenny (19). This regression-based approach may be applied to distinguish a mediated effect of the exposure (E) on an outcome (Y) through an intermediate (M) from both a consequential (reverse cause) effect and a common cause (confounding) effect (Fig. 3), through the application of four tests:

1) E is associated with Y
2) E is associated with M
3) M is associated with Y after adjusting for E
4) E is independent of Y after adjusting for M

The Sobel test may then be used to indicate whether the decrease in the effect of E on Y after adjusting for M is "statistically significant". If this test provides evidence for mediation, the proportion of the effect of E on Y that is mediated by M can be calculated.

While this approach is widely implemented, it is known to be problematic because it is highly dependent on a number of strong assumptions, the measurement characteristics of the variables and on reliable identification of causal effects. Some such often overlooked assumptions are that (i) both Y and M are continuous; (ii) there are no unmeasured confounders of E and Y or of M and Y; (iii) E must not cause a confounder of the M-Y association; (iv) the correct functional form has been specified for each model (e.g. linearity); (v) there are no interactions between E and M on Y; and (vi) there is no measurement error (20). Here, measurement error is the difference between a measured value of E, M or Y and its true value, which could be due to either imperfect measurement (e.g. measuring weight using a standard set of scales) or fluctuating about an underlying "true" value (e.g. day-to-day variation in weight about the individual's underlying average weight), or both. Furthermore, this method can only be used under the assumption of complete mediation as in a situation of partial mediation, the fourth condition will not hold.

Further methods have been developed to allow much more flexible modelling than the traditional Baron and Kenny approach and allow for a more general outcomes framework, distribution-free estimates of mediated effects, interactions and intermediate confounding (20,21). Such methods include linear equations, structural equation models, marginal structural models and G-computation. However, while these approaches offer more flexibility (e.g. allowing non-continuous variables and interactions between E and M in their effect on Y), they also require strong assumptions, specifically related to no measurement error or unmeasured confounding. If the assumptions are not satisfied, these methods may also lead to incorrect inference (22–26).

Nonetheless, some of these approaches have been readily applied in the setting of the molecular mediation without much consideration being given to the underlying assumptions and thus may have led to spurious results and interpretations. For example, large epigenome-wide association studies (EWAS) have identified associations between smoking and DNA methylation (27), and lung cancer and DNA methylation (15). Interestingly, CpG sites in the *AHRR* region have shown the largest signals of differential methylation in both these EWASs. These findings have driven subsequent analyses to investigate whether environmentally modified DNA methylation play an important role in the aetiology of cancer, through the use of mediation analysis.

One recent study used a causal mediation technique of G-computation to assert a mediating role of lower *AHRR* methylation in the association between smoking and lung cancer (15). Strikingly, the mediation analysis applied in this study identified that 32% of the total effect was mediated by differential methylation in the *AHRR* region. However, the study analysts also found that 31% of the total effect was mediated by methylation in a CpG in *F2RL3*, another site implicated in both smoking and lung cancer EWAS. Together, these two sites in *AHRR* and *F2RL3* explained a total of 37% of the total effect, which is lower than the proportion anticipated, given that these two genes are independent and act through different biological pathways.

One potential explanation for these findings is that the association between methylation and lung cancer might just reflect the known causal effect of smoking on lung cancer, as DNA methylation is a strong biomarker for smoking (28). Mediation analysis may lead to a spurious inference due to measurement error, in this instance in the exposure, whereby self-reported smoking is more error-prone than objectively-measured DNA methylation, leading to residual confounding of the intermediate – outcome association.

Genetic Variant – Outcome Mediation

A widely-used approach for establishing causal relationships with molecular intermediates is the causal inference test (CIT) (29). This test builds on the 'Causality Equivalent Theorem' (30) to infer causal indirect effects of a genetic variant on an outcome. It is analogous to the Baron and Kenny approach in its reliance on a series of models to statistically test conditional independencies between covariates in order to distinguish a mediated effect of the genetic variant (G) on an outcome (Y) through an intermediate (M) from a reverse cause and a common cause (pleiotropic) effect (7,29,31). Therefore, replacing E (exposure) in the Baron and Kenny approach with G (genetic variant), the required tests are:

1) G is associated with Y without adjusting for M
2) G is associated with M after adjusting for Y
3) M is associated with Y after adjusting for G
4) G is independent of Y after adjusting for M

This approach has typically been applied to understand the extent to which molecular processes mediate the effect of quantitative trait loci (QTLs) on the risk of a particular disease. By focusing on the assessment of the genetic component of the molecular intermediate, this avoids limitations in the observational epidemiology setting of potential confounding from environmental factors on the intermediate – outcome relationship and also the influence of reverse causation, whereby changes in the outcome may influence the intermediate factor. In addition, some other qualities of genetic variants which make them useful in the causal inference analysis are that they are not influenced by reporting bias and are subject to relatively little measurement error.

By using a genetic variant as a causal "anchor" to dissect causal relationships, the causal inference test has close links with Mendelian randomization (MR) (5,32), a method which will be discussed in more detail later in this review, although with the use of a different modelling approach. Given its ease of application, the causal inference test has been used to evaluate molecular mediation in a range of omics settings (6,11,16,33). However, this test is limited by its emphasis on an "omnibus statistical test" (29) which is reliant on a p-value for asserting a causal effect, rather than providing an estimate for the magnitude and precision of the true causal effect (34).

Furthermore, this approach is also vulnerable to measurement error, which can be in either the mediator or the outcome, in those steps 2) to 4) of the causal inference test (see above) which involve adjustment in the regression models.

In particular, the presence of measurement error can make it hard to delineate a situation of true mediation from that of horizontal pleiotropy (defined in Box 2), whereby the genetic variant influences the outcome through pathways other than those containing the mediator. The problem of incorrectly identifying pleiotropy can be illustrated with an example, whereby a genetic variant associated with cholesterol levels is strongly associated with coronary heart disease even after adjusting for measured cholesterol (35). While the causal inference test would infer from this that cholesterol is not fully mediating the effect of this genetic variant on risk of coronary heart disease, in reality this situation probably arises because the genetic variant represents a life-long elevated risk of cholesterol levels compared with a single, poor measure of cholesterol acting as the mediator in this situation (Table 1).

Measurement error can also lead to erroneous results from the causal inference test due to reverse cause, whereby the genetic variants influence the proposed mediated factor indirectly through an

Table 1. Causal inference test analysing the causal effect of total cholesterol on coronary heart disease (CHD) using the rs72658867 LDLR splice variant SNP as a causal anchor

Condition	P-value
CHD associated with LDLR	$4.33\text{x}10^{-3}$
CHD associated with total cholesterol given LDLR	$<1.00\text{x}10^{-5}$
Total cholesterol associated with LDLR given CHD	$<1.00\text{x}10^{-5}$
LDLR **independent** of CHD given total cholesterol	>0.99
Total cholesterol causes CHD (omnibus P-value)	>0.99

Analysis was performed using cholesterol and CHD data from the Copenhagen General Population Study using 95,275 individuals. Data described here http://www.nature.com/nature/journal/v526/n7571/full/nature14962.html.(36) Analysis performed using the R/cit package and the *cit.bp* function. *P*-values represent the strength of evidence for each condition of the causal inference test. In this situation, three of the four conditions of the causal inference test are satisfied although the fourth is not (*P* >0.99), as *LDLR* is associated with CHD even after adjusting for total cholesterol. The omnibus test therefore selects this largest *P*-value, which in this case is used to reject the hypothesis that total cholesterol causes CHD.

effect of the outcome (Fig. 3). This may occur in a situation where a genetic variant is found to be associated with a proposed mediator but is of unknown biology. Here it is possible that the genetic variant is directly related to the outcome and only indirectly to the proposed mediator through the causal effect of the outcome on the intermediate factor. For example, greater adiposity is known to have a causal effect on levels of the inflammatory biomarker, C reactive protein, but not vice versa (37,38). With adequate sample size and measurement precision, any genetic variant related to body mass index will be related to CRP because of this causal effect (37,38). Furthermore, in the absence of knowledge about the functionality of the genetic variants, true adiposity variants may be assumed to be directly related to CRP levels, which may lead to incorrect causal inference.

Use of Mendelian Randomization for Mediation Analysis

Solutions to analysing mediation which overcome unmeasured or residual confounding, reverse causation and measurement error include the use of instrumental variable methods (8), of which Mendelian randomization (MR) is a form. In MR, genetic variants robustly associated with modifiable exposures are used to infer causality (5,32,39) by serving as instrumental variables which are not associated with various confounders of the exposure-outcome association and are not directly influenced by the outcome of interest (40,41) (Fig. 4).

The assumptions and application of Mendelian randomization analysis were outlined in detail in a recent review (5). In addition, this review outlined how the MR approach may be adapted to the setting of appraising causality of molecular phenotypes. However, with specific reference to gene-outcome mediation analysis, using the MR approach has two potential advantages over the CIT: 1) it allows for a formal test of the direction and magnitude of causality, rather than a p-value driven assessment, and 2) by using the genetic variant as an instrumental variable for the mediator, the correct direction of causality can be inferred even in the presence of measurement error in the mediator, as genetic variants are typically measured with high accuracy and will typically proxy for lifetime differences in exposures (39).

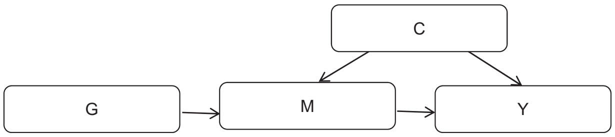

Figure 4. Schematic representation of Mendelian randomization to assess the causal effect of a molecular intermediate Mendelian randomization can be used to test the hypothesis that molecular intermediate M has a causal effect on outcome Y, given that the genetic variant G is associated with the intermediate phenotype of interest, has no association with the outcome except through the intermediate phenotype and is not related to measured or unmeasured confounding factors (C).

Limitations of MR have been discussed in detail in a recent review (5), which also highlighted some methodological developments to overcome these limitations. Specifically with respect to using MR to assess mediation, the main limitations of this approach are low power, potential pleiotropy of the genetic instruments and reverse cause.

While genetic instruments for molecular phenotypes often explain a large proportion of variance in these traits (Box 1), MR studies involving such intermediates are often of low power because of the availability of biological samples and the relative expense of measuring these phenotypes in large enough numbers. One recent means of enhancing power in Mendelian randomization analysis is with the use of a two-sample approach (50,51). This approach is particularly relevant for establishing the causal effect of a molecular intermediate, which only needs to be measured in a subset of individuals with genetic data, and then integrating these gene-exposure estimates with gene-outcome estimates obtained from larger studies to harness power. With respect to the latter, such estimates may be obtained from publicly available summary measures which are increasingly available for many large genome-wide association studies (51).

In situations such as that outlined with respect to evaluating the role of a molecular intermediate (e.g. DNA methylation) in a known exposure-outcome relationship (e.g. smoking-lung cancer), it may be possible to obtain causal estimates from MR studies for all steps in the chain, e.g. from smoking to DNA methylation, and from DNA methylation to lung cancer, in a two-step Mendelian randomization approach (Fig. 5). Here the logic of MR can be extended to interrogate causality of a mediating effect using one genetic instrument to estimate the causal effect of the exposure on DNA methylation, and a separate

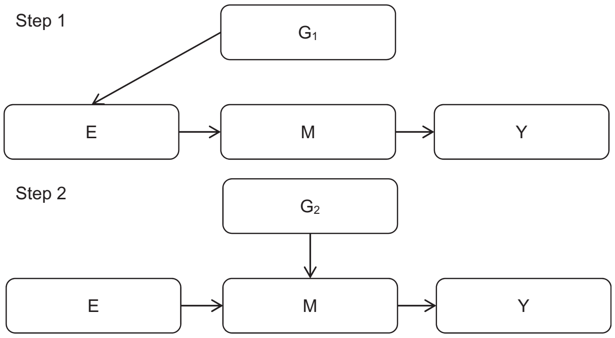

Figure 5. Schematic diagram of two-step Mendelian randomization In Step 1, a genetic variant, G_1, is used to proxy for the environmentally-modifiable exposure of interest, E, to examine how this exposure influences in the intermediate, M, e.g. DNA methylation. In Step 2, a different genetic variant unrelated to the exposure, G_2 is used to proxy for this specific difference in the intermediate, M, and relate this to the outcome of interest, Y.

Box 1. Identifying genetic proxies for ‘omics measures

Developments in genomics have driven the identification of many genetic variants associated with a wide range of exposures which have potential utility in MR analysis (5,52). As molecular intermediates are more proximal to genotype than downstream phenotypes (53) this boosts the statistical power to detect associations compared to more complex traits, as exemplified in genetic association studies.(54–56) Furthermore, studies have identified that many genetic effects on intermediates are highly stable across the life course (55) and between tissues (56). Such stability is useful when these genetic variants (or quantitative trait loci (QTLs)) are being applied as causal anchors. Catalogues of these QTLs are being made freely available.(55,56) independent instrument to estimate the causal effect of DNA methylation on the outcome. As variation in DNA methylation is associated with widespread local (*cis*) genetic variation, this provides the opportunity to use genetic proxies to probe causality between DNA methylation and particular outcomes using MR (5,42–44).

While the two-step MR method was initially posed for the delineation of mediation by specific epigenetic processes between environmental exposures and disease, it may equally be applied to a full range of potential mediators, such as transcriptomics, proteomics and metabolomics (45). In addition, while evidence of association in both steps of the two-step MR framework implies some degree of mediation, in its original form this method did not give a quantitative contribution of the mediator to the causal link explicitly. Extensions of network Mendelian randomization (10,45) allow for the magnitude of the direct and indirect effects to be estimated and can be used to obtain support for a number of testable hypotheses and degrees of association between increasingly complex networks. Such methods will be particularly useful for integrating omics data and challenging the "central dogma" of biological causation (46–49).

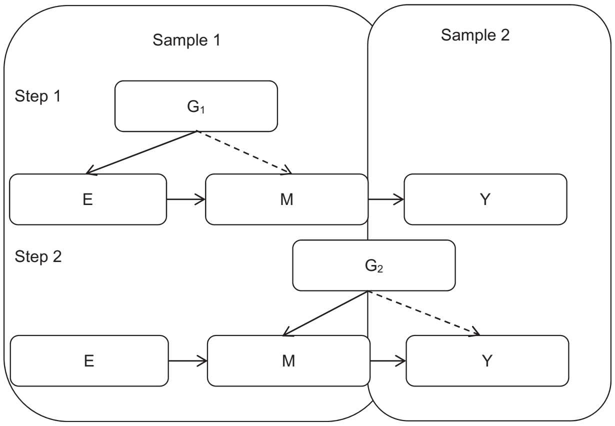

Figure 6. Schematic diagram of two-step, two-sample Mendelian randomization In the smaller Sample 1, the association of the exposure to the intermediate is established using an MR approach (using the exposure-related G_1); and the association of an additional variant (G_2, not related to the exposure) with the same intermediate is established. G_1 and G_2 should be identified in an independent study. In the larger Sample 2, the intermediate is shown to influence the outcome through the use of G_2, which is related to the outcome. N.B. the dotted arrows represent the fact that these genetic variants, G_1 and G_2, influence the intermediate or outcome indirectly through the exposure or intermediate, rather than having a pleiotropic effect. In theory, G_1 would also be found to influence the outcome indirectly through both the exposure and intermediate.

In addition, with regards to asserting mediation in an exposure-outcome setting, the two-step MR approach could be combined with the two-sample approach to powerfully and efficiently examine the extent of mediation in causal networks (5). First, the causal associations of both the exposure on the intermediate and of an independent variant on the intermediate could be established, and then in a larger population-based sample, the genetic associations with the disease outcome delineated (Fig. 6). This gives two-step MR an advantage over traditional mediation approaches which require the exposure, mediator and outcome to be measured in the same subset of individuals.

As with the causal inference test, complexity of associations between omic level intermediates and inadequate biological knowledge of the genetic variants associated with them pose a challenge to Mendelian randomization. Arguably, the biggest challenge to overcome is that of potential pleiotropy of the genetic instrument (Box 2). Approaches which have recently been developed to allow causal effect estimates in the presence of pleiotropy, and which are also particularly relevant to causal inference for molecular mediation, are described in more detail in Box 2.

In situations of reverse causation whereby a genetic variant may be causing the outcome which in turn causes the molecular phenotype, rather than vice versa, bidirectional Mendelian randomization using well characterized genetic variants for both the molecular intermediate and the outcome may be used to distinguish between these causal models (37,38). Alternatively, the use of the Steiger test may be able to provide evidence for the prevailing causal direction, based on the

Box 2. Consequences of pleiotropy and potential solutions for Mendelian randomization analysis

Pleiotropy is the phenomenon by which a genetic variant may affect more than one phenotypic characteristic (57–59). There are two main mechanisms by which pleiotropy occurs: 1) a single locus influences a cascade of events e.g. a variant influences a particular molecular intermediate which causes perturbation in another phenotype (vertical pleiotropy) 2) a single locus directly influences multiple phenotypes e.g. via more than one post-transcriptional process (horizontal pleiotropy). Vertical pleiotropy has also been referred to as 'mediated pleiotropy' (60) and is the very essence of the Mendelian randomization approach, in which the downstream effects of a phenotype are estimated through the use of genetic variants that relate to that phenotype. On the other hand, horizontal pleiotropy is more problematic as it violates the assumption that the genetic instrument has no association with the outcome except through the intermediate phenotype being investigated and its presence can bias Mendelian randomization effect estimates.

One potential means of investigating potential pleiotropy is with the use of multiple genetic instruments. With an amassing number of independent instruments, it would be increasingly improbable that they would result in the same conclusion regarding a causal effect if they were all pleiotropic variants. In particular, the finding that all genetic variants have an effect on the outcome to the extent expected given their effect on the exposure can be used to support an assertion of no horizontal pleiotropy (5,61). If some variants deviate from this proportional effect, then the extent of directional pleiotropy can be investigated with the use of an approach known as "MR Egger" (62), and further derivatives including a weighted median approach (63), which can also be used to provide valid causal estimates even in the presence of pleiotropy.

A further approach used to separate independent effects of risk factors when multiple phenotypes are correlated with a particular genetic variant or set of variants is with multi-variable Mendelian randomization analysis (64) which provides a more promising alternative to analyses which attempt to isolate the effects of correlated phenotypes using regression-based approaches (65).

estimated variance explained by the SNPs in the molecular phenotype and the outcome, as long as measurement error in the molecular phenotype is lower than the product of the measurement error in the outcome and the causal correlation between the molecular phenotype and the outcome (66).

Conclusions

Mediation analysis and causation are linked concepts and the former cannot be successfully applied without some consideration given to the latter (67). Care must be taken when conducting mediation analysis in making sure that the assumptions made in the causal model are justified (68). In particular, the assumptions of no unobserved confounding and no measurement error are often made in both conventional epidemiology (exemplified in the Baron and Kenny approach) and computational systems biology (69) (exemplified in the Causal Inference Test) which vitiate many of the models attempting to utilize measured phenotypes and which therefore can lead to erroneous inferences or conclusions being drawn.

Mendelian randomization approaches hold promise for investigating mediation without relying on such strict assumptions (10,43,45). Furthermore, such techniques focus on minimizing reliance on correlation statistics and maximizing quantitative causal interpretation by using genetic variants as causal anchors in situations of mediation. These approaches are increasingly being used in an automated, hypothesis-free fashion (51) and may be used to integrate multiple tiers of omics data in a causal framework. This offers potential for identifying novel risk factors and modifiable targets for intervention.

While MR is a helpful solution in some circumstances when considering molecular mediation, it is not a global solution and its application can be restricted due to the availability of genetic instruments. In addition, while MR makes less strict assumptions about confounding and measurement error, it does make other assumptions (39,70,71) which should be explored through sensitivity analysis (62,63). We would also advocate an integration of causal inference approaches and the triangulation of findings in the domain of high-dimensional molecular data to improve the identification of causal mediating effects (6).

Acknowledgements

We would like to thank Gemma Sharp and Emma Anderson for their valuable comments on an earlier draft of this paper. We would also like to acknowledge Børge Nordestgaard for allowing us to include data from the Copenhagen General Population Study.

Conflict of Interest statement. None declared.

Funding

All authors work in a Unit that receives funds from the University of Bristol and the UK Medical Research Council (MC_UU_12013/1, MC_UU_12013/2 and MC_UU_12013/9). This work was also supported by CRUK (grant number C18281/A19169) and the ESRC (grant number ES/N000498/1). Funding to pay for Open Access publication charges for this article was provided by CRUK, MRC and ESRC.

References

1. Vineis, P., van Veldhoven, K., Chadeau-Hyam, M. and Athersuch, T.J. (2013) Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ Mol Mutagen.*, **54**, 461–467.
2. Khoury, M.J. (2014) A primer series on -omic technologies for the practice of epidemiology. *Am. J. Epidemiol.*, **180**, 127–128.
3. Civelek, M. and Lusis, A.J. (2014) Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, **15**, 34–48.
4. Georgiadis, P., Hebels, D.G., Valavanis, I., Liampa, I., Bergdahl, I.A., Johansson, A., *et al.* (2016) Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. *Scientific Reports*, **6**, 20544.
5. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.*, **23**(R1):R89–98.
6. Shin, S.Y., Petersen, A.K., Wahl, S., Zhai, G., Romisch-Margl, W., Small, K.S., *et al.* (2014) Interrogating causal pathways linking genetic variants, small molecule metabolites, and circulating lipids. *Genome Med.*, **6**, 25.
7. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.
8. Preacher, K.J. (2015) Advances in Mediation Analysis: A Survey and Synthesis of New Developments. *Annu. Rev. Psychol.*, **66**, 825–852.
9. Davey Smith, G. (2016) A fatter, healthier but more unequal world. *Lancet*, **387**, 1349–1350.
10. Varbo, A., Benn, M., Davey Smith, G., Timpson, N.J., Tybjaerg-Hansen, A. and Nordestgaard, B.G. (2015) Remnant Cholesterol, Low-Density Lipoprotein Cholesterol, and Blood Pressure as Mediators From Obesity to Ischemic Heart Disease. *Circ. Res.*, **116**, 665–673.
11. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., 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**, 142–147.
12. Kupers, L.K., Xu, X.L., Jankipersadsing, S.A., 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**, 1224–1237.
13. Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W.H., *et al.* (2015) Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.*, **47**, 1282–1282.
14. Liang, L.M., Willis-Owen, S.A.G., Laprise, C., Wong, K.C.C., Davies, G.A., Hudson, T.J., *et al.* (2015) An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature*, **520**, 670–U188.
15. Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., *et al.* (2015) Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun*, **6**, 10192.
16. Hong, X.M., Hao, K., Ladd-Acosta, C., Hansen, K.D., Tsai, H.J., Liu, X., *et al.* (2015) Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat. Commun*, **6**, 6304.
17. Assi, N., Fages, A., Vineis, P., Chadeau-Hyam, M., Stepien, M., Duarte-Salles, T., *et al.* (2015) A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. *Mutagenesis*, **30**, 743–753.
18. Bellavia, A., Urch, B., Speck, M., Brook, R.D., Scott, J.A., Albetti, B., *et al.* (2013) DNA Hypomethylation, Ambient

Particulate Matter, and Increased Blood Pressure: Findings From Controlled Human Exposure Experiments. *J. Am. Heart. Assoc.*, **2**, e000212.

19. Baron, R.M. and Kenny, D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc Psychol.*, **51**, 1173–1182.
20. Hayes, A.F. (2009) Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. *Commun. Monogr.*, **76**, 408–420.
21. Imai, K., Keele, L. and Tingley, D. (2010) A General Approach to Causal Mediation Analysis. *Psychol. Methods*, **15**, 309–334.
22. VanderWeele, T.J., Valeri, L. and Ogburn, E.L. (2012) The Role of Measurement Error and Misclassification in Mediation Analysis Mediation and Measurement Error. *Epidemiology*, **23**, 561–564.
23. le Cessie, S., Debeij, J., Rosendaal, F.R., Cannegieter, S.C. and Vandenbroucke, J.P. (2012) Quantification of Bias in Direct Effects Estimates Due to Different Types of Measurement Error in the Mediator. *Epidemiology*, **23**, 551–560.
24. Blakely, T., McKenzie, S. and Carter, K. (2013) Misclassification of the mediator matters when estimating indirect effects. *J. Epidemiol. Community Health*, **67**, 458–466.
25. Tchetgen-Tchetgen, E.J. and Lin, S.H. Robust estimation of pure/natural direct effects with mediator measurement error. Harvard Univ Biostat Work Pap 152. 2012;http://biostats.bepress.com/harvardbiostat/paper152.
26. VanderWeele, T.J. (2013) A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, **24**, 224–232.
27. Zeilinger, S., Kuhnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., *et al.* (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *Plos One*, **8**, e63812.
28. Philibert, R., Hollenbeck, N., Andersen, E., Osborn, T., Gerrard, M., Gibbons, F.X., *et al.* (2015) A quantitative epigenetic approach for the assessment of cigarette consumption. *Front Psychol*, **6**, 656.
29. Millstein, J., Zhang, B., Zhu, J. and Schadt, E.E. (2009) Disentangling molecular relationships with a causal inference test. *BMC Genet*, **10**, 23.
30. Chen, L.S., Emmert-Streib, F. and Storey, J.D. (2007) Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.*, **8**, R219.
31. Millstein, J., Chen, G.K. and Breton, C.V. (2016) cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*, pii: btw135.
32. Davey Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, **32**, 1–22.
33. Koestler, D.C., Chalise, P., Cicek, M.S., Cunningham, J.M., Armasu, S., Larson, M.C., *et al.* (2014) Integrative genomic analysis identifies epigenetic marks that mediate genetic risk for epithelial ovarian cancer. *Cancer Res.*, **74**.
34. Sterne, J.A. and Davey Smith, G. (2001) Sifting the evidence-what's wrong with significance tests? *BMJ*, **322**, 226–231.
35. Khera, A.V., Won, H.H., Peloso, G.M., Lawson, K.S., Bartz, T.M., Deng, X., *et al.* (2016) Diagnostic yield of sequencing familial hypercholesterolemia genes in patients with severe hypercholesterolemia. *J. Am. Coll. Cardiol.*, **67**(22):2578–2589.
36. The UK10K Consortium UK (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.
37. Welsh, P., Polisecki, E., Robertson, M., Jahn, S., Buckley, B.M., de Craen, A.J.M., *et al.* (2010) Unraveling the Directional Link between Adiposity and Inflammation: A Bidirectional Mendelian Randomization Approach. *J. Clin. Endocr. Metab.*, **95**, 93–99.
38. Timpson, N.J., Nordestgaard, B.G., Harbord, R.M., Zacho, J., Frayling, T.M., Tybjaerg-Hansen, A., *et al.* (2011) C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int. J. Obes.*, **35**, 300–308.
39. Davey Smith, G. and Ebrahim, S. (2004) Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidemiol.*, **33**, 30–42.
40. Didelez, V. and Sheehan, N. (2007) Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.*, **16**, 309–330.
41. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. and Davey Smith, G. (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.*, **27**, 1133–1163.
42. Relton, C.L. and Davey Smith, G. (2010) Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med*, **7**, e1000356.
43. Relton, C.L. and Davey Smith, G. (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.*, **41**, 161–176.
44. Kirkbride, J.B., Susser, E., Kundakovic, M., Kresovich, J.K., Davey Smith, G. and Relton, C.L. (2012) Prenatal nutrition, epigenetics and schizophrenia risk: can we test causal effects? *Epigenomics*, **4**, 303–315.
45. Burgess, S., Daniel, R.M., Butterworth, A.S., Thompson, S.G. and Consortium, E.-I.A. (2015) Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidemiol.*, **44**, 484–495.
46. Zhou, X., Li, D.F., Zhang, B., Lowdon, R.F., Rockweiler, N.B., Sears, R.L., *et al.* (2015) Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.*, **33**, 345–346.
47. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
48. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
49. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
50. Pierce, B.L. and Burgess, S. (2013) Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *Am. J. Epidemiol.*, **178**, 1177–1184.
51. Haycock, P.C., Burgess, S., Wade, K.H., Bowden, J., Relton, C. and Davey Smith, G. (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am. J. Clin. Nutr.*, **103**, 965–978.
52. Brion, M.-J.A., Benyamin, B., Visscher, P.M. and Davey Smith, G. (2014) Beyond the single SNP: emerging developments in Mendelian randomization in the "Omics" era. *Curr. Epidemiol. Rep.*, **1**, 228–236.
53. Ala-Korpela, M., Kangas, A.J. and Soininen, P. (2012) Quantitative high-throughput metabolomics: a new era in epidemiology and genetics. *Genome Med.*, **4**, 36.

54. Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikainen, L.P., *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.*, **44**, 269–276.
55. Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., *et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.*, **17**, 61.
56. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
57. Gruneberg, H. (1938) An analysis of the “pleiotropic” effects of a new lethal mutation in the rat (Mus norvegicus). *Proc. R. Soc. Ser. B-Bio.*, **125**, 123–144.
58. Hodgkin, J. (1998) Seven types of pleiotropy. *Int. J. Dev. Biol.*, **42**, 501–505.
59. Wagner, G.P. and Zhang, J.Z. (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.*, **12**, 204–213.
60. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.
61. Ference, B.A., Yoo, W., Alesh, I., Mahajan, N., Mirowska, K.K., Mewada, A., *et al.* (2012) Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A Mendelian randomization analysis. *J. Am. Coll. Cardiol.*, **60**, 2631–2639.
62. Bowden, J., Davey Smith, G. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.*, **44**, 512–525.
63. Bowden, J., Davey Smith, G., Haycock, P.C. and Burgess, S. (2016) Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.*, **40**, 304–314.
64. Burgess, S. and Thompson, S.G. (2015) Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.*, **181**, 251–260.
65. Davey Smith, G. and Phillips, A.N. (1992) Confounding in epidemiological studies: why “independent” effects may not be all they seem. *BMJ*, **305**, 757–759.
66. Steiger, J.H. (1980) Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, **87**, 245–251.
67. Wright, S. (1920) Correlation and causation Part I. Method of path coefficients. *J. Agric. Res.*, **20**, 0557–0585.
68. Kenny, D.A. (2008) Reflections on mediation. *Organ. Res. Methods*, **11**, 353–358.
69. Lagani, V., Ball, G., Tegner, J. and Tsamardinos, I. Probabilistic computational causal discovery for systems biology In: Geris L, Gomez-Cabrero D, editors. *Uncertainty in Biology, A Computational Modeling Approach. Studies in Mechanobiology, Tissue Engineering and Biomaterials*. **17**. New York City: Springer; 2015.
70. VanderWeele, T.J., Tchetgen Tchetgen, E.J., Cornelis, M. and Kraft, P. (2014) Methodological challenges in mendelian randomization. *Epidemiology*, **25**, 427–435.
71. Pierce, B.L. and VanderWeele, T.J. (2012) The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *Int. J. Epidemiol.*, **41**, 1383–1393.