Published in final edited form as:

J Hum Genet. 2018 March ; 63(3): 339–348. doi:10.1038/s10038-017-0385-8.

Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer

Dongyan Wu^{1,2}, Haitao Yang³, Stacey J. Winham⁴, Yanina Natanzon⁴, Devin C. Koestler⁵, Tiane Luo¹, Brooke L. Fridley⁶, Ellen L. Goode⁴, Yanbo Zhang¹, and Yuehua Cui^{1,7,*}

¹Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi 030001, PR China

²Department of Medical Affairs, Yanqing Hospital of Peking University Third Hospital, Beijing 102100, PR China

³Division of Health Statistics, School of Public Health, Hebei Medical University, Shijiazhuang 050017, PR China

⁴Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, MN 55905, USA

⁵Department of Biostatistics, The University of Kansas Medical center, Kansas City, KS 66160, USA

⁶Department of Biostatistics & Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA

⁷Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

Abstract

Epigenetic factors and consumption of alcohol, which suppresses DNA methylation, may influence the development and progression of epithelial ovarian cancer (EOC). However, there is a lack of understanding whether these factors interact to affect the EOC risk. In this study, we aimed to gain insight into this relationship by identifying leukocyte-derived DNA methylation markers acting as potential mediators of alcohol-associated EOC. We implemented a causal inference test (CIT) and the VanderWeele and Vansteelandt multiple mediator model to examine CpG sites that mediate the association between alcohol consumption and EOC risk. We modified one step of the CIT by adopting a high-dimensional inference procedure. The data were based on 196 cases and 202 age-matched controls from the Mayo Clinic Ovarian Cancer Case-Control Study.

Implementation of the CIT test revealed two CpG sites (cg09358725, cg11016563), which represent potential mediators of the relationship between alcohol consumption and EOC case-control status. Implementation of the VanderWeele and Vansteelandt multiple mediator model further revealed that these two CpGs were the key mediators. Decreased methylation at both CpGs was more common in cases who drank alcohol at the time of enrollment versus those who did not.

*Corresponding author: cui@stt.msu.edu.

Conflict of interest

The authors declare no conflict of interests.

cg11016563 resides in *TRPC6* which has been previously shown to be overexpressed in EOC. These findings suggest two CpGs may serve as novel biomarkers for EOC susceptibility.

Keywords

Casual inference test; Epigenetic marker; Genetic association; High-dimensional testing; Mediation analysis

Introduction

As the country's fifth most common cause of cancer mortality in women, epithelial ovarian cancer (EOC) is the leading cause of gynecologic cancer death in the United States¹. The most common personal factors associated with increased risk of EOC include nulliparity or older age at first birth (> 35 years), family history (such as due to *BRCA1* and *BRCA2* mutation or Lynch syndrome). On the other hand, oophorectomy, younger age at pregnancy and first birth (< 25 years), the use of oral contraceptives, or breast-feeding are associated with a reduced risk of EOC¹. Although environmental exposures such as drinking alcohol could play a role on the development of EOC risk, no consistent result has been produced on an association between alcohol consumption and EOC risk. A meta-analysis by Bagnardi et al.² showed that a higher daily alcohol intake (100g/day) is a risk factor for EOC. While the majority of studies found no association^{3–8}, some identified that wine consumption is associated with a reduced risk of EOC^{9, 10} or of endometrioid EOC⁶.

Epigenetic mechanisms such as histone modifications and DNA methylations (DNAm) can lead to functional modifications of the genome by regulating gene expression changes which consequently influence the phenotypic outcome. They are typically not accompanied by nucleotide changes, but rather the structural changes in the underlying DNA sequence. It has been commonly recognized that DNAm patterns can be influenced by environmental factors, thus can change over time¹¹. Such a dynamic changing mechanism mediated by environmental stimuli regulates gene transcription, and has been thought to play a key role in adaptation of the genome to environmental changes. In a study of EOC, Cvetkovic¹² showed that DNAm alterations are an early step in carcinogenesis and could represent a mechanism of disease. Increasing evidence shows that alcohol consumption can induce epigenetic changes¹³. For example, Philibert et al.¹⁴ found that recent alcohol use is associated with widespread changes in DNAm in women, bringing the necessity for further study to confirm these findings and determine their relationship to somatic function. While much recent work has been focused on investigating associations between DNAm and EOC risk^{15–17}, few studies have examined the role of DNAm-associated environmental factors, and none on the role of alcohol consumption on EOC pathophysiology. Because of the findings about the relationship between alcohol consumption and EOC risk, and the known suppression of alcohol on DNAm mechanisms, we conducted a rigorous assessment of potential mechanisms underlying the potential relationship of alcohol consumption with EOC that may be mediated by DNAm.

Identifying the mediating effect of DNAm on EOCs associated with alcohol use is a focus of environmental epigenetics, which is a field for studying epigenetic mediation of exposure-

response relationships¹⁸. Given the empirical evidence, our hypothesis is that DNAm may act as an important causal mediator of alcohol consumption on EOC development. For causal inference, causal mediation modeling is a rigorous approach to identify and explain the mechanism by which an independent variable (e.g., alcohol consumption) affects the outcome (e.g., EOC case-control status) via a third explanatory variable, known as a mediator variable (e.g., DNAm). Specifically, using leukocyte-based epigenetic data collected on a total of 196 EOC cases and 202 controls enrolled in the Mayo Clinic Ovarian Cancer Case-Control Study, we aimed to explore the differences of alcohol consumption by EOC case-control status and further quantify whether such differences may be causally mediated through DNAm.

In this work, we adopted the causal inference test (CIT)¹⁹ to understand the causal mediation mechanism. In the CIT procedure, a series of regression models can be fitted and p-values can be obtained to make statistical inference. The original CIT procedure was developed for low dimensional mediators. In our analysis, the mediators (CpG sites) are in the order of 20K. Such high dimensional data can introduce spurious correlations in addition to the true correlations among the CpG sites²⁰. When high correlations occur, a marginal regression as implemented by the original CIT procedure could potentially inflate the type I errors. To overcome this problem, we modified the original CIT procedure by adopting a high-dimensional testing procedure²¹ which is theoretically justified for high dimensional data. This modification improves the original CIT for high-dimensional data and offers an alternative strategy for CIT.

Materials and Methods

Study population and risk factor data

Study participants were women of European ancestry who were invasive EOC cases and controls one-to-one matched with EOC cases on the basis age (within 1-year) who provided pre-treatment blood as a source of DNA for leukocyte-derived DNAm analysis. Readers are referred to Koestler et al.²² for a detailed description of this study population which included 214 cases and 214 controls. Participants were between the ages of 27 and 91 enrolled between year 1999 and year 2007 in the Mayo Clinic Ovarian Cancer Case-Control Study²³. Cases were enrolled within a year of diagnosis (median [range] days between diagnosis and consent: 2 [0–170] and mean (SD): 4.15 (12.7)). Information on alcohol use was obtained via written questionnaire asking “Do you currently drink alcoholic beverages?”, and was missing on 18 cases and 12 controls. Since alcohol consumption was a key variable in our analysis, we therefore focused on 196 cases and 202 controls.

DNA methylation data

Leukocyte-derived DNA was assayed with the Illumina Infinium HumanMethylation27 Beadchip platform and underwent quality control procedures at the Mayo Clinic Molecular Genome Facility (Rochester, MN). Readers are referred to Koestler et al.²² for a detailed description of the data processing procedures. The methylation level of each CpG locus was calculated in GenomeStudio Methylation module (V.1.9.0) and was scored as beta values (β), ranging from 0 (unmethylated) to 1 (methylated). DNAm values were logit-transformed

$(\log(\frac{\beta}{1-\beta}))$ to obtain the M-value for each CpG locus. A prior study²⁴ shows that M-value is more statistically accepted than beta value for the differential analysis of methylation levels. Thus, the M-values were used for the analysis and the corresponding beta values were used when reporting the results. After normalization and adjusting for any batch or plate effects, we utilized beta values and M-values for 25,926 CpG loci.

Leukocyte sub-types composition data

Koestler et al.²² shows that heterogeneity in the distribution of white blood cell types could be a key confounder when examining the association between DNAm and EOC status. Koestler et al.²² examined the difference in cell type proportions between ovarian cases and controls and measured the leukocyte sub-types composition for each of the study samples. They found that granulocyte fractions were higher in EOC cases, while CD8+ T lymphocyte and CD4+ T lymphocyte, and, to a lesser extent B cell and NK cell fractions, were lower in EOC cases compared to controls. A mixture deconvolution method by Houseman et al.²⁵ was applied to estimate the expected difference in cell type proportions between EOC cases and controls, and to predict the proportion of the leukocyte sub-types for each of the study samples. The DNAm signatures were finally corrected across different leukocyte subtypes, as well as the differences in white blood cell proportions by EOC case-control status²².

Clinical data

Of the EOC cases ($n = 196$), 131 had tumors of high-grade serous histology (67%), 45 tumors were endometrioid (23%), 10 were clear cell (5%), 4 were mucinous (2%) and 6 (3%) were other/unknown. Table 1 summarizes lifestyle and demographic characteristics of the study population and is similar to those of our original DNAm report²². Baseline characteristics are described as mean (SD) for continuous variables and number (%) for categorical variables. The Student *t*-test or Chi-square test were used for comparisons between groups for continuous or categorical variables, respectively. As can be seen in the table, alcohol use was significantly lower for cases compared to controls (although we cannot rule out whether this represents a risk factor or reflects changes due to disease status because the questionnaire was given at time of diagnosis). Next, we considered a causal inference test to assess how alcohol consumption modulates DNAm variation which further affect EOC case-control status.

Statistical Analyses

The causal mediation model—In casual inference, mediation modeling is a rigorous approach to identify and explain the mechanism by which an exposure (e.g., alcohol consumption) operates via another explanatory variable, known as a mediator variable (e.g., DNAm), to affect an outcome (e.g., EOC). Consider an environmental exposure (L), a potential mediator (G), and a disease trait (T). Figure 1 demonstrates the components of a mediation analysis: L, G and T. The effect of L on T is coined as the direct effect, while the path from L to G then to T is called the indirect effect. A complete mediation is observed if L affects T only through G, otherwise, a partial mediation is concluded²⁶. Mediation analysis has been broadly applied in social analyses and epidemiological studies. Different

from a traditional mediation analysis, our interest is to assess the causal effect of L on T mediated through G.

Statistical testing procedures—We implemented a causal inference test (CIT)¹⁹, which simultaneously evaluates multiple hypotheses known to be consistent with causal mediation. In CIT, four conditions are evaluated: 1) L is associated with T; 2) G is associated with T conditional on L; 3) L is associated with G conditional on T; and 4) L is independent of T conditional on G. A p-value can be computed for each condition based on a particular model, for example, a logistic or linear regression model depending on the distribution of the response. Then a maximum p-value is used to declare significance. This is similar to an omnibus hypothesis test for testing the intersection of the component rejection regions, termed as an ‘intersection/union test’ (IUT). The omnibus null hypothesis is defined as that at least one component null hypothesis is true.

We implemented the four step procedure corresponding to the four conditions of CIT as illustrated in Fig 2. All models were adjusted for the effects of estimated differential leukocyte cell counts, age (continuous), current smoking status (yes vs. no), study enrollment year (1999–2002, 2003, 2004, 2005, and 2006–2007), location of residence (MN vs. other), parity and age at first birth (nulliparous, 1–2 at ≤ 20 yrs, 1–2 at > 20 yrs, 3+ at ≤ 20 yrs, and 3+ > 20 yrs), and the first principal component representing within-European population sub-structure. All analyses were performed using R statistical software v3.3.2. In Step 1, we fit a logistic regression model to examine the association between alcohol use and EOC status, adjusting for other covariates’ effects. The odds ratio (OR) and its 95% CI were also recorded.

In Step 2, we evaluated the association between CpG sites and EOC status conditional on alcohol use, while adjusting for other covariates’ effects. Since we evaluated the effects of thousands of CpG sites, an optimal way is to fit all the CpG sites into one single multiple regression model. This, however, is infeasible as the number of variables is larger than the sample size. Alternatively, one can test each CpG site one at a time, then do a multiple testing correction using a method such as FDR. However, when fitting a regression models of case-control status with each methylation site conditional on alcohol assumption, this marginal inference procedure could potentially inflate the type I errors and cannot control the testing size if the CpG sites are correlated (true or spurious correlation). To get around this, we adopted the following high-dimensional testing method as an alternative to thousands of regression models. Penalized regression methods for high-dimensional variable selection have been extensively studied in statistics such as the famous LASSO method²⁷ and its various extensions, assuming the number of non-zero regression coefficients is sparse. However, such methods can only estimate the regression coefficients and cannot quantify the uncertainty of the estimated coefficients. Hence, they cannot be used for inference when the goal is to get p-values of the regression coefficients. Dezeure et al.²¹ did a comprehensive review of the methods for high-dimensional testing. Here we implemented a high dimensional inference procedure to test the regression coefficient of an individual CpG site in a high-dimensional regression setup. The method is developed under the generalized linear model setup, de-sparsifying the LASSO regression coefficients, and further quantify the uncertainty of an estimate. Original LASSO assumes sparsity of the

regression coefficients in a high-dimensional setting, i.e., many regression coefficients are zeros. By de-sparsifying the LASSO coefficients, one can reduce the estimation bias and obtain the asymptotic normality of the regression estimates. Specifically, we first did a LASSO type penalized logistic regression against all the CpG sites (total p CpG sites) with EOC status (denoted as Y) as the response, while adjusting for other covariates' effects as shown in model (1),

$$\text{logit}(p) = \alpha_{01} + \gamma_1 \text{Alcohol} + \delta_1^T \text{Covariate} + \sum_{j=1}^p \beta_j \text{CpG}_j. \quad (1)$$

Since the LASSO estimates $\hat{\beta}_j (j = 1, \dots, p)$ are biased, a de-sparsifying step is needed to obtain the bias-corrected regression coefficient. following van de Geer et al.²⁸, we could get a p-value for each CpG site based on the asymptotic normality of the de-sparsified LASSO estimates. Please see the Appendix for the details about the de-sparsified LASSO estimates. Compared with the coefficient estimated by fitting CpG sites one at a time, this new method gave the partial coefficient of a CpG site by holding other CpG sites constant, while adjusting for other covariates effects. In addition to provide meaningful coefficients' estimates, we could quantify the uncertainty of the estimate and got a p-value for testing each coefficient. After obtaining the p-values for all the CpG sites, we corrected for multiple testing by using the Bonferroni correction.

In the third step, the identified significant methylation CpGs (P-value < 0.05) in Step 2 were subsequently examined based on their association with alcohol use, conditional on EOC status. A linear regression model was fit with methylation as the response and EOC status and alcohol as the predictors, while adjusting for other covariates' effects (see model (2)). We tested $H_0: \eta_j = 0$. The p-value tells the strength of the association between the j th CpG site and alcohol consumption, conditional on the EOC status.

$$\text{CpG}_j = \alpha_{02} + \gamma_2 \text{EOC} + \delta_2^T \text{Covariates} + \eta_j \text{Alcohol} \quad (2)$$

The final step involved an examination to check if EOC is independent of alcohol use conditional on those k significant methylation CpGs identified in Step 3. The following logistic regression model (3) was fit for this purpose. We tested $H_0: \gamma_3 = 0$ vs $H_a: \gamma_3 \neq 0$. This test is different from a usual testing problem and the null distribution is estimated using a bootstrap type approach (see Millstein et al.¹⁹ for a detailed description). A p-value of less than 0.05 indicates independence. The model notation for this step is given as $L \perp T|G$.

$$\text{logit}(p) = \alpha_{03} + \gamma_3 \text{Alcohol} + \delta_3^T \text{Covariate} + \sum_{j=1}^k \lambda_j \text{CpG}_j \quad (3)$$

The VanderWeele and Vansteelandt multiple mediator model²⁹ was used for validation of the CIT results. In this model, causal inference is further extended to settings in which multiple

mediators may be of interest. One analytic approach based on regression is proposed to estimate the effect mediated through multiple mediators. The model estimates odds ratios (ORs) of the total, direct and indirect effects by combining logistic and linear regressions. Here we applied the multiple mediator model to evaluate the total effect, direct effect of alcohol consumption on EOC status as well as the indirect effect of alcohol consumption on EOC status mediated through multiple methylation change. The results were described by the total, direct and indirect ORs and the 95% confidence intervals (CIs). The log of ORs for the total effect (TE), controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) (Readers are referred to VanderWeele et al.²⁹ for a detailed description of these effects) are given by:

$$\begin{aligned}\log(OR_{CDE}) &= \theta_1 \\ \log(OR_{NDE}) &\approx \theta_1 \\ \log(OR_{NIE}) &\approx \beta_1^{(1)}\theta_2^{(1)} + \dots + \beta_1^{(K)}\theta_2^{(K)} \\ OR_{TE} &= OR_{NIE} \times OR_{NDE}\end{aligned}$$

where $\beta_1^{(1)}\dots\beta_1^{(K)}, \theta_2^{(1)}\dots\theta_2^{(K)}, \theta_1$ were estimated by the following two models:

$$\text{logit}(p) = \theta_0 + \theta_1 \text{Alcohol} + \theta_2^{(1)} \text{CpG}^{(1)} + \dots + \theta_2^{(K)} \text{CpG}^{(K)} + \theta_3^T \text{Covariates} \quad (4)$$

$$\text{CpG}^{(i)} = \beta_0^{(i)} + \beta_1^{(i)} \text{Alcohol} + \beta_2^{(i)T} \text{Covariates} \quad (i = 1, \dots, K) \quad (5)$$

In model (4), $\theta_1, \theta_2^{(i)} (i = 1, \dots, K)$ represent the coefficients of alcohol use and the i th DNAm respectively; θ_3 is a vector of coefficients for other covariates. In model (5), $\beta_1^{(i)}$ represents the coefficient of alcohol use, and $\beta_2^{(i)}$ is a vector of coefficients for other covariates. Model (5) is a series of linear regression models for the significant DNAm loci identified from the previous CIT method. Bootstrapping with 1,000 replicates was used to estimate the 95% CI for $\beta_1^{(i)}$.

Results

Identifying CpG dinucleotides that mediate the relationship between alcohol consumption and EOC status

We initially fit a logistic regression model, which modeled the EOC status as a function of alcohol consumption, while adjusting for other covariate effects such as differential leukocyte cell counts, and study enrollment year (OR=0.336; 95%CI [0.169, 0.651]; p-value=0.001). This shows that alcohol consumption was inversely associated with EOC status in these data. In an attempt to identify CpGs where DNAm might be associated with EOC conditional on alcohol consumption, Step 2 results show that 3,055 out of 25,926

(11.8%) CpGs (FDR q -value <0.05) were potentially associated with EOC status, conditional on alcohol consumption (see Fig. 3). We then fit a linear regression model expressed in model (2) to discern the relationship between these DNAm sites and alcohol consumption, conditional on EOC status (Step 3 in the CIT procedure). A total of 61 CpGs (p -value <0.05) were identified as potentially significantly associated with alcohol consumption conditional on EOC status (see Fig. 4). Note that there is no need to control the familywise error rate at this step, since the hypothesis is to assess the association of DNAm at each CpG site with alcohol consumption separately.

We then examined whether EOC status was independent of alcohol consumption conditional on those 61 DNAm loci (Step 4 in the CIT procedure). We identified two potential CpGs with statistical significance ($\mathbf{L\perp T|G}$, p -value < 0.05), which are shown in Table 2.

We next applied the VanderWeele and Vansteelandt multiple mediator model for a validation of the results obtained by using the CIT procedure. The results shown in Table 3 indicate that the total and direct effect of alcohol consumption on EOC status as well as the indirect effect mediated through multiple mediators (cg09358725 and cg11016563) are statistically significant, which represent a potential DNAm-mediated relationship between alcohol consumption and EOC risk. Also, after exploring the relationship between alcohol consumption and DNAm as it relates to EOC status, the estimated OR for the alcohol-EOC association was substantially increased after adjusting for these two CpGs simultaneously, suggesting that these CpGs act as potential mediators (see Fig. 5 for a comparison). Thus, alcohol use may associate with EOC status in these data by regulating CpG-specific DNAm patterns. As one reviewer pointed out, VanderWeele and Vansteelandt's method treats multiple mediators as joint mediators, hence it will be impossible to weigh the relative importance of the two mediators (cg09358725 and cg11016563). As such, the results should be interpreted with caution.

Discussion

Alcohol consumption was previously reported to be associated with epigenetic markers such as DNAm changes^{13, 14}. The effects of alcohol consumption on DNAm are likely modulated by its ability to interfere with the methyl donor, S-adenosylmethionine (SAdMe)³⁰. The role of DNAm at specific loci associated with alcohol consumption in human disease such as fetal alcohol spectrum disorders³¹, liver disease and the gastrointestinal tract³², colon cancer³³ and breast cancer³⁴, has been studied in recent years. Here, we focused our attention on the identification of epigenetic marks that are potential mediators of an association between alcohol consumption and EOC status in a clinic-based study population. Such analyses are critical to enhance our understanding of EOC pathogenesis. Moreover, the alcohol-DNAm markers identified through the analyses will contribute to the growing pool of risk-associated biomarkers for EOC.

In our analysis, we classified alcohol consumption as alcohol use at enrollment vs no use at enrollment. Our analysis identified cg11016563 as a potential mediator of alcohol consumption for EOC. In both cases and controls, people who drank alcohol at enrollment had lower methylation levels compared to those who did not drink. This locus resides in

CpG island region located within gene *TRPC6* (transient receptor potential cation channel, subfamily C, member 6) promoter region. TRPC channels are Ca^{2+} -permeable cationic channels controlling Ca^{2+} influx response to the activation of G protein-coupled receptors and protein tyrosine kinase pathways or the depletion of Ca^{2+} stores. Overexpression of *TRPC6* has been reported in gliomas and ovarian cancer cells^{35, 36}. Zeng et.al.³⁷ suggest that TRPCs and their spliced variants are important for human ovarian cancer development and alteration of the expression or activity of these channels could be a new strategy for anticancer therapy. The mRNA of *TRPC6* was detected in human ovarian adenocarcinoma. Overexpression of *TRPC6* increased the cancer cell colony growth. The overexpression of this gene might be due to the DNAm of *TRPC6*. Further biological evidence is needed to confirm this.

Our analysis also identified cg09358725, located in gene *LMO2* (LIM domain only 2) and was found to have low DNAm level in EOC cases in the present analysis (see Fig. 5B). *LMO2* was identified as the 5' gene in an interchromosomal fusion with *ACVR2A* (activin A receptor, type IIA). *LMO2* has been implicated in B-cell lymphoma³⁸ and prostate cancer³⁹ and is proposed to be a prognostic marker for survival in pancreatic cancer⁴⁰. In both cases and controls, people who drank alcohol at enrollment had lower DNAm levels at this locus compared to those who did not drink. Further investigation is needed to confirm its role in EOC.

Part of the data analyzed in this work was previously analyzed, focusing on the relationship between genetics and epigenetics as they relate to the risk of EOC²². The previous analysis aimed to identify DNAm markers that are mediators for the relationship between inherited variants and EOC risk, while we aimed to identify DNAm markers acting as potential mediators for alcohol-associated EOC risk. None of the two CpGs identified in our analysis have been previously identified as significant epigenetic markers differentiating case-control status²² and there is no direct evidence of these methylation sites associated with the incidence of EOC, highlighting the importance of the causal mediation analysis. Although direct casual effect of alcohol consumption on EOC risk is difficult to establish, it is possible that alcohol consumption may modify DNAm levels which may further affect EOC risk. The results thus emphasize the role of the identified DNAm sites and their potential role in the pathogenesis of EOC. In the meantime, we need to be cautious when we interpret the statistical results. Although alcohol consumption can affect DNAm levels at certain sites, such an effect may be temporary. Thus, the identified epigenetic markers may not be stable. Further investigations are needed to confirm the role of the identified CpG sites on EOC risk.

In this analysis, we modified the CIT procedure in Step 2 by applying a high-dimensional testing procedure. High-dimensional data are routinely generated in genetics and genomics with emerging biotechnologies. Classical linear regressions cannot be applied directly given that the number of variables could be much larger than the number of sample size. Although the recent developments in high-dimensional variable selection (e.g., LASSO) provide an alternative solution to marginal regression, these methods cannot quantify the uncertainty of coefficient estimates. Hence, one cannot get rigorous p-values that are more attractive to biologists from a practical perspective. In this work, we adopted a high-dimensional testing

procedure²¹ to test the regression coefficients in a high-dimensional linear model setup. This method has theoretical guarantees and is relatively easy to implement. We hope the modified CIT procedure by incorporating the high-dimensional testing step can be adopted by other researchers to improve their analysis.

Note that the data in this analysis were obtained based on a matched case-control design. A natural model choice for a matched case-control study is the conditional logistic regression model for a binary disease response. However, recent study revealed that it is not always the case that a matched design requires a matched analysis⁴¹. Pearce⁴¹ showed that a standard logistic regression model controlling for the matching factors should work, and in fact should be preferred in cases where matching could introduce confounding effects. We did try a conditional logistic regression model and got similar results for the low dimensional regression steps. For the high-dimensional testing procedure in Step 2, it is not clear how a conditional logistic regression can be applied. Thus, we focused our analysis with a regular logistic regression model in this work. Further methodology development is needed for high-dimensional testing in a conditional logistic regression setup.

Our study is based on a case-control design and there may be a non-causal association between alcohol consumption and EOC status. In a case-control study, it is impossible to directly establish causality on the basis of a purely retrospective observational study. With data collected with questionnaire about alcohol consumption at enrollment, the amount of DNA methylation may not represent stable epigenetic markers, making it difficult to establish the causality. Another limitation of this study is the difficulty in determining direction of the causal pathway. The results obtained by CIT test do not necessarily reflect a biological causal relationship. We cannot rule out the possibility that EOC development affects alcohol consumption, which in turn affects methylation, and this further complicates the problem with a retrospective case-control design. Furthermore, we were limited by sample size to combine across EOC histotype. Further molecular investigations are needed to confirm the mediation mechanisms of the identified DNAm sites associated with EOC. Thus, we hope that our findings can serve as a basis to generate meaningful biological hypotheses for future mechanistic and functional validation in independent study populations and/or prospective studies.

Acknowledgments

The authors wish to thank an associate editor and two anonymous referees for their insightful comments that greatly improved the manuscript. This work was supported in part by grants from NSFC (31371336) and NIH (R01-CA122443).

References

1. Morgan RJ Jr, Alvarez RD, Armstrong DK, Boston B, Burger RA, Chen LM, et al. Epithelial ovarian cancer. Journal of the National Comprehensive Cancer Network : JNCCN. 2011; 9:82–113. [PubMed: 21233246]
2. Bagnardi V, Blangiardo M, La Vecchia C, Corrao G. A meta-analysis of alcohol drinking and cancer risk. British journal of cancer. 2001; 85:1700–1705. [PubMed: 11742491]
3. Cancer, I.A.f.R.o. Alcohol consumption and ethyl carbamate. IARC Press, International Agency for Research on Cancer; 2010.

4. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose–response meta-analysis. *British journal of cancer*. 2015; 112:580–593. [PubMed: 25422909]
5. Yan-Hong H, Jing L, Hong L, Shan-Shan H, Yan L, Ju L. Association between alcohol consumption and the risk of ovarian cancer: a meta-analysis of prospective observational studies. *BMC public health*. 2015; 15:223. [PubMed: 25885863]
6. Rota M, Pasquali E, Scotti L, Pelucchi C, Tramacere I, Islami F, et al. Alcohol drinking and epithelial ovarian cancer risk. A systematic review and meta-analysis. *Gynecologic oncology*. 2012; 125:758–763. [PubMed: 22449732]
7. Kelemen LE, Bandera EV, Terry KL, Rossing MA, Brinton LA, Doherty JA, et al. Recent alcohol consumption and risk of incident ovarian carcinoma: a pooled analysis of 5,342 cases and 10,358 controls from the Ovarian Cancer Association Consortium. *BMC cancer*. 2013; 13:28. [PubMed: 23339562]
8. Genkinger J, Hunter D, Spiegelman D, Anderson K, Buring J, Freudenheim J, et al. Alcohol intake and ovarian cancer risk: a pooled analysis of 10 cohort studies. *British journal of cancer*. 2006; 94:757–762. [PubMed: 16495916]
9. Cook LS, Leung AC, Swenerton K, Gallagher RP, Magliocco A, Steed H, et al. Adult lifetime alcohol consumption and invasive epithelial ovarian cancer risk in a population-based case–control study. *Gynecologic oncology*. 2016; 140:277–284. [PubMed: 26691218]
10. Webb PM, Purdie DM, Bain CJ, Green AC. Alcohol, wine, and risk of epithelial ovarian cancer. *Cancer Epidemiology and Prevention Biomarkers*. 2004; 13:592–599.
11. Ramchandani S, Bhattacharya SK, Cervoni N, Szyf M. DNA methylation is a reversible biological signal. *Proceedings of the National Academy of Sciences*. 1999; 96:6107–6112.
12. Cvetkovic D. Early events in ovarian oncogenesis. *Reproductive Biology and Endocrinology*. 2003; 1:68. [PubMed: 14577833]
13. Umeora M. Epigenetics And Fetal Alcohol Spectrum Disorders (FASD). *Ebonyi Medical Journal*. 2012; 11:1–6.
14. Philibert R, Plume JM, Gibbons FX, Brody GH, Beach S. The impact of recent alcohol use on genome wide DNA methylation signatures. *Frontiers in genetics*. 2012; 3:54. [PubMed: 22514556]
15. Bauerschlag DO, Ammerpohl O, Bräutigam K, Schem C, Lin Q, Weigel MT, et al. Progression-free survival in ovarian cancer is reflected in epigenetic DNA methylation profiles. *Oncology*. 2011; 80:12–20. [PubMed: 21577013]
16. Watts GS, Futscher BW, Holtan N, DeGeest K, Domann FE, Rose SL. DNA methylation changes in ovarian cancer are cumulative with disease progression and identify tumor stage. *BMC medical genomics*. 2008; 1:47. [PubMed: 18826610]
17. Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, et al. Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. *Nature communications*. 2013; 4:1628.
18. Cortessis VK, Thomas DC, Levine AJ, Breton CV, Mack TM, Siegmund KD, et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Human genetics*. 2012; 131:1565–1589. [PubMed: 22740325]
19. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. *Bmc Genetics*. 2009; 10:23. [PubMed: 19473544]
20. Fan J, Han F, Liu H. Challenges of big data analysis. *National science review*. 2014; 1:293–314. [PubMed: 25419469]
21. Dezeure R, Bühlmann P, Meier L, Meinshausen N. High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science*. 2015; 30:533–558.
22. Koestler DC, Chalise P, Cicek MS, Cunningham JM, Armasu S, Larson MC, et al. Integrative genomic analysis identifies epigenetic marks that mediate genetic risk for epithelial ovarian cancer. *BMC medical genomics*. 2014; 7:8. [PubMed: 24479488]
23. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nature genetics*. 2010; 42:874–879. [PubMed: 20852632]

24. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*. 2010; 11:587. [PubMed: 21118553]
25. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*. 2012; 13:86. [PubMed: 22568884]
26. MacKinnon, DP. Introduction to statistical mediation analysis. Routledge; 2008.
27. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996;267–288.
28. Van de Geer S, Bühlmann P, Ritov Ya, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*. 2014; 42:1166–1202.
29. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiologic methods*. 2014; 2:95–115. [PubMed: 25580377]
30. Varela-Rey M, Woodhoo A, Martinez-Chantar ML, Mato JM, Lu SC. Alcohol, DNA methylation, and cancer. *Alcohol Res*. 2013; 35:25–35. [PubMed: 24313162]
31. Ungerer M, Knezovich J, Ramsay M. In utero alcohol exposure, epigenetic changes, and their consequences. *Alcohol Res*. 2013; 35:37–46. [PubMed: 24313163]
32. Shukla SD, Lim RW. Epigenetic effects of ethanol on the liver and gastrointestinal system. *Alcohol Res*. 2013; 35:47–55. [PubMed: 24313164]
33. Schernhammer ES, Giovannucci E, Kawasaki T, Rosner B, Fuchs CS, Ogino S. Dietary folate, alcohol and B vitamins in relation to LINE-1 hypomethylation in colon cancer. *Gut*. 2010; 59:794–799. [PubMed: 19828464]
34. Christensen BC, Kelsey KT, Zheng S, Houseman EA, Marsit CJ, Wrensch MR, et al. Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake. *PLoS Genet*. 2010; 6:e1001043. [PubMed: 20686660]
35. Yu Y, Fantozzi I, Remillard CV, Landsberg JW, Kunichika N, Platoshyn O, et al. Enhanced expression of transient receptor potential channels in idiopathic pulmonary arterial hypertension. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101:13861–13866. [PubMed: 15358862]
36. Thébault S, Zholos A, Enfissi A, Slomianny C, Dewailly E, Roudbaraki M, et al. Receptor - operated Ca²⁺ entry mediated by TRPC3/TRPC6 proteins in rat prostate smooth muscle (PS1) cell line. *Journal of cellular physiology*. 2005; 204:320–328. [PubMed: 15672411]
37. Zeng B, Yuan C, Yang X, Atkin LS, Xu SZ. TRPC channels and their splice variants are essential for promoting human ovarian cancer cell proliferation and tumorigenesis. *Current cancer drug targets*. 2013; 13:103–116. [PubMed: 22920441]
38. Lossos IS, Czerwinski DK, Alizadeh AA, Wechsler MA, Tibshirani R, Botstein D, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*. 2004; 350:1828–1837. [PubMed: 15115829]
39. Ma S, Guan X, Beh P, Wong K, Chan Y, Yuen H, et al. The significance of LMO2 expression in the progression of prostate cancer. *The Journal of pathology*. 2007; 211:278–285. [PubMed: 17167821]
40. Nakata K, Ohuchida K, Nagai E, Hayashi A, Miyasaka Y, Kayashima T, et al. LMO2 is a novel predictive marker for a better prognosis in pancreatic cancer. *Neoplasia*. 2009; 11:712–719. [PubMed: 19568416]
41. Pearce N. Analysis of matched case-control studies. *bmj*. 2016; 352:i969. [PubMed: 26916049]
42. Sun T, Zhang C-H. Scaled sparse linear regression. 2011 arXiv preprint arXiv:1104.4595.

Appendix

The LASSO estimates are biased and do not have the asymptotic normality property for testing. To make inference for high-dimensional data, van de Geer et al.²⁸ proposed the de-sparsified method. We first describe the de-sparsifying approach for a continuous response Y . A de-sparsifying step to obtain the j th bias-corrected regression coefficient as given as,

$$\hat{b}_j = \frac{Z_j^T Y}{Z_j^T X_j} - \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \hat{\beta}_{\text{Lasso}; k},$$

where $\hat{b}_j$ is the bias-corrected coefficient of the j th CpG site X_j ; X_k is the k th ($k \neq j$) CpG site; $\hat{\beta}_{\text{Lasso}; k}$ is the coefficient of the k th CpG site estimated by fitting the LASSO regression; Z_j is the “regularized” residuals obtained by $Z_j = X_j - X_{-j} \hat{\gamma}_{\text{Lasso}}^j$, where $\hat{\gamma}_{\text{Lasso}}^j$ is the regression coefficients obtained based on a LASSO regression by regressing X_j on all other CpG sites except the j th CpG site denoted as X_{-j} ²⁸. van de Geer et al.²⁸ proved the asymptotic normality of the de-sparsified estimate, i.e.,

$$u_j = \frac{\sqrt{n}(\hat{b}_j - \beta_j^0)}{\sigma_\varepsilon \sqrt{\Omega_{jj}}} \xrightarrow{d} N(0, 1) \text{ as } p \geq n \rightarrow \infty$$

where β_j^0 represents the true regression coefficient; σ_ε can be calculated by using the scaled LASSO algorithm⁴² and Ω_{jj} can be calculated by

$$\Omega_{jj} = \frac{n Z_j^T Z_j}{[X_j^T Z_j][X_j^T Z_j]}$$

Under the null that $H_0: \beta_j^0 = 0$, the statistic u_j follows a standard normal distribution based on which we can get a p-value for testing the j th coefficient.

For a binary response, van de Geer et al.²⁸ also proved the asymptotic normality for the de-sparsified estimates. Let $\rho_\beta(y, x) = \rho(y, x\beta)$ be a loss function, and define $\dot{\rho}_\beta = \frac{\partial}{\partial \beta} \rho_\beta$ and

$\ddot{\rho}_\beta = \frac{\partial^2}{\partial \beta \partial \beta^T} \rho_\beta$, and further define $P_{ng} = \sum_{i=1}^n g(y_i, x_i)/n$. The LASSO estimator for the CpG

coefficients β_1 is given as $\hat{\beta}_1 = \underset{\beta}{\operatorname{argmin}} \left(P_{ng} \rho_\beta + \lambda \|\beta\|_1 \right)$, where λ is a tuning parameter. Define $\hat{\Sigma} = P_{ng} \ddot{\rho}_{\hat{\beta}}$ and construct $\hat{\Theta} = \hat{\Theta}_{\text{LASSO}}$ by doing a nodewise LASSO with $\hat{\Sigma}$ as input. Then the de-sparsified LASSO estimator is given as $\hat{b}_1 = \hat{\beta}_1 - \hat{\Theta} P_{ng} \dot{\rho}_{\hat{\beta}}$. van de Geer et al.²⁸ gave the detailed algorithm for computing the de-sparsified LASSO estimators in a generalized linear model framework. They also proved the asymptotic normality of the de-sparsified estimate, i.e.,

$$u_j = \frac{\sqrt{n}(\hat{b}_j - \beta_j^0)}{\hat{\sigma}_j} \xrightarrow{d} N(0, 1) \text{ as } p \geq n \rightarrow \infty$$

where β_j^0 represents the true regression coefficient, and $\hat{\sigma}_j^2 = \left(\hat{\Theta}^P \begin{matrix} n\hat{\rho} \\ \hat{\beta} \end{matrix} \hat{\rho}^T \hat{\Theta}^T \right)_{j,j}$. Under the null that $H_0: \beta_j^0 = 0$, we could get a p-value for each CpG site based on the asymptotic normality property.

Figure 1.
Schematic representation of the causal mediation model.

Figure 2. Identification of epigenetically mediated alcohol consumption for EOC

Flow diagram illustrating the various testing steps, and ensuing results, used to identify methylation sites that are candidates for mediators of alcohol consumption for EOC. Step (1) – (4) are the four components of CIT, where L refers to alcohol use, G refers to DNAm and T refers to the EOC status.

Figure 3. Potential methylation sites between EOC status conditional on alcohol consumption
Left panel shows the histogram of p-values obtained from examining the association between DNAm and EOC status conditional on alcohol consumption. Right panel shows the Manhattan plot for 25926 methylation locus.

Figure 4. Potential methylation sites associated with alcohol consumption conditional on EOC status

Left panel shows the histogram of p-values obtained from examining the association between DNAm and alcohol consumption conditional on EOC status. Right panel shows the Manhattan plot for 3055 methylation sites. Red line indicates $-\log_{10}(0.05)$.

Figure 5. Alcohol-dependent candidate CpGs (cg09358725, cg11016563) that mediate alcohol consumption for EOC

The figure shows the OR and its 95% CI of the alcohol-EOC associations based on the logistic model: A (without CpGs, $T \sim L + \text{Covariates}$); B (with CpGs, $T \sim L + \text{Covariates} + G_1 + G_2$, where G_1 and G_2 represents cg09358725 and cg11016563, respectively).

Table 1

Partial list of covariates and their association with case/control status

	Case (N=196)	control (N=202)	Total (N=398)	p value
Age at Diagnosis/Interview				0.946
Mean (SD)	62.9 (12.7)	62.8 (12.4)	62.4 (12.7)	
Parity/Age first birth combo				0.050
Missing	0	2	2	
Nulliparous	36 (18.4%)	28 (13.9%)	64 (16.1%)	
1-2,<=20 yrs	9 (4.6%)	11 (5.4%)	20 (5.0%)	
1-2,>20 yrs	59 (30.1%)	53 (26.2%)	112 (28.1%)	
3+,<=20 yrs	41 (20.9%)	29 (14.4%)	70 (17.6%)	
3+,>20 yrs	51 (26.0%)	79 (39.1%)	130 (32.7%)	
Alcohol use at study enrollment				< 0.001
Yes	123(62.8%)	172(85.1%)	295(74.1%)	
No	73(37.2 %)	30(14.9%)	103(25.9%)	
Minnesota(MN) state				0.168
Other	93 (47.4%)	82 (40.6%)	175 (44.0%)	
MN	103 (52.6%)	120 (59.4%)	223 (56.0%)	
Smoking at study enrollment				0.099
No	178 (90.8%)	192 (95.0%)	370 (93.0%)	
Yes	18 (9.2%)	10 (5.0%)	28 (7.0%)	

Table 2

List of CpG sites that are potential mediators of alcohol consumption for EOC

CpG name	Chr	Gene	Position	CIT component p-values			
				T ~ L	G ~ T L	L ~ G T	L T TG
cg09358725	11	LMO2	33870664	0.001	0.001	0.031	0.022
cg11016563	11	TRPC6	100959836	0.001	0.016	0.027	0.009

L: Alcohol consumption, **G:** Methylation locus, **T:** EOC status.

Gene: Denotes the nearest gene for the given CpG sites.

Table 3

List of ORs and CIs for various effects analyzed by the VanderWeele and Vansteelandt's multiple mediator model

Effect	Odds ratio (OR)	95% CI	
		lower	upper
Control direct effect (CDE)	0.477	0.244	0.913
Natural direct effect (NDE)	0.477	0.244	0.913
Natural indirect effect (NIE)	0.789	0.724	0.828
Total effect (TE)	0.376	0.177	0.756