Assessing exposure effects on gene expression

Sarah A. Reifeis[1] | **Michael G. Hudgens**[1] | **Mete Civelek**[2] | **Karen L. Mohlke**[3] | **Michael I. Love**[1,3]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[2]Department of Biomedical Engineering, Center for Public Health Genomics, The University of Virginia, Charlottesville, Virginia

[3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Correspondence: Michael I. Love, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC27516.
Email: michaelisaiahlove@gmail.com

Funding information
National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Numbers: R01 DK093757, R01 DK118287; National Heart, Lung, and Blood Institute, Grant/Award Number: R21 HL135230; National Institute of Allergy and Infectious Diseases, Grant/Award Number: R01 AI085073; National Institute of Mental Health, Grant/Award Number: R01 MH118349; National Institute of Environmental Health Sciences, Grant/Award Number: P30 ES010126; Graduate School at UNC-Chapel Hill, Grant/Award Number: Chancellor's Fellowship; National Human Genome Research Institute, Grant/Award Number: R01 HG009937

Abstract

In observational genomics data sets, there is often confounding of the effect of an exposure on gene expression. To adjust for confounding when estimating the exposure effect, a common approach involves including potential confounders as covariates with the exposure in a regression model of gene expression. However, when the exposure and confounders interact to influence gene expression, the fitted regression model does not necessarily estimate the overall effect of the exposure. Using inverse probability weighting (IPW) or the parametric g-formula in these instances is straightforward to apply and yields consistent effect estimates. IPW can readily be integrated into a genomics data analysis pipeline with upstream data processing and normalization, while the g-formula can be implemented by making simple alterations to the regression model. The regression, IPW, and g-formula approaches to exposure effect estimation are compared herein using simulations; advantages and disadvantages of each approach are explored. The methods are applied to a case study estimating the effect of current smoking on gene expression in adipose tissue.

KEYWORDS

confounding, inverse probability weighting, observational genomics, parametric g-formula, regression

1 | INTRODUCTION

Increasing numbers of large-scale observational genomic data sets are available, in which tissue is collected from human donors sampled from a population. These donors differ in various ways, for example, they may differ by age, sex, and other demographic variables, as well as by clinical or exposure variables such as body-mass index (BMI), diet, level of physical activity, history of medicine use, history of smoking or alcohol use, or various environmental exposures. Investigators are often interested in assessing the effect of various exposures on genomic variables, such as gene expression, as this may be useful to generate hypotheses about potential cellular mechanisms through which exposures may influence the development of diseases in human populations. Gene expression is a common molecular measurement in the context of exposure effects, although additional genomic assays, such as methylation, metabolites, or protein abundance, may also be of interest.

A number of statistical methods have been proposed to address the problem of structural technical variation in gene expression measurements (Gagnon-Bartsch & Speed, 2012; Leek & Storey, 2007; Stegle, Parts, Piipari, Winn, & Durbin, 2012), where *structural* refers to variation in the measurements across samples that is common across many genes. These methods address sample non-independence with a focus on estimation of latent factors, orthogonal to the biological condition of the samples, to be included in a linear model framework as regressors to account for the technical variation in the measurements. These methods can account for differences in measurements among sample preparation batches that may otherwise impair correct inference of differences across the biological conditions. Additionally, methods have been proposed to address sample correlations that arise from biological sources, for example, repeated measures or genetically related individuals. Such sample non-independence can be addressed by explicit modeling of the known sample correlation structure as in a random effects framework (Cui, Ji, Li, Cheng, & Qiu, 2016); software for this approach include the *duplicateCorrelation* method (Smyth, Michaud, & Scott, 2005) in the *limma* R/Bioconductor package (Smyth, 2004), the *ShrinkBayes* R package (Van De Wiel et al., 2012), or the *MACAU* R package (Sun et al., 2017), all of which can be run from within the R environment (R Core Team, 2019).

Regression frameworks alone may not be able to properly address the problem of confounding variables, whether measured or unmeasured, in observational data sets. Confounding is an issue when estimating causal effects, and as such is a distinct problem from the technical and biological sources of correlations among samples described above. Confounding has received relatively less attention in computational genomics, compared to the problems of structural technical variance or repeated measures. Existing work addressing confounding in observational genomic data sets has focused on sample matching (Heller, Manduchi, & Small, 2008), the combination of targeted minimum loss-based estimation (TMLE) and empirical Bayes shrinkage estimation (Hejazi, Kherad-Pajouh, van der Laan, & Hubbard, 2017), and TMLE for differential methylation controlling for observed methylation at neighboring genomic sites (Hejazi, Phillips, Hubbard, & van der Laan, 2018).

Since the exposure in an observational study is not randomly assigned, there is often confounding of the effect of exposure on the outcome. In general, randomized clinical trials to assess how various exposures affect gene expression cannot be conducted in human populations for ethical or feasibility reasons. Similar randomized studies can be performed on model organisms, but there is unique value in understanding the mechanism of these exposures in humans, and human populations are readily available for observational studies. In light of the increasing number of observational genomics studies on humans and the anticipated presence of confounding in such studies, methods of exposure effect estimation are worth further investigation.

For many studies, it is often useful to assess exposure effects on gene expression *in the exposed individuals*. This may be the case when researchers have particular interest in the effect of an exposure only on those types of individuals who likely will experience the exposure. For example, when studying the effect of smoking, it is often most relevant to obtain effect estimates interpretable for those who actually smoke, as opposed to the effect smoking would have, averaging over all the persons in the general population. In these cases, the target estimand is referenced as the exposure effect in the exposed; this terminology will be used interchangeably with the average treatment effect in the treated (ATT) throughout. In contrast, the average treatment effect (ATE) is a different estimand and is interpretable in the context of the population as a whole, including both treated and untreated individuals. This paper will focus on obtaining estimates for the former, the ATT.

The conventional approach to quantifying the effect of exposure, while attempting to adjust for confounding, is to fit a linear model of gene expression with the exposure and various potential confounders as covariates. In this approach, the estimated coefficient of the exposure variable is often interpreted as an estimate of the exposure effect. However, fitting the conventional linear model falls short of our goal in two respects: (a) it does not directly produce an estimate of the effect of interest, the effect of the exposure *in the exposed individuals*, who may differ in various respects from unexposed individuals, and (b) it may not appropriately adjust for confounding, resulting in estimates that are not consistent and confidence intervals that do not provide their nominal coverage. For these reasons, this paper demonstrates existing causal inference methods that can be employed in these scenarios to adequately adjust for confounding and return consistent exposure effect estimates and valid confidence intervals.

Regression, inverse probability weighting (IPW), and the parametric g-formula are compared herein for obtaining exposure effect estimates. Both IPW and the parametric g-formula are methods established and widely applied to observational studies in the causal inference and epidemiology literature. This paper seeks to demonstrate that these methods also have utility in the space of observational genomics. In general, IPW uses weights to construct a "pseudo-population" in which there is no longer confounding of the effect of an exposure on the outcome of interest; simple linear regression is then applied to this "pseudo-population" to obtain consistent estimates of the exposure effect (Robins, Hernán, & Brumback, 2000). The parametric g-formula, also referred to as standardization, entails fitting an outcome regression model and then averaging the predicted outcomes across all individuals for a fixed level of the exposure. Both IPW and the parametric g-formula rely on standard assumptions of causal inference (conditional exchangeability, positivity, and consistency), but they differ in the modeling assumptions required (Naimi, Cole, & Kennedy, 2017). Statistical validity of the IPW and parametric g-formula methods rely on asymptotic justifications, and are not guaranteed to perform well for small sample sizes.

The following is an outline for the remaining sections of this paper. A brief summary is given of the models used, followed by more extensive description in Section 2 for both the simulation study and data analysis; formal definitions are left to the Supporting Information Methods. In Section 3, the methods are compared in simulations and a data analysis. Simulation studies are based on the Metabolic Syndrome in Men Finnish cohort (METSIM; Laakso et al., 2017) analysis data set ($n = 770$). In particular, scenarios falling into three categories for a binary exposure effect on gene expression are investigated: no confounding, and confounding both with and without interaction(s) between exposure and covariates. The METSIM cohort data is then analyzed to investigate the effect of current smoking on gene expression in adipose tissue. Section 4 concludes with reviewing and providing insight into the main results, and addresses limitations of and future directions for this study. The Supporting Information Methods gives details on the models for the regression, IPW, and g-formula approaches, as well as estimates of standard errors and assumptions required for each approach. In the Supporting Information Results, it is first established why the proposed methods are being compared to linear regression alone, and not to the linear model with empirical Bayes moderation of the standard errors (e.g., as implemented in the *limma* package; Smyth, 2004). This section also contains additional regression analysis results for both the simulated data sets and the METSIM cohort data. The Supplementary Results section concludes with reporting of the root mean square deviation for each estimation approach using the simulated data, and sensitivity analyses for the METSIM cohort data. Appendix A demonstrates the equivalence of the g-formula estimator presented in the main text and the g-computation algorithm of Snowden, Rose, and Mortimer (2010). Appendix B provides R markdown (Rmd) workflows showing the generation of the simulated data and performing the three methods on a simulated data set.

2 | METHODS

2.1 | Summary of models compared

Three exposure effect estimation approaches were assessed in the following evaluation: traditional linear regression, IPW, and the parametric g-formula. In fitting the models associated with each approach, it was assumed that the models were correctly specified, the set of confounders identified was sufficient to adjust for confounding, and the data were free from selection bias and systematic measurement error. The regression model with the exposure and potential confounders as predictors was fit using ordinary least squares for both the parameter estimates and their standard errors, in keeping with the conventional approach. For the IPW approach, the confounders were used as predictors in a logistic regression model of the exposure to obtain the weights, and the weights were used in the simple linear regression model of gene expression on the exposure using weighted least squares. In the g-formula approach, all potential confounders were centered at the mean value in the exposed, and the linear regression model with the exposure, centered confounders, and their interactions was fit using ordinary least squares. The standard errors for both the IPW and g-formula estimators were computed using stacked estimating equations (Stefanski & Boos, 2002). The details of using these methods with observational genomics data are described further in Section 2, formally defined in the Supporting Information Methods section, and demonstrated in the R code included in Appendix B.

2.2 | Simulation study

Performance of methods was first compared using a simulation. The simulated covariates and exposure were based on counterparts from the METSIM data analysis in the next section. Specifically, the simulated variables included a current smoking indicator and five variables considered to be potential confounders of the relationship between current smoking and gene expression. Table 1 below gives more details regarding variable distributions and dependencies.

Following the generation of the variables in Table 1, normalized gene expression values for various scenarios were simulated as well. For scenarios where no confounding was present, the mean of the expression values were dependent on only the exposure or none of the variables. When confounding was present, the mean expression values were dependent on both the exposure and the other covariates, with some scenarios including interactions between the exposure and the covariates.

Variable	Distribution	Dependencies
Age (*age*)	Normal	None
Alcohol consumption (*alc*)	Exponential	None
Vegetable consumption (*veg*)	Binary	None
Hobby exercise (*hex*)	Categorical (4 levels)	*veg*
BMI (*bmi*)	Normal	*veg*, *hex*
Current smoking (*smk*)	Binary	*alc*, *bmi*, *veg*, *hex*

Abbreviations: BMI, body-mass index; IPW, inverse probability weighting.

TABLE 1 Definitions of exposure and confounding variables for simulation study comparing regression, IPW, and the parametric g-formula

Expression values were generated with different means for each individual and each gene, according to the simulated exposure and covariates. The mean for gene g and individual i was

$$\begin{aligned}\mu_{gi} = {} & \beta_{g1} age_i + \beta_{g2} alc_i + \beta_{g3} veg_i + \beta_{g4} hex_i + \beta_{g5} bmi_i \\ & + smk_i \Big(\beta_{g6} + \beta_{g7} alc_i + \beta_{g8} veg_i + \beta_{g9} hex_i \\ & + \beta_{g10} bmi_i \Big),\end{aligned}$$

where the β_{gk}, $k = 1, ..., 10$, varied by gene and were restricted to $\beta_{gk} \in [-2,2]$ for this simulation study. Here *age*, *alc*, and *bmi* were each centered about their population mean and scaled. Note that there was no quadratic age term and no interaction term for age and smoking in the true model for mean expression, although the former was included in the analysis models; these terms can be thought of as not contributing to the true mean gene expression for any individual. The standard deviation of each gene was set to the same value to aid with comparing results for different genes, and was equal to 0.24.

The variables listed in Table 1 were generated for a population of 10 million individuals, from which the true ATT was calculated for each gene. From this population, 1,000 sets of 770 individuals were randomly selected with replacement to build the analysis data sets. The sampling algorithm allowed for the same individual to be present in more than one data set, but not more than once within a single data set.

In all instances the regression model had the same form shown in Equation S.1 of the Supporting Information Materials, namely the exposure and covariates (with both linear and quadratic age terms) were each present as main effects in the model and no interaction terms were included; for simulation study results of the regression analyses including interactions in the model, see Supporting Information Results. The covariates *age*, *age*^{2}, *alc* and *bmi* were centered at their sample mean and scaled for each data set, as would typically be done to avoid collinearity with the intercept. The standard errors used to construct the 95% confidence intervals were obtained through fitting the model with ordinary least squares.

For the IPW approach, first the logistic regression model in Equation S.2 of the Supporting Information materials was used to compute the components needed for the weights for each data set, with terms for the five covariates and a quadratic age term. Weights for the ATT were then constructed according to expression S.3 of the Supporting Information materials. Then the linear regression model of gene expression in S.4, with only the exposure and an intercept, was fit with the weights to obtain the effect estimate. Standard error estimates used to obtain the 95% confidence intervals were computed with the stacked estimating equations approach using the *geex* package (Saul & Hudgens, 2020) in R, taking into account estimation of the weights by stacking the estimating equations for the logistic regression model with those used in computing the IPW estimator.

The regression model given in Equation S.5 of the Supporting Information materials was used to obtain the g-formula estimate for each data set. It contained variables for the main effects of the exposure and covariates (including both linear and quadratic age) as well as interactions between *smk* and each of *bmi, veg, hex,* and *alc*. Since the ATT is being compared across methods in this simulation study, the non-exposure covariates were all centered at the sample mean in the exposed. The standard error estimates used for the 95% confidence intervals were computed with stacked estimating equations by passing *geex* the set of stacked estimating equations corresponding to the covariate means and the regression model parameters.

2.3 | METSIM smoking exposure effect

In the METSIM project data set, the goal was to obtain the estimated effect of current smoking on gene expression in the smokers, adjusting for the set of potential confounders: linear and quadratic age, BMI, alcohol consumption, vegetable consumption, and hobby exercise. The data consisted of adipose gene expression values and several covariates, measured for $n = 770$ individuals (details on data preprocessing in Supporting Information Methods). There were no missing outcome, exposure, or covariate values. This cohort was analyzed in Civelek et al. (2017), where BMI and linear and quadratic age were considered confounding variables for various phenotypic traits. This analysis, and consultations with a subject-matter expert, guided the choice of the confounding variable set. Note that current smoking was not examined by Civelek et al., so their analyses were not compared directly to those in this paper. Each of the three methods introduced above were implemented for this cohort in the analyses that follow.

Table 2 briefly summarizes the variables used in this analysis; the range, mean and standard deviation are reported for continuous variables and the levels and distribution are given for each categorical variable. For hobby exercise, higher levels denote increased activity levels.

Note that the models used for the data analysis had the same form as those fit for the simulation study, described in the section above. The estimated exposure effect was obtained using regression, IPW, and the g-formula for each of the 18,510 genes; the coefficient for current smoking represented the exposure effect in each model. The models were fit and standard errors computed, again using the same process as for the simulation study; for results of the regression analyses including interactions in the model with the METSIM data, see Supporting Information Results. With each approach, a t-test of no effect of exposure on gene expression was performed for each gene and the resulting p-values were adjusted using the correction from Benjamini and Hochberg (1995) to control the false discovery rate.

TABLE 2 Descriptive statistics for the METSIM cohort data

Variable	Range	Mean (*SD*)
Age (years)	45, 68	55 (5)
Alcohol consumption (g/week)	0, 1134	105 (119)
BMI (kg/m^2)	18.5, 48.1	26.6 (3.5)
	Level	**Proportion**
Vegetable consumption	Everyday	0.83
	Not everyday	0.17
Hobby exercise	1	0.05
	2	0.28
	3	0.18
	4	0.49
Current smoking	Yes	0.17
	No	0.83

To compute weights for the 770 individuals, the logistic regression model of current smoking was fit with the previously listed covariates as predictors. Before computing effect estimates and standard errors, it is good practice to check that the weights have a mean value close to the expected value (details in Supporting Information Methods) and that none of the weights are extreme. The weights had mean value 0.34, which was exactly what was expected for these data. There was one weight with a value of 5.26, substantially larger than the rest; for this reason, a sensitivity analysis was conducted in the Supporting Information Results section where the observation with this large weight was deleted and the same analysis was performed again to investigate the influence of this observation.

For both the g-formula and regression methods, leverage values were computed for each observation to determine if there were any influential points in these analyses; the same observation returned the largest leverage point for both the g-formula and regression, which took values 0.38 and 0.11, respectively. In both instances these leverage values were approximately twice the magnitude of the next largest value, so another sensitivity analysis was conducted in the Supporting Information Results where the observation with this large leverage value was deleted and the analysis was performed again to investigate its influence.

3 | RESULTS

3.1 | Simulation study

The empirical bias and confidence interval coverage and width for the regression, IPW, and g-formula estimators are shown in Table 3, averaging over 1,000 simulations per scenario. While there were instances where all estimators appeared to perform well, the IPW and g-formula estimators provided advantages over regression in a subset of the simulated scenarios. In particular, the IPW and g-formula estimators remain unbiased and meet nominal confidence interval coverage in all scenarios, but the regression estimator manifests bias and fails to meet nominal coverage in the presence of exposure-covariate interactions. The reported coverage represents the proportion of confidence intervals which included the true value of the ATT. The true ATT value is shown for each scenario, and is based on the original population of 10 million individuals. The null case where there was no

TABLE 3 Average empirical bias and 95% CI coverage and width for the regression (Reg), IPW, and g-formula (G-form) estimators

		Estimate bias			95% CI coverage			95% CI width		
Scenario	**True ATT**	**Reg**	**IPW**	**G-form**	**Reg**	**IPW**	**G-form**	**Reg**	**IPW**	**G-form**
Null case	0.00	0.00	0.00	0.00	0.96	0.96	0.95	0.10	0.10	0.10
No interactions 1	−2.00	−0.00	0.00	0.00	0.95	0.96	0.95	0.10	0.11	0.10
No interactions 2	2.00	0.02	0.01	−0.00	0.91	0.92	0.94	0.15	0.35	0.18
Interactions 1	1.59	0.12	0.00	0.00	0.34	0.96	0.96	0.18	0.39	0.39
Interactions 2	−0.36	−0.20	0.00	0.00	0.19	0.96	0.95	0.23	0.51	0.51
Interactions 3	−1.75	−0.20	0.01	0.01	0.37	0.95	0.95	0.31	0.70	0.70

Abbreviations: ATT, average treatment effect in the treated; CI, confidence interval; IPW, inverse probability weighting.

effect of current smoking on gene expression is shown in addition to several representative scenarios where confounding was present, two without and three with one or more interactions between exposure and confounders.

True ATT values in these simulations are in units of $\log_2$ fold change in gene expression. As normalized gene expression data are often on the $\log_2$ scale, linear effects are commonly interpreted as $\log_2$ fold changes with respect to the raw gene expression scale (Smyth, 2004). The most extreme scenario shown here, where the true ATT is 2.00, thus represents a fourfold change in gene expression attributable to smoking. Additional simulation studies were conducted that are not shown in this table, but the results were similar to those included. Across all simulations run, the average bias of the regression estimator took values in the range [−0.29, 0.29], whereas the IPW and g-formula average biases were contained to [−0.01, 0.01]. For this simulation study setup, the regression bias appears to be larger in magnitude when interaction terms involving alcohol and hobby exercise contributed to the true ATT.

When smoking and the confounders did not interact to influence gene expression, for example, the first three examples in Table 3, all methods met nominal coverage and yielded no bias on average. In scenarios where interactions existed between smoking and the confounders, the regression effect estimator had nonzero average bias and substantially below nominal confidence interval coverage. The IPW and g-formula estimators both resulted in very low or no bias on average, and both uniformly met (or very nearly met) the nominal confidence interval coverage. Except in the null case, the IPW and g-formula estimators were generally more variable than the regression estimator. When smoking and the confounders interacted, the IPW and g-formula estimates had confidence intervals that were on average approximately twice as wide as those for regression. Of the two methods that overall maintained the nominal coverage across scenarios, IPW and g-formula, they tended to have comparable interval width except in one of the no-interaction scenarios, in which IPW had nearly double the average interval width of g-formula.

3.2 | METSIM smoking exposure effect

Regression, IPW, and g-formula were applied to assess the effect of smoking on gene expression among smokers in the METSIM cohort. Estimates and confidence intervals for each of the three methods, for the top two genes as ranked by p-value are presented in Figure 1a,b (the top two genes are shown as their effect sizes and test statistics were appreciably larger compared to the other genes). The three methods were in agreement on the ranking for the top three genes, but beyond this the rankings were not consistent across method (Figure 1c). The top two genes in terms of estimated effect size and test statistic were *CYP1A1* and *CYP1B1*, which were expected as they play a role in metabolizing cigarette smoke (Nebert & Russell, 2002). The confidence intervals for the IPW and g-formula estimates of the top two genes were similar in width, and the regression confidence interval was substantially less wide. While the $-\log_{10}$ adjusted p-values for the top two genes were all large for each of the three methods (highly significant), those from regression were much larger than those from IPW and the g-formula (*CYP1A1*: R = 119, W = 44, G = 44; *CYP1B1*: R = 30, W = 18, G = 20). This is in accordance with the displayed difference in confidence interval widths for the smoking effect estimates of these two genes. Although the ATT estimates were similar across methods for the genes shown, instances of substantial differences in standard error produced vastly different confidence intervals and p-values.

IPW and the g-formula tended to produce larger p-values than regression for these data, though this was

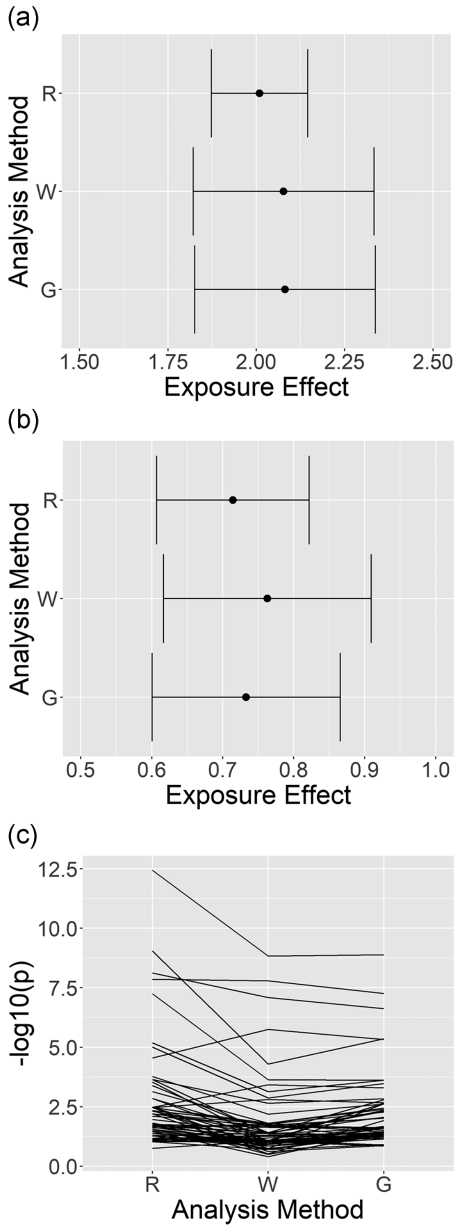

FIGURE 1 METSIM analysis results for the top 50 genes, ranked by *p*-value. (a,b) Estimates and 95% confidence intervals for the effect of current smoking in the smokers for genes *CYP1A1* and *CYP1B1*, respectively. Note that the null value for the smoking effect estimate ($\log_2$ fold change = 0) is not included on the *x*-axis. (c) $-\log_{10}$ adjusted *p*-values for the top 50 genes (omitting top 2), for each of R = regression, W = inverse probability weighting, and G = parametric g-formula

not the case for every gene (Figure 1c). Note that in this figure the top two genes were not represented, as their $-\log_{10}$ adjusted *p*-values were much larger than the others. Additionally, the smoking effect estimates for all genes except the top two were in the range (−0.5, 0.5), with many very close to zero.

4 | DISCUSSION

Geneticists and epidemiologists may analyze differential gene expression due to exposures in a population to generate hypotheses as to how those exposures may be related to health outcomes. Results of these analyses, that is lists of genes affected by the exposure under a false-discovery rate bound and their associated effect sizes, may be inaccurate and irreproducible across study populations unless potential confounders of the exposure and gene expression are properly adjusted for. Here, exposure effect estimates were compared using causal inference techniques such as IPW and the parametric g-formula, as well as with common practice techniques such as regression. Comparisons were performed across simulated data and a data analysis in which gene expression was measured in subcutaneous adipose tissue. Tissue donors also had various clinical and demographic covariates measured, and it was desired to adjust for differences among the exposed and unexposed donors when estimating the ATT. Analyses of the METSIM cohort found that estimation method did not make a substantial difference for the effect estimates for the top two genes, *CYP1A1* and *CYP1B1*. Simulations based on the METSIM data showed that there was potential for the regression estimate to be biased, but the effect biases observed in the data analyses were small. Differences between the methods were most pronounced when examining the standard errors and therefore also the resulting confidence intervals and *p*-values. In particular, simulations based on the METSIM data showed that if there were any interactions of the exposure with the confounder(s), the regression method produced confidence intervals that can have far below nominal coverage. Furthermore, what may appear to be modest changes in standard errors can produce a dramatically different set of adjusted *p*-values for a given set of genes.

In addition to the standard errors, regression as applied here differs from IPW and the parametric g-formula in that the regression estimates do not represent the effect of exposure in the exposed, but rather in the population as a whole. While IPW and the g-formula can be adapted to produce the exposure effect in the population or subpopulation of interest, regression estimates remain population-wide estimates—unless operating under the assumption that ATE and ATT are equal.

It should be mentioned that if all appropriate interaction terms were included in the regression model, the parameter estimates could be combined to yield consistent conditional exposure effect estimates. However, this approach was not taken here for two reasons: (a) the goal was to obtain one exposure effect estimate that could be read directly from software output without additional steps, and (b) the exposure effect estimate constructed via combination of exposure and interaction terms would be interpreted conditional on values of the covariates, whereas the desired exposure effect estimate has a marginal interpretation. Expanding on this second reason, regression with all appropriate interaction terms would result in a variety of exposure effects across combinations of covariates used for conditioning, as opposed to the causal approaches presented here which provide one exposure effect integrating over the exposed individuals. If all appropriate interaction terms were included in the regression model and centered at the mean in the population of interest, then the regression model would be equivalent to the parametric g-formula model.

Often in analyses of exposure effects on microarray gene expression, the *limma* method is used to fit the regression models and obtain a moderated t-statistic (Smyth, 2004), whereas here the ordinary t-statistic was used. The simulation results illustrating the rationale behind this choice are included in the Supporting Information Results section. In short, the sample size of the cohort analyzed here ($n = 770$) was sufficiently large that the ordinary and moderated t-statistics are practically equivalent. More recently, another method has been proposed involving the combination of TMLE and empirical Bayes shrinkage estimation, which has demonstrated utility with small and moderate sample sizes (Hejazi et al., 2017). The intended audience of this paper is working with larger data sets, allowing for reliance on large sample theory; for this reason the simpler and more readily available approach was used for the regression analysis. Here, prenormalized microarray gene expression, which takes continuous values, was analyzed, while RNA sequencing experiments result in count-valued observations for gene expression. In order for the causal inference approaches shown here to be applied to count data from RNA sequencing, it would be desirable to first perform library size scaling and apply a variance stabilizing transformation to the gene expression (Anders & Huber, 2010; Law, Chen, Shi, & Smyth, 2014), though such data sets and procedures were not evaluated in the present work.

The IPW and parametric g-formula approaches are both presented here as alternates to regression that adequately adjust for confounding in a wider variety of circumstances. While IPW and the g-formula both accomplish this goal, they require slightly different assumptions and they have different strengths and weaknesses. The IPW estimator relies on correct specification of the exposure model, which is often more plausible than correct specification of the outcome model. IPW can be sensitive to extreme weights, as shown in the sensitivity analysis results for the METSIM data, and can be more variable than the g-formula estimator. While the consistency of the g-formula estimator relies heavily on the correct specification of the outcome model, it appears to be less sensitive to extreme values of the covariates and can be less variable than the IPW estimator. Due to the limited overall differences in the bias and efficiency of the IPW and g-formula estimators, the researcher is encouraged to choose among methods based on their relative confidence in specification of the exposure or outcome models.

There are several assumptions made in these analyses which may be violated and deserve further exploration. First, the assumption of causal consistency states that there are not multiple ways to be a current smoker. This assumption is clearly not met since the amount of cigarettes smoked daily can vary from person to person, but this assumption can be replaced by another less stringent assumption. In particular, it can more reasonably be assumed that these different versions of exposure do not have any bearing on the causal effect; this is referred to as treatment variation irrelevance (Vander Weele, 2009). The data analyses above rely on the additional assumptions that the set of confounders L are sufficient to adjust for confounding. In addition to confounding, there may be other possible sources of bias when estimating exposure effects. For example, the methods here implicitly assume no systematic measurement error; for methods that account for measurement error see Hernán and Cole (2009), Hernán and Robins (2020), Kuroki and Pearl (2014), and Suzuki, Tsuda, Mitsuhashi, Mansournia, and Yamamoto (2016). In addition, the methods described here assume the data constitute a random sample from the target population. Nonrandom sampling from the population of interest may also introduce bias; for methods to accommodate biased sampling, see Buchanan et al. (2018), Lesko et al. (2017), Stuart, Cole, Bradshaw, and Leaf (2011). If any of these assumptions are unmet then the exposure effect estimates may be biased. Furthermore, formal arguments for the methods presented rely on large sample theory; while there is some empirical evidence suggesting that, for example, IPW can perform well with moderate sample sizes (Pirracchio, Resche-Rigon, & Chevret, 2012), these methods are not guaranteed to perform well for small or moderate samples.

The performance of doubly robust estimators for estimating exposure effects on gene expression could be investigated in future work. Doubly robust estimators have been shown to provide advantages over IPW or the g-formula (Lunceford & Davidian, 2004; Moodie, Saarela, & Stephens, 2018; Naimi & Kennedy, 2017), and could conceivably allow a relaxation of certain modeling assumptions in observational genomics analyses while maintaining the desirable properties of causal methods. Additionally, this paper focuses on binary exposures but future work could expand this to allow for continuous or longitudinal exposures.

ACKNOWLEDGMENTS

The authors thank the editor and reviewer for their feedback that improved the quality of the manuscript. In addition, the authors acknowledge the following individuals for useful discussions: Nathan Kallus, Melissa Troester, and the METSIM investigators. S. A. R. was supported by The Chancellor's Fellowship from the Graduate School at the University of North Carolina at Chapel Hill. M. G. H. was supported by R01 AI085073. M. C. was supported by R01 DK118287 and R21 HL135230. K. L. M. was supported by R01 DK093757. M. I. L. was supported by R01 HG009937, R01 MH118349, and P30 ES010126.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GEO under accession number GSE70353 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70353; Civelek et al., 2017).

ORCID

Sarah A. Reifeis https://orcid.org/0000-0002-6058-9422
Michael G. Hudgens https://orcid.org/0000-0002-9106-4194
Mete Civelek https://orcid.org/0000-0002-8141-0284
Karen L. Mohlke https://orcid.org/0000-0001-6721-153X
Michael I. Love http://orcid.org/0000-0001-8401-0545

REFERENCES

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., … Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *181*(4), 1193–1209.

Civelek, M., Wu, Y., Pan, C., Raulerson, C. K., Ko, A., He, A., … Lusis, A. J. (2017). Genetic regulation of adipose gene expression and cardio-metabolic traits. *American Journal of Human Genetics*, *100*(3), 428–443.

Cui, S., Ji, T., Li, J., Cheng, J., & Qiu, J. (2016). What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Statistical Applications in Genetics and Molecular Biology*, *15*(2), 87–105.

Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, *13*(3), 539–552.

Hejazi, N., Phillips, R., Hubbard, A., & van der Laan, M. (2018). methyvim: Targeted, robust, and model-free differential methylation analysis in R [version 1; peer review: 1 approved with reservations, 2 not approved]. *F1000Research*, *7*, 1424.

Hejazi, N. S., Kherad-Pajouh, S., van der Laan, M. J., & Hubbard, A. E. (2017). Variance stabilization of targeted estimators of causal parameters in high-dimensional settings. *arXiv e-prints*, *1710*, 05451. page arXiv.

Heller, R., Manduchi, E., & Small, D. (2008). Matching methods for observational microarray studies. *Bioinformatics*, *25*(7), 904–909.

Hernán, M. A., & Cole, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, *170*(8), 959–962.

Hernán, M. A., & Robins, J. (2020). *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC.

Kuroki, M., & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, *101*(2), 423–437.

Laakso, M., Kuusisto, J., Stancáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., … Boehnke, M. (2017). The Metabolic Syndrome in Men study: A resource for studies of metabolic & cardiovascular diseases. *Journal of Lipid Research*, *58*(3), 481–493.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, *15*(2), R29.

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, *3*(9), 1–12.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, *28*(4), 553–561.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Moodie, E. E. M., Saarela, O., & Stephens, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, *7*(1), e205.

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2017). An introduction to g methods. *International Journal of Epidemiology*, *46*(2), 756–762.

Naimi, A. I., & Kennedy, E. H. (2017). Nonparametric double robustness. *arXiv e-prints*, *1711*, 07137. page arXiv.

Nebert, D. W., & Russell, D. W. (2002). Clinical importance of the cytochromes P450. *Lancet*, *360*(9340), 1155–1162.

Pirracchio, R., Resche-Rigon, M., & Chevret, S. (2012). Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC Medical Research Methodology*, *12*(70), 70.

R Core Team (2019). R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Saul, B. C., & Hudgens, M. G. (2020). The calculus of M-Estimation in R with geex. *Journal of Statistical Software*, *92*(2), 1–15.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–26.

Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, *21*(9), 2067–2075.

Snowden, J. M., Rose, S., & Mortimer, K. M. (2010). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, *173*(7), 731–738.

Stefanski, L., & Boos, D. (2002). The calculus of M-Estimation. *The American Statistician*, *56*(1), 29–38.

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *174*(2), 369–386.

Sun, S., Hood, M., Scott, L., Peng, Q., Mukherjee, S., Tung, J., & Zhou, X. (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Research*, *45*(11), 1–15.

Suzuki, E., Tsuda, T., Mitsuhashi, T., Mansournia, M. A., & Yamamoto, E. (2016). Errors in causal inference: An organizational schema for systematic error and random error. *Annals of Epidemiology*, *26*(11), 788–793.

Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., & Van Wieringen, W. N. (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, *14*(1), 113–128.

Vander Weele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Reifeis SA, Hudgens MG, Civelek M, Mohlke KL, Love MI. Assessing exposure effects on gene expression. *Genetic Epidemiology*. 2020;44:601–610. https://doi.org/10.1002/gepi.22324