Substance Use Transmission and Outcomes: Using Genetically Informative Research Designs for Causal Inference with Observational Data

A Dissertation
SUBMITTED TO THE FACULTY OF
THE UNIVERSITY OF MINNESOTA
BY

Gretchen Ruth Baker Saunders

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisers: Matt McGue and Niels Waller

July 2019

Acknowledgements

The reported research was made possible by NIH grants AA009367, AA011886, DA005147, DA036216, DA024417, and MH066140. The author, Gretchen Ruth Baker Saunders, was supported by the Eva O. Miller Fellowship from the University of Minnesota Graduate School.

Thank you to my advisers, Dr. Matt McGue and Dr. Niels Waller, who have provided me constant support and encouragement. Thank you, Matt, for your patience, belief in my abilities, and the sheer amount of time you invested in my success. Your kindness has made an immeasurable impact on me and who I strive to be both in academia and in life. Thank you, Niels, for your empathy and willingness to listen and support me in times of difficulty and struggle. I am greatly appreciative that you believed I could do it all even when I was sure it was impossible. To both Matt and Niels, I have come to realize that the support I received from you was unusual, and for that I am eternally grateful. It kept me moving when I was utterly overwhelmed and sure of failure.

Thank you to Dr. James Lee and Dr. Nate Helwig for serving on my preliminary and final exam committees and for taking time to provide invaluable feedback. And thank you to many other who contributed to my success: Dr. Irene Elkins for giving me the opportunity to work as her research assistant over many years and for being an extraordinarily kind and generous person; Dr. Scott Vrieze and Mengzhen Liu for providing data and feedback to strengthen pieces of this dissertation; Dr. Steve Malone for ideas, comments, and kindness throughout the years; and to everyone at the MCTFR including participants, staff, and researchers.

Thank you to my husband, Nick, who has always thought I was smarter and more capable than could ever be true. My accomplishments would mean little to me without you to share them with. Thank you to my mother for raising me to be strong and critical but to never take myself too seriously, to my father for calling me ‘doctor’ since I was a child, and to my sister, nephews, and niece for providing joy and motivation. Finally, thank you to my son, Elliott. While it will be many years before you are able to read this, you have changed my life in a way that is not possible to describe. You are a silly, sensitive, empathetic, and curious observer, and have provided me constant joy and deep meaning in life. I love you.

Abstract

One of the most difficult, yet arguably the most important aspect of research is the issue of causal inference using observational data. For phenotypes like substance use, in which it is impractical or unethical to conduct randomized controlled trials, understanding the causal mechanisms that influence substance use behavior as well as the outcomes caused by these behaviors remains difficult. The current work explores how genetically related samples can be exploited to better understand the causal effects of environmental factors on adult outcomes related to early substance use. In Study 1, polygenic risk scores for alcohol and tobacco use are used to identify a genetic nurture effect of parental smoking initiation on offspring alcohol and tobacco use in a large parent-offspring sample. The effect of parental genotype on offspring use is mediated by parental socioeconomic status (SES), suggesting that rearing SES, or the resources higher SES provide, may causally influence substance use in adolescence. Study 2 is a methodological exploration of co-twin control (CTC) designs, in which an exposure-outcome effect is decomposed into a within-twin pair and between-twin pair effect. A limitation of the CTC design is that it cannot implicitly control for environmental factors that are not perfectly shared within a twin pair, the presence of which may bias CTC findings. We use analytical derivations and simulations to show that while inclusion of a covariate as a proxy measure of a non-shared environmental confounder will always reduce bias, results from CTC studies will continue to be biased away from the null to at least some extent in most practical situations. Interpretation and suggestions for use of CTC, and more generally between-within, models are discussed. Finally, in Study 3 we

use a large sample of twins to investigate the adult socioeconomic outcomes related to adolescent substance use. Using the co-twin control (CTC) design we find that within monozygotic (MZ) twin pairs, who share all genetic and common environmental factors, the twin who consumes more tobacco and alcohol in adolescence has lower educational attainment and occupational status in adulthood compared to their lesser using co-twin, consistent with a causal effect of early substance use on later socioeconomic outcomes. We focus on interpretation of these results in the context of findings from Study 2.

Table of Contents

List of Tables

List of Figures

Introduction

That ‘correlation does not imply causation’ is a mantra equally understood and readily stated by undergraduate students and established researchers alike. Every statistics course drills this idea into its students to avoid the logical fallacy of attributing cause and effect of two variables based only on the correlation between them. There are other explanations of such associations like reverse causality, bi-directional or cyclic causation, and confounding, in which both variables share a common cause but do not cause each other (Greenland, 2003; Rutter, 2007). While much of the existing research in psychology, and the social sciences more broadly, makes claims only of associations between variables, being very clear to avoid causal implications, this ignores the fact that the primary goal of conducting such research is to understand cause and effect. Disease prevention and treatment, public health recommendations, and policy advice is based on our understanding and assumptions of the causal effects of an exposure on outcome.

Causal inference with observational data, however, is a difficult problem. For many phenotypes we are interested in, like substance use behaviors, randomized controlled trials (RCTs), in which participants are randomly assigned to treatment or control conditions, are either impractical or unethical to conduct. Additionally, RCTs may not unequivocally answer causal questions as they can be biased by subject non-adherence to treatment, loss to follow-up (attrition), and differences between subjects who participate and those who refuse participation (West et al., 2008). In order to make causal claims, the effect of substance use on later outcomes would likely need to be

measured over a period of many years. In this case an RCT may not be feasible logistically or financially in the long-term.

Genetically informative samples can add information to the understanding of causal relationships and provide stronger tests of causal inference than using unrelated individuals alone. Two ways in which genetic relationships can be exploited to better understand causal effects are 1) directly genotyped parent-offspring samples and 2) large, longitudinal twin samples. Genotype data, measured as single genes or polygenic risk scores, may help us to understand the environmental transmission of behavior from parents to offspring. In other words, we are better able to identify shared environmental causes of behavior that could theoretically be modified to alter outcomes. Study 1 applies this design to substance use phenotypes by investigating the intergenerational transmission of tobacco and alcohol use behaviors through the indirect effect of untransmitted genomes from parents to offspring, previously called a genetic nurture effect (Kong et al., 2018). We measure the genotype directly using polygenic risk scores (PRSs), which are weighted sums of associated risk alleles for a phenotype, for some subset of ranked markers. Polygenic risk scores are computed for a sample of parents and offspring using weights derived from the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) for smoking initiation, cigarettes per day, and drinks per week (Liu et al., 2019). Parental PRS is used to predict offspring substance use behaviors controlling for offspring's own PRS. In this way, we test whether parental genomes are associated with offspring substance use independent of the alleles that are transmitted through inheritance (i.e., controlling for shared genes). We then explore whether any

identified genetic nurture effects are mediated by the shared environmental factors of parental socioeconomic status (SES), which would be consistent with a causal effect of rearing SES on offspring substance use.

Twin pair samples may also provide stronger tests of causal inference by examining differences in outcomes within exposure discordant twin pairs. This design is often called the co-twin design (CTC) or, more generally, between-within models. Perhaps the most famous example of the discordant twin design was in examining the association between smoking and lung cancer. Fisher, while rightfully famous for a large body of work, strongly, and erroneously, objected to the idea that smoking caused lung cancer arguing that the observed associations where due to genetic confounding (Stolley, 1991). Twin studies later provided evidence of a causal exposure effect by showing that within discordant twin pairs, the twin who smoked had higher rates of lung cancer than their abstaining co-twin (Carmelli & Page, 1996). The strength of the CTC design comes from the fact that monozygotic (MZ) twins share all of their genetic material and rearing environment. In MZ twin pairs any difference in outcomes within a pair must be due to differences in an unshared exposure. Thus, if the heavier smoking twin has increased risk of lung cancer compare to their lesser using co-twin, this provides stronger evidence that tobacco use contributes to lung cancer risk, which we now take to be fact (Sasco, Secretan, & Straif, 2004).

Co-twin control designs are increasingly being used in psychology and epidemiology though this design has not been fully explored methodologically. Previous simulation studies have shown that non-shared confounders within twin pairs can

increase the bias in CTC estimates (Frisell, Öberg, Kuja-Halkola, & Sjölander, 2012). It remains unclear, however, whether measured covariates, as proxies of unmeasured non-shared confounders, can reduce or eliminate this bias. Study 2 uses analytical proofs and simulations to determine under what conditions this bias may be reduced or eliminated with covariate inclusion, and whether these situations are encountered in practice. With the increasing use of the CTC design, understanding the methodological implications of model specification and correct interpretation is critical. Finally, Study 3 makes use of a genetically informative twin sample in combination with CTC models to explore the causal relationship between adolescent substance use and adult socioeconomic outcomes. We examine the relationship between adolescent alcohol and tobacco use and the adult outcomes of educational attainment and occupational status within twin pairs. If adolescent substance use is causally related to reduced SES attainment, for twins who are discordant on early substance exposure, we would expect the heavier using twin to have lower attainment than their lesser using co-twin. We additionally explore the impact of IQ and externalizing behavior as potential sources of non-shared environmental confounding. Interpretation of these results is discussed in light of findings from Study 2.

Study 1

It is widely accepted that alcohol and tobacco use behaviors follow both genetic and environmental transmission. A large meta-analysis concluded that approximately 50% of the variance in alcohol use disorder is attributable to genetic factors, while 10% is due to shared environmental factors (Verhulst, Neale, & Kendler, 2015). Among those who ever initiate, 46% – 75% of the variance in tobacco use is associated with genetic factors, while 0% - 28% of the variance is due to the shared environment (Li, Cheng, Ma, & Swan, 2003; Sullivan & Kendler, 1999; Vink, Willemsen, & Boomsma, 2005). The influence of genetic factors appears to increase from first exposure to alcohol and tobacco in adolescence through early adulthood, mirroring a decline in the shared environmental effect (Kendler, Schmitt, Aggen, & Prescott, 2008).

A goal of substance use research is in determining whether, and to what extent, environmental factors causally influence substance use outcomes. A problem in identifying causal exposure effects, however, is the fact that genetic and environmental influences are not independent of each other. The idea that environmental exposures are associated with genetic risk has been termed gene-environment correlation (rGE). There are three types of rGE: passive, evocative, and reactive (Jaffee & Price, 2008; Scarr & McCartney, 1983). We focus only on passive rGE here, which occurs when parents pass on genes to their offspring that influence substance use behaviors while also fostering an environment that influences risk.

Detecting the presence of passive rGE is of importance if we are interested in the causal effect of environmental exposures on offspring outcomes (Jaffee & Price, 2012).

The existence of passive rGE can imply that shared genes confound the relationship between a rearing environmental exposure and offspring substance use, inducing a spurious correlation between the two. In this case the environmental exposure is non-causal and modifying it will likely have no impact on substance use outcomes. Another possibility is that while there might be some confounding by genotype, there is also evidence of an environmental exposure effect that is independent of the genes transmitted by parents. This would be consistent with a causal effect of the environmental exposure on offspring substance use and would suggest that modifying this exposure may be a worthwhile goal in reducing substance use. To date, much of the evidence for passive rGE has been provided by twin and adoption studies.

Twin studies support the existence of passive rGE by showing that environmental exposures are themselves heritable, meaning there are genetic influences acting on these exposures. A large meta-analysis showed that a wide range of environmental measures (e.g., stressful life events, parenting behaviors, and family environments) are modestly to moderately heritable (Kendler & Baker, 2007). The overall weighted heritability, i.e., the proportion of variance attributable to genetic factors, of all included environmental measures was 27%. Similarly, a separate large twin study showed that warmth and conflict in parent-child relationships was influenced by genetic factors with heritability estimates ranging from 14% - 56% depending on the sex and age of offspring, and that the genetic influence appeared to increase over adolescence (Ludeke, Johnson, McGue, & Iacono, 2013). Adoption studies also provide evidence for passive rGE by showing correlations between adoptive parents and their adopted children despite not transmitting

any genes. For instance, adoption studies have found a correlation between parental smoking and adopted offspring tobacco use (Keyes, Legrand, Iacono, & McGue, 2008; Samek et al., 2014). There did not, however, appear to be evidence for passive rGE between parental drinking and adopted offspring alcohol use (Samek et al., 2014). Lastly, the children of twins design has also been used to show that the relationship between parental divorce and offspring drug use may be confounded by shared genetics, suggesting a lack of passive rGE (D'Onofrio et al., 2006).

These methods, however, have several limitations. Adoption studies can be difficult to conduct with large samples while twin studies can only investigate passive rGE with exposures that differ within a twin pair. The effect of family-level exposures (i.e., rearing socioeconomic status) on offspring developmental outcomes would be missed using a twin design (Plomin, 2014). More recently, molecular genetic approaches have been used to test for passive rGE. These designs measure the genotype directly, testing for environmental effects on offspring outcomes through the parent's genotype that is not passed to their offspring, termed a "genetic nurture" effect (Kong et al., 2018). For example, parents have genes that influence their risk of substance use, some of which will be randomly transmitted to their offspring. Other (non-transmitted) genes may influence some parental phenotype (environmental exposure) that causally influences offspring substance use. Studies using single genes have failed to find evidence of passive rGE between DAT1 and DRD4 as predictors of prenatal smoking and offspring externalizing behaviors (O'Brien, Mustanski, Skol, Cook, & Wakschlag, 2013) and between the COMT gene and schizotypal traits in offspring (Savitz, van der Merwe,

Newman, Stein, & Ramesar, 2010). These results suggest that the maternal genotype of DAT1 and DRD4 has a direct effect on offspring externalizing through genes that are passed from mother to offspring, but there is no evidence for an indirect effect of the genotype through maternal prenatal smoking. In this case, the evidence is inconsistent with a causal effect of prenatal smoking on offspring externalizing after controlling for shared genetics.

Going beyond single genes, polygenic risk scores may provide greater power in detecting the effects of passive rGE. A polygenic risk score (PRS) is a weighted sum of risk alleles associated with a phenotype. Scores are computed for each individual based on single-nucleotide polymorphisms (SNPs) associated with phenotypes of interest, like tobacco and alcohol use, in previous large-scale genome wide association studies (GWAS). Polygenic scores for behavioral undercontrol (parent genotype) have suggested a lack of passive rGE between parent monitoring (parent phenotype) and offspring affiliation with substance using peers (Elam et al., 2017). Polygenic scores for educational attainment, which so far have provided some of the greatest predictive accuracy for behavioral phenotypes, have been used to provide evidence that the relationship between socioeconomic status and offspring educational attainment may be causal in nature (Bates et al., 2018; Kong et al., 2018, Willoughby et al., in press).

The current study makes use of polygenic risk scores for tobacco and alcohol behaviors to test for the presence of a genetic nurture effect in a large parent-offspring sample. We first derive polygenic scores for smoking initiation, cigarettes per day, and drinks per week and assess their predictive power for all outcomes. We then examine

whether the influence of parental substance use on offspring substance use outcomes is due only to the genetic risk they pass (measured here by polygenic scores), consistent with the presence of passive rGE and a non-causal environmental exposure effect on offspring outcomes, or if there is an independent effect of the non-transmitted risk genes, through some form of parental phenotype, on offspring substance use. The latter case would suggest the presence of a genetic nurture effect and would be consistent with a causal environmental exposure effect on offspring substance use. Figure 1 illustrates the logic of the current analysis. There are two possible paths from parent genotype to offspring substance use: through the genes they transmit to offspring (denoted as T in Figure 1) and through an effect of the non-transmitted genes (NT) on some parental phenotype, or environmental mediator, to offspring outcomes. By controlling for offspring's own polygenic risk score, thereby controlling for genetic confounding, we are able to test whether there is evidence for an effect of parent phenotype through non-transmitted genetic risk (denoted by the path of dashed arrows in Figure 1). Our hypotheses are:

1. The derived PRSs will predict all substance use outcomes.
2. Parental PRSs will predict offspring substance use outcomes.
3. Parental PRSs will predict offspring substance use outcomes independent of the transmitted genetic risk (evidence for a genetic nurture effect).

Methods

Sample

The current sample includes parents and offspring from the Minnesota Center for Twin and Family Research (MCTFR), a longitudinal study of risk for substance misuse (Iacono, Carlson, Taylor, Elkins, & Mcgue, 1999; Keyes et al., 2009). MCTFR participants were ascertained through Minnesota state birth records from the birth years 1972 to 1994, and are largely representative of families in Minnesota at the time of recruitment (e.g., approximately 98% of the sample is Caucasian). The current sample is drawn from three cohorts of same-sex twin pairs: an older cohort (birth years of 1972-1979) first assessed at a target age of 17 (N=1,252; age at first assessment M=17.5, SD=0.5), a younger cohort (birth years of 1977-1984) first assessed at a target age of 11 (N=1,512; age at first assessment M=11.7, SD=0.4); and a third ('ES') cohort (birth years of 1988-1994) first assessed at a target age of 11 (N=998; age at first assessment M=11.9, SD=0.4). Each cohort of twins was assessed approximately every three years, with up to six assessments between the ages of 11 and 29. Through an overlapping cohort design, the older cohort contains data from assessment ages 17, 20, 24, and 29; the younger cohort for assessment ages 11, 14, 17, 20, 24, and 29; and the ES cohort for assessment ages 11, 14, 17, 20, and 24. Follow-up participation rates ranged from 88.5% to 94.0% across assessments.

Parents reported on their own substance use at their offspring's intake assessment (mean parental age of 42.8 years of age). For validating the derived PRSs, all individuals with genotype and phenotypic outcome data are included, resulting in samples sizes ranging from N=758-3,012 for offspring and N=2,618-2,880 for parents. Table 1 provides approximate sample sizes for each outcome. Of note, the sample size for

cigarettes per day (CPD) at age 29 is far smaller than at earlier ages because data on CPD was only collected from one cohort due to funding limitations at the time. In testing our core hypothesis of a genetic nurture effect, all offspring with genotype and phenotypic outcome data, as well as genotype data from at least one parent, are included. This results in sample sizes ranging from N=695-2,879 from N=435-1,491 families, which are slightly smaller than those shown in Table 1 due to the availability of parental genotype data.

The effects of sample attrition were evaluated by comparing participants and nonparticipants at a given assessment with their responses at the previous assessment. For the tobacco outcome, age 24 participants smoked fewer cigarettes per day than nonparticipants at their prior age 17 assessment (standardized mean difference of 0.10, p=.02), while age 29 participants smoked more cigarettes per day at their prior age 24 assessment than did nonparticipants (standardized mean difference of -0.14, p<.001). For alcohol use, age 24 participants had higher drinking index scores (defined below) than nonparticipants at their prior age 17 assessment (standardized mean difference of 0.18, p<.001). Age 29 participants also had higher drinking index scores at their prior age 24 assessment than did nonparticipants (standardized mean difference of 0.27, p=.01). Overall, all standardized mean differences were less than or equal to 0.27, indicating minimal bias in the sample, with a slight overrepresentation of those with higher levels of alcohol use.

Phenotypic Measures

Outcome measures were collected in offspring at the age 14, 17, 20, 24, and 29 assessments using either the Substance Abuse Module (SAM; Robins, Babor, & Cottler, 1987) of the Composite International Diagnostic Interview (CIDI; Robins et al., 1988) or the Computerized Substance Use Questionnaire (CSU), depending on the age of the participant. Parental substance use measures were collected at the offspring intake assessment. The tobacco related outcome variable is an ordinal measure of cigarettes per day (CPD), including all forms of tobacco, reported in the past 12 months. The CPD measure is coded only for offspring who have reported ever smoking and ranges from 1 (less than one cigarette/one cigar or pipe/one pinch per day) to 6 (2 or more packs/20 or more cigars or pipes/2 or more tins per day). In this way, we are interested in smoking progression among those who have ever been exposed, rather than smoking initiation. The current analysis uses offspring CPD at age 17, 24, and 29, while the parent measure of CPD is their lifetime maximum.

Alcohol related outcome variables include a composite measure of alcohol consumption (ALC) at ages 17, 24, and 29 in offspring and at the twin intake in parents (mean age of 42.8 years old). The drinking index incorporates measures of quantity, frequency, number of intoxications, and maximum number of drinks reported in the previous 12 months. Specifically, it consists of four self-report alcohol use items: frequency of alcohol use (scored from 0=never to 5=at least once per day), average number of drinks per drinking event (scored from 0=never drank to 6=30 or more), maximum number of drinks in a 24-hour period (scored from 0=never drank to 6=30 or more), and number of times intoxicated (scored from 0=never to 6=50 or more). This

composite index provides a more comprehensive view of overall alcohol exposure than a single component measure.

Covariate measures include sex, age at reported use (to account for the variation in actual age at each target assessment), year of birth (up to a cubic term), and rearing socioeconomic status (SES). Rearing SES was defined by the maximum education attained by either parent on a 5-point scale (1 = < High School, 2 = High School, 3 = Some College, 4 = College, 5 = Professional Degree) as well as the maximum occupational status reported by either parent (coded on a 1-7 Hollingshead scale only for those in a full-time job and reversed so that higher scores reflect higher occupational status). We additionally adjusted for the first five genetic principal components to better account for possible spurious associations due to population stratification (Price et al., 2006). Population stratification refers to allele frequency differences due to ancestral differences, and not the phenotype of interest. Failure to account for the correlations between variants and ancestry can lead to false positives in which we conclude there is an association between the genotype and phenotype when, in reality, there is not.

Genotypic Measures

Participants were genotyped on 527,829 single nucleotide polymorphism (SNP) markers using Illumina's Human660W-Quad array (Miller et al., 2012). Polygenic risk scores were then computed for parents and offspring for cigarettes per day among ever smokers (CPD PRS), smoking initiation (SI PRS), and drinks per week (DPW PRS). A polygenic risk score (PRS) is a weighted sum of alleles associated with a trait. In the standard way it is defined as $\hat{S} = \sum_{i=1}^{m} \hat{\beta}_i G_i$, where β_i is the regression coefficient of

SNP i and G_i is an individual's reference allele count at SNP i (Dudbridge, 2013). The regression coefficients (weights) used in computing the polygenic scores are taken from independent genome-wide association study summary statistics. For the current study, the weights for computing PRSs come from the GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN; Liu et al., 2019). Because the MCTFR contributed data to GSCAN, these participants were removed before deriving the weights.

Each PRS was computed using LDpred software which accounts for the effects of linkage disequilibrium (LD) thereby improving prediction accuracy (Vilhjálmsson et al., 2015). LDpred uses LD information as well as a prior for the marker effects to estimate mean SNP effect sizes. The fraction of causal variants was set to 1 in computing all polygenic scores, meaning that we are using information from every genetic variant. We do this to avoid cherry picking a value that results in the strongest predictive effect, which would increase the likelihood of false positive results.

Analysis

Data were analyzed in R using mixed effect models with family ID as a random effect to account for the correlated nature of the data. For the outcome of cigarettes per day ordinal logistic models (Archer, Hedeker, Nordgren & Gibbons, 2018) were used and for the drinking index linear models were used (Bates, Maechler, Bolker, & Walker, 2015). For ease of interpretation, drinking index scores and all polygenic risk scores were standardized to have a mean of 0 and a standard deviation of 1. Only participants of white European ancestry were included in the analysis as this was the population used in deriving the weights from the GSCAN consortium (Martin et al., 2017).

Results

Descriptive Statistics

Descriptive statistics for phenotypic outcomes are given in Table 1. As expected due to secular changes in tobacco use, offspring smokers at all ages report fewer cigarettes per day than smoking parents (because we do not have a continuous measure of cigarettes per day for all offspring, we used the midpoint of each CPD level when computing the means and standard deviations). The mean age of peak tobacco use in offspring is 20.5 years. We are unable to compare this to the parental generation as age at maximum tobacco use is not available. Mean drinking index scores increase from age 17 to age 24 before falling slightly at age 29, reaching a level comparable with what parents report at a mean age of 42.8 years old.

Correlations between offspring phenotypic outcomes and polygenic scores for both parents and offspring are given in Figure 2. While this figure contains a large amount of information, we highlight four patterns of results here. One, there are positive, non-zero correlations between all polygenic scores and offspring outcomes. Two, the correlations between SI PRS and all tobacco and alcohol outcomes are larger than with any other polygenic score. This holds for correlations between parental polygenic scores and offspring outcomes as well, meaning the correlations between parental SI PRS and all offspring outcomes are greater than for the other parental polygenic scores. Three, for parents and offspring, CPD and DPW polygenic scores are modestly correlated with SI PRS, but are not correlated with each other. Lastly, there appears to be a trend for an increasing correlation between polygenic scores and CPD outcomes over offspring

development. The same does not appear to be true for alcohol. The correlations between polygenic scores and alcohol outcomes appear to decline, or remain stable, from ages 17 to 29. Of additional note, there is little to no evidence for assortative mating on polygenic risk scores. None of the correlations between maternal and paternal PRSs are significantly different from zero.

Do polygenic risk scores predict own substance use?

Beginning with the tobacco outcome of cigarettes per day, we tested the predictive accuracy of all three polygenic risk scores at offspring ages 17, 24, and 29 as shown in Figure 3A. All models adjusted for sex, age at assessment, year of birth (up to a cubic term), and the first five genetic principal components, and all polygenic risk scores have been standardized. As expected, CPD PRS predicts offspring CPD at age 17 (*OR* = 1.58, 95% CI [1.29, 1.95], $p < .001$), age 24 (*OR* = 1.76, 95% CI [1.49, 2.08], $p < .001$), and age 29 (*OR* = 1.73, 95% CI [1.36, 2.22], $p < .001$). Interestingly, SI PRS has the strongest predictive effect for offspring CPD at all ages (Age 17: *OR* = 2.06, 95% CI [1.63, 2.62], $p < .001$; Age 24: *OR* = 2.02, 95% CI [1.69, 2.41], $p < .001$; Age 29: *OR* = 2.20, 95% CI [1.67, 2.89], $p < .001$). There is evidence of a small, but positive, predictive effect of DPW PRS on the CPD outcomes at age 17 (*OR* = 1.30, 95% CI [1.04, 1.62], $p = .02$) and age 29 (*OR* = 1.45, 95% CI [1.11, 1.89], $p = .01$). The effect DPW PRS at age 24 is negligible as evidenced by confidence intervals that cross an odds ratio of one.

Similar analysis for the alcohol outcome of drinking index scores is shown in Figure 3B (all reported beta coefficients are standardized). As expected, DPW polygenic scores predict drinking index scores at all ages (Age 17: $\beta = 0.09$, 95% CI [0.05, 0.13], p

< .001; Age 24: β = 0.09, 95% CI [0.05, 0.14], p < .001; Age 29: β = 0.11, 95% CI [0.07, 0.16], p < .001). Similar to tobacco, SI polygenic scores have the strongest predictive effect for offspring drinking, though this effect declines across offspring age from β = 0.17 at age 17 (95% CI [0.13, 0.20], p < .001), to β = 0.13 at age 24 (95% CI [0.08, 0.17], p < .001), to β = 0.11 at age 29 (95% CI [0.07, 0.15], p < .001). There is a near zero effect of CPD polygenic scores on alcohol outcomes at all ages as evidenced by confidence intervals that include zero.

Because we also have phenotypic outcome data for all parents, we then tested whether the predictive accuracy of the polygenic scores differed by generation. To do this we fit models in the combined parent and offspring sample that included an interaction between each polygenic score and an indicator variable of generation (parent vs. offspring). Parents reported the maximum number of cigarettes per day lifetime as well as their drinking at a mean age of 42.8 years old. To facilitate comparisons with the offspring generation, we use the maximum cigarettes per day reported by the offspring through age 29 as well as their drinking index scores from the age 29 assessment. For the outcome of maximum CPD there was an interaction between SI polygenic score and the indicator variable of generation (OR = 1.15, 95% CI [1.02, 1.30], p = .02). The direction of the interaction indicates that the effect of smoking initiation scores on maximum CPD is stronger in the parent generation compared to their offspring but the effect size is small. There was no evidence of interactions between the generation indicator variable and CPD or DPW polygenic scores, or for any interactions in relation to drinking index scores. This suggests that for the most part the effect of the polygenic scores remains

relatively consistent over time despite changes in consumption patterns between generations.

Does parental PRS predict offspring outcomes controlling for shared genetics?

Parental polygenic risk scores alone predict offspring outcomes following similar patterns of own PRS prediction. Parent CPD and SI PRSs predicts offspring CPD outcomes at all ages, while parent DPW and SI PRSs predict alcohol use. The effect of parental SI PRS on outcomes is stronger than for the other parent polygenic scores. After controlling for offspring's own polygenic scores, parental SI PRS remains a significant predictor of offspring's CPD outcomes at ages 17 and 24 (*OR* = 1.42, 95% CI [1.07, 1.88], p = .02 and *OR* = 1.33, 95% CI [1.08, 1.63], p = .01, respectively) but no longer has an effect at age 29 (*OR* = 1.02, 95% CI [0.72, 1.44], p = .92). While the effect of parental CPD polygenic scores on offspring CPD, controlling for shared genetics, increases over development (Age 17 *OR* = 1.09, 95% CI [0.82, 1.44], p = .55), there is only slight evidence of an independent predictive effect at age 24 (*OR* = 1.23, 95% CI [1.01, 1.49], p = .04). The size of the parental CPD PRS increases at age 29 but the small offspring sample size at this age results in wide confidence intervals spanning an odds ratio of 1 (*OR* = 1.30, 95% CI [0.95, 1.78], p = .11). There is no evidence of an independent predictive effect of parental DPW scores on offspring CPD after controlling for offspring's own PRS. Full results for CPD are shown in Table 2.

For offspring alcohol outcomes, after controlling for offspring's own polygenic scores, parental SI polygenic scores predict offspring's alcohol use at age 17 (β = 0.08, 95% CI [0.02, 0.13], p = .005) but no longer have an effect at ages 24 and 29 (β = 0.03,

95% CI [-0.03, 0.09], p = .37, and β = 0.03, 95% CI [-0.03, 0.09], p = .36, respectively). There is no evidence of an independent predictive effect of parental DPW or CPD scores after controlling for offspring's own PRS. Full results for alcohol use are shown in Table 3.

Because rearing socioeconomic status (SES) has previously been linked with both tobacco and alcohol use, we tested whether SES mediates the relationship between parental SI polygenic scores and offspring outcomes, shown in Figure 4. Adding parental education and occupation as covariates diminishes the effect of parent SI scores on offspring CPD at ages 17 and 24 (*OR* = 1.30, 95% CI [0.94, 1.78], p = .11 and *OR* = 1.20, 95% CI [0.95, 1.51], p = .14, respectively). The effect at age 29 remains negligible with or without the inclusion of rearing SES covariates. After inclusion of parent education and occupation, the effect of parent SI scores on offspring alcohol use at age 17 also diminishes (β = 0.06, 95% CI [-0.01, 0.12], p = .10) while the effect at ages 24 and 29 remains close to zero. Overall, after adjustment for rearing SES, there is no evidence of an effect of parental PRS on any offspring outcome.

Discussion

Using molecular genetic methods, we have provided evidence of a genetic nurture effect for offspring tobacco use in late adolescence and early adulthood, and alcohol use in late adolescence only. This methodology can separate environmental influences on outcomes from shared genetic transmission, allowing us to disentangle the potentially confounded relationship between rearing environmental exposures and offspring outcome. Our results show that there is an effect of parental genotype on offspring

smoking and alcohol use in early development independent of the genetic risk that is transmitted. This effect of parental genes must operate through the environment, causing some form of parental phenotype, to influence offspring behaviors. We find that the relationship between parental genotype and offspring substance use is environmentally mediated by parental education and occupation, evidence consistent with a causal effect of rearing SES on offspring tobacco and alcohol outcomes.

Our first hypothesis, that polygenic scores derived from the largest GWAS of alcohol and tobacco phenotypes to date would predict outcomes, was largely supported. There was strong predictive accuracy for same-substance outcomes, meaning SI and CPD PRSs predicted CPD outcomes and DPW PRSs predicted drinking index scores, and for cross-substance prediction where SI PRSs predicted drinking index scores, often to a larger extent than DPW PRSs. These results, coupled with the pattern of correlations between PRSs reported in Figure 2 and genetic correlations reported in the original GSCAN GWAS from which the present PRS weights were taken (Liu et al., 2019), support the idea that there are genetic variants common to the risk of both alcohol and tobacco use as well as risk variants specific to each substance.

We also found that the magnitude of the predictive effects is relatively stable across offspring development and over time. Figure 3, which displays the effects of the polygenic scores on tobacco and alcohol outcomes in offspring from age 17 to age 29, shows some variability in the predictive accuracy of each PRS, but there is no clear pattern of increasing or decreasing prediction over developmental age. The same is mostly true when comparing the parent and offspring generations. While we find an

interaction between SI scores and generation, such that the effect of SI PRS on maximum CPD is stronger in the parent generation, the effect size is very small. This finding is also in contrast to recent work showing an increase in the genetic influence on smoking behaviors over birth cohorts (Domingue, Conley, Fletcher, & Boardman, 2016). In general, despite large secular changes tobacco use over time, and to a lesser extent alcohol use, as well as normative increases in substance use from first exposure to young adulthood, the predictive accuracy of all polygenic risk scores appears to remain stable.

Interestingly, a common pattern that appears is the strong predictive effect of SI polygenic scores for both tobacco and alcohol outcomes across age and generation. In most cases, the SI PRS effect on drinking index scores is larger than that of the DPW PRS. We believe there are two possible explanations for this that are not mutually exclusive. One, the effect of SI PRSs is strongest because of greater precision, owing to a much larger sample size and number of detected variants in the original GSCAN GWAS (Liu et al., 2019). The GWAS of smoking initiation included N = 1,232,091 participants and identified 378 associated variants compared to N = 337,334 participants and 55 variants for cigarettes per day and N = 941,280 participants and 99 associated variants for drinks per week. Consistent with this idea, in the current sample we find that the variance explained in offspring cigarettes per day by CPD PRS ranges from 3.2%-3.6%, from 2.6-2.9% for DPW PRS, and 3.9-4.1% for SI PRS using McFadden's pseudo R^2 statistic (McFadden, 1973). We find a similar pattern for variance explained in offspring drinking index scores by CPD PRS, R^2 = 1.9%-2.8%, DPW PRS, R^2 = 2.3%-3.0%, and SI PRS, R^2 = 2.2%-3.6%. It may be the case that as sample sizes continue to increase and a greater

number of genetic variants are identified, the polygenic scores for CPD and DPW will reach a predictive accuracy closer to that of smoking initiation. A second possible explanation is that what the SI PRSs are capturing is simply a better predictor of both tobacco and alcohol use, and that this is unique to smoking initiation. The correlations between SI PRSs and CPD PRSs are r = 0.16-0.19 and between SI PRSs and DPW PRSs are r = 0.23, so it is clear that each polygenic score is indexing somewhat different genetic risk. In this case, even with increasing sample sizes and identified variants, SI PRSs will continue to have the strongest predictive accuracy for tobacco and alcohol outcomes.

Our second hypothesis, that parental PRS would predict offspring outcomes, was also supported. Parental CPD and SI polygenic scores predict offspring cigarettes per day at all ages, while parental SI and DPW polygenic scores predict offspring alcohol outcomes at all ages. There are trends for differential effect of parent polygenic scores by offspring age (e.g., a declining effect of parental SI PRS and increasing effect of parental CPD PRS on offspring CPD from ages 17 to 29) but all confidence intervals are overlapping suggesting the effect of parental PRS on offspring substance use remains relatively stable over offspring development, following a similar pattern of the effect of offspring's own PRS. Another similarity between parent and offspring PRS prediction on offspring outcomes is the strength of the effect for SI polygenic scores. This similarity does not clarify why this is the case, but we highlight that it remains when predicting offspring outcomes only from parental polygenic risk scores.

Finally, our third hypothesis of genetic nurture effects was partially supported. It was hypothesized that parental polygenic scores would predict offspring tobacco and alcohol outcomes after controlling for offspring's own PRS (and thereby controlling for shared genetic confounding). We found evidence for a genetic nurture effect of parental SI PRS on offspring CPD at ages 17 and 24 and offspring drinking index scores at age 17 suggesting that parental SI polygenic scores influence offspring outcomes in late adolescence and early adulthood independent of shared genes. The size of the parental SI PRS on offspring CPD, after controlling for offspring PRS, effect is about 61% of the size of offspring's own SI PRS effect at age 17 and 52% the size of offspring SI PRS at age 24. The effect of parental SI PRS on offspring drinking at age 17 is approximately 62% of the size of offspring's own SI PRS effect. These findings support the presence of passive rGE and a potentially causal effect of some parenting phenotype, influenced by the genetic risk variants associated with smoking initiation, on offspring substance use and are consistent with prior work in adopted families that provides evidence for environmental mediation of parental smoking on adopted offspring use (Keyes et al., 2008; Samek et al., 2014). There was also a trend for an independent effect of parental CPD PRS on offspring CPD as evidenced by an increasing effect over offspring age after accounting for offspring's own CPD PRS. The reduced offspring CPD sample size at age 29, however, results in wide confidence intervals making it difficult to determine whether this is a real effect.

Given the evidence supporting genetic nurture effects for both substances, we then explored whether rearing SES explained the relationship between parental SI PRS

and offspring outcomes. After adjusting for the parental phenotypes of educational attainment and occupational status, as measures of childhood SES, we find that the effect of parental SI PRS declines significantly. This is consistent with a causal effect of higher rearing SES on lower smoking and alcohol use in late adolescence and into young adulthood for tobacco use, and mirrors prior research finding associations between childhood SES and adolescent smoking (Borland & Rudolph, 1975; Poonawalla, Kendzor, Owen, & Caughy, 2014; Soteriades & DiFranza, 2003; Tyas & Pederson, 1998) as well as smoking into young adulthood (Jefferis, Power, Graham, & Manor, 2004; Melchior, Moffitt, Milne, Poulton, & Caspi, 2007; Patrick, Wightman, Schoeni, & Schulenberg, 2012). Our findings also support the prior relationships found between rearing SES and adolescent alcohol use (Lemstra et al., 2008), though there are inconsistent findings in the literature (Goodman & Huang, 2002; Kendler et al., 2014). We see a decline in the independent prediction of parental SI PRS as offspring reach adulthood (age 29). Presumably, the influence of parental SES on substance use would be strongest while offspring live in the home and would diminish once they leave. This lack of a genetic nurture effect on offspring in later adulthood is somewhat consistent with the prior literature showing that the relationship between rearing SES and adult substance use is weaker than in adolescence (Wiles et al., 2007). Importantly, however, in contrast to the associational literature we are controlling for genetic confounding here to provide stronger evidence of a causal relationship.

Why we find a genetic nurture effect for parental SI PRS on both tobacco and alcohol use, and not for CPD and DPW scores, may be for similar reasons as to why own

SI PRSs have the strongest predictive effect for all outcomes. As mentioned above, one reason may be greater precision in parental SI polygenic scores due to the larger discovery sample size. Larger sample sizes for CPD and DPW phenotypes may result in evidence in support of a genetic nurture effect through these polygenic scores as well. A second possibility is that there is something unique about parental SI PRSs not captured by the other scores. Smoking initiation PRSs differ from the CPD and DPW scores in that they index the genetic risk for smoking initiation rather than progression or consumption measures of substance use. It may be the case that the genetic variants associated with smoking initiation influence parental phenotypes, like attained educational and occupational status, to a greater extent than those associated with progression, like cigarettes per day and drinks per week. Indeed, the correlations between parental PRSs and their own educational attainment are r = -0.15 for SI PRS, r = -0.06 for CPD PRS, and r = 0.04 for DPW PRS (similar patterns are found for occupational status: r = -0.07 for SI PRS, r = -0.02 for CPD PRS, and r = 0.04 for DPW PRS). These correlations are consistent with prior research showing an inverse relationship between smoking initiation and educational attainment (r = -0.27) but no relationship between smoking progression and educational outcomes in those who have already initiated (r = -0.05; McCaffery et al., 2008). The same study also found a significant genetic correlation between initiation and educational attainment was r_A = -0.30 (95% CI [-0.57, -0.07] but a non-significant genetic correlation between progression and education in ever smokers (r_A = -0.21, 95% CI [-0.80, 0.23]). Taken together, this suggests that the genetic variants that influence smoking initiation influence attained SES to a greater extent than the variants influencing

substance use/progression (i.e., CPD and DPW) which may explain, in part, why we find a genetic nurture effect of parental SI PRS and not for CPD or DPW scores. This does not necessarily rule out the possibility that the effect only for smoking initiation is also due to the imprecision of DPW and CPD polygenic scores.

While the current study uses weights from the largest GWAS of alcohol and tobacco phenotypes to derive the polygenic scores, these scores only explain a small portion of variance in each outcome. The variance explained in tobacco and alcohol use by each polygenic score is similar to what was reported in the original GWAS (Liu et al., 2019) despite our use of different outcome phenotypes. As we continue to increase GWAS sample sizes and are able to explain more of the variance in tobacco and alcohol use behaviors we will have an increased ability to identify genetic nurture effects. This will also help to clarify whether the strong effect of SI polygenic scores in all models is due to precision, or something unique about these scores.

A second limitation in the interpretation of the current results is that the causal mechanisms by which rearing socioeconomic status influences offspring early tobacco and alcohol use are difficult to fully explain. Rearing SES, as measured here by parental education and occupation, is likely a marker for other factors that may be contributing to the causal effect. For instance, parents with higher SES may provide resources for their offspring, like afterschool activities and sport participation, have higher levels of parental monitoring, or influence attitudes about acceptable substance use that all contribute to lower offspring smoking and drinking. Additionally, we used only one measure each for tobacco and alcohol outcomes. It would be of interest whether the genetic nurture effects

seen for CPD and drinking index scores remain for other measures of use and abuse. It may be that there is evidence for genetic nurture effects only for certain types, or severities, of substance use behavior rather than a more broad tobacco and alcohol effect. It would also be of interest whether the genetic nurture effects are substance specific or if they index externalizing behaviors more broadly. Lastly, the vastly reduced CPD sample size at age 29 limits our ability to identify the possible genetic nurture effect of parental CPD polygenic scores. While there is trend for an increasing effect over offspring age, the confidence intervals at age 29 preclude us from identifying an independent genetic effect.

Molecular genetic methods are becoming increasingly useful tools for testing a variety of genetic and environmental effects. These may include exploring how the strength of genetic effects change across time, environmental mediation of genetic effects over the lifespan, gene-by-environment interactions, and reciprocal sibling effects. Here we show how summary statistics from large genome-wide associations, coupled with a genetically informative sample, provide insights on the causal relationship between a shared environmental factor (socioeconomic status) and offspring alcohol and tobacco use. Understanding causal relationships in observational data, as we have attempted to do here, is of primary importance for selecting interventions to reduce substance use. As we continue to increase GWAS sample sizes and better understand the effect of genetic variation on outcomes, the better we are able to elucidate the causal structure between genes, environment, and substance use.

Table 1-1. Descriptive statistics for each outcome, split by age and generation. Sample sizes include all participants with phenotypic outcome and genotype data.

	N	Female (%)	M	SD
Cigarettes per day (among smokers)				
Offspring (at age 17)	1,425	46%	6.95	9.69
Offspring (at age 24)	1,676	41%	8.19	10.20
Offspring (at age 29)	758	38%	8.55	10.74
Parents (maximum lifetime)	2,618	53%	15.50	15.46
Drinking index				
Offspring (at age 17)	3,012	53%	1.08	1.00
Offspring (at age 24)	2,251	55%	2.02	0.75
Offspring (at age 29)	2,311	53%	1.72	0.73
Parents (at twin intake)	2,880	55%	1.76	0.70

Note: M = mean; SD = standard deviation; mean values for cigarettes per day are computed by assigning the midpoint of each level as we do not have a continuous measure of cigarettes per day available at each age.

Table 1-2. Association between parental polygenic scores and offspring CPD controlling for shared genetics (own PRS).

	Age 17 (N = 1,347)	Age 24 (N = 1,567)	Age 29 (N = 695)
	OR (95% CI)	OR (95% CI)	OR (95% CI)
CPD polygenic score model:			
Parental CPD PRS	1.09 (0.82, 1.44)	1.23 (1.01, 1.49)	1.30 (0.95, 1.78)
Own CPD PRS	1.51 (1.16, 1.96)	1.55 (1.27, 1.89)	1.60 (1.18, 2.18)
SI polygenic score model:			
Parental SI PRS	1.42 (1.07, 1.88)	1.33 (1.08, 1.63)	1.02 (0.72, 1.44)
Own SI PRS	1.77 (1.33, 2.37)	1.73 (1.40, 2.14)	2.24 (1.60, 3.15)
DPW polygenic score model:			
Parental DPW PRS	0.89 (0.65, 1.20)	0.94 (0.76, 1.17)	0.74 (0.52, 1.04)
Own DPW PRS	1.41 (1.06, 1.87)	1.11 (0.90, 1.36)	1.67 (1.19, 2.36)

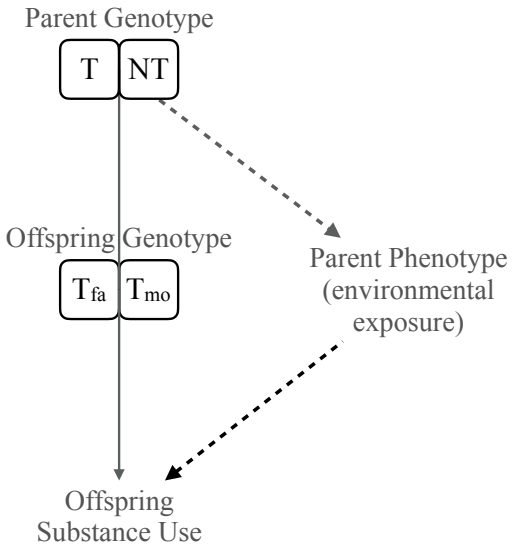
Note. All models adjust for the covariates of sex, age at assessment, year of birth (up to a cubic term), and the first 5 genetic principal components. OR = odds ratio; CI = confidence interval; SI = smoking initiation; DPW = drinks per week; CPD = cigarettes per day; prs = polygenic risk score.

Table 1-3. Effect of parental polygenic scores on offspring drinking index scores controlling for shared genetics.

	Age 17 (N = 2,879)	Age 24 (N = 2,121)	Age 29 (N = 2,178)
	OR (95% CI)	OR (95% CI)	OR (95% CI)
CPD polygenic score model:			
Parental CPD PRS	0.02 (-0.03, 0.07)	0.01 (-0.05, 0.07)	-0.01 (-0.07, 0.05)
Own CPD PRS	0.02 (-0.03, 0.06)	0.01 (-0.05, 0.06)	0.05 (-0.01, 0.11)
SI polygenic score model:			
Parental SI PRS	0.08 (0.02, 0.13)	0.03 (-0.03, 0.09)	0.03 (-0.03, 0.09)
Own SI PRS	0.13 (0.08, 0.17)	0.11 (0.05, 0.17)	0.10 (0.04, 0.16)
DPW polygenic score model:			
Parental DPW PRS	0.01 (-0.05, 0.06)	0.03 (-0.03, 0.09)	-0.01 (-0.07, 0.05)
Own DPW PRS	0.09 (0.04, 0.14)	0.07 (0.02, 0.13)	0.12 (0.06, 0.18)

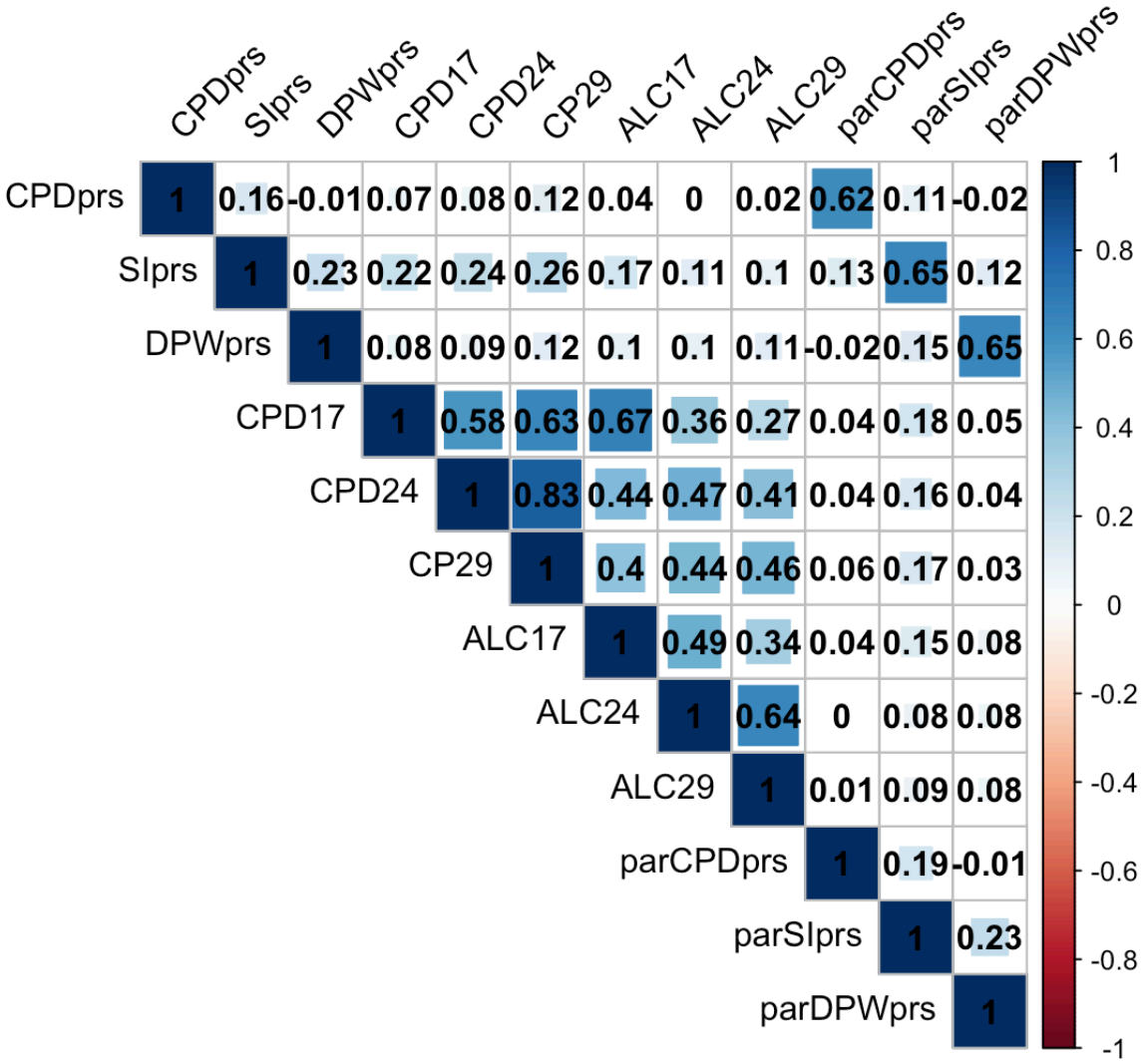
Note. All models adjust for the covariates of sex, age at assessment, year of birth (up to a cubic term), and the first 5 genetic principal components. OR = odds ratio; CI = confidence interval; SI = smoking initiation; DPW = drinks per week; CPD = cigarettes per day; prs = polygenic risk score.

Figure 1-1. Example of passive gene-environment correlation where parental genotype can confound the relationship between a rearing environmental exposure and offspring substance use.



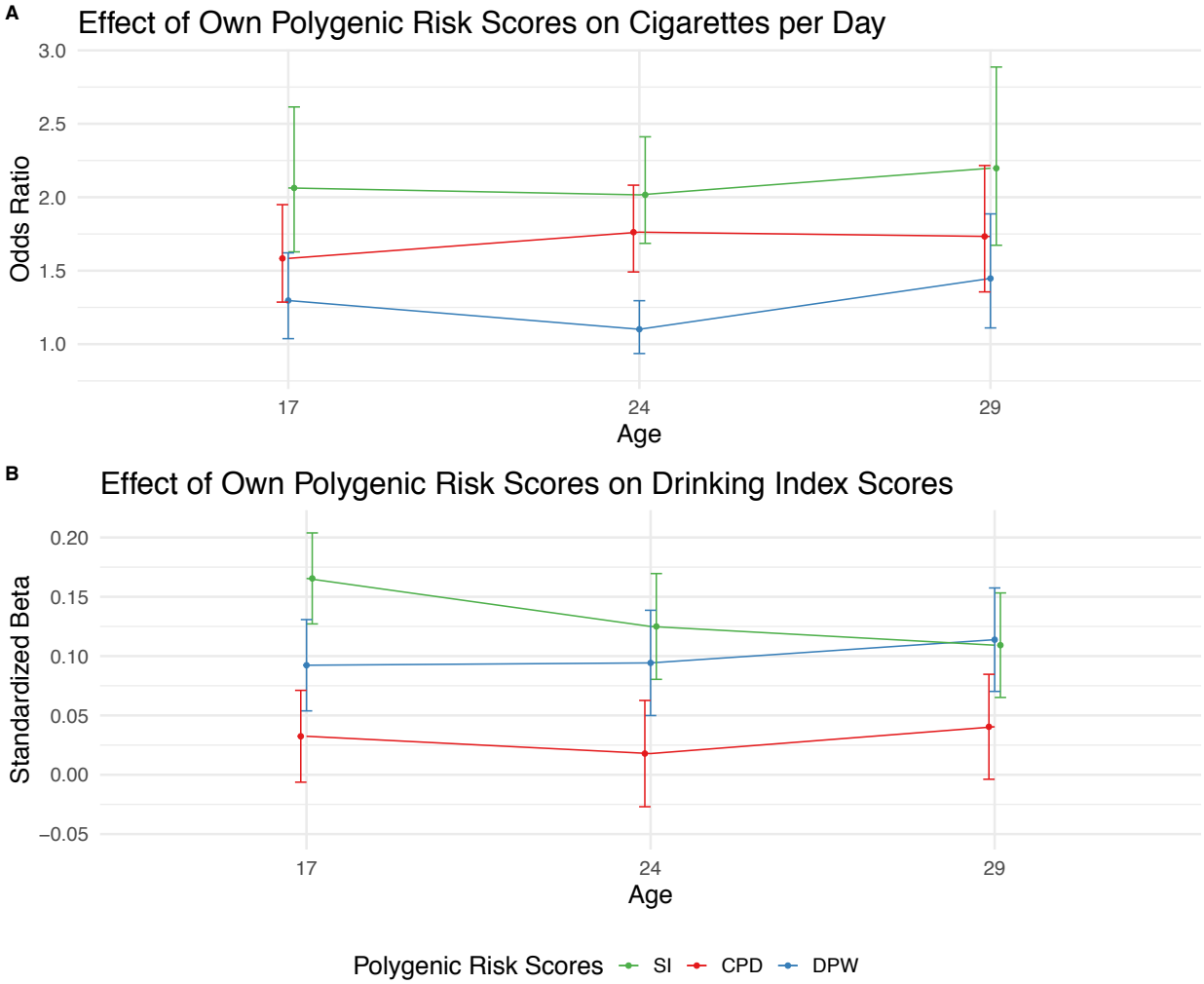
Note. T = transmitted genes (e.g., T_{fa} denotes the genes passed from father to offspring); NT = non-transmitted genes. Solid lines represent substance use risk genes that are passed from parent to offspring. Dashed lines represent an indirect effect of the non-transmitted risk genes on offspring substance use through a parent phenotype.

Figure 1-2. Correlations between parent and offspring polygenic risk scores and offspring alcohol and tobacco outcomes across development.



Note. CPD = cigarettes per day; SI = smoking initiation; DPW = drinks per week; ALC = drinking index score (e.g., ALC17 refers to offspring drinking index scores at age 17); prs = polygenic risk score; par = parent. Includes all parents in computing the mean parental PRS, not just complete parent pairs.

Figure 1-3. Prediction of offspring polygenic scores on own A) cigarettes per day and B) drinking index scores across age.



Note. CPD = cigarettes per day; SI = smoking initiation; DPW = drinks per week. Error bars denote 95% confidence intervals.

Figure 1-4. Effect of parental smoking initiation polygenic score (parental SIprs) on offspring A) cigarettes per day and B) drinking index scores.

Note. Parental PRS (polygenic risk score) only models include parent SI polygenic scores, parent and offspring PRS models additionally include offspring SI polygenic scores, and add SES additionally includes parental education and occupation level as covariates. Error bars denote 95% confidence intervals. SIprs = smoking initiation polygenic risk score; CPD = cigarettes per day.

Study 2

Co-twin control (CTC) or discordant twin models, are a special case of what are commonly referred to as between-within models (Begg & Parides, 2003; Carlin, Gurrin, Sterne, Morley, & Dwyer, 2005; McGue, Osler, & Christensen, 2010). CTC models make use of the genetic and environmental relationships within twin pairs to estimate an exposure effect controlling for all factors shared within a pair. Monozygotic (MZ) twins share all genetic factors and rearing environment, so any difference in outcome must be due to factors not shared within the twin pair. If an exposure has a causal effect on an outcome, the outcome levels will differ within exposure discordant twin pairs. In this way, the unexposed twin acts as the counterfactual to their exposed co-twin; they are an approximation of what the twin would have looked like had they not been exposed. The same logic can be extended to genetic relationships other than twins as in sibling-comparison designs (Lahey & D'Onofrio, 2010).

The power in the CTC design lies in its ability to implicitly control for all factors shared within a twin pair even when they are unmeasured (McGue et al., 2010). For this reason, CTC designs are widely used as a stronger method of causal inference than using genetically unrelated individuals (Donovan & Susser, 2011). Examples of their use range from the effects of cannabis on intelligence and educational attainment (Meier et al., 2018; Verweij, Huizink, Agrawal, Martin, & Lynskey, 2013) to alcohol's effect on stroke risk (Kadlecová, Andel, Mikulík, Handing, & Pedersen, 2015) or hippocampal volume (Wilson, Malone, Hunt, Thomas, & Iacono, 2018) to how lifestyle factors influence cancer risk (Hübinette, Lichtenstein, Ekbom, & Cnattingius, 2001; Milán, Verkasalo,

Kaprio, & Koskenvuo, 2003; Swerdlow, De Stavola, Swanwick, Mangtani, & Maconochie, 1999). Despite the increasing popularity of the CTC design it has not been fully methodologically explored. Work by Frisell and colleagues has shown that bias can be introduced in the CTC estimates in the presence of non-shared confounding (Frisell et al., 2012). The magnitude of this bias is a function of the within-twin pair correlation in the exposure and the confounder. This work also shows that measurement error in the exposure will bias the CTC estimate toward the null.

The current study builds on these findings by testing whether inclusion of a measured covariate can counteract the non-shared confounding bias. In other words, can the bias induced by a non-shared confounder be reduced when a measured covariate is included in the CTC model? Incorporating potential confounders as covariates in a regression model is a popular way of controlling for confounding bias (Greenland & Morgenstern, 2001). If the covariate is a perfect measure of the confounder, doing so will eliminate all confounding bias. Most often, however, the covariate measures the confounder with some error, resulting in residual confounding bias (Becher, 1992). Using analytic derivations and simulations, we investigate whether covariate inclusion will reduce the bias in the CTC model estimates more than in a model treating the twins as individuals, and explore what parameters affect the bias reduction in this scenario. Lastly, the impact of measurement error in not only the exposure, but also the measured covariate, is investigated. The interpretation of CTC model estimates is discussed in light of our findings.

Co-Twin Control Model

A generalized linear regression model, treating twins as individuals (the individual-level model), is given by

$$g\left\{E\left(Y_{ij} \mid X_{ij}\right)\right\} = \beta_0 + \beta X_{ij}, \quad (1)$$

where X_{ij} is the exposure of person j in twin pair i, Y_{ij} is their outcome, and $g\{\ \}$ is a link function allowing the generalized linear model to be extended to different forms of regression, like linear or logistic regression. For example, in a linear regression model Y follows a normal distribution with the identity link function $\left(g\{\mu\} = \mu\right)$.

The CTC model decomposes the exposure effect from the individual-level model $\left(\beta\right)$ into a within-twin pair and between-twin pair effect by incorporating the twin pair mean. The CTC model is given as

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \overline{X}_i\right)\right\} = \beta_0 + \beta_W\left(X_{ij} - \overline{X}_i\right) + \beta_B \overline{X}_i, \quad (2)$$

where $\overline{X}_i$ is the mean exposure of twin pair i. The within-twin pair estimate $\left(\beta_W\right)$ is the estimate of the exposure effect controlling for all shared genetic and environmental factors. The between-twin pair estimate $\left(\beta_B\right)$ is an estimate of the magnitude of confounding due to shared factors. The interpretation of β_B depends on the specification of the CTC model and can be difficult in some cases (i.e., the interpretation differs based on whether the twin deviation from the pair mean is included or only the individual's exposure level; Begg & Parides, 2003). In general, the within-twin pair effect is of more interest to researchers than the between-pair effect.

Interpretation of the within-pair effect is commonly made by comparing the magnitude of β_W from the CTC model to the magnitude of β from the individual-level model (McGue et al., 2010). When these estimates are not significantly different from one another, $\beta = \beta_W$, this would suggest that the observed association is not due to confounding factors, consistent with a causal effect of exposure on outcome. When β_W is significantly different from β but is not zero, $\beta \neq \beta_W > 0$, this suggest that the observed association is partially due to confounding factors. And, finally, when the within-pair effect is not significantly different from zero, $\beta_W = 0$, this would suggest that the entire association is due to confounding and is not consistent with a causal interpretation.

Bias due to Non-shared Confounding and Measurement Error

Prior statistical analysis of CTC models by Frisell et al. (2012) has shown that bias is induced in the within-twin pair estimate in the presence of factors that are not perfectly shared within a twin pair. Environmental confounding within twin pairs will increase bias in the within-twin pair term as a function of the degree to which such confounding reflects influences that are unshared within a pair. If all confounding variables are perfectly shared within a twin pair, the estimate of the effect of the exposure $\left(\beta_W\right)$ will be unbiased. As the correlation between confounding variables decreases within a twin pair, the estimate of the effect of the exposure $\left(\beta_W\right)$ will be biased upwards. In some cases this bias will exceed that of the individual-level effect $\left(\beta\right)$. To illustrate this, we assume that the confounding variable affects both the exposure and the

outcome, but that the exposure does not have a causal effect on the outcome. If we select twin pairs in which the members of the pair are discordant on the exposure, they will also likely be more discordant on the confounding variable than unselected twin pairs (the correlation of the confounding variable between members of a pair will be reduced). This will in turn increase the correlation between the confounder and exposure variables and create a spurious relationship between the exposure and outcome. The impact of non-shared confounders on the bias of β and β_W depends on the ratio of the within-pair correlation of the confounding variable $\left(\rho_C\right)$ to the within-pair correlation of the exposure variable $\left(\rho_X\right)$. If the correlation between confounders is greater than the correlation between exposures, the within-twin pair term is less biased than the individual-level term (if $\rho_C > \rho_X$ then $bias\left(\beta_W\right) < bias\left(\beta\right)$). If the correlation between confounders is less than the correlation between exposure, the within-twin pair term is *more* biased than the individual-level term (if $\rho_C < \rho_X$ then $bias\left(\beta_W\right) > bias\left(\beta\right)$). If the correlations are equal, both estimates will have the same amount of bias. Unless $\rho_C < 1$, however, bias will always exist in the within-pair estimate.

Additionally, random measurement error in the exposure can lead to twin pairs being incorrectly classified as concordant or discordant, which is important given that only discordant twin pairs are informative for the within-pair effect in CTC models. As measurement error increases, the within-twin pair estimate increasingly underestimates the true effect. Both biases due to confounding and measurement error affect the estimates from CTC models as well as more general between-within models.

Inclusion of a Measured Covariate to Reduce Bias

While non-shared confounding may induce bias in the within-twin pair effect, most researchers attempt to control for this by including covariates in the CTC regression model. The rationale is that the covariates incorporated into the model are an imperfect measure of unmeasured confounding variables, and by controlling for them bias due to confounding is thereby reduced. Figure 1 shows a causal diagram where the exposure-outcome relationship is confounded by an unmeasured variable, C, that also affects the measured covariate, Z.

A standard way to include covariates in CTC models is given by

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \overline{X}_{i}, Z_{ij}\right)\right\} = \beta_0 + \beta_W\left(X_{ij} - \overline{X}_{i}\right) + \beta_B \overline{X}_{i} + \beta_Z Z_{ij}, \quad (3)$$

where Z is a measured covariate. Sjölander and colleagues, however, show that this model specification does not properly adjust for the covariate and causes β_W to lose its causal interpretation (Sjölander, Frisell, & Öberg, 2012). Briefly, by conditioning on $\overline{X}_i$ a spurious association is induced between the exposure of twin 1 $\left(X_{i1}\right)$ and the covariate of their co-twin $\left(Z_{i2}\right)$ and between the outcome of twin 1 $\left(Y_{i1}\right)$ and the covariate of twin 2 $\left(Z_{i2}\right)$. Essentially, $\left(Z_{i2}\right)$ becomes a collider variable, a common effect of two or more variables (Greenland, 2003), and an artificial confounder of the exposure-outcome relationship. Given this model specification, even in the absence of a true causal effect $\left(\beta_{YX} = 0\right)$ β_W will not equal zero. The authors show that a simple modification of the model can recapture the causal interpretation of β_W:

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \overline{X}_{i}, Z_{ij}, \overline{Z}_{i}\right)\right\} = \beta_0 + \beta_W\left(X_{ij} - \overline{X}_{i}\right) + \beta_B \overline{X}_{i} + \beta_Z\left(Z_{ij} - \overline{Z}_{i}\right), \quad (4)$$

where $\overline{Z}_i$ is the mean covariate value of twin pair i (Sjölander et al., 2012). The current study explores both forms of covariate inclusion to evaluate whether confounding bias can be reduced, with particular interest in bias reduction in β_W. We focus on whether, or to what extent, bias remains in the within-pair estimate even if the causal interpretation is retained as in equation 4.

Bias Reduction with a Covariate Under a Linear Model

Assuming that all effects in the causal diagram (Figure 1) are linear and that all variables are continuous, we are able to derive the exact mathematical formula for the regression coefficients. We further assume, without loss of generality, that all variables other than error terms are standardized (mean of zero and standard deviation of one). We can then ignore the intercept term so that the true causal model is given by

$$Y_{ij} = \beta_{YX} X_{ij} + \beta_{YC} C_{ij} + \varepsilon_{Y_{ij}}, \quad (5)$$

$$X_{ij} = \beta_{XC} C_{ij} + \varepsilon_{X_{ij}}, \quad (6)$$

$$Z_{ij} + \beta_{ZC} C_{ij} + \varepsilon_{Z_{ij}}. \quad (7)$$

With this data generating structure, all confounding between X and Y is due to C, with Z being a measure of C that has no direct effect on X or Y. We let $\text{var}\left(C\right) = \sigma_C^2 = 1$, $\text{var}\left(\varepsilon_{Y_{ij}}\right) = \sigma_{Y_{ij}}^2$, $\text{var}\left(\varepsilon_{X_{ij}}\right) = \sigma_{X_{ij}}^2$, and $\text{var}\left(\varepsilon_{Z_{ij}}\right) = \sigma_{Z_{ij}}^2$. Because the causal diagram assumes twin pairs, we have $\text{cov}\left(C_{i1}, C_{i2}\right) = \rho_C \sigma_C^2$, $\text{cov}\left(\varepsilon_{Y_{i1}}, \varepsilon_{Y_{i2}}\right) = \rho_{\varepsilon_Y} \sigma_{\varepsilon_Y}^2$,

$\text{cov}\left(\varepsilon_{X_{i1}}, \varepsilon_{X_{i2}}\right) = \rho_{\varepsilon_X}\sigma^2_{\varepsilon_X}$, and $\text{cov}\left(\varepsilon_{Z_{i1}}, \varepsilon_{Z_{i2}}\right) = \rho_{\varepsilon_Z}\sigma^2_{\varepsilon_Z}$. Further, each twin's error terms (ε) are independent of all other variables and there is no correlation between the error terms of different variables within a twin pair.

We are interested in the true causal effect of X on Y (β_{YX}). Regressing Y on X and C would result in an unbiased estimate of the exposure effect. However, C is unmeasured and leaving it out results in a biased estimate of the exposure effect. We explore the bias when regressing Y on X and Z instead. Because Z is a measure of C, including it in the regression model may reduce the confounding bias induced by the unmeasured confounder, C. Further, we are interested in whether inclusion of Z reduces the bias more for the within-twin pair effect (β_W) than the individual-level effect (β).

Confounding Bias with Covariate Inclusion

The derived estimate of the exposure effect from the individual-level model without adjusting for a covariate (equation 1) is:

$$\beta = \beta_{YX} + \beta_{YC}\beta_{XC}\,. \tag{8}$$

The derived estimate of the exposure effect from the CTC model without adjusting for a covariate (equation 2) is:

$$\beta_W = \beta_{YX} + \frac{\beta_{YC}\beta_{XC}}{\left(\frac{1-\rho_{\varepsilon_X}}{1-\rho_C}\right)}\,. \tag{9}$$

The full derivation steps can be found in Frisell et al. (2012). It is clear that both estimates are a function of the true causal effect (β_{YX}) plus a bias term. Because the

within-twin pair correlation in the exposure, ρ_X, is a linear combination of ρ_{ε_X} and ρ_C, the difference in bias between the β and β_W is a function of the relative magnitudes of ρ_X and ρ_C. When $\rho_X = \rho_C$ then, by definition, $\rho_{\varepsilon_X} = \rho_C$ resulting in $\beta = \beta_W$. Following similar reasoning, when $\rho_X > \rho_C$, ρ_{ε_X} will be greater than ρ_C resulting in $\left(\frac{1-\rho_{\varepsilon_X}}{1-\rho_C}\right) > 1$. This illustrates how bias in β_W will be larger than bias in β when the within-pair correlation in the exposure is greater than the within-pair correlation in the confounder.

After inclusion of a covariate, Z, the derived exposure estimate from the individual-level model becomes (see Appendix for full derivation):

$$\beta_{\text{cov}} = \beta_{YX} + \frac{\beta_{YC}\beta_{XC}\left(1-\beta_{ZC}^2\right)}{1-\beta_{ZC}^2\beta_{XC}^2}. \tag{10}$$

The bias term now additionally depends on how well Z measures C (the magnitude of β_{ZC}), which confirms our intuition. The estimate for the within-pair effect when adjusting for a covariate the standard way (equation 3) is given by:

$$\beta_{W_{\text{cov std}}} = \frac{\beta_{YX}\left(1-\beta_{XC}^2\rho_C-\sigma_{\varepsilon_X}^2\rho_{\varepsilon_X}\right)+\beta_{YC}\beta_{XC}\left(1-\rho_C\right)-\beta_{zc}\beta_{XC}\left(1-\rho_C\right)\left(\beta_{YX}\beta_{XC}\beta_{ZC}+\beta_{YC}\beta_{ZC}\right)}{2\left(1-\beta_{XC}^2\rho_C-\sigma_{\varepsilon_X}^2\rho_{\varepsilon_X}\right)-\beta_{ZC}^2\beta_{XC}^2\left(1-\rho_C\right)^2}.$$

(11)

The estimate for the within-pair effect when adjusting for a covariate in a way that retains the correct causal interpretation (equation 4) becomes:

$$\beta_{W_{\text{cov}}} = \frac{\left[2\left(1-\beta_{ZC}^2\rho_C-\rho_{\varepsilon_Z}\sigma_{\varepsilon_Z}^2\right)\right]\left[\beta_{YX}\left(1-\beta_{XC}^2\rho_C-\rho_{\varepsilon_X}\sigma_{\varepsilon_X}^2\right)+\beta_{YC}\beta_{XC}\left(1-\rho_C\right)\right] - \left[2\beta_{ZC}\beta_{XC}\left(\beta_{YX}\beta_{XC}\beta_{ZC}+\beta_{YC}\beta_{ZC}\right)\left(1-\rho_C\right)^2\right]}{\left[2\left(1-\beta_{ZC}^2\rho_C-\rho_{\varepsilon_Z}\sigma_{\varepsilon_Z}^2\right)\left(1-\beta_{XC}^2\rho_C-\rho_{\varepsilon_X}\sigma_{\varepsilon_X}^2\right)\right]-4\beta_{ZC}^2\beta_{XC}^2\left(1-\rho_C\right)^2} . \quad (12)$$

The interpretation of this estimate is not intuitively clear, though it must depend on the within-twin pair correlation in exposure (ρ_X), the confounder (ρ_C), and the covariate (ρ_Z). Like the individual-level estimate, it also depends on the magnitude of β_{ZC} , i.e., how well the covariate measures the confounder.

Measurement Error with Covariate Inclusion

It is likely that an exposure and covariate are measured with some amount of error. It is well documented that measurement error in an exposure will attenuate the exposure effect estimate in a simple linear regression (Hutcheon, Chiolero, & Hanley, 2010; Liu, 1988; Spearman, 1904). Additionally, it has been shown that the estimate from CTC models will be attenuated more than individual-level models (Frisell et al., 2012; McGue et al., 2010). In the case of multiple regression, where covariates are also subject to measurement error, the estimated exposure effect may under- or overestimate the true causal effect (Liu, 1988; Rosner, Spiegelman, & Willett, 1990).

Let X_{ij}^* denote the observed exposure, which is equal to the true exposure (X) plus measurement error:

$$X_{ij}^* = X_{ij} + \varepsilon_{X_{ij}^*} \quad (13)$$

And let Z_{ij}^{*} denote the observed covariate, which is similarly equal to the true covariate (Z) plus measurement error:

$$Z_{ij}^{*} = Z_{ij} + \varepsilon_{Z_{ij}^{*}} \tag{14}$$

In the presence of measurement error in both the exposure and covariate, the individual-level exposure effect can be shown to be

$$\beta_{ME} = \frac{\gamma_X \operatorname{cov}\left(Y_{ij}, X_{ij}\right) - \gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij}, X_{ij}\right) \operatorname{cov}\left(Z_{ij}, Y_{ij}\right)}{1 - \gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij}, X_{ij}\right)^2}, \tag{15}$$

where γ_X and γ_Z are the reliabilities of the measures of X and Z, respectively. When both reliabilities equal 1, this simplifies to the estimate given in equation 10. As the reliability of the exposure decreases, holding γ_Z constant, the coefficient estimate will be attenuated as would be expected based on prior work (Frisell et al., 2012). As the reliability of the covariate decreases, holding γ_X constant, the coefficient estimate will be increased. This is for the same reason that small values of β_{ZC} result in increased bias; γ_Z affects the coefficient estimate in the same way β_{ZC} does.

The CTC exposure estimate based on equation 4 becomes

$$\beta_{W_{ME}} = \frac{\begin{array}{c} 2\left[\gamma_X - \gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij}, Z_{ij'}\right)\right] \gamma_X \gamma_Z \operatorname{cov}\left(Y_{ij}, X_{ij} - X_{ij'}\right) \\ -\gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right) \operatorname{cov}\left(Z_{ij} - Z_{ij'}, Y_{ij}\right) \end{array}}{\begin{array}{c} 4\left[\gamma_X - \gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\left[\gamma_Z - \gamma_X \gamma_Z \operatorname{cov}\left(X_{ij}, X_{ij'}\right)\right] \\ -\gamma_X \gamma_Z \operatorname{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)^2 \end{array}}. \tag{16}$$

Despite the increased complexity of the within-pair estimate, the same conclusion holds as above: decreasing values of γ_Z result in increased bias in the same way that small values of β_{ZC} do. Because γ_Z functions as a measure of β_{ZC} we do not include simulation results for this type of measurement error. These results would mirror the impact of β_{ZC} shown in Figure 3. Importantly, holding γ_Z fixed, $\beta_{W_{ME}}$ will be attenuated faster, as a function of γ_X, than β_{ME}, as would be expected based on prior simulation work and discussion (Frisell et al., 2012; McGue et al., 2010).

Simulation Study

To help interpret how covariate inclusion affects bias in CTC models, we simulated paired data according to the data generating structure in Figure 1. Details of the simulation set-up are included in the Appendix. The values chosen for each parameter were mostly arbitrary though we attempted to choose practical values (R code is included in the Appendix if readers wish to test other parameter combinations). The general pattern of results holds for all values chosen though in some cases a particular combination of parameters is not possible (e.g., low ρ_Z, high ρ_C, and high β_{ZC}). For this reason some lines in the figures illustrating the results may abruptly cut off when an inadmissible situation occurs. Figure 2 essentially recapitulates the work of Frisell et al. (2012) whereas Figure 3 extends this to a variety of situations. In both figures simulation results (solid lines) are overlaid on the derivation results (dashed lines) to show their concordance. In Figure 3, darker lines denote the exposure effect estimate with covariate inclusion while lighter shaded lines denote the same estimate without covariate inclusion

to better show the change in bias between these models. The true causal exposure effect was zero for all simulations ($\beta_{YX} = 0$).

Confounding Bias

Figure 2 shows how non-shared confounding induces bias in both the individual-level and within-pair exposure effect, and how the bias is affected by the relationship between the within-pair correlation in the exposure and confounder in the absence of covariates (Frisell et al., 2012). The blue line indicates the estimated exposure effect from the individual-level model while the red line indicates the within-pair effect from the CTC model. Because no covariates are included in either model bias does not depend on the magnitude of β_{ZC}. Each panel shows the bias under the possible relationships between ρ_X and ρ_C: $\rho_X < \rho_C$, $\rho_X = \rho_C$, and $\rho_X > \rho_C$. As was found in previous work, when $\rho_X > \rho_C$ the β_W estimate from CTC models is a more biased estimate of the exposure effect than the individual-level β.

We now consider each relationship between ρ_X and ρ_C separately. Figure 3A illustrates the bias when the twin correlation is greater for the covariate than the exposure ($\rho_C > \rho_X$) with the inclusion of a covariate. In this case, based on findings from Frisell et al. (2012), we expect that β_W will be less biased than β. We do indeed see that for most values of ρ_Z and β_{ZC}. As β_{ZC} increases, meaning the covariate is an increasingly accurate measure of the confounder, the bias decreases in both β_W and β, as would be

expected. The magnitude of ρ_Z, the within-pair correlation in the covariate, affects the rate at which the bias decreases in the β_W coefficients only. When ρ_Z is high the rate of decrease in bias of the β_W estimate is the highest. Comparing both forms of covariate inclusion, when β_{ZC} is low $\beta_{W_{\text{cov}}}$ and $\beta_{W_{\text{cov std}}}$ perform similarly. As the value of β_{ZC} increases, $\beta_{W_{\text{cov std}}}$ shows less bias at low values of ρ_Z while $\beta_{W_{\text{cov}}}$ shows less bias at high values of ρ_Z.

Figure 3B illustrates the bias with the inclusion of a covariate when $\rho_X = \rho_C$. In this case, we expect that β_W will have the same amount of bias as β. This occurs only when ρ_Z is also the same (i.e., $\rho_X = \rho_C = \rho_Z$). When ρ_Z is low the within-pair effect is less biased than the individual-level effect. The reverse is true when ρ_Z is high. As in the previous scenario, as ρ_Z increases in magnitude the rate of bias reduction also increases but only for the within-pair effect. Comparing both forms of covariate inclusion in this scenario, $\beta_{W_{\text{cov std}}}$ shows similar bias to β across all values of β_{ZC} and ρ_Z. As the value of β_{ZC} increases, $\beta_{W_{\text{cov}}}$ shows increased bias at low values of ρ_Z but reduced bias at high values of ρ_Z.

Finally, Figure 3C illustrates the bias with the inclusion of a covariate when $\rho_X > \rho_C$. This is the "worst case" scenario where we expect that β_W will have more bias

than β. As β_{ZC} increases the bias in both estimates decreases. Additionally, as ρ_Z increases there comes a point at which β_W is less biased than β. It is clear, however, that this only occurs when ρ_Z is high and for narrow ranges of β_{ZC}. Finally, comparing both forms of covariate inclusion, we see a similar relationship between $\beta_{W_{\text{cov}}}$ and $\beta_{W_{\text{cov std}}}$ as in Figure 3A. When β_{ZC} is low $\beta_{W_{\text{cov}}}$ and $\beta_{W_{\text{cov std}}}$ perform similarly. As the value of β_{ZC} increases, $\beta_{W_{\text{cov std}}}$ shows less bias at low values of ρ_Z while $\beta_{W_{\text{cov}}}$ shows less bias at high values of ρ_Z. Interestingly, $\beta_{W_{\text{cov std}}}$ never results in less bias than β even at very high values of β_{ZC} and ρ_Z.

Discussion

The current study extends work by Frisell et al. (2012) by showing that inclusion of a covariate as a proxy measure of a confounder reduces bias in CTC exposure effect estimates in only a limited set of circumstances. It remains that in most situations likely encountered in practice β_W will be a biased estimate of the true causal exposure effect. This result has important implications for the use and interpretation of CTC, and more broadly between-within, models.

As previously shown in CTC models, when the within-twin pair correlation in the exposure is greater than the within-pair correlation in the confounder (i.e., $\rho_X > \rho_C$), β_W will be more biased than the individual-level β. In this "worst case scenario" one may chose to include a covariate measure as a proxy of the confounder in order to reduce this

bias. While covariate inclusion reduces bias in β_W more than in β as illustrated in Figure 3, the current work shows that β_W will be less biased than β only when the within-pair correlation in the covariate (ρ_Z) is high and the covariate is an accurate measure of the confounder (β_{ZC} is large). In comparing forms of covariate inclusion, $\beta_{W_{\text{cov std}}}$ generally shows less bias than $\beta_{W_{\text{cov}}}$ when ρ_Z is low but shows greater bias at high values of ρ_Z. While it may be the case that using $\beta_{W_{\text{cov std}}}$ results in the greatest bias reduction in the exposure effect estimate, this form of covariate inclusion does not retain its assumed causal interpretation (Sjölander et al., 2012). The increased bias reduction in select scenarios is not sufficient to justify its use over $\beta_{W_{\text{cov}}}$ which does retain the correct causal interpretation.

The effect of β_{ZC} on these results is intuitive. If the covariate is an accurate measure of the confounder, including it in the model will clearly reduce confounding bias. The effect of ρ_Z on bias reduction is less intuitive. Across all relationships between ρ_X and ρ_C, increasing values of ρ_Z decrease the bias in the within-pair estimate, as illustrated in Figure 3. In other words, holding ρ_X and ρ_C constant, increasing ρ_Z will reduce bias in β_W (the individual-level estimate, β is not affected by the value of ρ_Z). This occurs for the same reason that increasing ρ_C, holding ρ_X constant, results in lower levels of bias in β_W as discussed in Frisell et al. (2012). When twins are less discordant

on the confounder, meaning that ρ_C is larger, they are also likely to be less discordant on the covariate (ρ_Z is larger). This decreases the correlation between the covariate and exposure variables resulting in less bias. Importantly, the within-pair estimate is only unbiased when all confounders are perfectly shared within a twin pair.

The current results have important implications for the interpretation of CTC results. As described above, interpretation of the within-pair effect is commonly made by comparing β_W from the CTC model to β from the individual-level model. We show that in the presence on non-shared confounding CTC results can support a causal effect of exposure on outcome even when the true causal effect is zero ($\beta_W = \beta \neq 0$). This will occur even if a covariate is included in the CTC model as a proxy measure of the confounder.

Additionally, the within-pair estimate between monozygotic ($\beta_{W_{MZ}}$) and dizygotic ($\beta_{W_{DZ}}$) twin pairs is usually compared to identify whether genetic or shared environmental factors confound the exposure-outcome relationship. For instance, when $\beta_{W_{MZ}} < \beta_{W_{DZ}} < \beta$ this suggests that the observed relationship is confounded by genetic factors (McGue et al., 2010). This is because MZ twin pairs share all genetic factors, while DZ twin pairs shared approximately 50% of these factors. Both types of twin pairs share all common (rearing) environmental factors. Given heritable phenotypes, the within-pair correlation in exposure, confounder, and covariate will be greater for MZ compared to DZ twins influencing the comparison of $\beta_{W_{MZ}}$ and $\beta_{W_{DZ}}$. Even in the case of a true, non-zero effect

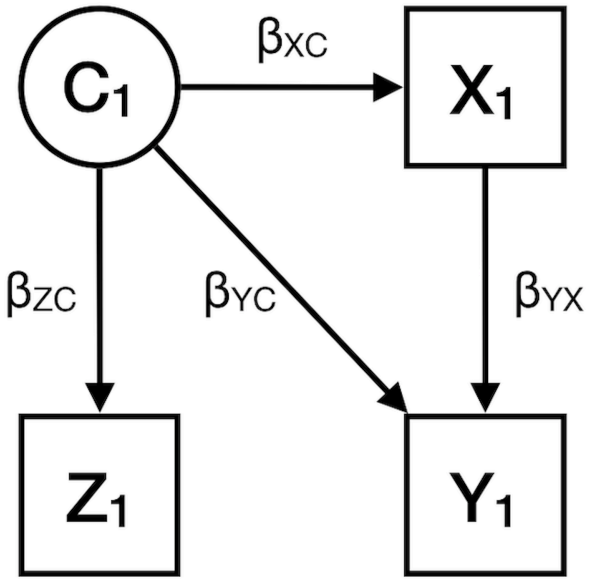
of exposure on outcome, it would be possible to conclude that genetic factors confound the causal relationship ($\beta_{W_{MZ}} < \beta_{W_{DZ}} < \beta$) when, in reality, they do not. This point has been made previously (Frisell et al., 2012), but we highlight that it continues to hold in the context of the current results.

While we show that exposure effect estimates from CTC designs are likely to be biased, we maintain that the CTC design can provide useful information when used appropriately. Results from CTC studies can often be used to argue that an observed relationship is not consistent with a causal exposure effect. For instance, when $\beta_W = 0$ and the expected level of measurement error does not likely account for this magnitude of attenuation, it would suggest that shared environmental confounders explain at least part of the exposure-outcome relationship. Results may also suggest that an observed association cannot be due entirely to shared confounders within a twin pair. When $\beta_W \neq 0$, this suggests that some influence beyond shared confounders is contributing to the observed relationship.

The best case for bias reduction in CTC model estimates occurs when the within-twin pair correlation in the exposure is less than the within-twin pair correlation in the confounder, when the within-twin pair correlation in the covariate is high, and the covariate is an accurate measure of the confounder. Of these pieces of information, only ρ_X and ρ_Z are known in practice. These values should always be reported and a case should be made about the likely relationships to the possible confounders to determine whether CTC models are appropriate for a given situation. Lastly, there are additional

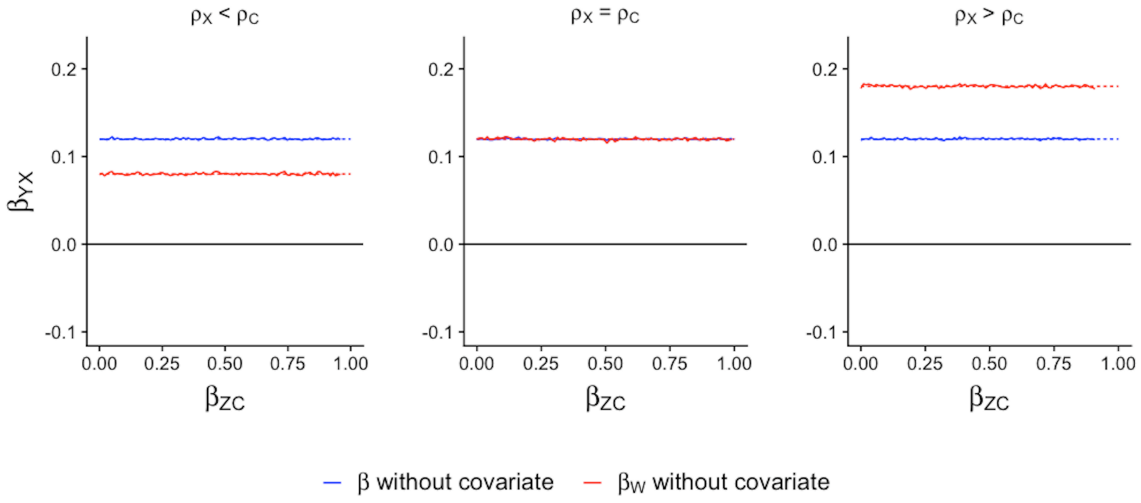
limitations of the CTC design that the current study does not address, like reverse causality and the potential causal influence of non-shared environmental factors not included in the models (McGue et al., 2010). Future methodological work should be focused on the extent to which these factors affect exposure effect estimates from CTC models.

Figure 2-1. Causal diagram shown for only one twin in a pair for simplicity.



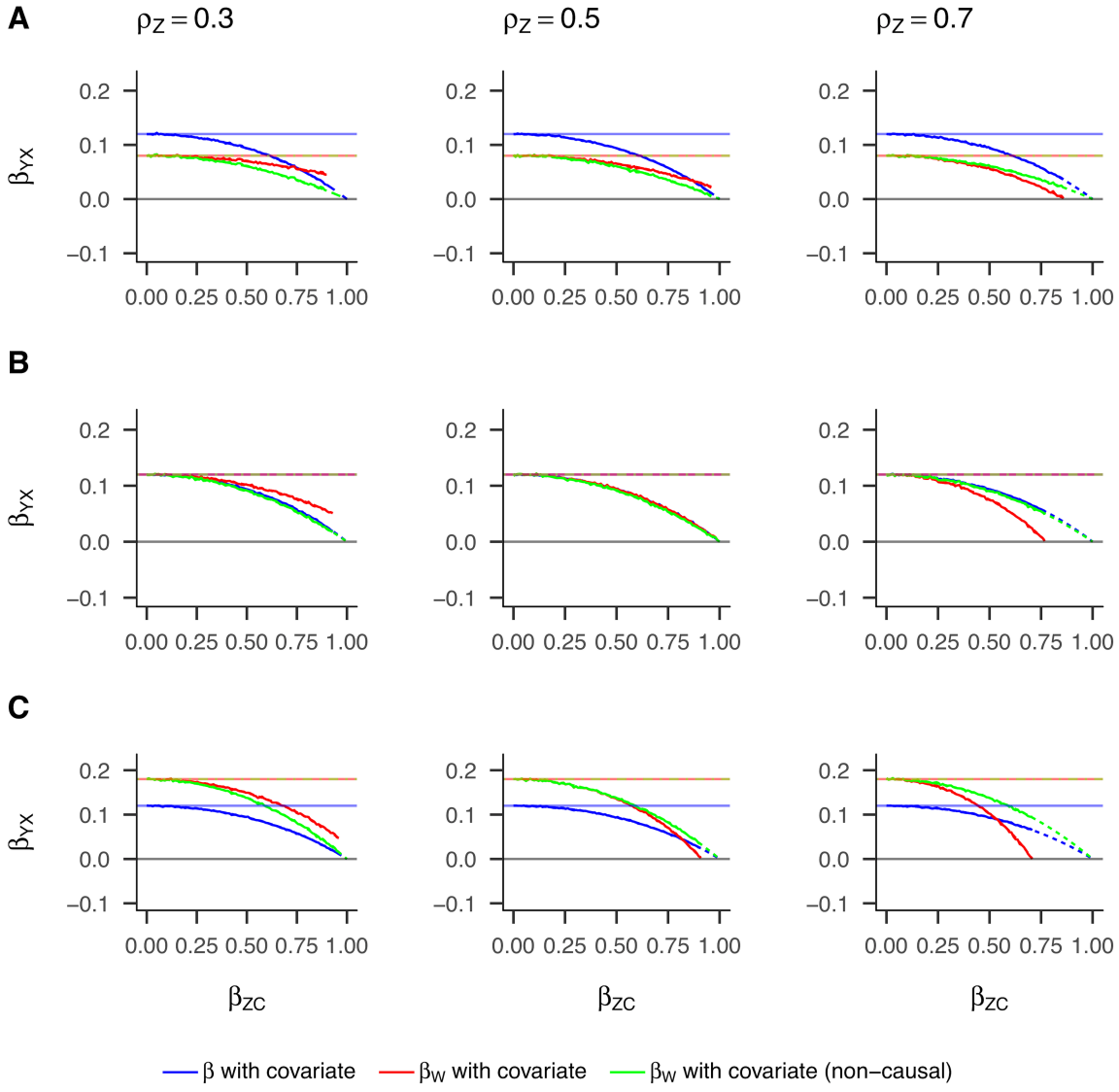
Note. Variables X, Y, C represent an exposure, outcome, and unmeasured confounder, respectively. Z represents a measured covariate. β_{YX} is the true causal effect of exposure on outcome. β_{ZC} is the effect of the confounder on the covariate.

Figure 2-2. Results from Frisell et al. (2012).



Note. Blue lines denote the exposure estimate from individual-level models while red lines denote the exposure estimate from CTC models. Simulation results (solid lines) are overlaid on the derivation results (dashed lines) to show their concordance. The true causal effect is zero ($\beta_{YX} = 0$). The within-twin pair correlations in the exposure and confounder are ρ_X and ρ_C, respectively. The bias in the individual-level effect and the within-twin pair effect does not depend on β_{ZC}, the effect of the confounder on the covariate, because the covariate is not included in these models.

Figure 2-3. Exposure effect estimates with the inclusion of a covariate from individual-level and within-pair models.



Note. Exposure effect estimates with the inclusion of a covariate from individual-level and within-pair models when A) the within-pair correlation in the exposure ($\rho_X = 0.4$) is less than the within-pair correlation in the confounder ($\rho_C = 0.6$); B) the within-pair correlation in the exposure ($\rho_X = 0.5$) equals the within-pair correlation in the confounder ($\rho_C = 0.5$); C) the within-pair correlation in the exposure ($\rho_X = 0.6$) is less

than the within-pair correlation in the confounder ($\rho_C = 0.4$). β_{ZC} is the effect of the confounder on the covariate, and ρ_Z is the within-pair correlation in the covariate. Blue lines denote the exposure estimate from individual-level models, red lines denote the exposure estimate from CTC models as specified in equation 4, and green lines denote the exposure estimate from CTC models as specified in equation 3. Simulation results (solid lines) are overlaid on the derivation results (dashed lines) to show their concordance. Darker lines denote the exposure effect estimate with covariate inclusion while lighter shaded lines denote the same estimate without covariate inclusion. The true causal exposure effect is zero ($\beta_{YX} = 0$).

Study 3

Alcohol consumption and tobacco use have been associated with the socioeconomic outcomes of lower educational attainment, occupational status, and income (Casswell, Pledger, & Hooper, 2003; Townsend, Flisher, & King, 2007). An important public health question is whether use in adolescence is associated with negative adult socioeconomic outcomes, and whether this relationship is independent of adult substance use. Cross-sectional studies suggest an association between adolescent substance use and rearing, or parental, socioeconomic status (SES). Higher levels of rearing SES have been consistently linked to decreased tobacco use (Borland & Rudolph, 1975; Goodman & Huang, 2002; Soteriades & DiFranza, 2003; Tyas & Pederson, 1998) but less robustly related to increased alcohol use and binge drinking behaviors in adolescence (Goodman & Huang, 2002; Humensky, 2010; Kendler et al., 2014; Melchior et al., 2007). Findings from research by Kendler et al. (2014) suggest that adolescents who grow up in higher SES households consume more alcohol, both in frequency and binge drinking behaviors, but have fewer alcohol-related problems than those in lower SES homes.

Adolescent substance use has also been associated with adult attained socioeconomic status, including educational outcomes. For instance, early tobacco use predicts greater academic problems and lower educational attainment in adulthood (Ellickson, Tucker, & Klein, 2001; Kandel, Davies, Karus, & Yamaguchi, 1986; Latvala et al., 2014). Prior research has shown similar associations with alcohol use, suggesting a link between adolescent alcohol exposure and lower educational achievement and

attainment (Hicks, Iacono, & McGue, 2010; Latvala et al., 2014; Staff & Maggs, 2017; Viner & Taylor, 2007), though a relatively large summary of cohort studies did not find a link between early use and adult educational outcomes (McCambridge, McAlaney, & Rowe, 2011). It may be the case that the effect of adolescent substance exposure on adult SES is merely a function of rearing SES, however. In other words, childhood SES acts as a confounder in the relationship between early substance exposure and adult SES outcomes. The prior research is inconsistent in this area with some studies finding an attenuated association between adolescent substance use and adult SES after controlling for rearing SES (Staff & Maggs, 2017; Wells, Horwood, & Fergusson, 2004) while another study finds that the effect of early exposure remains after controlling for this source of confounding (Viner & Taylor, 2007).

An alternative method to address sources of shared environmental confounding, like rearing SES, is through a discordant twin pair design, often called the co-twin control (CTC) design (Carlin et al., 2005; McGue et al., 2010). The CTC design seeks to strengthen causal inference in observational research through using a twin with less exposure as the control for their co-twin, thus controlling for everything the twins are matched on including genetic and shared environmental factors, even if those factors are not directly measured (Carlin et al., 2005). In other words, within discordant twin pairs, the twin who uses less should show better socioeconomic outcomes than their heavier using co-twin if substance use is causal. In monozygotic (MZ) twin pairs, who share all genetic and common environmental factors, any difference in outcomes within a pair must be due to differences in exposure. Thus, if the heavier using twin has poorer

outcomes than their lesser using co-twin, this provides stronger evidence that substance use contributed to these negative outcomes. Two previous studies of alcohol use in discordant twin pairs did not find an effect of early use on adult financial independence or truncated education within twin pairs (Irons, Iacono, & McGue, 2015; Rose, Winter, Viken, & Kaprio, 2014) while two other studies similar in design did find an effect within discordant MZ twin pairs of early alcohol consumption on later educational attainment and occupation stagnation (Grant et al., 2012; Waldron, Malone, McGue, & Iacono, 2018). There have been three twin studies looking at tobacco use in discordant twin pairs, two of which did not find an effect of educational attainment on risk of smoking in MZ twins (Gilman, Abrams, & Buka, 2003; Gilman et al., 2008) while one study did find an effect of regular nicotine use on educational attainment (Grant et al., 2012). Importantly, however, all of these studies assessed the relationship between educational attainment and adult smoking rather than use in adolescence.

A major strength of the CTC design lies in the inherent ability to control for all factors, measured or unmeasured, that are shared within a twin pair. By definition, however, environmental confounders that are not perfectly shared within a twin pair are not implicitly controlled for and can lead to bias in within-pair estimates (Frisell et al., 2012, see also Study 2). Prior research suggests that cognitive ability and externalizing behaviors, factors that are not necessarily shared within twin pairs, may account for a significant portion of the variance in social mobility (McGue et al., in press). We explore these potential sources of non-shared confounding by additionally incorporating measures of IQ and externalizing behavior into CTC models.

In the current study we aim to address limitations of the existing literature by evaluating the relationship between adolescent substance use and adult socioeconomic outcomes in a longitudinal, genetically informative sample, strengthening the basis for causal inference of these relationships by utilizing powerful CTC models that can control for both genetic and environmental confounds. We first examine the relationship between rearing SES, as measured by parental education and occupation, and offspring alcohol and tobacco use. Next, in examining the socioeconomic consequences of adolescent substance use, we evaluate the relationship between adolescent use and attained socioeconomic status and social mobility (i.e., attained SES relative to rearing SES) at age 29, and explore whether the associations depend on initial SES status in childhood. We then make use of our genetically informative sample in combination with CTC models to explore the potentially causal relationship between adolescent substance use and adult attained SES. If adolescent substance use is causally related to reduced attained SES, for twins who are discordant on early substance exposure, we would expect the heavier using twin to have lower adult SES than their lesser using co-twin. The interpretation of the CTC results after inclusion of potential non-shared confounders will be discussed in the context of findings from Study 2. Our specific hypotheses are:

1. Higher rearing SES will be associated with lower tobacco use in adolescence. The pattern of associations with alcohol use will be less clear.
2. Adolescent substance use will be associated with lower attained SES and social mobility (attained relative to rearing SES) at the individual-level (i.e., treating twins as unrelated individuals).

3. Within twin pairs discordant for adolescent substance use, the heavier using twin will have a lower attained SES compared to their lesser using co-twin. This would be consistent with a causal effect of early substance exposure and would mean that the effect on attained SES is independent of rearing SES level.

Methods

Sample

Similar to the Study 1 sample, participants include twin pairs from the Minnesota Twin Family Study (MTFS) from two birth cohorts: an older cohort (birth years of 1972-1979) and a younger cohort (birth years of 1977-1984). Through an overlapping cohort design, the younger cohort participants completed assessments at the target ages of 11, 14, 17, 20, 24, and 29, while the older cohort completed assessments at the target ages of 17, 20, 24, and 29. Inclusion criteria for the current study include alcohol and tobacco use data at age 17, attained SES data at age 29, and information on rearing SES from at least one parent. This results in an approximate sample size of N = 2,424 for individual-level tobacco analyses and N = 2,421 for individual-level alcohol analyses. For CTC models we have information on approximately N = 811 complete MZ twin pairs and N = 425 complete DZ twin pairs for both tobacco and alcohol analyses. Actual sample sizes vary due to availability of exposure, outcome, and covariate measures.

The follow-up participation rate from age 17 to age 29 was approximately 91%. The effects of sample attrition were evaluated by comparing participants and nonparticipants at the age 29 assessment with their responses at the age 17 assessment. There were no differences in tobacco and alcohol use or externalizing behaviors between

those who participated in both age 17 and 29 assessments and those who only participated in the age 17 assessment. Those who participated in both assessments, however, had higher IQ scores than those only participating in the age 17 assessment (p = 0.01; Cohen's d = 0.21). This suggests that the current sample is unlikely to be biased by attrition due to tobacco and alcohol use, but may represent those with slightly higher IQ (though the standardized mean difference is small).

Measures

Substance use

Tobacco use was measured on an ordinal scale of cigarettes per day (CPD; including all forms of tobacco). The CPD measure ranges from 0 (never used tobacco) to 6 (2 or more packs/20 or more cigars or pipes/2 or more tins per day). Alcohol use was measured as a composite measure of alcohol consumption (ALC). The drinking index incorporates measures of quantity, frequency, number of intoxications, and maximum number of drinks. Specifically, it consists of four self-report alcohol use items: frequency of alcohol use (scored from 0=never to 5=at least once per day), average number of drinks per drinking event (scored from 0=never drank to 6=30 or more), maximum number of drinks in a 24-hour period (scored from 0=never drank to 6=30 or more), and number of times intoxicated (scored from 0=never to 6=50 or more). This composite index provides a more comprehensive view of overall alcohol exposure than a single component measure. Adolescent alcohol use was defined as reported values at the age 17 assessment.

Socioeconomic status and social mobility

Rearing socioeconomic status (SES) was defined by the maximum education attained by either parent on a 5-point scale (1 = < High School, 2 = High School, 3 = Some College, 4 = College, 5 = Professional Degree) as well as the maximum occupational status reported by either parent (coded on a 1-7 Hollingshead scale but reversed so that higher scores reflect higher occupational status and only coded for those in a full-time job). Social mobility of the offspring was defined as the difference between rearing SES and their own attained SES at age 29, measured the same as for parents. Each measure of SES was examined separately as they may capture different aspects of socioeconomic status (Krieger, Williams, & Moss, 1997).

Potential non-shared confounders

The potential non-shared environmental confounders of IQ and externalizing symptoms were selected based on prior research showing that these variables explain a substantial portion of the variability in social mobility (McGue et al., in press). IQ was measured at twin intake using the Wechsler Adult Intelligence Scale (WAIS-R) for older cohort participants and the Wechsler Intelligence Scale for Children (WISC-R) for the younger cohort. We define externalizing behaviors as the sum of attention-deficit/hyperactivity disorder (ADHD), conduct disorder (CD), and oppositional defiant disorder (ODD) symptom counts. For comparability across the younger and older cohorts who were different ages at the intake assessment, we use lifetime DSM-III-R symptoms of CD and DSM-IV symptoms of ODD reported before age 15 as well as lifetime DSM-IV ADHD symptoms reported before age 12.

Statistical analysis

For individual-level analyses mixed effects models, allowing for random intercepts by family ID, are used to account for the correlated nature of the twin data. For the outcomes of CPD, educational attainment, and occupational status we use ordinal mixed effects models. For drinking index scores, and educational and occupational mobility outcomes, we use linear mixed effects models. To assess whether the effect of early substance use on adult SES is a function of initial, rearing SES, we include an interaction term in the appropriate models. All models include the base covariates of sex, year of birth, and age at assessment. Models that estimate the association between adolescent tobacco use and adult SES outcomes also control for adult tobacco use at ages 20 and 24 (age 29 is not included as the CPD measure was only collected in one cohort at this assessment age due to funding limitations). Models that estimate the association between adolescent alcohol use and adult SES outcomes also control for adult alcohol use at ages 20, 24, and 29. We control for adult substance use in order to focus on the effect of early substance exposure, prior to when an individual usually terminates formal education (only 1.5% of the sample reports less than a high school degree).

In CTC analyses the exposure effect is decomposed into a between- and within-twin pair effect (Begg & Parides, 2003). At the individual level, the regression of adult attained SES on adolescent substance use can be modeled as $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$, where Y_{ij} and X_{ij} are the outcome (attained SES) and exposure (early substance use) respectively, of individual j in twin pair i, β_0 is the intercept term, and $\beta_1$is the individual-level effect of the exposure on the outcome. Extending this to the CTC framework gives $Y_{ij} = \beta_0 + \beta_W(X_{ij} - \bar{X}_i) + \beta_B \bar{X}_i$, where $\bar{X}_i$ is the i^{th} twin pair mean of

adolescent substance use. The within-twin pair term is an estimate of the effect of early substance exposure on adult SES controlling for shared genetic and environmental confounding, even if the confounding is unmeasured, while the between-twin pair effect is an estimate of the effect of genetic and/or shared environmental confounding, which would include rearing SES (Carlin et al., 2005; McGue et al., 2010). Non-shared environmental confounders are incorporated into the CTC framework as covariates, giving the model $Y_{ij} = \beta_0 + \beta_W\left(X_{ij} - \bar{X}_i\right) + \beta_B\overline{X}_\iota + \beta_{Z1}(Z_{ij} - \bar{Z}_i) + \beta_{Z2}\overline{Z}_\iota$, where Z_{ij} is the IQ score, for example, of individual j in twin pair i and $\overline{Z}_\iota$ is the i^{th} twin pair mean of IQ.

Results

Descriptive statistics

Table 1 shows means and standard deviations of exposure, outcome, and covariate variables in the full, individual-level sample and in complete MZ and DZ twin pairs. Mean levels of educational mobility differ significantly between groups such that DZ pairs have significantly lower attainment relative to their parents than MZ twin pairs (p = 0.03). Additionally, DZ twin pairs have significantly lower mean values of externalizing behaviors than MZ twins (p = 0.001). Twin pair correlations in substance use, SES outcomes, and non-shared confounding variables are shown in Table 2. As would be expected for variables with non-zero heritability, all MZ twin pair correlations are significantly greater than DZ twin pair correlations as evidenced by non-overlapping confidence intervals.

Associations between rearing SES and substance use

The associations between rearing SES, as measured by maximum parental educational attainment and occupational status, and offspring tobacco and alcohol use across developmental age are shown in Figures 1A and 1B, respectively. In general, there is a protective effect of rearing SES on smoking across all ages. The pattern is very similar for both measures of parental SES, but is stronger for parent educational attainment. The association between rearing SES and alcohol use across age is less clear. There is some evidence of a protective effect at ages 14 and 17, little to no association at ages 20 and 24, and some evidence for an inverse relationship again at age 29. Only the association between parental educational attainment and alcohol use at age 17 is significantly different from zero. The associations between both measures of rearing SES and alcohol use follow generally the same pattern over developmental age as well.

Associations between adolescent substance use and SES outcomes

The associations between early substance exposure and adult SES outcomes, treating twins as individuals, are shown in Table 3. Adolescent tobacco use is associated with lower adult educational attainment (OR = 0.69, 95% CI [0.61, 0.77], $p < 0.001$) and occupational status (OR = 0.82, 95% CI [0.74, 0.90], $p < 0.001$). We find similar, and slightly stronger, associations between adolescent alcohol use and adult educational (OR = 0.55, 95% CI [0.47, 0.64], $p < 0.001$) and occupational outcomes (OR = 0.68, 95% CI [0.60, 0.78], $p < 0.001$). In examining the outcomes of educational and occupational mobility, without controlling for rearing SES (i.e., before adding a substance use by rearing SES interaction term), we find a significant association for adolescent tobacco use on educational (β = -0.11, 95% CI [-0.16, -0.05], $p < 0.001$) but not on occupational

mobility (β = -0.05, 95% CI [-0.11, 0.01], p = 0.08). Similar associations are found between adolescent alcohol use and educational (β = -0.09, 95% CI [-0.14, -0.04], p < 0.001) and occupational mobility (β = -0.07, 95% CI [-0.12, -0.01], p = 0.02).

Associations between adolescent substance use and SES outcomes after controlling for rearing socioeconomic status

We then tested whether the effect of adolescent substance use on adult attained SES was purely due to initial status of childhood SES in two ways: 1) incorporating an interaction between substance use and parental SES in individual-level models of social mobility outcomes, and 2) using CTC models which implicitly control for the shared environmental confounder of rearing SES. The associations between adolescent use, as well as the interaction between use and rearing SES, and social mobility outcomes are shown in Table 4. While there is possibly weak evidence of an interaction between adolescent tobacco use and parental education such the effect of early smoking on educational mobility is weaker in those who come from higher SES household (β = -0.09, 95% CI [-0.19, 0.01], p = 0.06), we do not find strong evidence for an interaction between adolescent use and rearing SES for other outcomes suggesting that the effect of adolescent smoking and alcohol use on social mobility largely does not depend on the level of childhood SES. Interestingly, after accounting for parental SES with inclusion of an interaction term (see the 'Unadjusted' columns in Table 4), the association between adolescent tobacco use and educational mobility (β = -0.06, 95% CI [-0.16, 0.04], p = 0.22) are attenuated but the associations with early alcohol use remain significantly

different from zero (β = -0.11, 95% CI [-0.21, -0.003], p = 0.04; β = -0.15, 95% CI [-0.28, -0.03], p = 0.01, respectively).

Results for CTC models are shown in Table 3 and are largely consistent with results from the individual-level models that control for parental SES. We find evidence for an effect of adolescent smoking on adult educational attainment within MZ twin pairs only (DZ twins: *OR* = 0.88, 95% CI [0.72, 1.08], p = 0.22; MZ twins: *OR* = 0.78, 95% CI [0.64, 0.95], p = 0.01). We do not find an effect of adolescent smoking on occupational status in discordant twin pairs, though the within-MZ twin pair point estimates are very similar to individual-level estimates but with much wider confidence intervals (DZ twins: *OR* = 1.03, 95% CI [0.86, 1.25], p = 0.73; MZ twins: *OR* = 0.85, 95% CI [0.69, 1.05], p = 0.13). We also find associations between adolescent alcohol use and educational attainment within twin pairs (DZ twins: *OR* = 0.65, 95% CI [0.47, 0.91], p = 0.01; MZ twins: *OR* = 0.70, 95% CI [0.51, 0.95], p = 0.02) and occupational status within MZ twin pairs only (DZ twins: *OR* = 0.92, 95% CI [0.66, 1.27], p = 0.60; MZ twins: *OR* = 0.71, 95% CI [0.52, 0.98], p = 0.03).

Associations between adolescent substance use and SES outcomes after controlling for potential non-shared environmental confounders

Given the within-pair effects for adolescent substance use on adult SES, as well as similar effects of alcohol use on social mobility measures controlling for initial SES status, we explored whether the observed exposure-outcome relationships are further confounded by the non-shared environmental factors of IQ and externalizing behavior. Results are shown in Tables 3 and 4 under the 'adjusted' columns, indicating that these

models additionally controlled for IQ and externalizing. We find that effect of adolescent alcohol use on educational and occupational mobility is now attenuated to near zero after inclusion of these confounding sources (β = -0.06, 95% CI [-0.16, 0.04], p = 0.22; β = -0.10, 95% CI [-0.21, 0.02], p = 0.11, respectively). Within twin pair results (Table 3) show only slightly attenuated results after the inclusion of IQ and externalizing. The effects of adolescent tobacco and alcohol use on adult educational attainment remains in discordant MZ twins (*OR* = 0.78, 95% CI [0.64, 0.94], p = 0.01; *OR* = 0.73, 95% CI [0.53, 0.99], p = 0.04, respectively). The effect of early alcohol use on later occupational status no longer reaches statistical significance at the $p < 0.05$ level (*OR* = 0.73, 95% CI [0.53, 1.01], p = 0.05), but the point estimate remains very similar to the effect prior to adjusting for non-shared confounders as well as the point estimate for the outcome of educational attainment. Additionally, the within-MZ twin pair point estimates of adolescent tobacco and alcohol use on later occupational status, while containing relatively wide confidence intervals, are actually stronger in magnitude than those from individual-level models.

Discussion

The current study explored the exposure effect of adolescent substance use as well as its relationship to rearing SES and attained socioeconomic outcomes using a longitudinal twin sample. We first examine the associations between parental education level and occupational status and tobacco and alcohol use in offspring across developmental age. We find an inverse association between parental SES and offspring tobacco use at all ages and alcohol use in adolescence only. This is generally consistent

with results from Study 1, that suggest genetic nurture effects found for tobacco and alcohol use are mediated by parental SES (measured in the same way as the current study), and is in line with prior research finding similar associations across offspring age (Humensky, 2010; Jefferis et al., 2004; Soteriades & DiFranza, 2003; Wiles et al., 2007).

Treating twins as individuals (i.e., ignoring the paired nature of the data) we also find associations between adolescent substance use and adult educational attainment and occupational status such that increased use in adolescence is associated with reduced attained SES at age 29. These associations also hold for educational and occupation mobility, meaning that early substance exposure is associated with lower SES attainment relative to parents. There is little to no evidence that these associations are merely capturing the effect of rearing SES as evidenced by non-significant interaction terms between substance use and parental SES in individual models of both educational and occupational mobility, as well as significant within-pair effects in CTC models. Co-twin control models implicitly control for levels of parental SES, as well as all other genetic and environmental factors common to twins within a pair. Our findings of significant within-pair effects of adolescent tobacco and alcohol use on educational outcomes, and alcohol use only on occupational status, suggest that these relationships are independent of initial SES status in childhood. Prior research in this area is mixed but the current findings are generally consistent with earlier CTC studies finding significant within-pair effects of tobacco and alcohol use on educational outcomes (Grant et al., 2012; Waldron et al., 2018). That we do not find a significant within-pair effect of adolescent tobacco use on occupational status implies that this relationship may be confounded by genetic

and shared environmental factors (though point estimates of this relationship within-MZ twin pairs is very similar to estimates from the individual-level model). Given the lack of interaction between tobacco use and rearing SES in the model of occupational mobility, we do not believe that childhood SES is a major source of the shared environmental confounding in the relationship between early tobacco use and adult occupational outcomes.

The significant within-MZ twin pair effects of early substance use on attained SES from CTC models are also consistent with a causal exposure effect of adolescent use on both outcomes of educational attainment and occupation status. This effect must be independent of all genetic and shared environmental factors because these factors are identical within MZ twin pairs. Importantly, however, the CTC design can only control for non-shared confounders that are measured and included in the model. As found in previous simulation work, failure to control for environmental factors that are not shared within a twin pair biases the within-pair effect away from the null (Frisell et al., 2012). The amount of bias in the CTC estimates, relative to estimates from individual-level models, is a function of the relationship between the within-pair correlation in the exposure and the within-pair correlation in the confounder. The within-MZ correlation for both CPD and alcohol use is $r = 0.72$ (Table 2). The findings from Frisell and colleagues show that if the twin correlation in an unmeasured confounder, which cannot be known, is less than 0.72 in this case, the within-MZ twin pair estimate shown in the unadjusted models in Table 3 is more a biased estimate of the true causal exposure effect than the individual-level estimate. If the correlations are equal, both individual-level and

within-pair estimates are equally biased and finally, if the twin-pair correlation in the confounder is greater than 0.72, in this case, the within-MZ twin pair estimate is less biased than the individual-level estimate.

To address this possibility we included the potential non-shared confounders of IQ and externalizing behavior as covariates in the CTC models with the idea that these variables index, however imperfectly, the underlying (unmeasured) confounders of cognitive ability and general externalizing liability. The within-MZ twin pair effects of adolescent substance use on educational attainment and occupational status are only slightly attenuated with the inclusion of IQ and externalizing behaviors, suggesting that this is not a major source of non-shared environmental confounding. These results are in contrast to individual-level models of social mobility and attainment in which inclusion of IQ and externalizing attenuate the substance use effects to a greater extent. The apparent disagreement between model results may be due to the different outcomes (i.e., attainment vs. mobility) and levels of confounding control. In other words, within-MZ twin pair models are able to control for many more sources of confounding than individual-level models, so it may be that IQ and externalizing have little effect on SES outcomes after shared genetic and environmental confounding is already adjusted for.

In the current sample, we find that the within-MZ correlation in IQ is $r = 0.79$ and in externalizing is $r = 0.77$ (Table 2). If the IQ and externalizing variables used here are perfect measures of the underlying, unmeasured confounders, these correlations will equal the twin-pair correlations in the unmeasured confounders (i.e., $\rho_Z = \rho_C$ from Study 2). In this case, because the twin-pair correlations in the unmeasured confounders are

equal to or greater than the twin-pair correlations in exposure (i.e., $\rho_C \geq \rho_X$; this scenario is illustrated in Study 2 Figure 2-3A), the within-MZ twin pair estimates, while still biased estimates of the true causal exposure effect, are likely to be less biased then the individual-level effect. In reality, both IQ and externalizing as measured here are imperfect measures of the underlying non-shared confounding factors, which cannot be known or measured in practice. In this situation, based on findings from Study 2, if we can argue that IQ and externalizing measures collected here very accurately capture the underlying confounders and the true within-pair correlation in the unmeasured confounder is similar to the observed within-twin pair correlation in the measured covariates, the within-MZ twin pair estimates in the adjusted models from Table 3, while still likely to be biased, are closer to the true causal exposure effect than individual-level estimates. As these variables more accurately index the underlying confounders, the bias in within-pair estimates will converge towards zero.

We illustrate this scenario in Figure 2 based on the derivations from Study 2. To better depict the possible bias in the context of the current study, we define the parameters as such: $\beta_{YX} = 0$, $\beta_{YC} = 0.3$, $\beta_{XC} - 0.1$, $\rho_X = 0.72$, $\rho_Z = 0.75$, and $\rho_C = 0.75$. While there is subjectivity in the selected parameter values, we believe that the ones chosen sufficiently approximate the pattern of potential bias in the within-pair and individual-level analyses, though the magnitude of bias shown in the figure may be inaccurate. We see that as IQ and externalizing better measure the true underlying confounders (i.e., as β_{ZC} increases) the individual-level and within-pair effects become less biased. Overall the within-pair estimate is always less biased than the individual-

level estimate and all estimates are at least somewhat biased unless the covariate is a perfect measure of the confounder, which is very unlikely in practice. This means that the within-MZ twin pair estimates shown in Table 3 are likely to be at least somewhat biased away from the null, suggesting that the true causal effect of adolescent alcohol use on adult attained SES is less than what is estimated here, but that these estimates are closer to the true causal effect than the individual-level estimates. Importantly, we have not considered the effect of measurement error which has been shown to attenuate the within-pair estimate more than the individual-level estimate (Frisell et al., 2012). Additionally, we have only included two possible non-shared confounders. In reality, there are likely to be many more variables that confound the alcohol-SES relationship that may additionally induce bias in the within-pair estimates presented here like academic problems or other psychiatric disorders.

In additional to the issue of non-shared confounding, the current study is subject to other limitations as well. A perennial problem with co-twin control models is that of limited statistical power, which is a function of the level of exposure discordance within twin pairs as well as the sample size. Given our sample size of N = 811 MZ twin pairs and a twin-pair correlation in substance use of $r = 0.72$, we have 80% power to detect effects that account for approximately 1.7% of the variance in outcomes. For DZ-twin pair analyses, given a sample size of N = 425 twin pairs and a twin-pair correlation of $r = 0.55$, we have 80% power to detect effects that account for around 2.9% of the variance. It may be the case that substance use effects are very small and we do not have sufficient power to detect them. The wide confidence intervals, particularly for DZ twin pairs,

suggest that we may be underpowered to detect these effects. Co-twin control models also do not implicitly address reverse causation. While we make use of longitudinal data in which adolescent substance use is measured before the vast majority of participants end formal education, it could be the case that the underlying processes that lead to lower education attainment and reduced occupational status have begun prior to initiation of substance use and are impacting the level of adolescent use. Lastly, our findings may not generalize to other populations as the current sample was almost entirely made up of White twins from the United States.

Finally, we highlight two additional patterns of results that emerge: 1) there are consistently stronger effects of adolescent substance use on educational attainment outcomes compared to occupational status, and 2) there are stronger effects of adolescent alcohol, compared to tobacco, use on all outcomes. It may be that the relationship between smoking and adult SES is more heavily affected by confounding, or it may be that the associations between early alcohol use and later SES outcomes are simply stronger than for early tobacco use. It is also likely the case that our measure of educational attainment better indexes attained socioeconomic outcomes than occupational status. By age 29, when adult SES was measured, the vast majority of individuals have completed their highest educational degree, but social class may not yet be reflected in the occupational status measure. Additionally, in the context of occupational mobility outcomes, we are comparing offspring age 29 occupational status to parental occupational status measured at an average age of 42.8 years old. This may not be a fair comparison in that offspring are not yet expected to have reached comparable

occupational outcomes. Increasing lengths of formal education in younger generations may also be delaying maximum attained occupational status in offspring. Finally, the Hollingshead occupational status scale has been criticized for the subjectivity of occupational class assignment and for being outdated.

Overall, the present study extends the existing literature on the relationship between adolescent substance use and socioeconomic outcomes by utilizing a longitudinal, genetically informative sample to control for sources of genetic and shared environmental confounding. After additionally controlling for two sources of non-shared environmental confounding in IQ and externalizing behavior, as well as measures of adult substance use, we continue to find within-MZ twin pair effects of adolescent tobacco and alcohol use on educational attainment, as well as alcohol use on adult occupational status. These findings are consistent with, but do not prove, a causal exposure effect such that increased substance use in adolescence is associated with reduced educational and occupational attainment independent of genetic and shared environmental confounding, like rearing SES, as well as IQ, externalizing, and adult substance use. In relating the current results to those from Study 2, we conclude that the estimates presented here are likely to be biased away from the null, to at least a small degree. So, while we provide evidence consistent with a causal substance use exposure effect we also highlight the potential bias. Replication of these findings in larger samples of twins, incorporating other possible sources of non-shared environmental confounding, would be useful in providing additional evidence in support of the causal relationship between adolescent substance exposure and adult socioeconomic outcomes.

Table 3-1. Descriptive statistics of substance use exposures and adult outcomes.

	Total sample M (SD) or %	Complete DZ pairs M (SD) or %	Complete MZ pairs M (SD) or %
	N = 1,787 – 2,079	N = 313 – 408 pairs	N = 588 – 765 pairs
Female (%)	52%	51%	54%
Cigarettes per day[1] (age 17)	3.71 (8.0)	3.92 (7.9)	3.61 (8.0)
Drinking index (age 17)	1.15 (1.0)	1.20 (1.0)	1.13 (1.0)
Educational outcomes			
Educational attainment (age 29)	3.32 (1.1)	3.29 (1.1)	3.34 (1.1)
Educational mobility	0.38 (1.3)	0.30 (1.2)[2]	0.43 (1.3)[2]
Occupational outcomes			
Occupational status (age 29)	4.51 (1.6)	4.50 (1.6)	4.52 (1.6)
Occupational mobility	-0.23 (2.0)	-0.28 (2.0)	-0.21 (2.0)
IQ	102.01 (14.2)	101.99 (14.2)	102.05 (14.2)
Externalizing behavior	5.28 (5.6)	5.00 (5.4)[2]	5.81 (5.9)[2]

Note. [1]Mean values for cigarettes per day are computed by assigning the midpoint of each level as we do not have a continuous measure of cigarettes per day available for both cohorts. [2]Denotes significant differences in mean levels of educational mobility between MZ and DZ twin pairs. Educational and occupation mobility are defined as the difference between parent and offspring values. MZ = monozygotic; DZ = dizygotic.

Table 3-2. Twin pair (intra-class) correlations in exposure and outcomes, split by zygosity.

	DZ Twins (95% CI)	MZ Twins (95% CI)
Cigarettes per day (age 17)	0.46 (0.39, 0.54)	0.72 (0.69, 0.75)
Drinking index (age 17)	0.55 (0.48, 0.61)	0.72 (0.69, 0.75)
Educational outcomes		
Educational attainment (age 29)	0.48 (0.41, 0.55)	0.62 (0.58, 0.66)
Educational mobility	0.54 (0.47, 0.60)	0.73 (0.70, 0.76)
Occupational outcomes		
Occupational status (age 29)	0.14 (0.03, 0.25)	0.50 (0.44, 0.55)
Occupational mobility	0.46 (0.37, 0.54)	0.69 (0.65, 0.73)
IQ	0.54 (0.47, 0.60)	0.79 (0.76, 0.81)
Externalizing behavior	0.50 (0.43, 0.57)	0.77 (0.74, 0.80)

Note. Educational and occupation mobility are defined as the difference between parent and offspring values. MZ = monozygotic; DZ = dizygotic.

Table 3-3. Effect of adolescent substance use on educational attainment and occupational status at age 29.

	Educational Attainment			Occupational Status		
	Individual-level OR (95% CI)	Within-DZ OR (95% CI)	Within-MZ OR (95% CI)	Individual-level OR (95% CI)	Within-DZ OR (95% CI)	Within-MZ OR (95% CI)
Unadjusted:						
Cigarettes per day	0.69 (0.61, 0.77)	0.88 (0.72, 1.08)	0.78 (0.64, 0.95)	0.82 (0.74, 0.90)	1.03 (0.86, 1.25)	0.85 (0.69, 1.05)
Adjusted for confounders:						
Cigarettes per day	0.76 (0.68, 0.84)	0.90 (0.73, 1.11)	0.78 (0.64, 0.94)	0.89 (0.81, 0.97)	1.04 (0.86, 1.25)	0.87 (0.71, 1.08)
Unadjusted:						
Drinking index	0.55 (0.47, 0.64)	0.65 (0.47, 0.91)	0.70 (0.51, 0.95)	0.68 (0.60, 0.78)	0.92 (0.66, 1.27)	0.71 (0.52, 0.98)
Adjusted for confounders:						
Drinking index	0.66 (0.57, 0.77)	0.70 (0.50, 0.99)	0.73 (0.53, 0.99)	0.78 (0.69, 0.90)	0.95 (0.68, 1.32)	0.73 (0.53, 1.01)

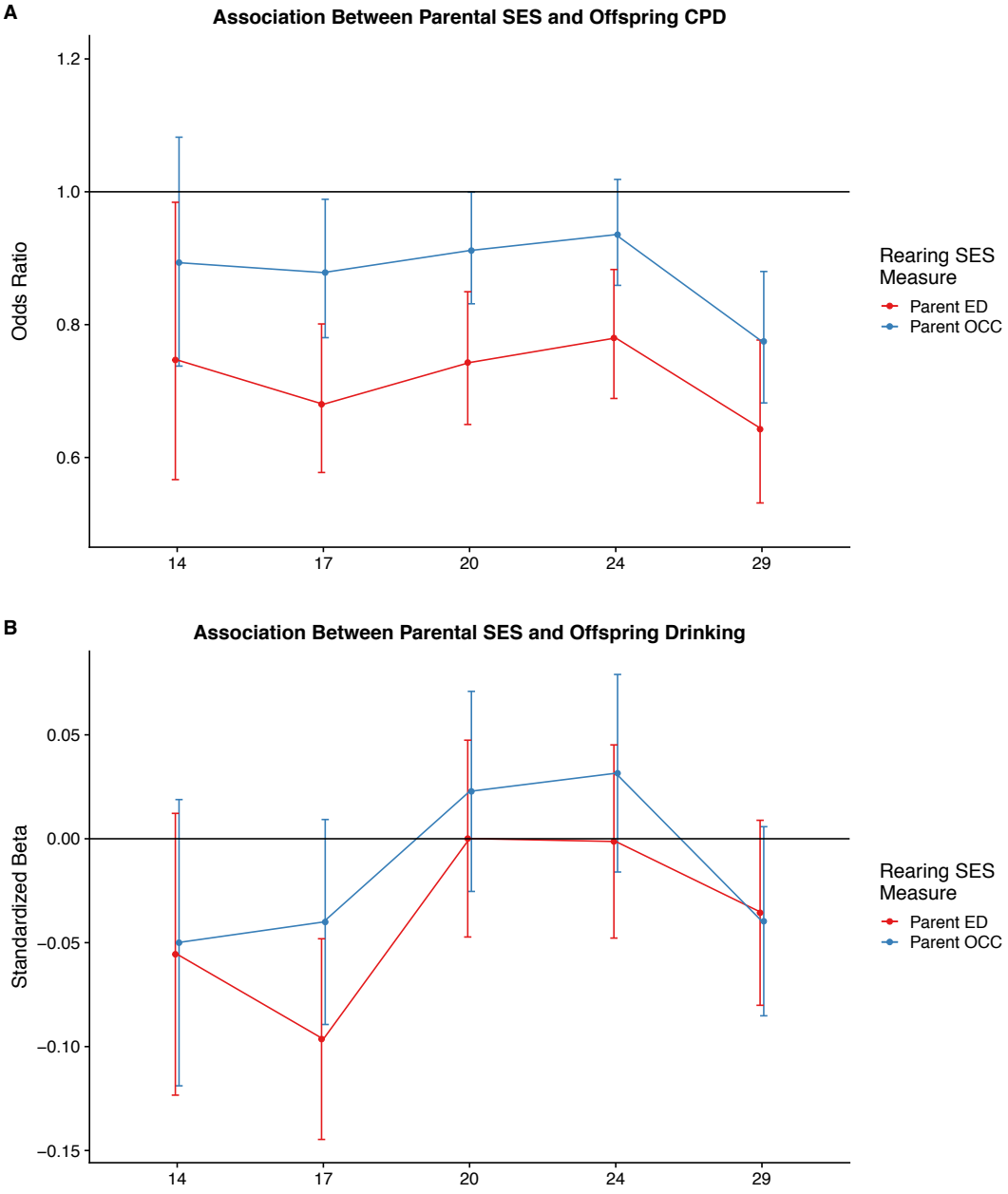
Note. All models control for the covariates of sex, year of birth, age at assessment, and adult substance use. Adjusted models differ from unadjusted models with the inclusion of IQ and externalizing behaviors. OR = odds ratio; CI = confidence interval; DZ = dizygotic; MZ = monozygotic.

Table 3-4. Effect of adolescent substance use on educational and occupational mobility outcomes (the difference between parent and offspring attainment).

	Educational Mobility		Occupational Mobility	
	Unadjusted	Adjusted	Unadjusted	Adjusted
	β (95% CI)	β (95% CI)	β (95% CI)	β (95% CI)
Cigarettes per day (CPD)	-0.06 (-0.16, 0.04)	-0.02 (-0.12, 0.08)	-0.08 (-0.19, 0.03)	-0.04 (-0.15, 0.07)
CPD x rearing SES	-0.09 (-0.19, 0.01)	-0.10 (-0.19, -0.01)	-0.01 (-0.12, 0.10)	-0.02 (-0.12, 0.09)
Drinking index	-0.11 (-0.21, -0.003)	-0.06 (-0.16, 0.04)	-0.15 (-0.28, -0.03)	-0.10 (-0.21, 0.02)
Drinking index x rearing SES	-0.04 (-0.15, 0.06)	-0.04 (-0.14, 0.05)	0.04 (-0.08, 0.16)	0.02 (-0.09, 0.13)

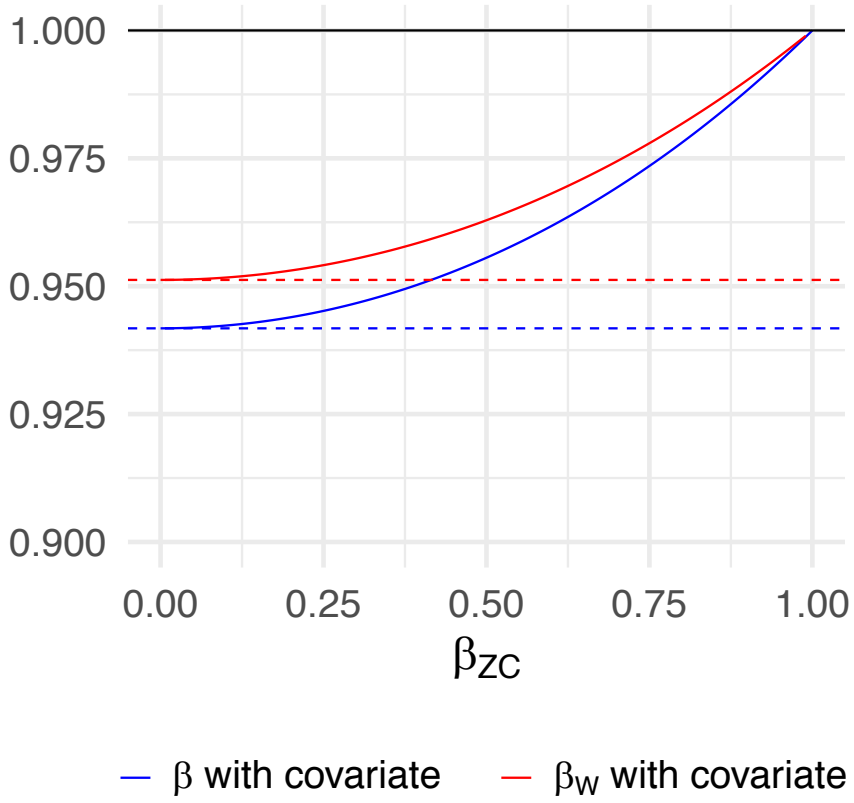
Note. All models control for the covariates of sex, year of birth, age at assessment, and adult substance use. Adjusted models differ from unadjusted models with the inclusion of IQ and externalizing behaviors. β = standardized beta; CI = confidence interval.

Figure 3-1. Associations between rearing socioeconomic status and substance use across developmental age.



Note. SES = socioeconomic status; CPD = cigarettes per day; ED = educational attainment; OCC = occupational status.

Figure 3-2. Depiction of potential bias in individual-level and within-MZ twin pair estimates from Study 2 derivations.



Note. The true causal effect (β_{YX}) is set to zero. Solid lines denote model estimates with inclusion of covariates; dashed lines denote model estimates without inclusion of covariates. Blue lines denote individual-level models; red lines denote the within-MZ pair estimates from CTC models. Parameters are defined as: $\beta_{YX} = 0$, $\beta_{YC} = 0.3$, $\beta_{XC} = -0.1$, $\rho_X = 0.72$, $\rho_Z = 0.75$, and $\rho_C = 0.75$.

Conclusions

The current dissertation investigates the transmission of, and outcomes related to, tobacco and alcohol use. We use genetically informative samples and research designs to provide stronger evidence of causal effects using observational data. Study 1 uses a directly genotyped parent-offspring sample to clarify the causal relationship between a shared environmental confounder, in parental educational attainment and occupational status, and adolescent alcohol and tobacco use. We find that genetic variants associated with smoking initiation in parents, measured as a polygenic risk score, predict offspring tobacco and alcohol use in adolescent offspring even after controlling for the risk genes they transmit (i.e., controlling for shared genetic confounders). This effect of untransmitted smoking initiation genotype is mediated by parental educational attainment and occupational status, consistent with a causal effect of parental SES on offspring substance use in adolescence. The underlying mechanisms by which rearing SES may causally influence offspring substance exposure are not clarified here. It is likely that parental SES is capturing other causal influences like parental monitoring or access to resources in adolescence (Bradley & Corwyn, 2002). Why we find a genetic nurture effect of parental genotype for smoking initiation (SI), and not for cigarettes per day (CPD) or drinks per week (DPW) polygenic scores, may be due to imprecision in the polygenic scores for CPD and DPW, or it may be that SI scores are indexing something unique that impacts offspring use through parent's attained SES. It will likely take much larger GWAS sample sizes of alcohol and tobacco use to address this. Results from Study 1 have important implications on reduction and prevention of adolescent substance use,

which is a strong predictor of adult use and abuse (Hawkins et al., 1997). The ability to detect sources of shared environmental influence on adolescent substance use has not been previously possible using classic twin designs. This is the first study, that we are aware of, that addresses genetic nurture effects and potential shared environmental influences of both tobacco and alcohol use with polygenic risk scores. Our findings, which are consistent with a causal effect of rearing SES on early substance use, have important societal and public health implications. Socioeconomic status in childhood is a potentially modifiable risk factor for adolescent tobacco and alcohol use, increases in which may causally reduce early substance use, and in turn, later use and abuse as well.

We explore an alternative method of causal inference in observational data using discordant twin pairs in Study 2. Co-twin control (CTC) models are increasingly being used because of their ability to implicitly control for all genetic and shared environmental factors, measured or unmeasured. The presence of unmeasured non-shared environmental confounders, however, can bias CTC estimates away from the null, in some cases to a larger extent than individual-level estimates (Frisell et al., 2012). Study 2 examines bias reduction with inclusion of a measured covariate, as a proxy for the unmeasured confounder. We show that inclusion of the covariate is always better, and better yet when the covariate is a good measure of the non-shared confounder, but that CTC estimates will remain biased, at least to some extent, in most cases. The best-case scenario for bias reduction in CTC models occurs when the within-twin pair correlation in an exposure is less than the within-twin pair correlation in the confounder, when the within-twin pair correlation in the covariate is high, and the covariate is an accurate measure of the

confounder. Despite the likely bias, we maintain that CTC models provide valuable information on causal relationships when specified and interpreted correctly, with clear acknowledgement of the method's limitations.

Study 3 then examines the association between adolescent substance use and adult socioeconomic outcomes using CTC models, with a discussion of the results in the context of findings from Study 2. We find within-MZ twin pair effects of adolescent tobacco and alcohol use on educational attainment, and to a lesser extent alcohol use on occupational status in adulthood. These results are consistent with a causal substance use exposure effect on adult attained SES controlling for all genetic and shared environmental sources of confounding, as well as the non-shared confounders of IQ, externalizing behavior, and adult substance use. In the context of Study 2, we conclude that the reported point estimates are likely biased to a small degree but are closer to the true causal effect than estimates from individual-level models. While the CTC design cannot prove causal relationships, that we find evidence consistent with a causal tobacco and alcohol use effect after controlling for many sources of potential confounding, in addition to using longitudinal data in an attempt to rule out reverse causality, suggests relatively strong evidence in support of an underlying causal relationship.

The use of genetically informative samples, including parent-offspring and twin samples, coupled with causally informative research designs, like CTC models and genotype information, can provide insights into causal relationship in observational data that cannot be elucidated using unrelated individuals alone. Each method and analysis presented here controls for different sources and amounts of confounding, which almost

surely surpasses confounding control in associational studies of unrelated individuals. Importantly, each of these methods and analyses have real limitations. It is not possible to control for all sources of true confounding in practice and no method to date using observational data, with genetically related individuals or not, can unequivocally establish causality. The problem of causal inference in observational data remains. Further information on genetic variants associated with substance use with increasingly large sample sizes will continue to expand the potential causal questions that can be asked and addressed. Appropriate use and interpretation of CTC models will also help in more accurately weighing the current evidence for or against causal exposure-outcome relationships. Because even randomized controlled trials, in the situations in which they can be feasibly and ethically conducted, may not prove an unequivocal answer to causal questions, we must rely on a combination of observational methods, each with their own strengths and limitations, to best address causal inference.

References

Archer, K. J., Hedeker, D., Nordgren, R., & Gibbons, R. D. (2018). mixor: Mixed-effects ordinal regression analysis. R package version 1.0.4.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Bates, T. C., Maher, B. S., Medland, S. E., McAloney, K., Wright, M. J., Hansell, N. K., … Gillespie, N. A. (2018). The nature of nurture: Using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families. *Twin Research and Human Genetics*, *21*(2), 73–83.

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, *11*(13), 1747–1758.

Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, *22*(16), 2591–2602.

Borland, B. L., & Rudolph, J. P. (1975). Relative effects of low socio-economic status, parental smoking and poor scholastic performance on smoking among high school students. *Social Science & Medicine*, *9*(1), 27–30.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*(1), 371–399.

Carlin, J. B., Gurrin, L. C., Sterne, J. A., Morley, R., & Dwyer, T. (2005). Regression models for twin studies: A critical review. *International Journal of Epidemiology*, *34*(5), 1089–1099.

Carmelli, D., & Page, W. F. (1996). Twenty-four year mortality in World War II US male veteran twins discordant for cigarette smoking. *International Journal of Epidemiology*, *25*(3), 554–559.

Casswell, S., Pledger, M., & Hooper, R. (2003). Socioeconomic status and drinking patterns in young adults. *Addiction*, *98*(5), 601–610.

Domingue, B. W., Conley, D., Fletcher, J., & Boardman, J. D. (2016). Cohort effects in the genetic influence on smoking. *Behavior Genetics*, *46*(1), 31–42.

D'Onofrio, B. M., Turkheimer, E., Emery, R. E., Heath, A. C., Madden, P. A., Slutske, W. S., & Martin, N. G. (2006). A genetically informed study of the processes underlying the association between parental marital instability and offspring adjustment. *Developmental Psychology*, *42*(3), 486–499.

Donovan, S. J., & Susser, E. (2011). Commentary: Advent of sibling designs. *International Journal of Epidemiology*, *40*(2), 345–349.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*(3), e1003348.

Elam, K. K., Chassin, L., Lemery-Chalfant, K., Pandika, D., Wang, F. L., Bountress, K., … Agrawal, A. (2017). Affiliation with substance-using peers: Examining gene-environment correlations among parent monitoring, polygenic risk, and children's impulsivity. *Developmental Psychobiology*, *59*(5), 561–573.

Ellickson, P. L., Tucker, J. S., & Klein, D. J. (2001). High-risk behaviors associated with early smoking: Results from a 5-year follow-up. *Journal of Adolescent Health*, *28*(6), 465–473.

Frisell, T., Öberg, S., Kuja-Halkola, R., & Sjölander, A. (2012). Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology*, *23*(5), 713–720.

Gilman, S. E., Abrams, D. B., & Buka, S. L. (2003). Socioeconomic status over the life course and stages of cigarette use: Initiation, regular use, and cessation. *Journal of Epidemiology & Community Health*, *57*(10), 802–808.

Gilman, Stephen E., Martin, L. T., Abrams, D. B., Kawachi, I., Kubzansky, L., Loucks, E. B., … Buka, S. L. (2008). Educational attainment and cigarette smoking: A causal association? *International Journal of Epidemiology*, *37*(3), 615–624.

Goodman, E., & Huang, B. (2002). Socioeconomic status, depressive symptoms, and adolescent substance use. *Archives of Pediatrics & Adolescent Medicine*, *156*(5), 448–453.

Grant, J. D., Scherrer, J. F., Lynskey, M. T., Agrawal, A., Duncan, A. E., Haber, J. R., … Bucholz, K. K. (2012). Associations of alcohol, nicotine, cannabis, and drug use/dependence with educational attainment: Evidence from cotwin-control analyses. *Alcoholism: Clinical and Experimental Research*, *36*(8), 1412–1420.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, *14*(3), 300–306.

Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health*, *22*(1), 189–212.

Hawkins, J. D., Graham, J. W., Maguin, E., Abbott, R., Hill, K. G., & Catalano, R. F. (1997). Exploring the effects of age of alcohol use initiation and psychosocial risk

factors on subsequent alcohol misuse. *Journal of Studies on Alcohol*, *58*(3), 280–290.

Hicks, B. M., Iacono, W. G., & McGue, M. (2010). Consequences of an adolescent onset and persistent course of alcohol dependence in men: Adolescent risk factors and adult outcomes. *Alcoholism: Clinical and Experimental Research*, *34*(5), 819–833.

Hübinette, A., Lichtenstein, P., Ekbom, A., & Cnattingius, S. (2001). Birth characteristics and breast cancer risk: A study among like-sexed twins. *International Journal of Cancer*, *91*(2), 248–251.

Humensky, J. L. (2010). Are adolescents with high socioeconomic status more likely to engage in alcohol and illicit drug use in early adulthood? *Substance Abuse Treatment, Prevention, and Policy*, *5*, 5-19.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ*, *340*, c2289.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., & Mcgue, M. (1999). Behavioral disinhibition and the development of substance-use disorders: Findings from the Minnesota Twin Family Study. *Development and Psychopathology*, *11*(4), 869–900.

Irons, D. E., Iacono, W. G., & McGue, M. (2015). Tests of the effects of adolescent early alcohol exposures on adult outcomes. *Addiction*, *110*(2), 269–278.

Jaffee, S. R., & Price, T. S. (2008). Genotype–environment correlations: Implications for determining the relationship between environmental exposures and psychiatric illness. *Psychiatry*, *7*(12), 496–499.

Jaffee, S. R., & Price, T. S. (2012). The implications of genotype–environment correlation for establishing causal processes in psychopathology. *Development and Psychopathology*, *24*(4), 1253–1264.

Jefferis, B. J. M. H., Power, C., Graham, H., & Manor, O. (2004). Effects of childhood socioeconomic circumstances on persistent smoking. *American Journal of Public Health*, *94*(2), 279–285.

Kadlecová, P., Andel, R., Mikulík, R., Handing, E. P., & Pedersen, N. L. (2015). Alcohol consumption at midlife and risk of stroke during 43 years of follow-up: Cohort and twin analyses. *Stroke*, *46*(3), 627–633.

Kandel, D. B., Davies, M., Karus, D., & Yamaguchi, K. (1986). The consequences in young adulthood of adolescent drug involvement: An overview. *Archives of General Psychiatry*, *43*(8), 746–754.

Kendler, K. S., & Baker, J. H. (2007). Genetic influences on measures of the environment: A systematic review. *Psychological Medicine*, *37*(5), 615–626.

Kendler, K. S., Gardner, C. O., Hickman, M., Heron, J., Macleod, J., Lewis, G., & Dick, D. M. (2014). Socioeconomic status and alcohol-related behaviors in mid- to late adolescence in the Avon Longitudinal Study of Parents and Children. *Journal of Studies on Alcohol and Drugs*, *75*(4), 541–545.

Kendler, K. S., Schmitt, E., Aggen, S. H., & Prescott, C. A. (2008). Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. *Archives of General Psychiatry*, *65*(6), 674–682.

Keyes, M. A., Legrand, L. N., Iacono, W. G., & McGue, M. (2008). Parental smoking and adolescent problem behavior: An adoption study of general and specific effects. *American Journal of Psychiatry*, *165*(10), 1338–1344.

Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L. N., McGue, M., & Iacono, W. G. (2009). The Enrichment Study of the Minnesota Twin Family Study: Increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Research and Human Genetics*, *12*(5), 489–501.

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., … Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428.

Krieger, N., Williams, D. R., & Moss, N. E. (1997). Measuring social class in U.S. public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health*, *18*(1), 341-378.

Lahey, B. B., & D'Onofrio, B. M. (2010). All in the family: Comparing siblings to test causal hypotheses regarding environmental influences on behavior. *Current Directions in Psychological Science*, *19*(5), 319–323.

Latvala, A., Rose, R. J., Pulkkinen, L., Dick, D. M., Korhonen, T., & Kaprio, J. (2014). Drinking, smoking, and educational achievement: Cross-lagged associations from adolescence to adulthood. *Drug and Alcohol Dependence*, *137*, 106–113.

Lemstra, M., Bennett, N. R., Neudorf, C., Kunst, A., Nannapaneni, U., Warren, L. M., … Scott, C. R. (2008). A meta-analysis of marijuana and alcohol use by socio-economic status in adolescents aged 10–15 years. *Canadian Journal of Public Health*, *99*(3), 172–177.

Li, M. D., Cheng, R., Ma, J. Z., & Swan, G. E. (2003). A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction*, *98*(1), 23–31.

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology*, *127*(4), 864–874.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., … Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, *51*(2), 237-244.

Ludeke, S., Johnson, W., McGue, M., & Iacono, W. G. (2013). Genetic amplification and the individualization of the parent–child relationship across adolescence. *Psychological Medicine*, *43*(2), 413–422.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., … Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics*, *100*(4), 635–649.

McCaffery, J. M., Papandonatos, G. D., Lyons, M. J., Koenen, K. C., Tsuang, M. T., & Niaura, R. (2008). Educational attainment, smoking initiation and lifetime nicotine dependence among male Vietnam-era Twins. *Psychological Medicine*, *38*(9), 1287–1297.

McCambridge, J., McAlaney, J., & Rowe, R. (2011). Adult consequences of late adolescent alcohol consumption: A systematic review of cohort studies. *PLoS Medicine*, *8*(2), e1000413.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (pp. 105–142). New York: Academic Press.

McGue, M., Osler, M., & Christensen, K. (2010). Causal inference and observational research: The utility of twins. *Perspectives on Psychological Science*, *5*(5), 546–556.

Meier, M. H., Caspi, A., Danese, A., Fisher, H. L., Houts, R., Arseneault, L., & Moffitt, T. E. (2018). Associations between adolescent cannabis use and neuropsychological decline: A longitudinal co-twin control study. *Addiction*, *113*(2), 257–265.

Melchior, M., Moffitt, T. E., Milne, B. J., Poulton, R., & Caspi, A. (2007). Why do children from socioeconomically disadvantaged families suffer from poor health when they reach adulthood? A life-course study. *American Journal of Epidemiology*, *166*(8), 966–974.

Milán, T., Verkasalo, P. K., Kaprio, J., & Koskenvuo, M. (2003). Lifestyle differences in twin pairs discordant for basal cell carcinoma of the skin. *The British Journal of Dermatology*, *149*(1), 115–123.

Miller, M. B., Basu, S., Cunningham, J., Eskin, E., Malone, S. M., Oetting, W. S., … McGue, M. (2012). The Minnesota Center for Twin and Family Research genome-wide association study. *Twin Research and Human Genetics*, *15*(6), 767–774.

O'Brien, T. C., Mustanski, B. S., Skol, A., Cook, E. H., & Wakschlag, L. S. (2013). Do dopamine gene variants and prenatal smoking interactively predict youth externalizing behavior? *Neurotoxicology and Teratology*, *40*, 67–73.

Patrick, M. E., Wightman, P., Schoeni, R. F., & Schulenberg, J. E. (2012). Socioeconomic status and substance use among young adults: A comparison across constructs and drugs. *Journal of Studies on Alcohol and Drugs*, *73*(5), 772–782.

Plomin, R. (2014). Genotype-environment correlation in the era of DNA. *Behavior Genetics*, *44*(6), 629–638.

Poonawalla, I. B., Kendzor, D. E., Owen, M. T., & Caughy, M. O. (2014). Family income trajectory during childhood is associated with adolescent cigarette smoking and alcohol use. *Addictive Behaviors*, *39*(10), 1383–1388.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Robins, L., Babor, T., & Cottler, L. (1987). *Composite International Diagnostic Interview: Expanded Substance Abuse Module.* St. Louis, MO: Washington University.

Robins LN, Wing J, Wittchen H, & et al. (1988). The Composite International Diagnostic Interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, *45*(12), 1069–1077.

Rose, R. J., Winter, T., Viken, R. J., & Kaprio, J. (2014). Adolescent alcohol abuse and adverse adult outcomes: Evaluating confounds with drinking-discordant twins. *Alcoholism: Clinical and Experimental Research*, *38*(8), 2314–2321.

Rosner, B., Spiegelman, D., & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology*, *132*(4), 734–745.

Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science*, *2*(4), 377–395.

Samek, D. R., Keyes, M. A., Hicks, B. M., Bailey, J., McGue, M., & Iacono, W. G. (2014). General and specific predictors of nicotine and alcohol dependence in early adulthood: Genetic and environmental influences. *Journal of Studies on Alcohol and Drugs*, *75*(4), 623–634.

Sasco, A. J., Secretan, M. B., & Straif, K. (2004). Tobacco smoking and cancer: A brief review of recent epidemiological evidence. *Lung Cancer*, *45*, S3–S9.

Savitz, J., van der Merwe, L., Newman, T. K., Stein, D. J., & Ramesar, R. (2010). Catechol-o-methyltransferase genotype and childhood trauma may interact to impact schizotypal personality traits. *Behavior Genetics*, *40*(3), 415–423.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype greater than environment effects. *Child Development*, *54*(2), 424–435.

Sjölander, A., Frisell, T., & Öberg, S. (2012). Causal interpretation of between-within models for twin research. *Epidemiologic Methods*, *1*(1), 217–237.

Soteriades, E. S., & DiFranza, J. R. (2003). Parent's socioeconomic status, adolescents' disposable income, and adolescents' smoking status in Massachusetts. *American Journal of Public Health*, *93*(7), 1155–1160.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Staff, J., & Maggs, J. (2017). Alcohol and cigarette use from ages 23 to 55: Links with health and well-being in the Long-Term National Child Development Study. *Journal of Studies on Alcohol and Drugs*, *78*(3), 394–403.

Stolley, P. D. (1991). When genius errs: R. A. Fisher and the lung cancer controversy. *American Journal of Epidemiology*, *133*(5), 416–425.

Sullivan, P. F., & Kendler, K. S. (1999). The genetic epidemiology of smoking. *Nicotine & Tobacco Research*, *1*(S2), S51–S57.

Swerdlow, A. J., De Stavola, B. L., Swanwick, M. A., Mangtani, P., & Maconochie, N. E. (1999). Risk factors for testicular cancer: A case-control study in twins. *British Journal of Cancer*, *80*(7), 1098–1102.

Townsend, L., Flisher, A. J., & King, G. (2007). A systematic review of the relationship between high school dropout and substance use. *Clinical Child and Family Psychology Review*, *10*(4), 295–317.

Tyas, S., & Pederson, L. (1998). Psychosocial factors related to adolescent smoking: A critical review of the literature. *Tobacco Control*, *7*(4), 409–420.

Verhulst, B., Neale, M. C., & Kendler, K. S. (2015). The heritability of alcohol use disorders: A meta-analysis of twin and adoption studies. *Psychological Medicine*, *45*(05), 1061–1072.

Verweij, K. J. H., Huizink, A. C., Agrawal, A., Martin, N. G., & Lynskey, M. T. (2013). Is the relationship between early-onset cannabis use and educational attainment causal or due to common liability? *Drug and Alcohol Dependence*, *133*(2), 580–586.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., … Price, A. L. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *97*(4), 576–592.

Viner, R. M., & Taylor, B. (2007). Adult outcomes of binge drinking in adolescence: Findings from a UK national birth cohort. *Journal of Epidemiology and Community Health*, *61*(10), 902–907.

Vink, J. M., Willemsen, G., & Boomsma, D. I. (2005). Heritability of smoking initiation and nicotine dependence. *Behavior Genetics*, *35*(4), 397–406.

Waldron, J. S., Malone, S. M., McGue, M., & Iacono, W. G. (2018). A co-twin control study of the relationship between adolescent drinking and adult outcomes. *Journal of Studies on Alcohol and Drugs*, *79*(4), 635–643.

Wells, J. E., Horwood, L. J., & Fergusson, D. M. (2004). Drinking patterns in mid-adolescence and psychosocial outcomes in late adolescence and early adulthood. *Addiction*, *99*(12), 1529–1541.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., … Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*(8), 1359–1366.

Wiles, N. J., Lingford-Hughes, A., Daniel, J., Hickman, M., Farrell, M., Macleod, J., … Lewis, G. (2007). Socio-economic status in childhood and later alcohol use: A systematic review. *Addiction*, *102*(10), 1546–1563.

Willoughby, E. A., McGue, M., Iacono, W. G., Rustichini, A., & Lee, J. J. (In press). The role of parental genotype in predicting offspring years of education: Evidence for genetic nurture. *Molecular Psychiatry*.

Wilson, S., Malone, S. M., Hunt, R. H., Thomas, K. M., & Iacono, W. G. (2018). Problematic alcohol use and hippocampal volume in a female sample: Disentangling cause from consequence using a co-twin control study design. *Psychological Medicine*, *48*(10), 1673–1684.

Proof of equation 10:

$$\beta_{cov} = \frac{\text{var}(Z_{ij})\,\text{cov}(X_{ij}, Y_{ij}) - \text{cov}(Z_{ij}, X_{ij})\,\text{cov}(Z_{ij}, Y_{ij})}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \text{cov}(Z_{ij}, X_{ij})^2}$$

$$= \frac{\begin{array}{l}\text{var}(Z_{ij})\,\text{cov}\left(X_{ij}, \beta_{YX}X_{ij} + \beta_{YC}C_{ij} + \epsilon_{Y_{ij}}\right) \\ \quad - \text{cov}\left(\beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}, \beta_{XC}C_{ij} + \epsilon_{X_{ij}}\right)\text{cov}\left(\beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}, \beta_{YX}X_{ij} + \beta_{YC}C_{ij} + \epsilon_{Y_{ij}}\right)\end{array}}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \text{cov}\left(\beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}, \beta_{XC}C_{ij} + \epsilon_{X_{ij}}\right)^2}$$

$$= \frac{\begin{array}{l}\text{var}(Z_{ij})\left[\beta_{YX}\text{var}(X_{ij}) + \beta_{YC}\text{cov}(X_{ij}, C_{ij})\right] \\ \quad - \beta_{ZC}\beta_{XC}\text{var}(C_{ij})\left[\beta_{ZC}\beta_{YX}\text{cov}(C_{ij}, X_{ij}) + \beta_{ZC}\beta_{YC}\text{var}(C_{ij})\right]\end{array}}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \left[\beta_{ZC}\beta_{XC}\text{var}(C_{ij})\right]^2}$$

$$= \frac{\begin{array}{l}\beta_{YX}\text{var}(Z_{ij})\,\text{var}(X_{ij}) + \beta_{YC}\text{var}(Z_{ij})\,\text{cov}\left(\beta_{XC}C_{ij} + \epsilon_{X_{ij}}, C_{ij}\right) \\ \quad - \beta_{ZC}\beta_{XC}\text{var}(C_{ij})\left[\beta_{ZC}\beta_{YX}\text{cov}\left(C_{ij}, \beta_{XC}C_{ij} + \epsilon_{X_{ij}}\right) + \beta_{ZC}\beta_{YC}\text{var}(C_{ij})\right]\end{array}}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \left[\beta_{ZC}\beta_{XC}\text{var}(C_{ij})\right]^2}$$

$$= \frac{\begin{array}{l}\beta_{YX}\text{var}(Z_{ij})\,\text{var}(X_{ij}) + \beta_{YC}\beta_{XC}\text{var}(Z_{ij})\,\text{var}(C_{ij}) \\ \quad - \beta_{ZC}\beta_{XC}\text{var}(C_{ij})\left[\beta_{ZC}\beta_{YX}\beta_{XC}\text{var}(C_{ij}) + \beta_{ZC}\beta_{YC}\text{var}(C_{ij})\right]\end{array}}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \left[\beta_{ZC}\beta_{XC}\text{var}(C_{ij})\right]^2}$$

$$= \frac{\beta_{YX}\text{var}(Z_{ij})\,\text{var}(X_{ij}) + \beta_{YC}\beta_{XC}\sigma_C^2\text{var}(Z_{ij}) - \beta_{YX}\beta_{ZC}^2\beta_{XC}^2(\sigma_C^2)^2 - \beta_{ZC}^2\beta_{XC}\beta_{YC}(\sigma_C^2)^2}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \beta_{ZC}^2\beta_{XC}^2(\sigma_C^2)^2}$$

$$= \beta_{YX} + \frac{\beta_{YC}\beta_{XC}\sigma_C^2\text{var}(Z_{ij}) - \beta_{ZC}^2\beta_{XC}\beta_{YC}(\sigma_C^2)^2}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \beta_{ZC}^2\beta_{XC}^2(\sigma_C^2)^2}$$

$$= \beta_{YX} + \frac{\beta_{YC}\beta_{XC}\sigma_C^2\left[\text{var}(Z_{ij}) - \beta_{ZC}^2\sigma_C^2\right]}{\text{var}(Z_{ij})\,\text{var}(X_{ij}) - \beta_{ZC}^2\beta_{XC}^2(\sigma_C^2)^2}$$

$$= \beta_{YX} + \frac{\beta_{YC}\beta_{XC}\sigma_C^2\left[\left(\beta_{ZC}^2\sigma_C^2 + \sigma_{\epsilon_Z}^2\right) - \beta_{ZC}^2\sigma_C^2\right]}{\left(\beta_{ZC}^2\sigma_C^2 + \sigma_{\epsilon_Z}^2\right)\left(\beta_{XC}^2\sigma_C^2 + \sigma_{\epsilon_X}^2\right) - \beta_{ZC}^2\beta_{XC}^2(\sigma_C^2)^2}$$

If $\text{var}(C_{ij}) = \text{var}(X_{ij}) = \text{var}(Z_{ij}) = 1$, then

$$\beta_{cov} = \beta_{YX} + \frac{\beta_{YC}\beta_{XC}\left(1 - \beta_{ZC}^2\right)}{1 - \beta_{ZC}^2\beta_{XC}^2}$$

Proof of equation 11:

β_B may be excluded from Eq. 3 without changing the value of β_W[1]:

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \bar{X}_i, Z_{ij}\right)\right\} = \beta_0 + \beta_W\left(X_{ij} - \bar{X}_i\right) + \beta_Z Z_{ij},$$

We can then modify this model as:

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \bar{X}_i, Z_{ij}\right)\right\} = \beta_0 + \beta_W^*\frac{1}{2}\left(X_{ij} - X_{ij'}\right) + \beta_Z Z_{ij} \text{ so that } \beta_W^* = \tfrac{\beta_W}{2}.$$

We then have:

$$\beta^*_{W_{covstd}} = \frac{\text{var}\left(Z_{ij}\right)\text{cov}\left[Y_{ij},\left(X_{ij}-X_{ij'}\right)\right] - \text{cov}\left[Z_{ij},\left(X_{ij}-X_{ij'}\right)\right]\text{cov}\left(Y_{ij},Z_{ij}\right)}{\text{var}\left(Z_{ij}\right)\text{var}\left(X_{ij}-X_{ij'}\right) - \text{cov}\left[Z_{ij},\left(X_{ij}-X_{ij'}\right)\right]^2}$$

Plugging in pieces from the derivations below gives:

$$\beta^*_{W_{covstd}} = \frac{\begin{array}{r}\left(\beta^2_{ZC}\sigma^2_C + \sigma_{\epsilon_Z}\right)\left[\beta_{YX}\left(1-\beta^2_{XC}\sigma^2_C\rho_C - \sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right) + \beta_{YC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right] \\ -\beta_{ZC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\left(\beta_{YX}\beta_{XC}\beta_{ZC}\sigma^2_C + \beta_{YC}\beta_{ZC}\sigma^2_C\right)\end{array}}{2\left(\beta^2_{ZC}\sigma^2_C + \sigma_{\epsilon_Z}\right)\left(1-\beta^2_{XC}\sigma^2_C\rho_C - \rho_{\epsilon_X}\sigma^2_{\epsilon_X}\right) - \left[\beta_{ZC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right]^2}$$

If $\text{var}\left(C_{ij}\right) = \text{var}\left(X_{ij}\right) = \text{var}\left(Z_{ij}\right) = 1$, then this simplifies to

$$\beta^*_{W_{covstd}} = \frac{\beta_{YX}\left(1-\beta^2_{XC}\rho_C - \sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right) + \beta_{YC}\beta_{XC}\left(1-\rho_C\right) - \beta_{ZC}\beta_{XC}\left(1-\rho_C\right)\left(\beta_{YX}\beta_{XC}\beta_{ZC} + \beta_{YC}\beta_{ZC}\right)}{2\left(1-\beta^2_{XC}\rho_C - \rho_{\epsilon_X}\sigma^2_{\epsilon_X}\right) - \left[\beta_{ZC}\beta_{XC}\left(1-\rho_C\right)\right]^2}$$

Proof of equation 12:

Similarly, β_B may be excluded from Eq. 4 without changing the value of β_W[1]:

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \bar{X}_i, Z_{ij}, \bar{Z}_i\right)\right\} = \beta_0 + \beta_W\left(X_{ij} - \bar{X}_i\right) + \beta_Z\left(Z_{ij} - \bar{Z}_i\right)$$

We can then modify this model as:

$$g\left\{E\left(Y_{ij} \mid X_{ij}, \bar{X}_i, Z_{ij}, \bar{Z}_i\right)\right\} = \beta_0 + \beta^*_W\frac{1}{2}\left(X_{ij} - X_{ij'}\right) + \beta^*_Z\frac{1}{2}\left(Z_{ij} - Z_{ij'}\right) \quad \text{so that } \beta^*_W = \tfrac{\beta_W}{2} \text{ and } \beta^*_Z = \tfrac{\beta_Z}{2}.$$

We then have:

$$\beta^*_{W_{cov}} = \frac{\text{var}\left(Z_{ij} - Z_{ij'}\right)\text{cov}\left[Y_{ij},\left(X_{ij}-X_{ij'}\right)\right] - \text{cov}\left[\left(Z_{ij} - Z_{ij'}\right),\left(X_{ij}-X_{ij'}\right)\right]\text{cov}\left[Y_{ij},\left(Z_{ij}-Z_{ij'}\right)\right]}{\text{var}\left(Z_{ij} - Z_{ij'}\right)\text{var}\left(X_{ij}-X_{ij'}\right) - \text{cov}\left[\left(Z_{ij} - Z_{ij'}\right),\left(X_{ij}-X_{ij'}\right)\right]^2}$$

Plugging in pieces from the derivations below gives:

$$\beta^*_{W_{cov}} = \frac{\begin{array}{r}\left[2\left(1-\beta^2_{ZC}\sigma^2_C\rho_C-\rho_{\epsilon_Z}\sigma^2_{\epsilon_Z}\right)\right]\left[\beta_{YX}\left(1-\beta^2_{XC}\sigma^2_C\rho_C-\sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right)+\beta_{YC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right] \\ -\left[2\beta_{ZC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right]\left[\left(\beta_{YX}\beta_{XC}\beta_{ZC}+\beta_{YC}\beta_{ZC}\right)\sigma^2_C\left(1-\rho_C\right)\right]\end{array}}{\left[2\left(1-\beta^2_{ZC}\sigma^2_C\rho_C-\rho_{\epsilon_Z}\sigma^2_{\epsilon_Z}\right)\right]\left[2\left(1-\beta^2_{XC}\sigma^2_C\rho_C-\rho_{\epsilon_X}\sigma^2_{\epsilon_X}\right)\right]-\left[2\beta_{ZC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right]^2}$$

If $\text{var}\left(C_{ij}\right)=\text{var}\left(X_{ij}\right)=\text{var}\left(Z_{ij}\right)=1$, then this simplifies to

$$\beta^*_{W_{cov}} = \frac{\begin{array}{r}\left[2\left(1-\beta^2_{ZC}\rho_C-\rho_{\epsilon_Z}\sigma^2_{\epsilon_Z}\right)\right]\left[\beta_{YX}\left(1-\beta^2_{XC}\rho_C-\sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right)+\beta_{YC}\beta_{XC}\left(1-\rho_C\right)\right] \\ -\left[2\beta_{ZC}\beta_{XC}\left(1-\rho_C\right)\right]\left[\left(\beta_{YX}\beta_{XC}\beta_{ZC}+\beta_{YC}\beta_{ZC}\right)\left(1-\rho_C\right)\right]\end{array}}{\left[2\left(1-\beta^2_{ZC}\rho_C-\rho_{\epsilon_Z}\sigma^2_{\epsilon_Z}\right)\right]\left[2\left(1-\beta^2_{XC}\rho_C-\rho_{\epsilon_X}\sigma^2_{\epsilon_X}\right)\right]-\left[2\beta_{ZC}\beta_{XC}\left(1-\rho_C\right)\right]^2}$$

Derivation pieces:

The components of β^*_W are derived as:

$$\begin{aligned}
\text{cov}\left[Y_{ij},\left(X_{ij}-X^{'}_{ij}\right)\right] &= \text{cov}\left(Y_{ij},X_{ij}\right)-\text{cov}\left(Y_{ij},X_{ij'}\right) \\
&= \text{cov}\left(\beta_{YX}X_{ij}+\beta_{YC}C_{ij}+\epsilon_{Y_{ij}},X_{ij}\right)-\text{cov}\left(\beta_{YX}X_{ij}+\beta_{YC}C_{ij}+\epsilon_{Y_{ij}},X_{ij'}\right) \\
&= \beta_{YX}\text{var}\left(X_{ij}\right)+\beta_{YC}\text{cov}\left(C_{ij},X_{ij}\right)-\beta_{YX}\text{cov}\left(X_{ij},X_{ij'}\right)-\beta_{YC}\text{cov}\left(C_{ij},X_{ij'}\right) \\
&= \beta_{YX}\left[\text{var}\left(X_{ij}\right)-\text{cov}\left(X_{ij},X_{ij'}\right)\right]+\beta_{YC}\left[\text{cov}\left(C_{ij},X_{ij}\right)-\text{cov}\left(C_{ij},X_{ij'}\right)\right] \\
&= \beta_{YX}\left[\text{var}\left(\beta_{XC}C_{ij}+\epsilon_{X_{ij}}\right)-\text{cov}\left(\beta_{XC}C_{ij}+\epsilon_{X_{ij}},\beta_{XC}C_{ij'}+\epsilon_{X_{ij'}}\right)\right] \\
&\quad +\beta_{YC}\left[\text{cov}\left(C_{ij},\beta_{XC}C_{ij}+\epsilon_{X_{ij}}\right)-\text{cov}\left(C_{ij},\beta_{XC}C_{ij'}+\epsilon_{X_{ij'}}\right)\right] \\
&= \beta_{YX}\left(\beta^2_{XC}\sigma^2_C+\sigma^2_{\epsilon_X}-\beta^2_{XC}\sigma^2_C\rho_C-\sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right)+\beta_{YC}\left(\beta_{XC}\sigma^2_C-\beta_{XC}\sigma^2_C\rho_C\right) \\
&= \beta_{YX}\left[\beta^2_{XC}\sigma^2_C\left(1-\rho_C\right)+\sigma^2_{\epsilon_X}\left(1-\rho_{\epsilon_X}\right)\right]+\beta_{YC}\left[\beta_{XC}\sigma^2_C\left(1-\rho_C\right)\right] \\
&= \beta_{YX}\beta^2_{XC}\sigma^2_C\left(1-\rho_C\right)+\beta_{YX}\sigma^2_{\epsilon_X}\left(1-\rho_{\epsilon_X}\right)+\beta_{YC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right) \\
&= \beta_{YX}\left(1-\beta^2_{XC}\sigma^2_C\rho_C-\sigma^2_{\epsilon_X}\rho_{\epsilon_X}\right)+\beta_{YC}\beta_{XC}\sigma^2_C\left(1-\rho_C\right)
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(Y_{ij}, Z_{ij}) &= \operatorname{cov}\left(\beta_{YX}X_{ij} + \beta_{YC}C_{ij} + \epsilon_{Y_{ij}}, Z_{ij}\right) \\
&= \beta_{YX}\operatorname{cov}(X_{ij}, Z_{ij}) + \beta_{YC}\operatorname{cov}(C_{ij}, Z_{ij}) \\
&= \beta_{YX}\operatorname{cov}\left(\beta_{XC}C_{ij} + \epsilon_{X_{ij}}, \beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}\right) + \beta_{YC}\operatorname{cov}\left(C_{ij}, \beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}\right) \\
&= \beta_{YX}\beta_{XC}\beta_{ZC}\sigma_C^2 + \beta_{YC}\beta_{ZC}\sigma_C^2 \\
\operatorname{cov}\left[Z_{ij}, (X_{ij} - X_{ij'})\right] &= \operatorname{cov}(Z_{ij}, X_{ij}) - \operatorname{cov}(Z_{ij}, X_{ij'}) \\
&= \operatorname{cov}\left(\beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}, \beta_{XC}C_{ij} + \epsilon_{X_{ij}}\right) - \operatorname{cov}\left(\beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}, \beta_{XC}C_{ij'} + \epsilon_{X_{ij'}}\right) \\
&= \beta_{ZC}\beta_{XC}\sigma_C^2 - \beta_{ZC}\beta_{XC}\sigma_C^2\rho_C \\
&= \beta_{ZC}\beta_{XC}\sigma_C^2(1 - \rho_C) \\
\operatorname{cov}\left[Y_{ij}, (Z_{ij} - Z_{ij'})\right] &= \operatorname{cov}(Y_{ij}, Z_{ij}) - \operatorname{cov}(Y_{ij}, Z_{ij'}) \\
&= \operatorname{cov}\left(\beta_{YX}X_{ij} + \beta_{YC}C_{ij} + \epsilon_{Y_{ij}}, Z_{ij}\right) - \operatorname{cov}\left(\beta_{YX}X_{ij} + \beta_{YC}C_{ij} + \epsilon_{Y_{ij}}, Z_{ij'}\right) \\
&= \beta_{YX}\operatorname{cov}(X_{ij}, Z_{ij}) + \beta_{YC}\operatorname{cov}(C_{ij}, Z_{ij}) - \beta_{YX}\operatorname{cov}(X_{ij}, Z_{ij'}) - \beta_{YC}\operatorname{cov}(C_{ij}, Z_{ij'}) \\
&= \beta_{YX}\left[\operatorname{cov}(X_{ij}, Z_{ij}) - \operatorname{cov}(X_{ij}, Z_{ij'})\right] + \beta_{YC}\left[\operatorname{cov}(C_{ij}, Z_{ij}) - \operatorname{cov}(C_{ij}, Z_{ij'})\right] \\
&= \beta_{YX}\left[\operatorname{cov}\left(\beta_{XC}C_{ij} + \epsilon_{X_{ij}}, Z_{ij}\right) - \operatorname{cov}\left(\beta_{XC}C_{ij} + \epsilon_{X_{ij}}, Z_{ij'}\right)\right] \\
&\quad + \beta_{YC}\left[\operatorname{cov}\left(C_{ij}, \beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}\right) - \operatorname{cov}\left(C_{ij}, \beta_{ZC}C_{ij'} + \epsilon_{Z_{ij'}}\right)\right] \\
&= \beta_{YX}\left[\beta_{XC}\operatorname{cov}\left(C_{ij}, \beta_{ZC}C_{ij} + \epsilon_{Z_{ij}}\right) - \beta_{XC}\operatorname{cov}\left(C_{ij}, \beta_{ZC}C_{ij'} + \epsilon_{Z_{ij'}}\right)\right] \\
&\quad + \beta_{YC}\left(\beta_{ZC}\sigma_C^2 - \beta_{ZC}\sigma_C^2\rho_C\right) \\
&= \beta_{YX}\left(\beta_{XC}\beta_{ZC}\sigma_C^2 - \beta_{XC}\beta_{ZC}\sigma_C^2\rho_C\right) + \beta_{YC}\left(\beta_{ZC}\sigma_C^2 - \beta_{ZC}\sigma_C^2\rho_C\right) \\
&= \beta_{YX}\beta_{XC}\beta_{ZC}\sigma_C^2(1 - \rho_C) + \beta_{YC}\beta_{ZC}\sigma_C^2(1 - \rho_C) \\
&= \left(\beta_{YX}\beta_{XC}\beta_{ZC} + \beta_{YC}\beta_{ZC}\right)\sigma_C^2(1 - \rho_C)
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}\left[\left(Z_{ij}-Z_{ij'}\right),\left(X_{ij}-X_{ij'}\right)\right] &= \operatorname{cov}\left(Z_{ij},X_{ij}\right)-\operatorname{cov}\left(Z_{ij},X_{ij'}\right)-\operatorname{cov}\left(Z_{ij'},X_{ij}\right)+\operatorname{cov}\left(Z_{ij'},X_{ij'}\right) \\
&= \operatorname{cov}\left(\beta_{ZC}C_{ij}+\epsilon_{Z_{ij}},X_{ij}\right)-\operatorname{cov}\left(\beta_{ZC}C_{ij}+\epsilon_{Z_{ij}},X_{ij'}\right) \\
&\quad -\operatorname{cov}\left(\beta_{ZC}C_{ij'}+\epsilon_{Z_{ij'}},X_{ij}\right)+\operatorname{cov}\left(\beta_{ZC}C_{ij'}+\epsilon_{Z_{ij'}},X_{ij'}\right) \\
&= \beta_{ZC}\operatorname{cov}\left(C_{ij},X_{ij}\right)-\beta_{ZC}\operatorname{cov}\left(C_{ij},X_{ij'}\right)-\beta_{ZC}\operatorname{cov}\left(C_{ij'},X_{ij}\right)+\beta_{ZC}\left(C_{ij'},X_{ij'}\right) \\
&= \beta_{ZC}\operatorname{cov}\left(C_{ij},\beta_{XC}C_{ij}+\epsilon_{X_{ij}}\right)-\beta_{ZC}\operatorname{cov}\left(C_{ij},\beta_{XC}C_{ij'}+\epsilon_{X_{ij'}}\right) \\
&\quad -\beta_{ZC}\operatorname{cov}\left(C_{ij'},\beta_{XC}C_{ij}+\epsilon_{X_{ij}}\right)+\beta_{ZC}\operatorname{cov}\left(C_{ij'},\beta_{XC}C_{ij'}+\epsilon_{X_{ij}}\right) \\
&= \beta_{ZC}\beta_{XC}\sigma_C^2-\beta_{ZC}\beta_{XC}\sigma_C^2\rho_C-\beta_{ZC}\beta_{XC}\sigma_C^2\rho_C+\beta_{ZC}\beta_{XC}\sigma_C^2 \\
&= 2\beta_{ZC}\beta_{XC}\sigma_C^2-2\beta_{ZC}\beta_{XC}\sigma_C^2\rho_C \\
&= 2\beta_{ZC}\beta_{XC}\sigma_C^2\left(1-\rho_C\right)
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}\left(X_{ij}-X_{ij'}\right) &= 2\left[\operatorname{var}\left(X_{ij}\right)-\operatorname{cov}\left(X_{ij},X_{ij'}\right)\right] \\
&= 2\left(1-\beta_{XC}^2\sigma_C^2\rho_C-\rho_{\epsilon_X}\sigma_{\epsilon_X}^2\right)
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}\left(Z_{ij}-Z_{ij'}\right) &= 2\left[\operatorname{var}\left(Z_{ij}\right)-\operatorname{cov}\left(Z_{ij},Z_{ij'}\right)\right] \\
&= 2\left(1-\beta_{ZC}^2\sigma_C^2\rho_C-\rho_{\epsilon_Z}\sigma_{\epsilon_Z}^2\right)
\end{aligned}$$

Proof of equation 15:

$$\begin{aligned}
\beta_{ME} &= \frac{\operatorname{var}\left(Z_{ij}^*\right)\operatorname{cov}\left(Y_{ij},X_{ij}^*\right)-\operatorname{cov}\left(Z_{ij}^*,X_{ij}^*\right)\operatorname{cov}\left(Z_{ij}^*,Y_{ij}\right)}{\operatorname{var}\left(Z_{ij}^*\right)\operatorname{var}\left(X_{ij}^*\right)-\operatorname{cov}\left(Z_{ij}^*,X_{ij}^*\right)^2} \\
&= \frac{\left(\frac{\operatorname{var}(Z_{ij})}{\gamma_Z}\right)\operatorname{cov}\left(Y_{ij},X_{ij}\right)-\operatorname{cov}\left(Z_{ij},X_{ij}\right)\operatorname{cov}\left(Z_{ij},Y_{ij}\right)}{\left(\frac{\operatorname{var}(Z_{ij})}{\gamma_Z}\right)\left(\frac{\operatorname{var}(X_{ij})}{\gamma_X}\right)-\operatorname{cov}\left(Z_{ij},X_{ij}\right)^2} \\
&= \frac{\gamma_X\operatorname{var}\left(Z_{ij}\right)\operatorname{cov}\left(Y_{ij},X_{ij}\right)-\gamma_X\gamma_Z\operatorname{cov}\left(Z_{ij},X_{ij}\right)\operatorname{cov}\left(Z_{ij},Y_{ij}\right)}{\operatorname{var}\left(Z_{ij}\right)\operatorname{var}\left(X_{ij}\right)-\gamma_X\gamma_Z\operatorname{cov}\left(Z_{ij},X_{ij}\right)^2}
\end{aligned}$$

If $\operatorname{var}\left(C_{ij}\right)=\operatorname{var}\left(X_{ij}\right)=\operatorname{var}\left(Z_{ij}\right)=1$, then this simplifies to

$$\beta_{ME}=\frac{\gamma_X\operatorname{cov}\left(Y_{ij},X_{ij}\right)-\gamma_X\gamma_Z\operatorname{cov}\left(Z_{ij},X_{ij}\right)\operatorname{cov}\left(Z_{ij},Y_{ij}\right)}{1-\gamma_X\gamma_Z\operatorname{cov}\left(Z_{ij},X_{ij}\right)^2}$$

Proof of equation 16:

$$\beta_{W_{ME}} = \frac{\text{var}\left(Z^*_{ij} - \bar{Z}^*_i\right)\text{cov}\left(Y_{ij}, X^*_{ij} - \bar{X}^*_i\right) - \text{cov}\left(Z^*_{ij} - \bar{Z}^*_i, X^*_{ij} - \bar{X}^*_i\right)\text{cov}\left(Z^*_{ij} - \bar{Z}^*_i, Y_{ij}\right)}{\text{var}\left(Z^*_{ij} - \bar{Z}^*_i\right)\text{var}\left(X^*_{ij} - \bar{X}^*_i\right) - \text{cov}\left(Z^*_{ij} - \bar{Z}^*_i, X^*_{ij} - \bar{X}^*_i\right)^2}$$

$$= \frac{\text{var}\left(\frac{Z^*_{ij} - Z^*_{ij'}}{2}\right)\text{cov}\left(Y_{ij}, \frac{X^*_{ij} - X^*_{ij'}}{2}\right) - \text{cov}\left(\frac{Z^*_{ij} - Z^*_{ij'}}{2}, \frac{X^*_{ij} - X^*_{ij'}}{2}\right)\text{cov}\left(\frac{Z^*_{ij} - Z^*_{ij'}}{2}, Y_{ij}\right)}{\text{var}\left(\frac{Z^*_{ij} - Z^*_{ij'}}{2}\right)\text{var}\left(\frac{X^*_{ij} - X^*_{ij'}}{2}\right) - \text{cov}\left(\frac{Z^*_{ij} - Z^*_{ij'}}{2}, \frac{X^*_{ij} - X^*_{ij'}}{2}\right)^2}$$

$$= \frac{2\left[\text{var}\left(Z^*_{ij}\right) - \text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\text{cov}\left(Y_{ij}, X_{ij} - X_{ij'}\right) - 2\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)\text{cov}\left(Z_{ij} - Z_{ij'}, Y_{ij}\right)}{2\left[\text{var}\left(Z^*_{ij}\right) - \text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]2\left[\text{var}\left(X^*_{ij}\right) - \text{cov}\left(X_{ij}, X_{ij'}\right)\right] - 2\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)^2}$$

$$= \frac{2\left[\frac{\text{var}(Z_{ij})}{\gamma_Z} - \text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\text{cov}\left(Y_{ij}, X_{ij} - X_{ij'}\right) - 2\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)\text{cov}\left(Z_{ij} - Z_{ij'}, Y_{ij}\right)}{4\left[\frac{\text{var}(Z_{ij})}{\gamma_Z} - \text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\left[\frac{\text{var}(X_{ij})}{\gamma_X} - \text{cov}\left(X_{ij}, X_{ij'}\right)\right] - 2\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)^2}$$

$$= \frac{\begin{array}{r}2\left[\gamma_X\text{var}\left(Z_{ij}\right) - \gamma_X\gamma_Z\text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\gamma_X\gamma_Z\text{cov}\left(Y_{ij}, X_{ij} - X_{ij'}\right) \\ - 2\gamma_X\gamma_Z\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)\text{cov}\left(Z_{ij} - Z_{ij'}, Y_{ij}\right)\end{array}}{\begin{array}{r}4\left[\gamma_X\text{var}\left(Z_{ij}\right) - \gamma_X\gamma_Z\text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\left[\gamma_Z\text{var}\left(X_{ij}\right) - \gamma_X\gamma_Z\text{cov}\left(X_{ij}, X_{ij'}\right)\right] \\ - 2\gamma_X\gamma_Z\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)^2\end{array}}$$

If $\text{var}\left(C_{ij}\right) = \text{var}\left(X_{ij}\right) = \text{var}\left(Z_{ij}\right) = 1$, then this simplifies to

$$\beta_{W_{ME}} = \frac{\begin{array}{r}2\left[\gamma_X - \gamma_X\gamma_Z\text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\gamma_X\gamma_Z\text{cov}\left(Y_{ij}, X_{ij} - X_{ij'}\right) \\ - 2\gamma_X\gamma_Z\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)\text{cov}\left(Z_{ij} - Z_{ij'}, Y_{ij}\right)\end{array}}{4\left[\gamma_X - \gamma_X\gamma_Z\text{cov}\left(Z_{ij}, Z_{ij'}\right)\right]\left[\gamma_Z - \gamma_X\gamma_Z\text{cov}\left(X_{ij}, X_{ij'}\right)\right] - 2\gamma_X\gamma_Z\text{cov}\left(Z_{ij} - Z_{ij'}, X_{ij} - X_{ij'}\right)^2}$$

R code to replicate findings (Study 2):

```
## Estimate byx with confounding
##
##  Arguments
##    byx, effect of x on y
##    byc, effect of c on y
##    bxc, effect of c on x
##    bzc, effect of c on z
##    rho.X, twin correlation in X
##    rho.C, twin correlation in C
##    rho.Z, twin correlation in Z
##    n, number of twin pairs per replication
##    reps, number of replications
##
##  Returns crude = (y ~ x)
##          crude.cov = (y ~ x + c)
##          within = (y ~ (x - x.mean))
##          within.cov = (y ~ (x - x.mean) + (z - z.mean))
##          within.cov2 = (y ~ (x - x.mean) + z)

sim.confounding.lm = function(byx, byc, bxc, bzc, rho.X, rho.C, rho.Z, n, reps)
{
  # Define rho.Y
  rho.Y = 0.2

  # Error variances
  var.epsY = 1 - (byx ^ 2 + byc ^ 2 + 2 * byx * byc * bxc)
  var.epsX = 1 - bxc ^ 2
  var.epsZ = 1 - bzc ^ 2
  var.C = 1.0

  # Correlations (sibling)
  rho.epsY = (rho.Y - byx ^ 2 * rho.X - byc ^ 2 * rho.C - 2 * byx * byc *
              bxc * rho.C) / var.epsY
  rho.epsX = (rho.X - bxc ^ 2 * rho.C) / var.epsX
  rho.epsZ = (rho.Z - bzc ^ 2 * rho.C) / var.epsZ

  # Covariances (sibling)
  cov.epsY = var.epsY * rho.epsY
  cov.epsX = var.epsX * rho.epsX
  cov.epsZ = var.epsZ * rho.epsZ
  cov.C = rho.C
```

```
# Error covariance matrix
A = matrix(c(var.epsY,0,0,0,
             0,var.epsX,0,0,
             0,0,var.epsZ,0,
             0,0,0,1), nrow=4, byrow=T)

B = matrix(c(cov.epsY,0,0,0,
             0,cov.epsX,0,0,
             0,0,cov.epsZ,0,
             0,0,0,cov.C), nrow=4, byrow=T)

M = rbind(cbind(A,B),
          cbind(B,A))

# Cholesky decomposition
# Prints NA for all estimates when M matrix is singular
if (is.na(tryCatch(chol(M), error=function(err) NA))){
  crude.est = NA
  crude.cov.est = NA
  within.est = NA
  within.cov.est = NA
  within.cov.est2 = NA
} else {
  L = chol(M)
  nvars = dim(L)[1]

  # Random variables that follow an M error covariance matrix
  r = t(L) %*% matrix(rnorm(nvars*n), nrow=nvars, ncol=n)
  r = t(r)
  rdata = as.data.frame(r)
  names(rdata) = c('epsY1','epsX1','epsZ1','epsC1','epsY2','epsX2','epsZ2','epsC2')
  round(cov(rdata), 2)

  # Generate Y,X,Z,C for each member of a twin pair
  C1 = rdata$epsC1
  C2 = rdata$epsC2

  Z1 = bzc*C1 + rdata$epsZ1
  Z2 = bzc*C2 + rdata$epsZ2

  X1 = bxc*C1 + rdata$epsX1
  X2 = bxc*C2 + rdata$epsX2
```

```
  Y1 = byx*X1 + byc*C1 + rdata$epsY1
  Y2 = byx*X2 + byc*C2 + rdata$epsY2

  # Set up data
  id <- c(1:n, 1:n)
  y <- c(Y1,Y2)
  x <- c(X1,X2)
  c <- c(C1,C2)
  z <- c(Z1,Z2)
  mx = with(data.frame(cbind(X1,X2)), (X1+X2)/2)
  x.mean <- c(mx,mx)
  mz = with(data.frame(cbind(Z1,Z2)), (Z1+Z2)/2)
  z.mean <- c(mz,mz)
  dat <- data.frame(cbind(id, y, x, c, z, x.mean, z.mean))

  # The crude estimate w/o covariate
  crude = summary(tryCatch(lm(y ~ x, data=dat), error=function(err) NA))
  crude.est = crude$coef[2,1]

  # The within-pair estimate w/o covariate
  within = summary(tryCatch(lm(y ~ I(x - x.mean), data=dat), error=function(err) NA))
  within.est = within$coef[2,1]

  # The crude estimate w/covariate
  crude.cov = summary(tryCatch(lm(y ~ x + z, data=dat), error=function(err) NA))
  crude.cov.est = crude.cov$coef[2,1]

  # The within-pair estimate w/covariate
  within.cov = summary(tryCatch(lm(y ~ I(x - x.mean) + I(z - z.mean),
                          data=dat), error=function(err) NA))
  within.cov.est = within.cov$coef[2,1]
  within.cov2 = summary(tryCatch(lm(y ~ I(x - x.mean) + z, data=dat),
                        error=function(err) NA))
  within.cov.est2 = within.cov2$coef[2,1]
 }
 return(list(crude=crude.est, crude.cov=crude.cov.est,
          within=within.est,  within.cov=within.cov.est,
          within.cov2=within.cov.est2))
}
```