2022

Detailing the Genetic and Environmental Influences Shared between Conventional and Electronic Cigarette Use Across Measures of Initiation and Past 12-Month Use

James Clifford
Virginia Commonwealth University

Detailing the Genetic and Environmental Influences Shared Between Conventional and Electronic Cigarette Use Across Measures of Initiation and Past 12-Month Use

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

James Samuel Clifford

Bachelor of Science, Psychology, Virginia Tech, 2005
Master of Science, Experimental Psychology, Radford University, 2007

Director:

Elizabeth C. Prom-Wormley, MPH PhD
Assistant Professor
Division of Epidemiology
Department of Family Medicine and Population Health

Virginia Commonwealth University
Richmond, VA
April 2022

Dissertation Committee:

Elizabeth Prom-Wormley, MPH PhD (Committee Chair)
Assistant Professor
Division of Epidemiology
Department of Family Medicine and Population Health
Virginia Commonwealth University, Richmond, VA

Juan Lu, MPH MD PhD
Associate Professor
Division of Epidemiology
Department of Family Medicine and Population Health
Virginia Commonwealth University, Richmond, VA

Silviu Bacanu, PhD
Associate Professor
Department of Psychiatry
Virginia Commonwealth University, Richmond, VA

Alexis Edwards, PhD
Associate Professor
Department of Psychiatry
Virginia Commonwealth University, Richmond, VA

Alison Breland, PhD
Assistant Professor
Department of Psychology
Virginia Commonwealth University, Richmond, VA

ACKNOWLEDGEMENTS

I would like to thank my exceptional Advisor, Dr Elizabeth Prom-Wormley for giving me this opportunity. It has been rough at times, it has been excellent at times, but I can't say thank you enough for teaching me, tutoring me, and believing in me even when I didn't believe in myself. I appreciate all that you've done for me and know that you have been, and are, the best role model that I could have ever imagined.

I'd also like to thank my committee members: Dr Juan Lu, Dr Alexis Edwards, Dr Alison Breland, and Dr Silviu Bacanu for all your dedication, patience, and tutelage. Know that every time you read something and gave me comments, I took them to heart and am happy that you have given me your time and energy over the last five years.

I thank everyone in the Division of Epidemiology, from the faculty members who taught classes to the staff who have assisted me in every way. Thank you!

To the students from Epidemiology and beyond, I appreciate you all. Thank you for always making it fun to learn with you. A big thank you to Dr Courtney Blondino who always pushed me to go a little further when I thought I was done. Mariam Sankoh who took up that torch after Courtney graduated, you taught me so much about perseverance and I appreciate our writing sessions.

I'd also like to thank individuals from other places: Dr Hermine Maes and Dr Roseann Peterson for looking over work and making me uncomfortable in meetings. That made me learn all that much harder. Dr Elizabeth Do for always being there when I needed something, even to talk about twin models. Dr Jeffrey Aspelmeier from Radford University, my first mentor; thank you. Drs George Kenna, Bob Swift, and Peter Monti from Brown University, along with Dr Lorenzo Leggio now at NIH, thank you for giving me a chance.

Dr Megan Cooke, who always has a smile and is one of the warmest colleagues I could ever hope for. It's been a long time coming, but I have arrived and I thank you for your help in getting here.

My family who have been supportive of their "professional student" for many years, thank you for sticking it out with me.

Ana Gordon, who has helped prop me up during my studies, I couldn't have done it without your love and support, thank you. Also to Hank, Lucille, and Pierce, thanks for the pets and listening to me rant during the writing process.

Finally, to my dog Scully. She has seen more ups and downs than anyone else and has never faltered in her support. You are truly my best friend, and I thank you.

There are too many people to thank to be listed, so let me close by saying that I am sorry if I missed you, but know that I am grateful.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

GLOSSARY OF ABBREVIATIONS

A – Additive Genetic effects
ALSPAC – Avon Longitudinal Study of Parents And Children
AUC – Area Under the Curve
C – Shared Environmental effects
CI – Confidence Interval
CIG – Conventional Cigarette
DAVID - The Database for Annotation, Visualization, and Integrated Discovery
E – Unique Environmental effects
ECIG – Electronic Cigarette
G4G – Genes for Good
GPS – Genome-Wide Polygenic Score
GSCAN – GWAS and Sequencing Consortium of Alcohol and Nicotine
GWAS – Genome Wide Association Study
GxE – Gene by Environment Interaction
HWE – Hardy-Weinberg Equilibrium
ISC - International Schizophrenia Consortium
LD – Linkage Disequilibrium
MAF – Minor Allele Frequency
MGS - Molecular Genetics of Schizophrenia
ND – Nicotine Dependence
OR – Odds Ratio
PC – Principal Component
PCA – Principal Components Analysis
PRS-CS – Polygenic Risk Score with Continuous Shrinkage
QC – Quality Control
ROC – Receiver Operating Characteristic
S4S – Spit for Science
SI – Smoking Initiation
SIA – Self Identified Ancestry
SNP – Single Nucleotide Polymorphism
TAG - Tobacco Genetics Consortium

ABSTRACT

Detailing the Genetic and Environmental Influences Shared Between Conventional and Electronic Cigarette Use Across Measures of Initiation and Past 12-Month Use

By James Samuel Clifford, MS, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University 2021

Director: Elizabeth Prom-Wormley, MPH, Ph.D.
Assistant Professor
Division of Epidemiology
Department of Family Medicine and Population Health

Introduction. Tobacco use continues to be a public health crisis with nearly 500,000 Americans suffering premature mortality directly attributable to tobacco use in 2014. Tobacco use, particularly among those who are nicotine dependent, has been associated with a host of negative health outcomes such as various cancers and cardiovascular system deficits. New research and development efforts have created new nicotine delivery systems whose health consequences are not yet fully understood such as electronic cigarettes. It is possible there are shared genetic and environmental factors that influence an individual's liability to initiate cigarette or electronic cigarette use, as both systems are designed to deliver nicotine.

Objective. The purpose of this study is to detail the shared genetic and environmental liability toward electronic and conventional cigarette initiation, or current use, and how to best measure these concepts to ensure consistency and reliability of results.

Methods. Four studies were used to help resolve the genetic and environmental influences that underlie cigarette and electronic cigarette initiation. The first study

(Adolescents and Young Adult Twin Study) to estimate the degree to which genetic and environmental factors for ECIG and CIG initiation were shared. Chapter 4 examined specific variants in the form of genome-wide association analysis (Genes for Good). Chapter 4 also addressed overlap via the use of genome-wide polygenic scores to quantify the degree of molecular overlap between these phenotypes (GWAS and Sequencing Consortia of Alcohol and Nicotine). The third study quantified a known environmental exposure for both CIG and ECIG use while also probing a potential environmental moderator (Population Assessment of Tobacco and Health). Meanwhile, the fourth study examines how genetically informative samples have measured cigarette use and shows the heterogeneity of results as a function of measure used. Further, this fourth study offers advice for future studies of electronic cigarettes and how best to quantify electronic cigarette use.

Results. The first study detected significant contributions of shared genetic and environmental factors shared between CIG and ECIG initiation. The twin study suggested there was significant overlap between cigarette and electronic cigarette initiation in regards to additive genetic variance. The scoping review of tobacco use measures in genetically informative samples reported that how individual studies measured different aspects of tobacco use lead to different genome-wide significant results. Aggregating genetic effects by biological function lead to greater consistency of results. Replication of GWAS results at a gene or biological function level rather than replicating individual SNPs lead to more consistent results. The third study did not detect any genome-wide significant association for ECIG initiation in self-identified white participants. Genome-wide polygenic scores reported no association between

conventional cigarette initiation and electronic cigarette initiation. Statistical evidence of a weak interaction between electronic cigarette coupon receipt, income level, and conventional cigarette use was reported.

Conclusions. These analyses showed there is genetic and environmental overlap between CIG and ECIG initiation.

CHAPTER 2: USE OF A TWIN STUDY TO QUANTIFY THE GENETIC AND ENVIRONMENTAL INFLUENCES SHARED BETWEEN ELECTRONIC CIGARETTE AND CIGARETTE INITIATION

Introduction. The use of electronic cigarettes (ECIG) continues to rise in the United States, especially among adolescents and young adults. Therefore, it is necessary to better understand factors associated with ECIG initiation. However, it is unclear whether genetic and environmental factors influence the initiation of ECIGs. Further, the degree to which genetic and environmental factors influences are shared between ECIG initiation and conventional cigarette (CIG) initiation is unknown.

Methods. A sample of young adult twins ages 15-20 (N = 858 individuals; 421 complete twin pairs) was used to estimate the genetic and environmental influences on the liability of initiation unique to ECIG and CIG as well as the degree to which these factors are shared between the two. Approximately 20% of participants ever initiated ECIG use and 19% initiated CIG. 11% of the sample had initiated dual use of both products.

Results. The combined contributions of additive genetic and shared environmental influences were non-significant, while unique environmental influences were significant, for CIG (A_{CC} = 0.19 [95% CI = 0-0.79], p = 0.57; C_{CC} = 0.42 [95% CI = 0-0.70], p = 0.13; E_{CC} = 0.39 [95% CI = 0.18-0.57], p < 0.001) and ECIG (A_{EC} = 0.25 [95% CI = 0-0.83, p = 0.44; C_{EC} = 0.42 [95% CI = 0-0.73], p = 0.12; E_{EC} = 0.32 [95% CI = 0.14-0.56], p < 0.001). There was a significant phenotypic correlation between ECIG and CIG initiation (r = 0.72, p < 0.001). This correlation was due to significant contributions due to unique environmental factors shared between ECIG and CIG initiation (r_e = 0.87, p = 0.01, 95%

CI = 0.50-0.99). However, genetic factors shared between ECIG and CIG initiation were not statistically significant.

Conclusions. These results suggest that both genetic and environmental influences are important for ECIG initiation among adolescents and young adults.

CHAPTER 3: TOBACCO USE MEASURES IN GENETICALLY INFORMATIVE SAMPLES: HETEROGENITY AND RECOMMENDATIONS FOR FUTURE STUDIES

Introduction. Many genetically-informative studies, (e.g., twin study designs and genome-wide association studies [GWAS}), have been conducted to examine a variety of phenotypes. Though twin studies consistently report significant additive genetic effects for tobacco use, GWAS have been plagued with inconsistent results. This may be due in part to the heterogeneity of measures for tobacco use. A scoping review was evaluated how tobacco use has been measured in previously published studies using genetically-informative samples.

Methods. Four databases (PubMed, EMBASE, PsychINFO, and CINAHL) were searched with terms from three concepts (tobacco use, genetically-informative designs, tobacco measurement), producing 310 articles. Of those, 87 directly used a twin design or GWAS (or a variation of GWAS) to examine tobacco use. Articles were then classified as one of five tobacco use classifications: initiation, quantitative measures of smoking, nicotine metabolism, nicotine dependence, or smoking cessation. Biological relevance of significant GWAS results was assessed and summarized using the Database for Annotation, Visualization, and Integrated Discovery (DAVID).

Results. Variants within genes responsible for nicotinic acetylcholine receptor function (e.g., *CHRNA3, CHRNB4, CHRNA5*) as well as nicotine metabolism (*CYP2A6*) were consistently associated with most measures of tobacco use.

Conclusions. Although GWAS results were highly variable, gene-level reporting of results informed by biological function produced greater consistency and improved interpretation across studies.

CHAPTER 4: ELECTRONIC CIGARETTE GENOME-WIDE ASSOCIATION AND POLYGENIC SCORES AMONG SELF-IDENTIFIED WHITE PARTICIPANTS: TEST OF OVERLAPPING GENETIC INFLUENCES WITH CONVENTIONAL CIGARETTE INITIATION

Introduction. Three studies have used genome-wide association data to produce polygenic scores (GPS) or CIG initiation to test its relationship with ECIG use. However, these studies have mainly focused on young adults (age 18-25 years old). Nevertheless, ECIG initiation occurs across adulthood, but the role of genetic factors associated with this outcome remains inconsistent. Some of the genetic variants associated with CIG initiation are also expected to influence ECIG initiation since both products contain nicotine. Tests for genetic association of ECIG initiation that take advantage of the genetic factors associated with CIG initiation may help to clarify the etiology of nicotine dependence which begins with initiation of products containing nicotine. In particular, specific genetic variants associated with ECIG initiation in adults have not yet been identified. Similarly, genetic variants contributing to ECIG and CIG initiation across adulthood have also not been identified although this information is needed to understand the etiology of nicotine dependence.

Methods. Data from the Genes for Good (G4G) study, a population-based sample of American adults aged 18-93 (N =15,881) were used. Two GWAS were conducted on lifetime CIG and ECIG initiation to test for genetic associations across all loci in the genome. Additionally, a GPS for CIG initiation was generated and used to test the degree to which there was polygenic overlap between CIG and ECIG initiation.

Results. No genome-wide significant associations were detected for ECIG or CIG lifetime initiation. However, there were four SNPs for ECIG lifetime initiation which approached genome-wide significance (locations summarized as chromosome number: base pair number- 13:32403784, OR = 0.62, p = 7.49 x 10^{-7}; 2:115364757, OR = 1.28, p = 5.19 x10^{-7}; 15:49010393, OR = 0.44, p = 8.01 x 10^{-7}; 6:33902823, OR = 0.79, p = 1.75 x 10^{-7}). No significant polygenic association was detected between polygenic scores calculated for CIG lifetime initiation in GSCAN with ECIG lifetime initiation as measured in G4G.

Conclusions. This first-ever GWAS of ECIG lifetime initiation identified two SNPs in novel genes for tobacco use. There was no evidence for overlapping genetic factors for CIG and ECIG lifetime initiation. Replication is strongly encouraged because these results have low power to detect statistically significant genetic effects.

CHAPTER 5: THE EFFECT OF COUPON RECEIPT ON THE RELATIONSHIP BETWEEN INCOME AND PAST 12-MONTH ELECTRONIC AND CONVENTIONAL CIGARETTE USE IN ADULTS

Introduction**.** Lower household income levels have been associated with electronic cigarette (ECIG) or conventional cigarette (CIG) use. However, it is unclear whether this relationship changes with the receipt of ECIG or CIG coupons in adults.

Methods**.** Data from Wave 3 of the Population Assessment of Tobacco and Health (N = 28,148) was used to test the association between tobacco use and income in adults. Associations were tested using multinomial logistic regression.

Results**.** Income level was significantly associated with CIG-exclusive and dual use, but not ECIG-exclusive use. Receipt of ECIG coupons was associated with past 12-month ECIG use (aOR = 1.40; 95% CI = 1.05-1.88), CIG use (aOR = 5.69; 95% CI = 5.08-6.38), and dual use (aOR = 7.61; 95% CI = 6.75-8.58). Receipt of CIG coupons was an independent risk factor for ECIG-exclusive (aOR = 2.32, 95% CI = 1.74-3.10) or dual use (aOR = 2.62, 95% CI = 2.10-3.28), but protective against CIG-exclusive use (aOR = 0.74, 95% CI = 0.59-0.92). There was evidence of weak moderation between receipt of ECIG coupons and CIG-exclusive use. Individuals with household incomes between $50,000 and $99,999 (aOR = 2.51; 95% CI = 1.50-4.16) were more likely to be CIG users if they received ECIG coupons relative to those who do not receive ECIG coupons.

Conclusions**.** Individuals with lower levels of income may be at greater risk for dual use of ECIG and CIG as well as CIG-exclusive use. Additionally, receipt of CIG and ECIG coupons appears to be an independent risk factor for past 12-month use of tobacco.

Chapter 1: General introduction

The Importance of the Problem

Nicotine consumption through electronic cigarette (ECIG) use is an ongoing public health issue in adults that contributes to morbidity and premature mortality. In 2014, there were 500,000 American adults ages 18 and over whose deaths were attributed to nicotine use [1]. ECIG use is also associated with mortality and morbidity [2]. From 2015 to 2017, 2,035 emergency room visits were due to explosions or burns from ECIGs [3]. Further, 68 deaths have been attributed to the e-cigarette, or vaping, product use associated lung injury (EVALI) as of 2020 [4]. ECIGs contain many chemicals including propylene glycol (a respiratory irritant), volatile organic compounds (VOC, associated with greater cancer risk) [5], and polycyclic aromatic hydrocarbons (PAH, mutagenic and carcinogenic properties) [6]. Compared to CIG users, ECIG-exclusive users showed lower levels of the 5 major classes of tobacco product constituents (tobacco-specific nitrosamines, metals, PAHs, VOCs, and nicotine) though still more than non-smokers [7]. Nevertheless, epidemiologic studies of ECIGs have reported that the odds of chronic obstructive pulmonary disease (COPD) and asthma increase with ECIG use [8]. ECIGs also contain nicotine, which has been associated with many negative health outcomes cardiovascular disease [9], psychiatric disorder [10], and substance use disorder [11]. Nicotine may also be associated with numerous types of cancers (e.g., gastrointestinal, pancreatic, breast, and lung) [12–15]. Further, nicotine is an addictive substance, which encourages continued use of ECIGs. [7]

What are Electronic Cigarettes and How Do They Work?

ECIGs are a nicotine delivery system that aerosolizes an e-liquid cartridge containing nicotine and a flavor (such as mint, tobacco, or candy) using a power source and an electronic heating element [16]. An ECIG is activated by pulling air through the mouthpiece or pressing an activation button. Upon activation, a battery engages the electronic heating element which aerosolizes the e-liquid. Therefore, ECIGs were previously marketed and are sometimes considered by conventional cigarette (CIG) users as "healthier" products because additional chemicals such as tar are not produced or are present in lower concentrations compared to CIGs [17,18].

ECIGs have undergone several alterations since they were first sold in the United States in 2007 (Figure 1.1). First generation ECIGs were designed to have the look and feel of a conventional cigarette, leading to the term "cig-a-like". Most of these first-generation devices were designed to be used once and then discarded [19]. Second generation devices allowed users to refill the tanks for additional e-liquid, leading to reusable ECIGs. Third generation ECIGs were designed to be the most accessible and allow the user to have the greatest opportunity for personalization and customization. Users can vary the voltage, wattage, and power of the device along with additional peripheral enhancements such as being able to charge a cell phone with the device [19]. The most recent version of ECIG devices, the fourth generation, are designed to use pods (i.e., cartridges containing e-liquid) and dock with the device (i.e., adjoining the e-liquid containing pod with the mouthpiece).

ECIG devices are broken down into two general categories of devices, open and closed systems (Figure 1.1) [20]. Open system devices allow the user to reload the device

with their choice of e-liquid, which may contain different levels of nicotine (including having no nicotine) and flavors. Importantly, open systems allow for additional user modifications. Batteries can also be modified creating more aerosolization per pull (i.e., low-ohm device) [21]. These modifications make it harder to standardize the amount of nicotine consumed in open systems. Therefore, open system ECIG use may lead to greater nicotine consumption and an increased likelihood of nicotine dependence [22]. Closed systems cannot generally be modified by the user. Generally, these are single use products designed to be discarded after use and not refilled with new liquids. Single use, preloaded, "cig-a-likes" are examples of closed systems (Figure 1.1). Many first-generation devices are closed systems, though closed systems are not limited to first-generation, having gained in popularity in recent years. The most popular closed system ECIG to date in the United States is JUUL (Figure 1.1). JUUL administers nicotine through "pods", small cartridges that contain standardized amounts of nicotine and a heating coil that cannot be replaced [23]. JUUL use has increased dramatically, doubling from 2018 to 2019, and several reports of harmful use in adolescents and young adults were reported at that time [24]. Further, some reports indicate the average JUUL user was exposed to much larger amounts of nicotine compared to CIG users [25], leading the CDC and FDA to carefully review the product in 2019 [26].

Figure 1.1. Common ECIG Devices Showing the Evolution from First Generation (Cig-A-Likes) to Fourth Generation Cartridge Devices (JUUL).

The Etiology of Nicotine Dependence for Combustible Cigarette Use Motivates Study of Initiation and Dependence due to Electronic Cigarette Use

The Diagnostic and Statistical Manual, 4th Edition (DSM-IV) defines nicotine dependence (ND) as a maladaptive pattern of nicotine use that leads to clinically relevant impairment or distress. There were four criteria that contributed to ND: 1) tolerance, 2) withdrawal, 3) consuming a larger amount of nicotine than originally planned, and 4) difficulty in quitting use [27]. The DSM-5 was released in 2013 and renamed ND as tobacco use disorder (TUD) [28]. The DSM-5 definition of TUD, while significantly overlapping with the DSM-IV definition of ND, removed the criteria of "difficulty in quitting" while adding an item related to craving [11]. As there is significant overlap between these definitions, and most research to date has been conducted using the DSM-IV definition, this dissertation will continue to use the term "nicotine dependence" rather than the updated "tobacco use disorder".

Nicotine dependence is considered an acquired disease of the brain since nicotine can cross the blood brain barrier [29]. Further, nicotine can activate several neurobiological pathways that function across several brain structures once in the brain, including: reward/saliency, inhibitory control and executive function, and motivation (Volkow 2014). Nicotine activates the dopaminergic pathway which has been reported to be associated with craving as well as activates the reward pathway associated with dopamine leading to feelings of euphoria. Rewarding feelings created via activation of this pathway create a feedback loop as the use of nicotine increases to create pleasurable feelings, which leads to increased use of nicotine.

Several additional neurotransmitters contribute to nicotine dependence. Nicotine activates neurotransmitters that influence arousal, attention, and motivations, including acetylcholine [30] and norepinephrine [31] which explains the motivating aspects of nicotine use. Neurotransmitters that engage the learning and memory systems are also influenced by nicotine. These neurotransmitters including serotonin [32], glutamate [33], and gamma-aminobutyric acid (GABA) [34], influence ND by helping the body remember the rewarding feelings from nicotine use.

The development of ND has generally been studied across several stages of CIG use behaviors (Figure 1.2). All stages of smoking behaviors and the development of ND require an individual to engage in *smoking initiation* before transitioning to other stages. After initiation, individuals may transition into *regular smoking* wherein they smoke consistently. Regular smoking has many definitions with some definitions being tied to the number of cigarettes smoked in one's lifetime, daily cigarette use, or most commonly asking participants to self-identify as a regular or current smoker [35,36]. Some regular smokers become *nicotine dependent*, as defined above and indicated via reliable and validated measures of ND, from CIG use due to consistent exposure to nicotine. Finally, individuals may transition to *smoking cessation*, a process whereby a smoker transitions into becoming a non-smoker. Successful cessation is defined as the abstinence of further tobacco use. As cessation is a process, individuals can relapse and return to regular smoking or nicotine dependence. Cessation is typically defined in research studies via self-identified smoking status (i.e., current or former smoker) [37]. A time frame is sometimes attached to smoking status (e.g., "Have you smoked a cigarette in the past 6 months?") [38]. Individuals who achieve cessation may relapse into

CIG use after abstaining for a period of time and then begin the cessation process again. Common measures of these nicotine phenotypes are presented in Table 1.1. Each phenotype has multiple measures which are used to assess these conceptual behaviors.

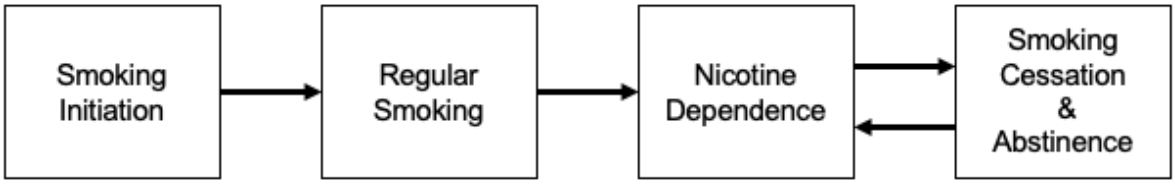


Figure 1.2. Stages of Commonly Measured Smoking Behaviors Corresponding to the Development of Nicotine Dependence and Smoking Abstinence

Nicotine dependence is common in American adults. A study of the NESARC-III (National Epidemiologic Survey on Alcohol and Related Conditions, 3rd Wave; data collected 2012-2013) estimated that 14% of Americans were currently nicotine dependent [39]. NESARC-III did not include questions on ECIGs, but rather exposure to nicotine via combustible (CIG or cigar use) or oral administration (e.g., snuff, dip) only. To date, most work detailing the progression of nicotine use and dependence has been conducted on CIG use behaviors [40–45]. However, the prevalence of CIG use is at its lowest point among American adults and is increasing for ECIGs. Between 2010 and 2013, the prevalence of adult CIG initiation decreased from 8.9% to 5.4% and ECIG initiation rates increased from 1.8% to 13% [46]. Rates of ND have been stable over the same time frame. These trends suggest that any population-level reduction in CIG use may reflect a shift to ECIG use.

Table 1.1. Common Measures of Cigarette Use Behaviors

Conceptual Measure	Operational Measures
Initiation	• "Have you ever smoked cigarettes?" (Rhee, 2003) • "At what age did you start smoking?" (Heath, 1999)
Regular or Current Smoking	• "Do you smoke cigarettes now?" (True 1997 • "During the last 30 days, how many (if any) have you used electronic cigarettes?" (McCabe, 2017) • "During the past 30 days, on how many days did you use e-cigarettes?" (Selya, 2017)
Nicotine Dependence	• Fagerström Test for Nicotine Dependence (FTND) • Nicotine Dependence Syndrome Scale (NDSS) • DSM-III/DSM-IV Nicotine Dependence Symptom Count or Diagnosis
Smoking Cessation & Abstinence	• Current smoking status (Current vs former) • eCO (expired carbon monoxide) verified • Abstinence for a period of time (e.g., 6 months)

ECIG use is also associated ND, although there are currently few studies that have established this conclusion in comparison with CIG use [47]. The prevalence of ND due specifically to ECIG use in adults is currently unclear in the US. However, a study of twelfth-grade students in Los Angeles, California reported that 16.7% of ECIG-exclusive users reported some level of ND dependence arising from ECIG use [48]. Additionally, a recent study of ND using a measure of "time to first cigarette" reported that ND in participants who use ECIGs was less than that of CIG users (i.e., participants could wait longer to use their ECIG than those who used CIGs). Past 30-day JUUL use was associated with at least some level of dependence in a sample of college students [22]. The mean score for nicotine dependence, using the Penn State Electronic Cigarette Dependence Index (PSECDI), was 7.8. Individuals who score below a 4 on the PSECDI are not considered to be nicotine dependent. Most individuals in this study had at least

low levels (i.e., scores greater than 4) of dependence. This study also reported that the level of dependence was unrelated to other tobacco product use, such as CIGs. Further, replacing CIG use with ECIG did not change the overall dependence level of the user among dual users [49]. A systematic review reported increased levels of ND in adolescent participants who used fourth generation pod-based ECIGs. While these studies report lower levels of ND are associated with ECIG use compared to CIG use, other work has suggested ECIGs are as addictive, if not more so, as CIGs [50]. These results highlight the inconsistency of research results for ND as arises from ECIG use.

The most common ECIG use measurement that parallel those of CIG use and relate to the development of ND are summarized in Table 1.2. ECIG initiation is measured similarly to CIG initiation, by asking participants if they have ever used an ECIG. One notable difference is in asking individuals if they own an ECIG, as users must purchase a device to use (though this method does not capture individuals who may have used a peer's device). Current ECIG use is also measured similarly to CIG use. Most often this is accomplished by asking participants to report if they had used ECIGs in the past 30 days. However, ECIGs present novel challenges compared to CIG measurement. CIGs are unable to be user modified so there is less variability, whereas ECIG measurement should consider characteristics of the device as well as nicotine concentration of e-liquid used. Utilizing standardized ECIGs (i.e., giving research participants the same device and e-liquid) will reduce variability due to user modifications, particularly in randomized controlled trials. Further, puffing behaviors may be investigated with attachable mouthpieces for ECIGs which measure puff duration, volume, puff count, flow rate, and inter-puff intervals. These measurements require

additional hardware and software to gather data. While measurement of ND due to ECIG use is still being investigated, several measures (e.g., PSECDI, e-FTND) have been validated.

Table 1.2. Common Measures of Electronic Cigarette Use Behaviors

Conceptual Measure	Operational Measures
Lifetime ECIG Initiation	• "Have you ever used an electronic cigarette, even one or two puffs?" (O'Loughlin 2017) • "Have you ever tried one of the following substances, devices, etc.?" followed by a list, including ECIG (Hammond 2017)
Regular or Current ECIG Use	• "In the last 30 days, did you use any of the following? (Mark all that apply)" (Hammond, 2017) • "During the last 30 days, how many (if any) have you used electronic cigarettes?" (McCabe, 2017) • "During the past 30 days, on how many days did you use e-cigarettes?" (Selya, 2017)
Nicotine Dependence	• Penn State Electronic Cigarette Dependency Index (PSECDI) • Electronic Wisconsin Inventory of Smoking Dependence Motives (eWISDM) • Electronic Fagerström Test for Nicotine Dependence (eFTND) • E-Cigarette Dependence Scale (EDS) (Morean, 2018)

The Epidemiology of Adult ECIG Initiation Emphasizes the Need for Study with CIG Initiation

This dissertation focuses on the study of ECIG and CIG initiation for several reasons. First, ECIG initiation often co-occurs with CIG initiation in adults. In 2017, 20.3% of US adults (aged 18+) initiated ECIG use [51]. Of those individuals, 21.6% reported currently using ECIGs at least some days [52]. Daily ECIG use has increased between 2016 (6.8%) and 2017 (7.5%) among adults in the United States [52]. Further, the lifetime prevalence of ECIG initiation among young adults (18-24) doubled between

2013 and 2014 from roughly 7% to nearly 14%. ECIG initiation is associated with subsequent CIG initiation among never smokers (OR = 3.62, 95%CI = 2.42-5.41) [46,53], and the reverse is also true: previous CIG initiation is associated with subsequent ECIG initiation (OR = 3.54, 95% CI = 1.68-7.45) [54].

Second, there are yet undetailed dynamic patterns related to the dual use of ECIG and CIGs. These two products are commonly used together, and users may switch between the two products. Consequently, current studies of ECIG use necessitate the inclusion of CIG use and vice versa. Similarly, studies that have not modeled dual use accurately may have produced biased estimates, limiting the knowledge of studies that have examined these phenotypes. Further, it is unclear if ECIGs are used by individuals who would have otherwise remained tobacco-naïve.

Third, as summarized above there is substantial overlap in the associations between ND with ECIG and CIG initiation. This is likely due to an etiology that is shared between ECIG and CIG that begins with the initiation of these products and that may continue into the development of ND. Consequently, it is important to first detail the factors involved with ECIG initiation and to identify those factors that may overlap with CIG initiation to fully understand the etiology of ND resulting from ECIG use in the future. Understanding the genetic and environmental influences that are shared between CIG and ECIGs will ensure a greater understanding of why they are used together.

Environmental and Genetic Factors Associated with CIG Use May Also Influence ECIG Use

To date, genome-wide association studies report significant genetic associations between multiple CIG use behaviors with loci in several genes. One set of genes, the nicotinic acetylcholine receptor gene cluster (e.g., *CHRNA2, CHRNA5, CHRNB4*), are consistently associated with multiple phenotypes of smoking behavior including CIG initiation, cigarettes per day (CPD), current smoking, and ND [55]. These genes encode for receptor polypeptides that respond to acetylcholine, but for which nicotine is also an agonist. In addition, *CYP2A6* is often associated with quantitative measures of smoking (e.g., CPD) and is involved with the metabolism of nicotine via oxidation [56]. The serotonin transporter genes (e.g., *SLC03A1*) are also associated with ND. These genes are responsible for transporting serotonin from the synaptic cleft back into the presynaptic neuron [57]. Further, preliminary evidence from polygenic association studies of ECIG and CIG initiation in young adults (aged 18 to 25 years) suggests that similar genes impact ECIG and CIG initiation. Genome-wide polygenic scores (i.e., the aggregate molecular genetic effect from measured and imputed markers) associated with CIG initiation were also associated with ECIG initiation (OR = 1.24, 95% CI = 1.14-1.34, $p < 0.001$) [58]. Thus, identifying the genetic factors and biological pathways associated with ECIG initiation could guide future studies on ECIG behaviors including ND.

Several environmental factors have also been implicated in increasing the risk of CIG initiation. These include positive peer opinions towards smoking [59,60], income levels [61], low parental monitoring [62,63], and high exposure to tobacco advertising [64,65]. These factors have also been reported to increase the likelihood of ECIG initiation [66–68].

Economic factors are particularly important in the development of CIG and ECIG initiation. Economic factors such as income, educational attainment, and health insurance status may disproportionately impact individuals in society, causing more harm to individuals with fewer resources [69,70]. For instance, tobacco users with less income spend a larger portion of their resources on tobacco compared to people with more income. Consequently, the economic factors of household income and coupon use are particularly useful for understanding ECIG and CIG initiation.

Income has previously been associated with CIG use [46]. Previous research suggests that the association between ECIG and income is similar. ECIG initiation is more expensive than CIG initiation as one must purchase the device prior to initiating use, which may prevent some lower income individuals from starting to use the device. However, mediating or moderating factors may impact this relationship, such as coupon receipt or use to reduce the initial cost burden.

There are several avenues tobacco companies use to increase the usage of their products. Coupons are a known method to reduce the cost of cigarette use. Previous research indicates that individuals who receive coupons are more likely to start using CIGs and ECIGs [71]. Further studies have reported that possessing any type of promotional material (e.g., ball caps, t-shirts, or posters) for alternative tobacco products, such as ECIGs, is associated with increased odds of initiation of those products [72]. However, it remains unclear how coupon receipt of tobacco products impacts the initiation of different tobacco products (i.e., is receipt of CIG coupons associated with ECIG initiation and vice versa).

Knowledge Gaps

To date, there are two important gaps in knowledge that are necessary to reduce the impact of ECIGs on ND. First, *the relative degree to which genetic and environmental influences ECIG lifetime initiation is currently unknown*, particularly among adults. Preliminary research has shown there may be genetic overlap between CIG and ECIG initiation using genome-wide association data, but this work has only been performed in young adults (those aged 24 or younger). It remains unknown if these influences are also relevant for a sample of adults. Adults remain an understudied population in genetically-informed studies as well as epidemiological studies of ECIG or CIG initiation. Prior studies of ECIG or CIG initiation have only assessed age of initiation. Further, most studies of ECIG and CIG use/initiation focus on adolescent samples, perhaps due to the more malleable nature of this stage of development (i.e., environmental factors are more influential for younger ages compared to older ages when genetic factors are more influential). However, the rates of tobacco naïve adults CIG and ECIG initiation remain above 10% for those aged 18-21. In contrast, there has been a slight increase in the prevalence in initiation those between the ages of 20-29 [73] indicating that individuals will still initiate use beyond adolescence. Further, many adults who initiate ECIGs may do so to address and reduce CIG use. Consequently, factors that influence ECIG initiation in adulthood may likely be similar to those of CIG initiation.

Second, *few specific genetic and environmental factors influencing CIG and ECIG lifetime initiation have been identified*. This is in part due to: 1) inconsistency of genetic epidemiology studies for CIG use, 2) a lack of synthesis of results across these study designs, and 3) few genetic epidemiology studies of tobacco products beyond

CIGs. Operational measures of tobacco use vary between studies, which contributes to the inconsistency of results. Without a synthesis of the available results, it is unknown which specific genetic variants are relevant for CIG and ECIG initiation. Additionally, specific environmental factors also remain undetailed for ECIG and CIG initiation.

Using Multiple Study Designs to Detail the Shared Genetic and Environmental Factors for ECIG and CIG Lifetime Initiation

Table 1.3. Knowledge Gaps, Research Questions, and the Chapters Addressed in the Dissertation

Knowledge Gap 1: Quantify the relative contribution of genetic and environmental factors toward ECIG and CIG initiation	• Chapter 2: Are there shared latent genetic and environmental factors for CIG and ECIG initiation in a sample of twins? • Chapter 4: Are the genetic factors that are associated with CIG initiation also associated with ECIG initiation?
Knowledge Gap 2: Which specific genetic and environmental factors are associated with CIG and ECIG initiation	• Chapter 3: Are there overlapping results in genetically informed studies of tobacco use? Are there ways to synthesize results which will lead to more consistent results? • Chapter 5: To what degree are environmental factors (receipt of coupons and income) associated with CIG use also associated with ECIG and dual use?

This dissertation will address the aforementioned knowledge gaps by characterizing the relative contribution of genetic and environmental factors associated with lifetime ECIG initiation as well as those shared with lifetime CIG initiation using three different study designs. For this dissertation lifetime CIG and ECIG initiation is

defined as having ever used either product at any time throughout an individual's life. These studies utilized secondary data and therefore, operationalization of these conceptual variables may differ between original studies (i.e., ever use may actually have been assessed with, "Have you smoked at least 100 cigarettes in your lifetime?"). A summary of the knowledge gaps as well as specific research questions and the chapters where they are addressed are summarized in Table 1.3.

Chapter 2 uses a sample of adolescent and young adult twins (mean age = 19.2, SD = 1.3, age range = 17.6-22.4) to assess the degree to which genetic and environmental effects influence ECIG initiation and to what degree are these influences shared with CIG initiation. Chapter 4 expands on results from Chapter 2 by testing for genetic association with CIG and ECIG initiation across loci through the genome using genome-wide polygenic scores in a community-based sample of adults (ages 18-93). This chapter uses a genome-wide association study (GWAS) approach to study ECIG and CIG initiation and answer the following questions: 1) Are there any genetic loci that are associated with ECIG initiation? and 2) Do the genetic factors that contribute to CIG initiation also contribute to ECIG initiation in a sample of adults? A GWAS is a study design that utilizes genetic data from a genetic marker to test for a statistical association between the marker and a phenotype. A significant association suggests that the genotype co-occurs with the phenotype more often than expected by chance[41]. Chapter 5 also builds on results from Chapter 2 by estimating the degree to which environmental factors (e.g., income and coupon use) are associated with CIG and ECIG past 12-month use. Past 12-month use is a relevant phenotype to study due to its proximity to CIG initiation as shown in figure 1.2. Chapter 5 uses an epidemiological sample of

American adults (age: 18-99) to detail a shared environmental factor (income). Chapter 5 answers the questions: 1) are ECIG-exclusive and dual users similar to CIG-exclusive users in terms of income and tobacco use? and 2) do coupons, a known method for reducing the cost of tobacco use, moderate the relationship between income and tobacco use? Taken together, these results will begin to detail the etiology of CIG and ECIG use in developmental stages beyond adolescence.

To address the second knowledge gap, Chapter 3 uses a scoping review approach to reflect on the current state of measurement and conclusions on tobacco products for use in genetic epidemiology studies. A scoping review is a review designed to examine the body of literature. This aim will also address gaps in environmental influences toward ECIG initiation in Chapter 5. Accurately modeling the outcome variable will detail how this environment changes across products. A potential moderator will further characterize how this environmental influence changes across a second environmental factor. This knowledge is expected to make the production of knowledge for ECIG use more efficient in the future.

CHAPTER 2: USE OF A TWIN STUDY TO QUANTIFY THE GENETIC AND ENVIRONMENTAL INFLUENCES SHARED BETWEEN ELECTRONIC CIGARETTE AND CIGARETTE INITIATION[1]

INTRODUCTION

Electronic cigarette (ECIG) initiation is associated with conventional cigarette (CIG) initiation in adolescents and young adults [74–77]. A recent meta-analysis reported that individuals who engaged in any ECIG use were 3.5 times more likely to initiate CIG use compared to those who did not use ECIG [53]. This is a major public health concern as both CIG and ECIG use exposes individuals to nicotine, the addictive component of both smoke and ECIG aerosol, which may lead to nicotine dependence. In addition to the addictive nature of nicotine, tobacco use is also associated with several negative health outcomes such as cancer and cardiovascular impairments [78].

Epidemiological studies have identified several factors, such as peer group use, that are associated with both CIG initiation [79] and ECIG initiation [80]. However, the degree to which these factors are shared between ECIG and CIG initiation remains unresolved. If overlap exists in risk factors for CIG and ECIG initiation, similar public health messaging and interventions may influence the initiation of both products. Genetic epidemiological study designs have the potential to provide clarity on the influence of shared risk factors for CIG and ECIG initiation.

[1] 1 This chapter has been modified from the original manuscript published in *Nicotine and Tobacco Research*, DOI: 10.1093/ntr/ntaa201

Genetic Epidemiology of Electronic Cigarette and Conventional Cigarette Initiation

Twin Concordance Studies and Adoption Studies. Early studies of twins calculated twin concordance rates to quantify familial aggregation of smoking. Concordance rates are a measure of probability, asking the question if one twin starts smoking, what is the probability that the other twin begins using [81]? Early twin studies on smoking initiation (SI) reported higher concordance rates in monozygotic (MZ; identical) twin pairs compared to dizygotic (DZ; fraternal) pairs, which suggested that genetic influences play a role in SI [82]. However, the degree to which these factors influenced smoking was not able to be estimated.

Two adoption studies have examined smoking initiation [83,84]. Adoption studies use data from adoptive children and examine their similarity to biological parents versus adoptive parents. Another possible comparison is between adoptees and their biological or adoptive siblings. Children adopted away from biological parents still resembled their biological parents (r = 0.21) more closely than their adoptive parents (r = -0.02) in regard to smoking behaviors. While these designs are powerful to untangle genetic and shared environmental effects, a major limitation of these early studies was their small sample size [83]. Furthermore, it may be difficult to ascertain biological parents for participation. Therefore, while these study designs are powerful, it is unlikely they will have appropriate sample sizes to provide stable estimates.

Classical Twin Studies. Since the 1970s, the "classical twin study design" (CTD) has been used to estimate the magnitude of genetic and environmental effects on SI of CIG. The CTD uses data from monozygotic (MZ) and dizygotic (DZ) twin pairs to

partition the total variance of a phenotype into the proportion of the contribution due to genetic and environmental influences. MZ twins share 100% of their genes while DZ twins share, on average, 50% of their segregating genes. This design can be used to estimate additive genetic influences (A - effects of alleles at every contributing locus); shared environmental effects (C - influence of all the environmental effects shared by twin pairs); and unique environmental effects (E - influence of all the environmental effects not shared by members of twin pair, which make the twins less similar and includes measurement error) [85].

The CTD is subject to the following conceptual assumptions. First, twins are assumed to be subject to the equal environments assumptions (i.e., both twins experience the environment in the same manner) [86,87]. Second, the CTD assumes random mating of adults in the population. Random mating is defined as choosing a partner not based on any sort of identifying characteristic (such as political preference) that may also be due to genetic factors [81]. Third, the CTD assumes that there is no influence due to gene by environment interaction or gene-environment correlation [88].

To date, 16 twin studies have investigated the genetic and environmental contributions toward smoking initiation of CIG.

Lifetime Ever Smoking Initiation. Twin studies most consistently (8 out of 16) studied SI, by asking participants to self-identify as initiators (e.g., “Have you ever tried a cigarette”). These studies reported a significant contribution of A to the total variance of SI (36-78%). However, when asked whether they smoked one or two puffs (e.g., “Have you ever tried a cigarette, even one or two puffs”) in addition to the ever use

question, the estimate of A dropped to 15%. Shared environmental effects (C) ranged from 7% to 24% of the total variance when measured with a self-reported ever use question. If 'even one or two puffs' was added to the question probing lifetime ever use, the estimate for C increased to 70% of the total variance.

*Initiation of Regular Smoking***.** Three twin studies examined the initiation of regular smoking via self-report (3/16). Overall, there was a significant effect of A on the initiation of regular smoking. One estimate of A was 49% of the total variance when measuring initiation of regular smoking as self-reported regular smoking initiation (e.g., "Have you ever been a regular smoker?"). Two twin studies[89,90] examined SI by asking if participants identified as a regular smoker ("Have you smoked cigarettes regularly for at least one month?"), which reported A to be between 62%-72%, slightly higher than the previous estimate, but still within the 95% confidence interval of the first estimate.

*Age of Smoking Initiation***.** Four additional studies examined the age of SI. These studies estimated substantial contribution due to A, ranging between 51% and 62% [91–94]. Further, estimates of C were between 31% to 53%.

Genetic and Environmental Influences on Smoking Initiation Vary Over Time. Twin studies of CIG tobacco, and other substances, have shown that the effect of additive genetic and shared environmental factors changes over time [95,96]. In general, shared environmental factors played a large role in SI during adolescence, and were less of a factor in young adulthood (i.e. college-aged individuals, 18-22), and even less

of a factor as one entered adulthood (age 23+) suggesting that genetic influences may be more prominent in young adults. In one of the first studies to examine this effect, Tully and colleagues reported a five-fold reduction in the proportion of variance due to C in nicotine dependence symptoms while the proportions due to A increased by a factor of two between the ages of 15 and 21 [97]. Other mega-analyses have reported C was most influential around ages 14-15 with A steadily increasing from age 15 onwards [95].

Sex Differences in Smoking Initiation. There are consistent sex differences in the prevalence of CIG initiation which may be due in part to sex differences in the contribution of genetic and environmental factors. There are differences by sex in the prevalence of SI [98–100]. Phenotypically, women over the age of 16 more often initiate cigarette use (59.8%) compared to men (50.3%) [101]. Additionally, prior twin studies indicate genetic and environmental differences also exist by sex. A meta-analysis of 17 twin studies reported a larger contribution of additive genetic effects in women compared to men (A_{women} =0.55, A_{men} = 0.37). This study reported a significantly larger contribution of shared environmental factors in men compared to women (A_{men} = 0.49, A_{women} = 0.24) [102]. These estimates were significantly different from one another, suggesting quantitative sex differences (i.e., the magnitude of the effect of additive genetic influence differs between men and women) in sources of variation for SI.

Additional studies also report differences in the magnitude of ACE estimates for SI by sex. A study of Australian twins reported similar results with smaller estimates of A for men (A_{men} = 0.22) compared to women (A_{women} = 0.63 for women; at the same time, C was larger for men (C_{men} = 0.42) compared to women (C_{women} = 0.11) [103]. In another

study of Australian adult twins, Morley and colleagues estimated similar estimates of A for men (A_{men} = 0.63) and women (A_{women} = 0.54) for SI [104]. Research utilizing adults from the United States has also provided additional evidence of the sex differences (A_{men} ranging 0.48-0.72, while A_{women} estimates ranged from 0.32-0.63), continuing to suggest there are differences between men and women in the US [105–107]. This pattern is also present in other countries as reported by a study of Finish adult twins which reported larger estimates of A for men (A_{men} = 0.59) compared to women (A_{women} = 0.35) [108]. More recent mega-analysis of 11 studies consisting of data from 19,313 twin pairs analyses have reported sex differences for genetic and environmental effects (for age 12 twins, A_{women} = 0.60, Amen = 0.40; C_{women} = 0.05, C_{men} = 0.10). However, these estimates were not statistically significant (18-year olds A_{women} = 0.30, A_{men} = 0.45; C_{women} = 0.10, C_{men} = 0.10) until after puberty (i.e., ages 15 and older). Consequently, although there are sex differences in estimates of genetic and environmental effects, they are only significant at later ages. Estimates of genetic and environmental contributions at young ages (ages 12, 13, and 15) produced non-significant estimates while older teens (14-18, excluding age 15) reported significant effects. Further complicating the results, this sex difference is no longer present as teens transition into young adulthood (age = 19)

Genetic and Environmental Factors Are Shared Between Initiation of CIG and Other Substances. Twin studies of SI and the use of other tobacco products report significant genetic overlap. When examining SI and the initiation of snus (a variant of dry snuff popular in parts of Europe where it is legal, similar to dipping tobacco in

America) data from Norwegian twins suggested similar, but not identical, genetic influences for both delivery systems (r_g = 0.82, A_{CIG} = 0.77 A_{SNU} = 0.54) [109]. CIG use has also been associated with other substances beyond tobacco. Alcohol and SI have been identified as sharing genetic overlap with the genetic correlation being estimated at 0.68 (95% CI = 0.61-0.74) in one study of adult twins [110,111]. Additional studies reported a significant genetic correlation between ND and cannabis use, such that roughly 50% of the genetic variance was shared between those phenotypes [112]. Therefore, these studies suggest the importance of genetic and environmental factors shared across tobacco products and possibly other substances, suggesting there may be common factors for substance use, in addition to unique factors for each substance.

In all, prior studies on the genetic and environmental contributions toward CIG initiation indicated: (1) significant additive genetic (i.e., the effect of alleles at every contributing locus) and shared environmental influences (i.e., the effect of environmental factors that increase similarity between members of twin pairs), (2) the magnitude of these influences changed across development (i.e., shared environmental influences have substantial contributions during adolescence and young adulthood which decrease into older adulthood), and (3) the presence of significant sex differences in genetic and environmental contributions [8-11].

To date, most twin studies of tobacco products focus on CIG use; however, recent population-level studies indicate that more adolescents and young adults using ECIGs over CIGs [5]. These tobacco products are often used together suggesting that using one product will increase the odds of using another product [113]. Not every individual who uses ECIGs will use CIGs, thus it is currently unclear whether different

genetic and environmental influences contribute to CIG and ECIG initiation or whether risk factors are shared between ECIG and CIG. We address this knowledge gap in a twin study of adolescents and young adults. Specifically, we explore: (1) the degree to which there is a correlation between ECIG and CIG initiation, (2) the degree to which the correlation between ECIG and CIG initiation is due to shared genetic and environmental influences, and (3) the degree to which genetic and environmental influences are specific to ECIG initiation.

METHODS

Data and Study Population.

Data were obtained from participants in the Adolescent and Young Adult Twin Study (AYATS), a US longitudinal cohort study of twins aged 15-20 (average age at wave 1 = 17.22, SD = 1.28; wave 2 = 19.23, SD = 1.33). Data were collected on 860 individuals via web-based questionnaires over two waves (Wave 1: March 2012 – December 2016; Wave 2: May 2016 – November 2019). A total of 858 individuals (421 complete twin pairs: 160 MZ pairs, 261 DZ pairs) who had tobacco use data were included in the analysis. The majority of the sample was female (56%), European-American (90%), had an average annual parental income greater than $35,000 per year (60%), and most parents (68%) had earned a Bachelor's degree or higher.

Tobacco Use Measures. Lifetime ECIG initiation was measured at both waves using a 4-item ordinal variable which asked, "*On how many occasions, in your lifetime, have you used an e-cigarette (assume one use is about 15 puffs or lasts around 10 minutes)?"* Participants indicating any level of use during Wave 1 or Wave 2 were

considered to have initiated ECIG at some point in their lifetime and were coded as 1. Participants who did not initiate in Wave 1 and Wave 2 were considered not to have engaged in lifetime initiation and were coded as 0. Lifetime CIG initiation was a binary variable that asked, *"Do you currently or have you ever smoked cigarettes?"* Participants indicating initiation during either Wave 1 or Wave 2 were considered to have initiated CIG at some point in their lifetime and were coded as 1. Participants who did not initiate in Wave 1 and Wave 2 were coded as 0.

Statistical Analysis. A CTD was used to study the contribution of genetic and environmental influences for binary traits. Twin modeling for binary traits builds on the principles of model development using continuous data. In a univariate genetic analysis of continuous data, the total phenotypic variance underlying the liability (V_P) of an outcome.

The expectations of twin member resemblance for MZ and DZ pairs can be summarized as a path diagram which can be used for the calculation of the phenotypic variance and variances due to genetic and environmental influences as well as the covariances between individual twins within pairs (Figure 2.1). The expected covariance in a twin model can be summarized with "Wright's Rules" for path analysis[114,115]. There are six rules: 1) the covariance is calculated as the sum of all possible paths between two variables, where each path represents the product of all path coefficients in the chain, 2) after moving forward along a single-headed arrow, moving backward is illegal, 3) a path can contain *at* most one double-headed arrow, 4) each variable can be crossed only once per path, 5) whenever changing direction (from upstream to

downstream) multiply the path by the variance of the upstream variable (in Figure 2.1, all variances are set to 1 and denoted with a double-headed arrow that returns to the latent variable), and 6) loops are not allowed.

The latent influences due to A, C, and E are represented by circles, which have a variance of one (denoted by double headed arrows above each circle). The paths a_{11}, c_{11}, and e_{11} represent the paths from the latent influences that contribute to the measured total phenotypic variance for each of the twins (represented as squares). The genetic correlation between members of an MZ pair is equal to 1 because MZ twin on average share 100% of their genetic material. The genetic correlation between members of DZ pair is equal to 0.5 because they share, on average, 50% of their genetic material. Therefore, the genetic covariance shared between members of a twin pair (represented as the double headed arrow between the A latent influences) is given a value of either 1 for MZ twins or 0.5 for DZ twins. Members of MZ and DZ twin pairs are assumed to share the same degree of common environmental influences and as such the correlation for C latent influences is one. Path tracing rules can be used to translate the path diagram into formulas to calculate means, variances, and covariances that can be used to estimate A, C, and E. First, the estimate of V_A for a single member of a twin pair is represented using the a_{11} path as: a*1* a, which is equal to a^2. VC and VE are similarly estimated as c^2 and e^2, respectively. The total phenotypic variance (V_P) is decomposed as the sum of the variances due to additive genetic (V_A), common environmental (V_C), and unique environmental (V_E) influences (Equation 2.1).

$$\sigma_P^2 = V_A + V_C + V_E \qquad \text{Eq. 2.1}$$

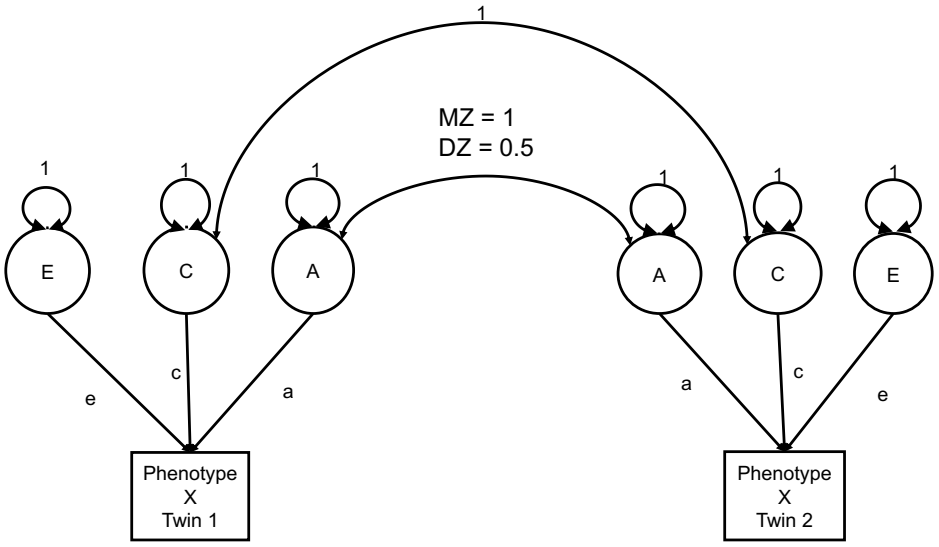


Figure 2.1. Univariate classical twin model used to estimate additive genetic (A), shared environmental (C), and unique environmental (E) influences.

The covariances between members of twin pairs are estimated as:

$$\text{covMZ} = a^2 + c^2 \quad (2.2)$$

$$\text{covDZ}_{\text{same-sex}} = 0.5a^2 + c^2 \quad (2.3)$$

$$\text{covDZ}_{\text{opposite-sex}} = 0.5r_g a^2 + c^2 \quad (2.4)$$

Estimation of A, C, and E from the CTD model using continuous data was modified for the study of the binary measures ECIG and CIG initiation. Such models adopt a threshold model approach, which describes discrete traits to have an underlying normal distribution of liability (e.g., susceptibility for endorsing an item measured as a probability with a Z-score distribution). Liability is measured as a series of ordered categories characterized by phenotypic discontinuities that occur when the liability reaches a given threshold. In other words, the thresholds differentiate those with and without the trait. Since the underlying trait is continuous in nature, the prevalence of the

trait under study can be used to estimate the threshold. For instance, if 1000 individuals were sampled and 120 had the trait of interest, a threshold could be put on the normal distribution so the area under the curve to the right of the threshold would be equal to 12% (120/1000; the prevalence of the trait in the sample). Therefore, modeling thresholds allows binary or ordinal data to be treated as continuous data [116].

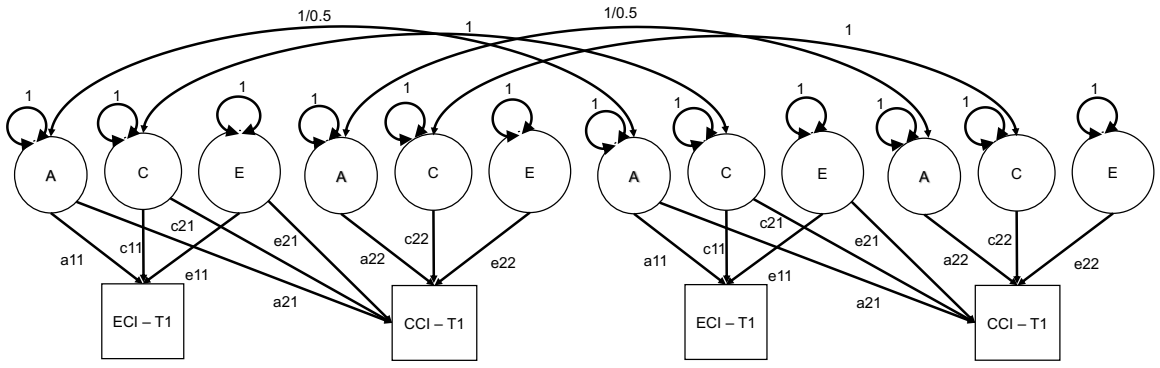


Figure 2.2. Bivariate genetic model used to estimate genetic and environmental contributions. a_{11} and a_{22} represent unique sources of additive genetic variance for electronic and conventional cigarette initiation respectively. a_{21} represents the overlapping additive genetic variance. ECI = ECIG Initiation, CCI = CIG Initiation.

Bivariate Genetic Modeling. The univariate twin model (Figure 2.1) was extended to evaluate two phenotypes simultaneously (a bivariate model; Figure 2.2). A bivariate model allows for possible overlap in genetic (A) and common (C) and unique (E) environmental factors between two traits and the estimation of genetic or environmental correlations.

A bivariate genetic model was implemented using a Cholesky factorization to determine how much of the covariance between ECIG and CIG initiation could be explained by shared genetic and environmental factors [85] (Figure 2.3). This is a method

of triangular decomposition of the genetic and environmental sources of variance where the first variable is assumed to be influenced by a latent factor that also explains some or all of the variance in the second variable. For each source of variance, the second variable is also explained by a latent factor that is uncorrelated with the first factor (path a_{22}; Figure 2.2). The diagonal elements in the genetic matrix (e.g., paths a_{11} and a_{22}; Figure 2.2) in a Cholesky factorization estimate the variances of a specific variable while the off-diagonal element (path a_{21}; Figure 2.2) estimates how much of the variance of the second variable is shared with the first variable.

The same path rules that guided the covariance in a univariate model apply in the bivariate model. Using Figure 2.3 as a guide, the expectation of A is slightly changed. Instead of one path to get between both latent additive genetic variables there are two. The first comes from the additive genetic variance alone: $a_{11}*1*a_{11} = a_{11}^2$ (in the case of MZ twins). However, there is now a secondary pathway starting from the second phenotype of twin 1 (here CCI – T1). Pathway a21 leads from CCI – T1 to A for ECI, then around the covariance between twin 1 and twin 2 and back down pathway a11, leading to a total expectation for MZ twins of $a_{11}^2 + a_{11}*a_{21}$. Similarly, the effect of C can be expected to consist of $c_{11}^2 + c_{11}*c_{21}$ for both MZ and DZ twins.

Genetic and environmental correlations between ECIG and CIG initiation were estimated to evaluate the degree to which genetic and environmental factors overlap between the initiation of ECIG and CIG. Standardized genetic and environmental covariances (cov_A, cov_C, cov_E) were also estimated to detail the degree to which genetic or environmental factors contributed to the phenotypic correlation (r_p) between ECIG and CIG initiation. The phenotypic correlation is equal to the sum of the standardized

genetic (cov_A) and environmental (cov_C and cov_E) covariances ($r_{pheno} = covA + covC + covE$). The standardized genetic correlation between two measures is defined in equation 2.5:

$$r_{x,y} = \frac{A_{xy}}{\sqrt{A_x \times A_y}} \quad \text{Eq. 2.5}$$

where A_{xy} is the genetic covariance between ECIG ever use and CIG ever use, and A_x and A_y represent the genetic variances of ECIG and CIG ever use.

The statistical significance of the genetic and environmental covariances was assessed by comparing the model fit of the full bivariate model to that of three submodels in which the genetic (pathway a_{21}) or environmental (pathway c_{21} or pathway e_{21}) paths between ECIG and CIG initiation were separately set to zero (difference in df = 1). Under certain conditions, such differences are asymptotically distributed as a chi-square distribution with one degree of freedom [85]. A fourth sub-model tested the significance of the phenotypic correlation by setting all genetic and environmental cross paths between ECIG and CIG initiation to zero (difference in df = 3).

All analyses were performed in R 3.4.1 [117] using the OpenMx package 2.8.3 [118], and missing data were addressed using full-information maximum-likelihood estimation. We chose *a priori* to retain and report all parameters in the model. Estimates from a full ACE model will be more accurate than simplified models. Further, attempts at parsimony result in oversimplification of the models rather than a more accurate representation of the data. Consequently, reporting a potentially oversimplified model might result in future research that may ignore an important source of variance [119].

RESULTS

Descriptive Statistics

Approximately 24% of the sample had initiated ECIG use while 19% had initiated CIG use, and 11% had initiated dual use. Males had a significantly higher prevalence of CIG (23.9%) and ECIG (22.5%) initiation compared to females (CIG- 14.2% and ECIG-12.5%). There was a moderate to large cross-twin correlation for ECIG initiation (rMZ = 0.65, 95% CI = 0.42-0.89; rDZ = 0.55, 95% CI = 0.33-0.77). A similar pattern was observed for CIG initiation (rMZ = 0.62, 95% CI = 0.38-0.86; rDZ = 0.52, 95% CI = 0.30-0.74).

Table 2.1. Summary Statistics of AYATS Sample

	Males N (%)	**Females N (%)**	**Total N (%)**
Tobacco Initiation			
ECIG	51 (22.5)	42 (12.5)	92 (20.2)
CIG	88 (23.9)	70 (14.2)	159 (18.5)
Race			
African-American	24 (6.4)	40 (8.0)	64 (7.3)
European-American	340 (90.4)	435 (87.3)	775 (88.7)
Latino	12 (3.2)	23 (4.6)	35 (4.0)
Parental Education			
HS/GED	47 (12.7)	85 (17.6)	132 (15.5)
Associate's	35 (9.5)	68 (14.1)	103 (12.1)
Bachelor's	145 (39.2)	182 (37.8)	327 (38.4)
Master's	111 (30.0)	105 (21.8)	216 (25.4)
Doctorate	16 (4.3)	24 (5.0)	40 (4.7)
Other	16 (4.3)	18 (3.7)	34 (4.0)

Tests of Twin Model Assumptions

Prior to genetic analysis, several data-related assumptions were tested to ensure that such genetic modeling would be plausible. Assumptions were tested to ensure that there were no significant differences in thresholds by twin order, zygosity

and sex prior to genetic modeling. Model assumptions were tested by equating threshold estimates to be the same for twin one and twin two, then equating threshold estimates to be the same for MZ and DZ twins, and finally making the thresholds equivalent between males and females. Each of these models was tested against a saturated model, or the model where all parameters could be estimated.

Submodel expectations were evaluated through a series of model fit comparisons against the saturated bivariate model. First, measures of model fit were estimated between the saturated model and each submodel. Model fit is measured by comparing measures of likelihood (i.e., negative two log-likelihood, -2LL). Model fit represents a measure of how well the model tested explains the data collected. Model fit comparisons between models are assessed by estimating the difference in model fit between two models, which produces a value that can be interpreted as a test with a Chi-square distribution and having degrees of freedom equal to the difference between the number of parameters between the two. A non-significant result would be interpreted as the submodel fitting the data equally well as the full model (i.e., there is no significant difference between the two models). The model with fewer parameters would be retained as this fits the criteria laid forth by Neale and Cardon for the best model. Second, a value of model parsimony estimated as AIC (Akaike Information Criterion) was used to examine which model is the most parsimonious [85]. The model with the lowest AIC value is considered to be the simplest model.

Evaluating models by comparing model fit and parsimony applies the four criteria of a good model as summarized by Neale and Cardon (1992): 1) a model provides a good fit to the data, 2) the model is consistent, 3) the model is simple, and 4)

the model has statistically significant parameter estimates. If the model does not provide a good fit to the data, then the model needs to be modified [85]. Further, a model that is not consistent with biometrical theory (such as a model that has dominance effects, but not additive genetic effects, Falconer 1990) may not be a good model. Simple models are easier to falsify and are more informative than complex models. Finally, any non-significant parameter should be removed from the model as it does not add to the model's ability to explain the data.

A saturated bivariate model was used as the base comparison model. A saturated model estimates the means, variances, and covariances for all variables (CIG and ECIG initiation in both members of a twin pair, both zygosity groups, and for males and females), and therefore each possible pathway is estimated in a saturated model.

Three bivariate models were fit to test the assumptions inherent in the twin study (Table 2.2). Model 1 tested a saturated model where all parameters were free to vary across twin order, zygosity, and sex. Model 2 equated the thresholds across twin order as a test of birth order (i.e., whether thresholds could be equated across twin order to assess the randomness of assigning twins to be the first or second twin). There was no significant difference in model fit between this model and the saturated model (p = 0.42). Model 3 equated the thresholds across the zygosities as a test of consistency between MZ and DZ twins on their ECIG and CIG initiation (i.e., whether thresholds could be equated across zygosity in same sex pairs). There was no significant difference between this model and the saturated model (p = 0.07).

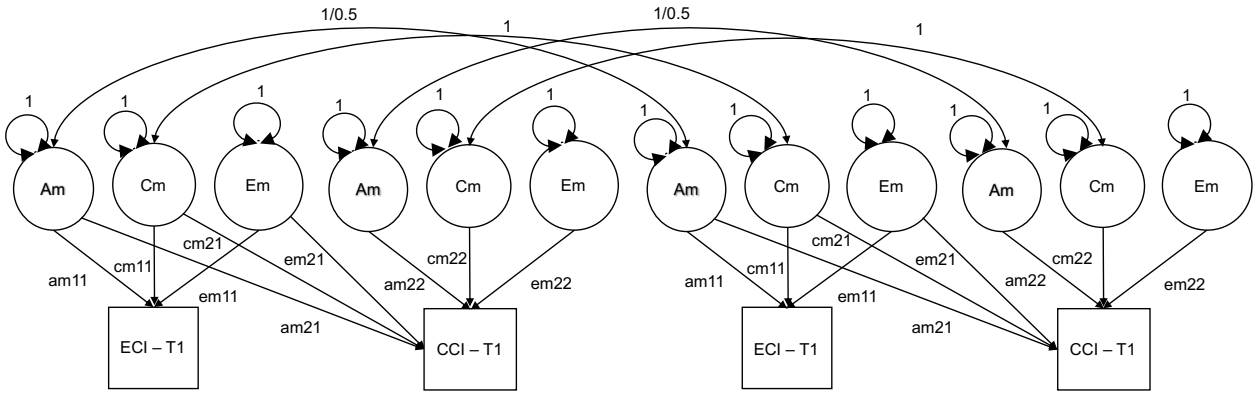
Table 2.2. Summary of Tests of Twin Model Assumptions

Model Number	Model Comparison	EP	DF	-2LL	AIC	Δ-2LL	*p*
1	-	50	1385	1110.16	-1659.84	-	-
2	2 vs 1	40	1395	1120.40	-1669.60	10.15	0.42
3	3 vs 1	32	1403	1137.71	-1668.29	27.55	0.07

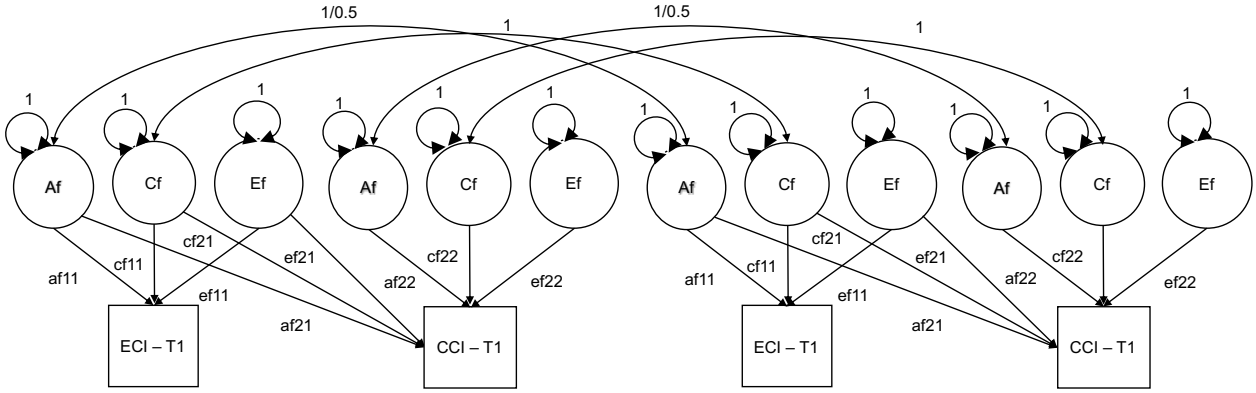
Model 1- Saturated
Model 2- Equate Twin 1/Twin 2 Thresholds
Model 3- Equate Twin1/Twin 2 Thresholds and MZ/DZ Thresholds
EP = Estimated Number of Model Parameters; DF = Degrees of Freedom; -2LL = Model Fit as Measured by the Negative Two Log Likelihood; AIC = Model Parsimony as Measured by Akaike Information Criteria

Tests of Sex Differences. A series of models were fitted to test the significance of sex differences in the magnitude and nature of genetic and environmental factors (Table 2.3). A full model included separate parameter estimates for each sex, as well as a parameter for the correlation between factors in males and females (Figure 2.3). There was no significant loss of model fit when the bivariate genetic model with all sources of genetic and environmental sex differences was compared to the saturated model (Table 2.3, Model 4, $p = 0.51$). Therefore, all subsequent tests of genetic and environmental influences were compared against this model. Model 5 (Table 2.3) tested the equivalence of thresholds across sex (i.e., male threshold is equal to the female threshold). The fit of Model 5 was not significantly different from that of Model 4 ($p = 0.63$). Model 7 tested the significance of the genetic correlation between additive genetic factors in males and females (r_g) by fixing it to 1, thus testing qualitative sex differences or whether the same set of genes contributes to liability in males and females. The fit of Model 7 was not statistically different from Model 4 ($p = 0.11$). Model 6 tested whether the A, C and E parameters could be equated for females and males as a test of quantitative sex differences to determine whether the magnitude of genetic and

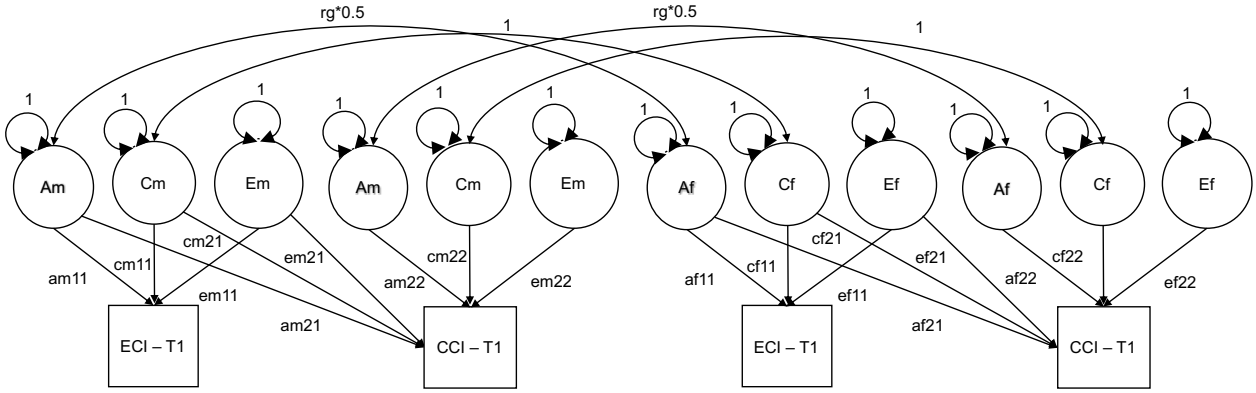
environmental influences is the same across the sexes. There was no significant loss of fit using this model (Table 2.3, $p = 0.06$) compared to a model that did estimate the sexes separately (Model 4). Finally, Model 8 tested the bivariate ACE model without any additional genetic and environmental sex differences (Figure 2.2). No significant differences in model fit were detected between Model 8 and either Model 4 (Table 2.3, $p = 0.51$) or Model 1 (saturated model, Table 2.3, $p = 0.20$). Consequently, a model without any sex differences was used to assess the magnitude of genetic and environmental effects on CIG and ECIG intiation.



(A)



(B)



(C)

Figure 2.3. Bivariate Models of Sex-Specific Models, Panel A Shows Male Pairs, Panel B Shows Female Pairs, Panel C Shows Opposite Sex Pairs

Table 2.3. Summary of Tests of Genetic and Environmental Sex Differences

Model Number	Model Comparison	EP	DF	-2LL	AIC	Δ-2LL	*p*
4	4 vs 1	23	1412	1136.35	-1687.65	26.19	0.51
5	5 vs 4	21	1414	1137.27	-1690.73	0.92	0.63
6	6 vs 4	12	1423	1155.65	-1690.35	19.31	0.06
7	7 vs 4	11	1424	1154.68	-1693.32	18.33	0.11
8	8 vs 1	13	1422	1154.03	-1689.97	43.87	0.20

Model 4: Bivariate genetic model estimating all sources of genetic and environmental sex differences
Model 5: Equate thresholds by sex
Model 6: Test of Quantiative Sex Differences- Equate path estimates by sex ($A_m/C_m/E_m = A_f/C_f/E_f$)
Model 7: Test of Qualitative Sex Differences- Set r_g to 1
Model 8: Bivariate ACE model without any genetic and environmental sex differences
EP = Estimated Number of Model Parameters; DF = Degrees of Freedom; -2LL = Model Fit as Measured by the Negative Two Log Likelihood; AIC = Model Parsimony as Measured by Akaike Information Criteria

Bivariate Genetic Model Testing. Tests of genetic and environmental covariance between CIG and ECIG initiation indicated no significant differences in models where a_{21} to zero (Model 2, Table 2.4, Figure 2.2) or setting c_{21} to zero (Model 3, Table 2.4, Figure 2.2). However, setting the e_{21} cross path to zero did result in a significant difference in model fit (Model 4, Table 2.4, $p = 0.01$). There was no significant difference when dropping either all A or all C influences (i.e., Models 5, paths a_{11}, a_{21}, and a_{22} were set to zero, for Model 6 paths c_{11}, c_{21}, and c_{22} were set to zero, Table 2.4). However, when both A and C influences were dropped simultaneously (i.e., all a and c paths were set to 0, Model 7), the models fit significantly worse when compared to the full bivariate ACE model (Model 1, Table 2.4; $p < 0.001$).

Table 2.4. Bivariate Modeling Fit Statistics

Model Number	Model Name	EP	DF	-2LL	AIC	Δ-2LL	*p*
1	Bivariate ACE	13	1422	1154.03	-1689.97	-	-
2	Test of CovA	12	1423	1155.18	-1690.82	1.15	0.28
3	Test of CovC	12	1423	1156.97	-1689.03	2.95	0.09
4	Test of CovE	12	1423	1160.46	-1685.54	6.43	0.01
5	CE Model	10	1425	1155.18	-1694.82	1.15	0.76
6	AE Model	10	1425	1159.71	-1690.29	5.68	0.13
7	E Model	7	1428	1213.18	-1642.82	59.15	< 0.001

Note. EP = number of estimate parameters; df = degrees of freedom; -2LL = negative two log likelihood, a measure of model fit; AIC = Akaike Information Criteria; Δ-2LL = difference of -2LL

Common environmental influences accounted for a non-significant proportion of the variance in the liability of ECIG and CIG initiation (ECIG: C = 0.42, 95% CI = 0-0.73, $p = 0.12$; CIG: C = 0.42, 95% CI = 0-0.70, $p = 0.13$). In addition, the contribution of additive genetic influences on the initiation of both delivery systems was non-significant (ECIG: A = 0.25, 95% CI = 0-0.83, $p = 0.44$; CIG: A = 0.19, 95% CI = 0-0.79, p = 0.57) (Table 2.4).

There was a strong phenotypic correlation between ECIG and CIG initiation (r = 0.77, p < 0.001). This phenotypic correlation was due to non-significant common environmental covariance ($cov_C = 0.23$, $p = 0.32$), non-significant additive genetic covariance ($cov_A = 0.23$, $p = 1$), and significant unique environmental covariance ($cov_E = 0.31$, $p = 0.01$). The unique environmental correlation ($r_E = 0.87$, $p = 0.01$) was significant between both delivery systems (Table 2.5)

Table 2.5. Standardized Genetic and Environmental Parameter Estimates for Electronic (ECIG) and Conventional Cigarette (CIG) Initiation

Parameter	Estimate (95% CI)	p-value
ECIG Initiation		
A	0.25 (0 – 0.83)	0.44
C	0.42 (0 – 0.73)	0.12
E	0.32 (0.14 – 0.56)	< 0.001
CIG Initiation		
A	0.19 (0 – 0.79)	0.57
C	0.42 (0 – 0.70)	0.13
E	0.39 (0.18 – 0.57)	< 0.001
Shared Parameters		
cov_A	0.23 (0 – 0.43)	1
cov_C	0.23 (0 – 0.52)	0.32
cov_E	0.31 (0.14 – 0.45)	0.01
r_g	0.76 (0 – 0.99)	1
r_c	0.68 (0 - 1.0)	0.32
r_e	0.87 (0.50 – 0.99)	0.01

cov_A- genetic covariance; cov_C- shared environmental covariance; cov_E- unique environmental covariance; r_g – genetic correlation; r_c- shared environmental correlation; r_e- unique environmental correlation

DISCUSSION

This is the first study to investigate the genetic and environmental contributions to the liability for ECIG initiation and explore the degree to which genetic and environmental factors influencing ECIG initiation are shared with CIG initiation. There was evidence for familial resemblance - likely a combination of additive genetic and shared environmental effects - on ECIG initiation. Additionally, there was substantial shared influences of unique environmental factors shared between both delivery systems.

The prevalence of ECIG initiation (~24%) in the current study was similar to those reported in other studies collected during a similar time frame (18.7% in 2014) [75]. However, the prevalence of CIG initiation was higher compared to national estimates

(19% vs. an average prevalence of 9.9% during 2011-2015) and may reflect regional preferences for CIG use (e.g., Virginia, North Carolina). There was also a strong association between ECIG and CIG initiation, which is supported by prior research which indicates that ECIG users are more likely to engage in CIG use [53,77,120].

Genetic and Environmental Factors Influence ECIG as well as CIG Initiation.

The magnitude of the estimates for the genetic and environmental contributions toward CIG initiation were similar to those previously reported in a mega-study of adolescents [95]. The magnitude of A from prior studies are generally smaller (range: 0.10 – 0.40) than estimates of C (range: 0.40 – 0.80) [95]. There was a similar pattern in the magnitude of genetic and environmental influences (A = 0.19 and C = 0.42) for CIG as well as ECIG initiation (A = 0.25 and C = 0.42).

Additionally, although genetic and shared environmental correlations (i.e. rG, rC) across ECIG and CIG initiation were individually not significant, their combined effects were. Similarly, though estimates of A and C were non-significant, models which did not include both sources of variance fit significantly worse than models that did, suggesting that both A and C are important factors of ECIG and CIG initiation.

Unique Environmental Factors Influence Electronic Cigarette Initiation and Have Overlap with Conventional Cigarette Initiation.

There were significant contributions of unique environmental factors specific to ECIG initiation, as well as significant shared influences of unique environmental factors contributing to ECIG and CIG initiation. Possible factors include peer smoking and

opinions towards nicotine products [62] as well as exposure to tobacco marketing [64]. Consequently, unique environmental factors are important for tobacco initiation and may be shared across delivery systems.

GWAS of CIG initiation has provided many insights into the molecular genetic architecture of CIG use. A recent meta-analysis of 1.2 million individuals reported 378 variants that were associated with CIG initiation [121]. However, a more recent GWAS of smoking initiation in ~800,000 individuals only reported 12 loci that were associated with CIG initiation [122]. There are several genes that are consistently reported in the molecular genetics literature of tobacco use. The nicotinic acetylcholine receptor (nAChR) genes (e.g., *CHRNA4*, *CHRNB2*) have consistently shown a significant association with SI [123]. These genes are able to be bound to and excited by nicotine, mimicking the action of choline.

Recent results using polygenic scores of cigarette use suggest that there is significant overlap in the genetic influences of ECIG and CIG initiation. For example, one study of participants in the Netherlands Twin Registry, reported a significant association between a polygenic risk score (PRS) for cigarettes per day, as calculated by summary statistics from the Tobacco and Genetics Consortium (TAG) and lifetime use of ECIGs such that ex-smokers had roughly 43% greater odds of initiating ECIGs [124]. Further investigation using the Avon Longitudinal Study of Parents and Children (ALSPAC) reported a positive association between a PRS for CIG initiation (as calculated from the GWAS and Sequencing Consortia for Alcohol and Nicotine) and ECIG ever use such that individuals were 24% more likely to be an ECIG initiator based on their PRS [58]. Both of these samples utilized adults (Allegrini mean age = 45; Khouja

age of data collection = 24) rather than adolescents or blended samples. There should be no bias for the changing effects of A and C across age due to the retrospective nature of the data (i.e., reporting on childhood behaviors as an adult). By reporting significant associations between PRS generated from CIG use and ECIG initiation, there appears to be confirmation that there is some genetic overlap between CIG and ECIG use.

These results should be evaluated in light of the following limitations. First, the majority of participants identified as European-American race/ethnicity, and results from this study may not generalize to other racial/ethnic populations. Second, this study used self-report data which may be subject to reporter bias. Third, we used measures of lifetime ECIG and CIG initiation which do not capture the complexities of long-term use (e.g., quantity/frequency). Nevertheless, many genetic epidemiological studies have focused on CIG initiation and have consistently identified similar results regarding the factors involved with this first step in nicotine dependence. Fourth, the power to detect significant additive genetic effects and genetic correlations was limited as a result of sample size and prevalence of ECIG/CIG initiation.

These results provide preliminary evidence for genetic and environmental influences involved in ECIG initiation as well as significant shared influences between ECIG and CIG initiation. More recent research has provided additional evidence from molecular genetics regarding the shared influences in genetic factors between ECIG and CIG. Two recent papers have reported the use of polygenic risk scores (PRS) generated for CIG initiation, that are significantly associated with ECIG initiation [58,124]. Thus, the evidence is mounting that genetic factors that influence the liability toward

CIG initiation may also play a role in ECIG initiation. However, there is not one hundred percent shared influences in the genetic variants that influence these phenotypes. Further research is needed to characterize which variants are unique and shared between ECIG and CIG initiation.

CHAPTER 3: SCOPING REIVEW OF TOBACCO USE MEASURES IN GENETICALLY-INFORMATIVE SAMPLES: RECOMMENDATIONS FOR FUTURE TOBACCO RESEARCH

INTRODUCTION

Tobacco use is a significant risk factor for several chronic conditions (e.g., lung cancer and cardiovascular disease) and remains a global public health concern [78]. Many individuals continue to engage in CIG use. However, the rates of CIG use in the US population have reached their lowest ever level[125]. CIG use often leads to the development of Tobacco Use Disorder (TUD), defined as tobacco use leading to dependence on nicotine—the addictive chemical in tobacco [126].

There is a large amount of heterogeneity in tobacco use measures. Several different instruments have measured the several stages of tobacco use: initiation of use, progression to regular smoking, and nicotine dependence [27]. The variation of measurement within a stage has led to several gaps in understanding across the etiology and remission of TUD as well as for each smoking behavior. In particular, measurement variation for tobacco product use is likely to have implications in the conclusions of recent and future genetic epidemiologic studies of tobacco use and TUD particularly for new and emerging tobacco products (such as electronic cigarettes, ECIGs). As inconsistent findings for CIG use may be due in part to measurement issues, future work in ECIGs (and other novel tobacco use products) should anticipate these concerns. Genetic epidemiology results may be inconsistent because of the measurement heterogeneity of tobacco use across studies. Meta-analysis of published

studies produces standardized results across studies. However, they have the potential to mask the degree of measurement variability and may not motivate careful consideration of the kinds of results that are produced specifically for a given measure. This concern is particularly important for researchers as they consider new data collection and are faced with making choices of the types of measurements to include in a study. Therefore, it is necessary to systematically evaluate the magnitude and quality of study results produced using genetic epidemiology study designs. Although some prior meta-analysis and systematic reviews of CIG use have been reported using twin studies, the most recent was conducted in 2017 [95] and have not evaluated genetic and environmental influences for other tobacco products.

The issue of measurement heterogeneity leading to variation in results may also have important implications for genome-wide association studies (GWAS). If the outcome, or phenotype of interest, is poorly defined two problems may occur. First, the parameter estimates of the magnitude of associations generated from GWAS may be biased. This would lead to differences in estimates of heritability from GWAS compared to those generated from twin studies. Second, heterogeneity in outcome measurement may lead to inconsistent results across GWAS. Further, this would lead to an inability to replicate GWAS results.

This study seeks to aggregate and summarize results across two sets of results. The first set focuses on results from specific studies of tobacco use. These studies include variance component estimates from twin studies, and single marker variants from GWAS. This will provide insight into the range of the influence of genetic factors across tobacco use phenotypes. The second set focuses on summarizing the biological

relevance of GWAS results. It is possible that conclusions about biological function derived from GWAS results may be dependent on phenotype measure. However, to date, there has been relatively little effort to broadly aggregate and reflect on the functional relevance of GWAS results across published studies of tobacco use. Consequently, it may be helpful to aggregate GWAS results by summarizing the known downstream biological products of genetic variants to establish the *functional relevance* of GWAS results for genetic variants (e.g., mRNA production, gene expression, and protein). Functional relevance may provide novel insights into the treatment of disorders by pharmacological means. Aggregating and synthesizing GWAS results within the context of functional relevance does not rely on the consistent replication of specific genetic variants of the genome. Instead, a summary of results related to functional relevance produces detail on biological pathways. For example, a study examined GWAS results for childhood onset asthma, reporting biological pathways involved with the immune system [127]. This type of information produces biological context to motivate future pathways for study.

There is fair amount of heterogeneity in the results of genetic epidemiologic studies of tobacco use because the measurement of tobacco use is heterogeneous. However, it is unclear if different operational measures of the same conceptual variable will lead to differing results (i.e., the detection of significant results across different SNPs or genes). Thus, the goal of this chapter is to summarize the genetic epidemiology results for two major study designs—twin studies and genome-wide association studies. These study designs were prioritized because twin studies provide the rationale for continuing to examine a phenotype for genetic variants. This will be done via first

providing twin study results as a basis for proceeding to GWAS designs. Secondly, functional relevance will be reported for significant published GWAS results. Recommendations for future studies will be outlined to conclude this chapter.

METHODS

A series of searches in four databases was performed (PubMed, EMBASE, PsychINFO, and CINAHL) to identify potential articles of interest. Article requirements included: (1) published in English, (2) published between the years 1982 (the first year a genetically-informative analysis was performed for tobacco use) and 2020, and (3) conducted using human subjects. The search was performed on 20 May 2020. Searches focused on articles satisfying three different concepts were created, each with unique search terms (tobacco product, tobacco use measurement, and genetically-informative studies).

The tobacco product concept reflects the different types of tobacco products that could be used by individuals. We wanted to capture all studies that examined tobacco products. The purpose of this concept was to capture each tobacco product to ensure that lesser known or used products (e.g., snus, hookahs, chewing tobacco) were included and not the most often studied products (e.g., combustible or conventional cigarettes). Tobacco products were searched using the following search terms: cigarette, conventional cigarette, combustible cigarette, electronic cigarette, chewing tobacco, hookah, cigar, snus, polytobacco use, comorbidity, and polysubstance. These terms returned 345,709 potential articles.

A tobacco use measurement concept was used to quantify the behaviors that are related to tobacco use. This concept was used to define the way researchers

previously conceptualized tobacco use behaviors to ensure all possible studies that may have studied tobacco use as an outcome measure (not a covariate) were found and included in the search. The search for this concept utilized the following terms: cigarettes per day, cessation, pack-years, initiation, ever use, current use, age of initiation, nicotine dependence, withdrawal symptoms, former smokers, current smoker, withdrawal, cravings, heaviness of smoking index, initial subjective experiences with tobacco use, number of puffs, over 100 cigarettes, smoking status, cotinine level, NNAL (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol) level, and Fägerström Test for Nicotine Dependence. This concept lead to 482,151 potential articles.

A genetically-informed studies concept was used to reflect studies that utilized genetically informative study designs, either through molecular, familial, or statistical genetic methods. This review is only interested in studies that addressed genetic influences (with or without environmental influences). Therefore, the search addressing this concept focused on the following terms: twin, family-based, genetic association, candidate gene, genotype, GWAS, co-twin design, consortia-based GWAS, genome wide association, genome wide, PheWAS, adoption studies, gene-environment interaction, epigenetic – methylation, epigenetic – miRNA, epigenetic – acetylation, heritability, multivariate, development. This concept returned 608,086 potential articles of interest. This study further focused on GWAS and twin studies for two reasons. First, twin studies are the first step for genetic epidemiologic studies, with results informing the degree to which latent genetic factors are associated with a phenotype. GWAS then examines measured genetic effects in an effort to pinpoint where in the genome the association comes from. Thus, these two designs form a logical dyad. Secondly,

tobacco use needed to be an outcome rather than an exposure or covariate. Many studies of more recent technologies (e.g., methylation or epigenetic studies) consider tobacco as an exposure rather than an outcome and are therefore not relevant for this review.

Utilizing the Concepts

While each concept returned hundreds of thousands of potential articles, they only returned articles based on their individual concept (e.g., the tobacco product concept only returned articles that addressed tobacco products). However, this strategy did not account for the overlap with the genetically-informative concept. Therefore, we combined each of the results from the three previous concepts and retained only those articles that were present in each of the concept searches. We retained 16,778 articles that were focused on tobacco products (i.e., from the tobacco product concept) and were genetically informed. Of those, 6,096 articles were present in all three concepts. The resulting 6,096 articles were then uploaded to the Rayyan application to facilitate the review of abstracts [128].

Article inclusion criteria included: written in the English language, use of a genetically-informative study design, and a tobacco use measure as the outcome (i.e., not as a covariate or main exposure variable). Agreement across two out of three research assistants was required to include an article into this review. Any disputes that could not be resolved by those authors were decided upon by the lead author (JSC). This resulted in 492 articles deemed relevant for data extraction. Fifty-five articles were removed from the data extraction process (Figure 1) as a result of duplications (i.e., published online the year prior to journal publication and were erroneously counted as

separate articles). An additional set of 127 articles were removed because they were abstracts from conference proceedings, were not available in English, or were review/commentary articles. This resulted in 310 articles relevant for data extraction (Figure 1). Some studies measured tobacco use multiple ways (e.g., initiation and cigarette quantity), therefore a single article may be referenced several times throughout this review. Further, results were limited to those articles that utilized a genome-wide association analysis or some variation of that design, leaving 87 articles.

Data Extraction

Data from the 87 articles were extracted with a team of researchers including undergraduate and graduate students. Articles were binned into sets of 20 by year of publication and disseminated to team members. In consultation with colleagues who had previously conducted systematic reviews, a data extraction spreadsheet was designed (Appendix S3.1) which detailed various aspects of the articles such as sample size, age range, any subsamples the study examined, the conceptual variable, and the operational variable. The spreadsheet was then refined to only include relevant information for this review (Appendix S3.2). After data extraction was completed, another team member reviewed to ensure all relevant data were captured during the extraction process. Finally, data were summarized based on the conceptual measure of nicotine use. In other words, all studies that examined smoking initiation were collated into a separate spreadsheet from studies that examined nicotine dependence. These conceptual measures are not mutually exclusive, and studies could be represented in multiple tobacco use behaviors.

Aggregating and Translating GWAS Results to Understand Biological Relevance

Following the extraction of relevant markers from the GWAS studies, gene lists were run through DAVID (The Database for Annotation, Visualization, and Integrated Discovery) to extract biological pathways which appeared more often than expected as denoted by Fisher's Exact test [129–131]. DAVID is a data-mining tool which extracts the biological pathways that are represented by the genes with which DAVID is given from the GWAS results. This organization of biological pathways is accomplished by examining the co-occurrences across multiple bioinformatic platforms that curate details on the functional annotation of genes relevant to a SNP-based result (i.e., gene name/aliases, molecular function, and biological role).

DAVID results produce four different classes of results. The first class is a "Category". This class provides information on the bioinformatic platforms from which the results were extracted. DAVID probes publically available bioinformatic platforms including (but not limited to): BioCarta (https://www.hsls.pitt.edu/obrc/index.php?page=URL1151008585), Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/), GO (GOTERM; http://geneontology.org), and UniProtKB keyword (UP_KEYWORD; UniProtKB keyword). The second class is "Term", which describes the actual functional relevance within the pathway, such as "cell signaling". Term details the pathway's relevance based on evaluation of results across bioinformatic platforms which are not mutally exlcusive (i.e., the term could appear in multiple platforms). The third class of results is "Genes". This column provides the genes that enriched, or occur more often than expected by chance, for that particular term, such as cell signaling. The final class for the results is

the *p* value calculated for the Fisher's Exact Test to determine if a gene or gene list is significantly associated with a term.

GWAS results for phenotypes were submitted to DAVID analysis if they were either (1) conducted in at least three independent published GWAS studies or (2) conducted in GWAS consortia consisting of at least three separate samples. DAVID reported pathways that were significant at 0.05 using a Benjamini-Hochberg False Disovery Rate. Single nucleotide polymorphisms (SNPs) that were reported to be genome-wide significant ($p < 10^{-8}$) in the original GWASs were included in DAVID analyses.

RESULTS

1. Smoking Initiation

Twin Studies

Twin studies give a broad estimate of the extent to which a phenotype is influenced by genetic variants. In brief (see Chapter 2), twin studies estimate additive genetic (A), shared environmental (C), and unique environmental (E) effects by decomposing the covariance shared between two twins. As monozygotic (MZ, or identical) twins are presumed to shared 100% of their genetic effects, as well as 100% of the shared environmental influences, the only differences that could arise between these twins are due to unique environmental influences or measurement error. However, dizygotic (DZ, or fraternal) twins are assumed to only share, on average, 50% of their genomes but 100% of the shared environment. Thus, any differences in DZ twins could arise from either genetic or unique environmental influences. Comparison of

the covariances from the different zygosity groups allows the decomposition of variance into A, C, and E estimates. Twin models may also be extended to model non-additive genetic variance (D) in lieu of C influences as D and C are confounded [132]. The choice to model C or D is driven by the correlation pattern of MZ and DZ twins [133]. An rMZ greater than two times that of rDZ would necessitate modeling D rather than C. In total, 14 studies examined smoking initiation (SI) within a twin design.

Lifetime Ever Smoking Initiation. A common way used to measure SI is by asking if the participant has ever smoked. Rhee and colleagues reported if 1,062 twin and sibling pairs had ever used cigarettes and reported significant additive genetic (A = 0.24), non-additive genetic (D = 0.08), and shared environmental influences (C = 0.34) [134]. Similarly, Maes et al. (2004) used data from 6,805 twins from the Virginia Twin Registry focused on whether they had ever used a cigarette [135]. They reported significant genetic effects (A = 0.75) and unique environmental effects (E = 0.25). Vink and colleagues also asked a sample of Finnish twins (N = 10,063) if they had ever smoked a cigarette [135,136]. They reported significant genetic (A = 0.36), shared environmental (C = 0.56), and unique environmental influences (E = 0.07) in this European sample.

Further research has examined SI through the use of lifetime ever smoking but stratified the results by different self-reported race/ethnicity groups. Sartor and colleagues studied SI from an American sample of 3,553 twins of European descent and 945 twins of African descent [137]. They reported significant genetic (A = 0.50) and unique environmental influences (E = 0.46) and nonsignificant shared environmental

effects (C = 0.04) in twins of African descent. In the twins of European descent, there were significant influences of A (0.51) and E (0.12) but nonsignificant C (0.24).

Some researchers examined lifetime ever smoking by adding a threshold of the number of cigarettes a participant needed to have smoked in their lifetime before classifying them as initiating smoking. Hamilton and colleagues asked if participants in their study had smoked more than 100 cigarettes in their lifetime, meaning if an individual had smoked less than 100 cigarettes they would be considered a non-smoker [107]. In this study of 32,359 American twin pairs, male pairs were reported to have significant genetic (A = 0.71), shared environmental (C = 0.12) and unique environmental influences (E = 0.17). Female twin pairs also reported significant genetic (A = 0.32), shared environmental (C = 0.48) and unique environmental influences (E = 0.21). Similarly, McCaffery, examining a group of 9,414 American twins who had served in the Vietnam War, categorized individuals as initiators if they had smoked more than 100 cigarettes in their lifetime [138]. They reported significant genetic (A = 0.49), shared environmental (C = 0.29) and unique environmental effects (E = 0.22) in this sample.

Finally, in a mega-analysis using raw data from multiple samples, Maes and colleagues defined SI as having ever had a cigarette, even one or two puffs [95]. In this multi-country study of adolescent (aged 10-19) twins (N = 19,313), the researchers reported changing impacts of additive genetic and shared environmental influences of the developmental period. Specifically, shared environmental influences played a larger role at younger ages (C = 0.70 at age 13 and 0.40 at age 19) while additive genetic influences played less of a role at younger ages (A = 0.15 at age 13 and 0.45 at age 19).

Initiation of Regular Smoking. One of the first twin studies of SI derived smoking initiation by asking an individual to categorize themselves as ever being a smoker or not. The research aggregated three twin cohorts (Australian twins, 3,808 pairs; Virginia twins, 2,145 pairs; and a study from the AARP consisting of 3,620 twin pairs) [139]. Researchers reported significant genetic effects (A for men = 0.83; for women A = 0.50) and nonsignificant shared environmental effects for men (C = 0.01), but significant for women (C = 0.29). A different study in 2006 examined SI as individuals having been a regular smoker (self-identified) versus those who did not identify as a regular smoker in 14,472 Australian twins and siblings [90]. Similar to the aforementioned study, Morley reported significant genetic effects (A = 0.59 for females; A = 0.63 for males) and unique environmental effects (E = 0.17 for female; E = 0.19 for males), as well as small but significant shared environmental effects (C = 0.04 for females; C = 0.07 for males). Finally, Kendler and colleagues asked participants to identify themselves as regular smokers by asking if they had regularly smoked for at least one month in their lifetime [109]. In this sample of 1,103 Norwegian twin pairs, there were significant additive genetic (A = 0.79) and unique environmental influences (E = 0.21) to the initiation of regular smoking.

Smoking Initiation Derived from Smoking Status. An initial study of SI used whether an individuals, 7,375 male-male twin pairs who served in the Vietnam War, was a current smoker or not as their initiation variable [92]. Participants were classified as smokers if they currently smoked, and had smoked at least 100 cigarettes in their lifetime, or if individuals had smoked at least 100 cigarettes but were not current smokers. These individuals were compared to individuals who reported not using 100

cigarettes in their lifetime. Researchers reported a significant effect of additive genetic influences (A = 0.70) and unique environmental influences (E = 0.30), but no effect due to shared environmental influences.

Wills and colleagues used a measure of smoking status that asked twins to report how many cigarettes they smoke in a day breaking up smoking by either 1-19 cigarettes per day or more than 20 per day [140]. They also asked twins to report if they used to or only occasionally smoke cigarettes. These groups were aggregated into a smoking initiator category and compared against individuals who never smoked. They reported significant genetic (A = 0.43), shared environmental (C = 0.39), and unique environmental influences (E = 0.18) in this sample of 850 American twin pairs.

Age of Smoking Initiation. An early twin study of SI assessed the age the twins started smoking by asking, "At what age did you start smoking?"[91] Using 3,810 adult twins from the Australian Twin Registry, Heath and colleagues reported significant genetic effects and unique environmental influences in both a sample of young twins (age at assessment was less than 30; A = 0.62, E = 0.38) and an older cohort (age at assessment greater than 30; A = 0.51, E = 0.49). During the same timeframe, Kendler and colleagues used a sample of 1,898 Virginian female twins to assess age of SI, though they used a slightly different question [141]. Kendler and colleagues asked when they started regular smoking with regular smoking defined as a pattern of cigarette use such that the participant smoked at least 7 cigarettes per week for one month. They reported significant genetic (A = 0.78), shared environmental (C = 0.07), and unique environmental influences (E = 0.15).

Genome-Wide Association Studies of Smoking Initiation

Genome-wide association studies (GWAS) examine a series of genetic markers across the genome. A genetic association is a single test of association which uses genotypic data from a genetic marker to test for statistical associations between a genetic variant (e.g., single nucleotide polymorphism, SNP) at a specific locus and CIG initiation. A GWAS expands this test to thousands of genetic markers that test for associations with CIG initiation with SNPs across located throughout the genome. These designs typically genotype a number of markers (generally between 500,000 and 1 million markers) and exploit linkage disequilibrium (the non-random segregation of alleles) to impute up to 30 million markers [142]. This study design allows researchers to identify specific genetic variants associated with phenotypes of interest. There were 12 GWAS that identified "Smoking Initiation" as an outcome and of these, 5 use data with relatively small sample sizes (N = 1,114-8,842). A known limitation of GWAS of complex polygenic phenotypes is that they require sample sizes of greater than 20,000 [143] because the effect sizes that are often detected are very small (e.g., OR < 1.3). Consequently, most single-sample GWAS generally do not have the statistical power necessary to detect significant genetic associations with genome-wide data, and only one was a single-sample study [144]. In response to this limitation, seven [122,145–150] studies used a multi-sample approach in which GWAS data was aggregated from a few samples [147] or a GWAS was conducted in a single discovery sample and tested for replication of results in additional samples [146,148]. Many of these studies have suffered from low power to detect statistically significant results at a genome-wide level. Nevertheless, these single-sample studies are the foundation by which GWAS on

tobacco use were expanded, and as such we briefly summarize published results using these approaches in addition to study designs.

Single- and Multi- Sample GWAS

Lifetime Ever Smoking Initiation. A study of Japanese participants ages 20-89 (Biobank Japan Projects, N = 165,456) used a measure of lifetime ever smoking initiation using an item on length of smoking in life ("How many years do/did you smoke?") [144]. Responses to this item were categorized as ever versus never smokers. One SNP with a genome-wide significant association was detected in men and women (rs117036946). This SNP is located in *DLC1.* Two additional genome-wide significant associations were detected in analyses of males only. These intergenic variants are located between *CXCL12* and *TMEM72-AS1* (rs117097449) as well as *GALR1* and *SALL3* (rs77105140) and their biological function is unclear.

Initiation of Regular Smoking. No single-sample GWAS reported results on initiation of regular smoking. Two multi-sample GWAS have been conducted. One multi-sample GWAS reported results on lifetime initiation of regular smoking [145] ("Have you ever smoked cigarettes/bidis regularly in the past?"). None of the reported results achieved genome-wide significance ($p <= 5 \times 10^{-8}$). However, seven loci with suggestive results ($p < 1 \times 10^{-6}$) were identified in men. Six of these loci were located in the *SLC39A11* region. This gene encodes a protein belonging to the ZIP, a Zinc transporter gene which is responsible for moving zinc in and out of the cell, transporter family [151]. In women, 14 loci with suggestive results were identified. Two of these loci were located

on the X chromosome and chromosome 15 in the region between *SLCO3A1* and *ST8SIA2* genes and as such biological function is as yet unknown.

A second study used a single smoking status item and coded two different ways to measure SI in the discovery sample (N = 8,842 Korean participants, ages 40-69). The first SI measure was a binary trait related to lifetime regular cigarette use (i.e., never versus having regular cigarette smoking experiences) [149].

Smoking Initiation Derived from Smoking Status. No single-sample GWAS has reported on smoking initiation measured as smoking status. Four multi-sample studies have reported lifetime ever smoking initiation using a measure of current smoking status and categorized participants as "ever versus never smokers". The first study used a discovery sample (N = 3,497) and three additional replication samples (N = 7,863) that measured SI in a similar fashion. No suggestive genetic associations at $p \leq 1x10^{-6}$ were detected. Twenty-two genetic associations at $p \leq 1x10^{-4}$ were detected. Of these, the single nucleotide polymorphisms (SNPs) with the lowest p-values were rs4423615 ($p = 5.3 \times 10^{-5}$). This SNP is located in *GRB14*, a gene involved in the tyrosine kinase signaling pathway. It is possible that GRB14 is involved in the development type 2 diabetes [152]. There was also an association with rs10794595 ($p = 4.3 \times 10^{-6}$). This SNP is not located in a gene nor in linkage disequilibrium with any other SNP [148].

The second study combined data from two samples using a measure of lifetime ever use as ever versus never use (N = 4,611) [147]. No suggestive genetic associations at $p \leq 1x10^{-6}$ were detected. Seven genetic associations at $p \leq 1x10^{-5}$ were detected, although none of the variants identified were located in known gene regions [148].

The third study used a single smoking status item in the discovery sample (N = 8,842 Korean participants, ages 40-69) and analyzed the original four-level smoking status, treating it as an ordinal categorical variable (i.e., “never”, “former”, “light”, and “habitual”) [149]. The replication samples assessed SI as age of smoking onset in the two replication samples (N = 402 African American participants, N = 200 European American participants, ages 21 and older). Although three loci located in *RGS17* were detected for SI derived from smoking status in the discovery sample, these associations did not extend to the replication samples.

A fourth multi-sample study of US military veterans (N = 286,118, Million Veterans Program, mean age = 64.4, 55.0% were between ages 50-69) used a measure of smoking status for being an ever smoker (past or current) versus never smoker as a single study and meta-analyzed their results with those from a consortium (GSCAN) [122]. Results from single study analyses identified 12 genome-wide significant associations. Of these, eight had been identified in previous GWAS studies of SI (rs12044362, rs1004787, rs11581459, rs1474011, rs6438208, rs11724738, rs78875955, and rs7126748). Additionally, three were also replicated in meta-analysis (rs11581459, rs1004787, rs6438208). These SNPs are located within or near genes responsible whose products are responsible for regulation of gene expression (*LINC01360,* long intergenic non-protein coding RNA; *CAMKMT* [153]*,* calmodulin-lysine N-methyltransferase; and *ZBTB20* [154]*,* zinc finger and BTB domain containing 20).

Age of Smoking Initiation. Matoba et al. 2019 studied common variants in 30,418 Japanese participants [144]. Age of smoking initiation was measured by subtracting the number of years the participant had smoked from the age at the time the interview or

the age at the time they quit smoking. This value was then log transformed. There were no significant associations when data from men and women were analyzed jointly. In females, there was a genome-wide significant association for one intergenic locus (rs6718569, $p = 3.6x10^{-9}$) between *LINC01793* and *MIR4432HG*.

Four multi-sample GWAS have reported this outcome. The first multi-sample GWAS used data from adults ages 44-67 (Finnish Twin Cohort Study, N = 1,114) and reported results on four measures of age of smoking initiation (age at first puff ["How old were you the very first time you smoked even a puff of a cigarette?"], age of first cigarette ["How old were you the first time you smoked a whole cigarette?"], age of onset of weekly smoking ["How old were you when you first smoked a cigarette at least once a week for at least two months in a row?"], age of onset of daily smoking ["How old were you when you first smoked cigarette every day or nearly every day for at least two months in a row?"]) [146]. Three SNPs located in the intergenic region between *NACKAP5* and *MGAT5* were associated with age of weekly smoking. However, these associations were not detected in the replication sample (Finnish twin study [FT12], N = 869 and an Australian twin family sample [NAG-OZALC], N = 4,425). A second study assessed as age of smoking onset in a discovery sample (N = 8,842 Korean participants) and two replication samples (N = 402 African American participants, N = 200 European American participants). No suggestive genetic associations at $p \leq 1x10^{-6}$ were detected.

Results from a meta-analysis of the three samples identified 6 SNPs with genetic associations ($p \leq 1x10^{-4}$) in males. Of these, two (rs7747583, $p = 2.03x10^{-5}$ and rs2349433, $p = 3.09x10^{-5}$) were in the regulator region of *RGS17* (G-protein Signaling 17) a gene whose product regulates gene expression [149].

The third study used data from two samples (Prostate, Lung, Colon, and Ovarian Trial and the Nurses Health Study, N = 4,611) of adults ages 55 and above [147]. No suggestive genetic associations at $p \leq 1x10^{-6}$ were detected. However, associations were detected with two SNPs (rs11082304, $p = 6.0x10^{-6}$ and rs17050782, $p = 8.4x10^{-6}$) in genes whose products are responsible for cell proliferation and/or differentiation (*CABLES1*) and gene expression (*SETD7*) [147].

Siedlinski et al. 2011 studied common variants in a four-cohort study of 3,397 patients of European ancestry with COPD (mean age = 65) [150]. Age of initiation was measured as using either a case Report Form or modified versions of the American Thoracic Society /Division of Lung Diseases Respiratory Disease Questionnaire [155]. Eight loci had suggestive associations at $p<1x10^{-6}$. Of these, three highly correlated SNPs (rs9380362, rs7743060, rs769051) were in an intergenic region between *BAK1* and *ZBTB9*.

Consortia-Based GWAS

Meta-analyses of consortia containing many individual samples with similarly measured smoking behaviors have been used to overcome issues of low statistical power to detect genome-side significant associations in GWAS. Consortia-based approaches typically differ from multi-sample approaches using pre-identified data analysis plans agreed upon by individual study teams as well as sharing and meta-analyzing results across studies within the consortium. Five consortia-based studies of smoking initiation have been reported. These represent the largest studies of tobacco use outcomes to date and have the strongest statistical power to detect significant

genetic associations. Below, we detail seven consortia-based studies and the results that were prioritized by smoking initiation phenotype. When reported, estimates of heritability are also included.

Lifetime Ever Smoking Initiation. No consortia-based GWAS has been conducted on a measure of lifetime ever smoking initiation.

Initiation of Regular Smoking. Five consortia-based GWAS studies measured initiation of regular smoking by distinguishing lifetime ever use of 100 or more cigarettes. One of the first studies to use a consortium-based used a consortium of European ancestry-based studies (Tobacco and Genetics Consortium, 16 studies, N = 74,053). Eight genome-wide significant associations were identified [156] (Table 4.1). All the SNPs identified were in *BDNF*, a gene whose product (Brain-Derived Neurotrophic Factor) is responsible for maintaining neuronal survival and neuronal plasticity. Although Brain-Derived Neurotrophic Factor was initially determined to function within the striatum, it has been reported to be widely expressed in cortical and subcortical regions of the brain [157–161].

A study of African American adults (Study of Tobacco in Minority Populations Genetics Consortium, 13 studies, N = 32,389) ages 20-75 detected no genome-wide significant associations for SI [162].

A study of European ancestry samples of adults (GWAS and Sequencing Consortium of Alcohol and Nicotine Use, 29 studies, N = 1,232,091) [163] identified 378 genome-wide significant variants. Smoking initiation was harmonized across studies from three types of measures (i.e., “Have you smoked over 100 cigarettes over the course of your life?”; “Have you ever smoked every day for at least a month?”; and

"Have you ever smoked regularly?"). This study reported the significant genetic variants that were identified explained 7.8% of the variance for SI. Of the variants identified, those with known biological function included a variant near *PPP1R1B* (protein phosphatase 1 regulatory subunit 1B). The product of *PPP1R1B* affects synaptic plasticity and reward-based learning in the striatum. Similarly, variants involved in the pathways related to glutamate processes were identified. Additionally, pathways involved in SI also included sodium-, potassium-, and calcium- voltage-gated channels which are important for neuronal excitability and signaling. Other identified variants were determined to be important in the pathway related to dopamine neurotransmitter release.

A study of European and African ancestry adults (17 studies, N = 146,117) focused on genome-wide associations with low-frequency, nonsynonymous and putative loss-of-function exonic variants [164]. A total of 93 loci associated with initiation of regular smoking and having genome-wide significance were detected. The heritability for initiation of regular smoking explained by all SNPs in the study was 14%. The variation due to rare coding variants explained 2.2% of the phenotypic variance. Therefore, rare coding variants explained 15.7% of the of the SNP heritability. Novel variants with the strongest statistical evidence were identified in rs2232423 (*ZSCAN12*), rs35891966 (*NAV2*), and rs6265 (*BDNF*). Additionally, a suggestive SNP association and tests of gene-based association detected a significant association with *HEATR5A*. *BDNF* is a protein coding gene which produces brain-derived neurotrophic factor (BDNF). BDNF promotes survival and differentiation of neuronal populations during mammalian development. It is widely expressed in the adult nervous system as well as

in non-neural tissues, including the thymus, heart, and lung [165,166]. Further, it acts as a regulator of activity-dependent neurotransmission and plasticity in adults [157,160,161,167,168]. *NAV2* is a protein coding gene whose product (neuron navigator 2) is responsible in part for neuronal development and neurite outgrowth [169–171]. It is highly expressed in the brain (frontal cortex), kidney and liver, as well as other locations. *ZSCAN12* is a protein coding gene whose product (zinc finger and SCAN domain containing 12) is involved in DNA binding and transcriptional regulation. It is moderately expressed throughout the body. *HEATR5A* is a protein coding gene whose product (HEAT repeat containing 5A) is moderately expressed throughout the body. Generally, a HEAT repeat refers to a protein tandem repeat structural motif. Further, this structural motif is commonly found in four proteins (huntingtin, elongation factor 3, protein phosphatase 2A, and yeast kinase TOR1, HEAT) [172,173]. HEAT repeats form extended super-helical structures which are involved in intracellular transport.

One study of adults ages 18 and over focused on associations with rare coding variants [174]. This study used data from a discovery cohort and 61 replication cohorts, consisting of three consortia (N = 346,813). The phenotypic variance explained by the rare variants were 0.53%. Further, out of 40 variants with genome-wide significant associations detected in a combined analysis of discovery and replication cohorts, novel associations with three SNPs were also detected in initial analysis in the discovery sample. These SNPs are in *BORCS7* (rs7096169), *SMG6* (rs216195), and *TMEM182* (rs6738833). An additional non-novel association was detected in both discovery and replication samples for a SNP (rs462779) in a gene (*REV3L*) that encodes a protein which protects DNA from damage. Similarly, the products of *SMG6* and *TMEM182* are

responsible for downstream regulation of gene expression for other genes. *BORCS7* is a protein coding gene and its product (BLOC-1 related complex subunit 7) is part of a multi-subunit complex that regulates lysosome positioning and cell function related to regulation of cell spreading and motility. Further, *BORCS7* is expressed throughout the body, including the brain. Prior studies report expression in adult neurons and astrocytes [175].

Smoking Initiation Derived from Smoking Status. No consortia-based studies of smoking initiation derived from smoking status were identified.

Age of Smoking Initiation. Four consortia-based GWAS measured age of smoking initiation. All studies used self-report.

Furberg et al (TAG consortium) 2010[163] studied common variants in 24,114 European ancestry participants. Age of smoking initiation measured the age the participant started smoking cigarettes. Some studies studied the age at which the participant first tried smoking, even one or two puffs. Others measured the age the participant began smoking regularly. No genome-wide significant associations were detected.

David et al. 2012 [162] studied common variants in 15,547 African American participants. Age of initiation was measured two different ways. Some studies measured as the age at which smoking was first attempted. Other samples measured as the age at which participants began smoking regularly. Three highly correlated SNPs (rs1678618, rs12445577, and rs1612028) located in *SPOCK2* had a suggestive association ($p < 1x10^{-6}$) with age of initiation. The product of *SPOCK2* (SPARC [Osteonectin], Cwcv And Kazal Like Domains Proteoglycan 2, also known as testican-2)

is a protein which binds with glycosaminoglycans to form part of the extracellular matrix. It is expressed across several tissues and has particularly high levels of expression in the brain and lung. Further, it is expressed prominently in normal brain, and its expression levels decrease as tumor grade in this area increases.

Liu et al 2019 [163] studied common variants in 341,427 European ancestry participants. Age of smoking initiation was measured as “At what age did you begin smoking regularly?” Alternatively, other studies used a combination of “How long have you smoked?” and “What is your current age?” to derive a continuous measure for age of regular smoking initiation. Ten loci were associated at a level of genome-wide significance. The total SNP heritability was 5%. Of the 10 loci identified, 7 were located within intronic regions of genes (rs72853300- *TEX41*, rs12611472- *CUL3*, rs7559982- *WDPCP*, rs11915747- *CADM2*, rs13136239- *MAML3*, rs624833- *ADD1*, rs1403174- *MAD1L1*). TEX41 is a long-noncoding RNA gene.

Brazel et al 2019 [164] focused on associations with rare coding variants in 124,590 adults. Age of smoking initiation was conceptualized as the age at which an individual started smoking cigarettes regularly. Studies used items such as: “At what age did you begin smoking regularly?” Alternatively, other studies used a combination of “How long have you smoked?” and “What is your current age?” to derive a continuous measure for age of regular smoking initiation. Three loci were associated at a level of genome-wide significance (rs12493563- *CADM2*, rs8082191- *ACCN1*, rs442467- *MACROD2*). The total SNP heritability was 6% and the heritability due to the detected rare variants was 1.1%. Therefore, rare variants accounted for 18% of the SNP heritability for age of smoking initiation.

Functional Analysis. DAVID analyses reported 15 biological pathways that were significantly associated with genome-wide significant SNPs for SI. Five pathways were associated with the synapse and cell junctions. These two pathways included *CHRNA4* among 14 other genes (*FOCAD, TENM2, GRID2, NLGN1, CADM2, NRXN1, MAGI2, DIXDC1, GRIK4, CABP1, GRIN2B, DPP4, OLFM1, TRIM9, SDK1, GRIN2A, ADAM15, DLC1, CTNNA2, ERC2, CBLN4, LRRC4C*). Five pathways were associated with the immune system, specifically immunoglobulins. These pathways impact the adhesion of the immunoglobulins rendering them less effective[176]. In total 14 genes were associated with this system (*ROBO2, LINGO1, NEGR1, CADM2, DCC, NTM, PXDNL, SEMA3F, PTPRF, IGSF11, PTPRD, SDK1, NCAM1, CNTN4, LRRC4C, OPCML*). Two other pathways related to transmembrane structure, proteins that span the membrane of the cell. More than 80 genes were associated with these pathways. Seven genes, including *CHRNA4* and *BDNF*, were associated with a pathway that control ligand-gated ion channels. Finally, three other pathways, with over 50 genes, involved in the nucleus and transcription within the nucleus were associated. These genes implicate either signaling between cells (i.e., synapse functions, cell junctions, transmembrane structure), transcription within the nucleus, or immune system response (Table 3.1).

Table 3.1. DAVID-Identified Gene Clusters and Biological Systems for Smoking Initiation

Category	Term	Genes	p
UP_KEYWORDS	Synapse	*TENM2, GRID2, NLGN1, CHRNA4, CADM2, NRXN1, MAGI2, GRIK4, CABP1, GRIN2B, OLFM1, TRIM9, SDK1, GRIN2A, ERC2, CBLN4, LRRC4C*	0.008
UP_KEYWORDS	Cell junction	*FOCAD, TENM2, GRID2, NLGN1, CHRNA4, CADM2, NRXN1, MAGI2, DIXDC1, GRIK4, CABP1, GRIN2B, DPP4, OLFM1, TRIM9, SDK1, GRIN2A, ADAM15, DLC1, CTNNA2, ERC2, CBLN4, LRRC4C*	0.022
INTERPRO	IPR013098: Immunoglobulin I-set	*ROBO2, LINGO1, NEGR1, DCC, NTM, PXDNL, PTPRF, PTPRD, SDK1, NCAM1, CNTN4, LRRC4C, OPCML*	0.002
INTERPRO	IPR003598: Immunoglobulin subtype 2	*ROBO2, LINGO1, NEGR1, CADM2, DCC, NTM, PXDNL, SEMA3F, PTPRF, IGSF11, PTPRD, SDK1, NCAM1, CNTN4, LRRC4C, OPCML*	0.005
UP_SEQ_FEATURE	domain: Ig-like C2-type 3	*ROBO2, PTPRD, SDK1, NEGR1, DCC, NTM, NCAM1, PXDNL, CNTN4, PTPRF, OPCML*	0.015
SMART	SM00408: IGc2	*ROBO2, LINGO1, NEGR1, CADM2, DCC, NTM, PXDNL, SEMA3F, PTPRF, IGSF11, PTPRD, SDK1, NCAM1, CNTN4, LRRC4C, OPCML*	0.037
UP_KEYWORDS	Transmembrane helix	*THSD7B, GIMAP2, XYLT1, TMEM261, IGF1R, HS6ST3, EDNRA, CHCHD3, NOMO2, ENTPD1, EPHA7, SEMA6D, VRK2, CDYL, HLA-G, INPP4B, ADAM15, ADGRB2, ADGRB3, YME1L1, GPM6A, TMEM161B, CHRNA4, PCDH15, CACNA1D, EFNA5, TMEM242, DPP4, RHOT2, GRIN2A, GALR1, RNF217, CDH23, LRRC4C, ST3GAL1, BTN2A2, BDNF, CADM2, PTCH1, ST8SIA2, EDEM1, SLC4A10, GRIN2B, IGSF11, PTPRD, SDK1, FAM163A, TBXAS1, SLC26A7, CSPG5, FAT3, TMEM18, XKR6, DDR1, ROBO2, ALK, RYR2, NRP1, TENM2, MCTP1, TENM3, RNF13, TMEM110-MUSTN1, GRIK4, TMEM182, SPPL3, ELFN1, PTPRF, CLYBL, PTPRG, IMMP2L, SGCD, GRM8, CNGA3, GRID2, EED, TRPC4, DCC, ELOVL3, SLC39A11, ELOVL7, SORCS3, CDH12, SMIM21, ST6GALNAC3, OR10A6, CHST3, TMEM55A, NFAT5, NLGN1, FOCAD, SLC24A4, SLC24A3, NRXN1, TMPRSS3, PTGER3, NRXN3, SEZ6, NRXN2, SPG7, ERBB3, HECTD4, NCAM1, CNNM2, KCNJ3, LINGO1, PCDH9, SYT14, SLC28A3*	0.037

Table 3.1 (continued). DAVID-Identified Gene Clusters and Biological Systems for Smoking Initiation

Category	Term	Genes	p
UP_KEYWORDS	Glycoprotein	*DDR1, ROBO2, ALK, NRP1, TENM2, TENM3, ITPRIP, RNF13, THSD7B, CPXM2, XYLT1, GRIK4, TMEM182, CLU, ELFN1, PTPRF, IGF1R, PTPRG, HS6ST3, EDNRA, SGCD, GRM8, NOMO2, POSTN, ENTPD1, GRID2, EPHA7, DCC, SEMA6D, ELOVL3, SORCS3, HLA-G, OLFM1, ADAM15, ADGRB2, ADGRB3, CDH12, MAPT, ST6GALNAC3, OR10A6, HIST1H2BD, CHST3, GPM6A, SLC24A4, TMEM161B, NLGN1, ITIH3, SLC24A3, CHRNA4, NRXN1, NTM, TMPRSS3, PTGER3, NRXN3, PCDH15, SEZ6, NRXN2, CACNA1D, SEMA3F, EFNA5, THBS4, DPP4, IGSF21, GRIN2A, BRINP1, ERBB3, GALR1, SPOCK2, CDH23, NCAM1, CNNM2, ST3GAL1, KCNJ3, OPCML, LINGO1, BTN2A2, NEGR1, PCDH9, BDNF, CADM2, ST8SIA2, PTCH1, EDEM1, PXDNL, GRIN2B, PTPRD, IGSF11, SDK1, CSPG5, FAT3, CNTN4, CBLN4*	0.030
GOTERM_MF_DIRECT	GO:0050839~cell adhesion molecule binding	*PTPRD, NLGN1, POSTN, TENM2, TENM3, CADM2, NRXN1, NRXN3, NRXN2*	0.004
GOTERM_BP_DIRECT	GO:0007416~synapse assembly	*GPM6A, SDK1, NLGN1, BDNF, NRXN1, SPOCK2, NRXN3, NRXN2*	0.042
UP_KEYWORDS	Ligand-gated ion channel	*RYR2, GRIN2A, GRID2, CHRNA4, GRIK4, CNGA3, GRIN2B*	0.030
UP_KEYWORDS	Nucleus	*ZBTB20, CLU, IKZF4, RPS6KA4, CHCHD3, PPP4R2, SALL3, GRAP2, CHEK2, ZNF207, MACROD2, BTRC, SMARCC1, EBF1, RFX3, ARID5B, HDGFRP2, VRK2, TNNI3K, CDYL, OVOL1, DGKZ, POU3F2, FOXP1, CAMKMT, MAML3, TOX, MAD1L1, HIST1H2BD, KHDRBS3, CUL3, TSHZ1, NOLC1, FOXO3, FHIT, PHF21A, ZSCAN12, ZNF789, CAMTA1, ZNF423, IP6K2, PHC2, ZFHX3, CPSF6, SMAD3, BCL11B, BCL11A, ZBTB16, CEP350, RANBP17, MICAL2, FOXN3, ZBTB9, ICK, DAZL, MLLT10, MSRA, LSM8, KIF4B, TMEM18, BRWD1, RERE, BARHL2, TENM2, BNC2, RNF13, BNC1, CELF2, SETD7, CHD3, AFF3, SMG6, ZNF407, SIX3, REV3L, ZNF644, EED, RMI1, SLC39A11, NAV2, MED27, MMS22L, NAV3, RARB, RBM20, NFAT5, CTDP1, ZBTB46, NOL4, FGD2, TRA2B, POLR2F, FPGT-TNNI3K, IGF2BP2, CTNNA2, ASCC3, RUNX1T1, MCRS1, CABLES1, WWP2, ZIC4, CHFR, REST, KANSL1*	0.030

Table 3.1 (continued). DAVID-Identified Gene Clusters and Biological Systems for Smoking Initiation

Category	Term	Genes	p
UP_KEYWORDS	Transcription	*RERE, BARHL2, TENM2, BNC2, BNC1, SETD7, ZBTB20, CHD3, AFF3, IKZF4, CHCHD3, SALL3, CHEK2, ZNF407, SIX3, ZNF644, SMARCC1, EED, LMO3, EBF1, ARID5B, RFX3, CDYL, OVOL1, MED27, POU3F2, FOXP1, RARB, MAML3, NFAT5, KHDRBS3, ZBTB46, TSHZ1, FOXO3, FHIT, PHF21A, ZSCAN12, ZNF789, POLR2F, CAMTA1, ASCC3, ZNF423, RUNX1T1, ZFHX3, SMAD3, BCL11B, BCL11A, ZBTB16, MCRS1, FOXN3, ZBTB9, MLLT10, REST, TMEM18, BRWD1*	0.030
UP_KEYWORDS	Transcription regulation	*RERE, BARHL2, TENM2, BNC2, BNC1, SETD7, ZBTB20, CHD3, AFF3, IKZF4, CHCHD3, SALL3, CHEK2, ZNF407, SIX3, ZNF644, SMARCC1, EED, LMO3, EBF1, ARID5B, RFX3, CDYL, OVOL1, MED27, POU3F2, FOXP1, RARB, MAML3, NFAT5, KHDRBS3, ZBTB46, TSHZ1, FOXO3, FHIT, PHF21A, ZSCAN12, ZNF789, CAMTA1, ASCC3, ZNF423, RUNX1T1, ZFHX3, SMAD3, BCL11B, BCL11A, ZBTB16, MCRS1, FOXN3, ZBTB9, MLLT10, REST, BRWD1*	0.037

Note. Category refers to the original database or resource where the term originates. Term refers to the detailed function in an annotation source.

Conclusions. SI is complex, polygenic trait with many genes influencing the phenotype. Genes that are responsible for acetylcholine receptor function (e.g., *CHRNA4*) were consistently associated via the DAVID analysis with SI.

2. Nicotine Metabolism

Almost 90% of nicotine is metabolized in the liver into six metabolites [177]. Approximately 70-80% of nicotine metabolizes to cotinine (the primary nicotine metabolite) and as such, most studies focus on this metabolite. The production of cotinine occurs in two steps: the first is metabolizing to nicotine-$\Delta^{1'(5')}$-iminium ion with cytochrome P450 2A6 (CYP2A6), and the second step involves catalyzing the nicotine iminium ion by a cytoplasmic aldehyde oxidase. These processes are predominately driven by the Cytochrome P450 enzymes (e.g., CYP2A6, Figure 3.1). Cotinine is excreted through urine, blood, and saliva. Salivary cotinine is highly correlated with blood cotinine making salivary cotinine measurement is a less invasive and cost-effect method of measuring cotinine [178]. Urinary cotinine is generally four to six times more concentrated, making urinary cotinine a more sensitive measure of cotinine [179]. Cotinine has a half-life of roughly 20 hours but is detectable for up to a week after exposure to nicotine. The process of metabolizing cotinine occurs across three separate pathways (Figure 3.1). Although there are 6 metabolites derived from nicotine, genetically-informed studies of metabolism have focused on cotinine, 3'-hydroxycotinine and its downstream metabolite 3'-hydroxycotinine glucuronide, 5'-hydroxycotinine, and cotinine glucuronide. The first step utilizes CYP2A6 and CYP2A13 enzymes to produce 3'-hydroxycotinine; and 5

enzymes from the UDP-glucuronosyltransferase family (UGT1A4, UGT1A9, UGT2B7, UGT2B4, and UGT2B15) to produce 3'-hydroxycotinine glucuronide. The highest concentrations of this metabolite are found in urine. A third metabolite, cotinine glucuronide, is produced when UGT1A1, UGT1A4, UGT1A9, and UGT2B10 enzymes act on cotinine to produce cotinine glucuronide. 5'-hydroxycotinine is produced from CYP2A6 and is frequently detected in urine. These products are all metabolites of cotinine, which allows for an estimation of the rate of nicotine metabolism. A nicotine metabolite ratio (NMR) is also able to be calculated as the ratio of cotinine to 3'hydroxycotinine.

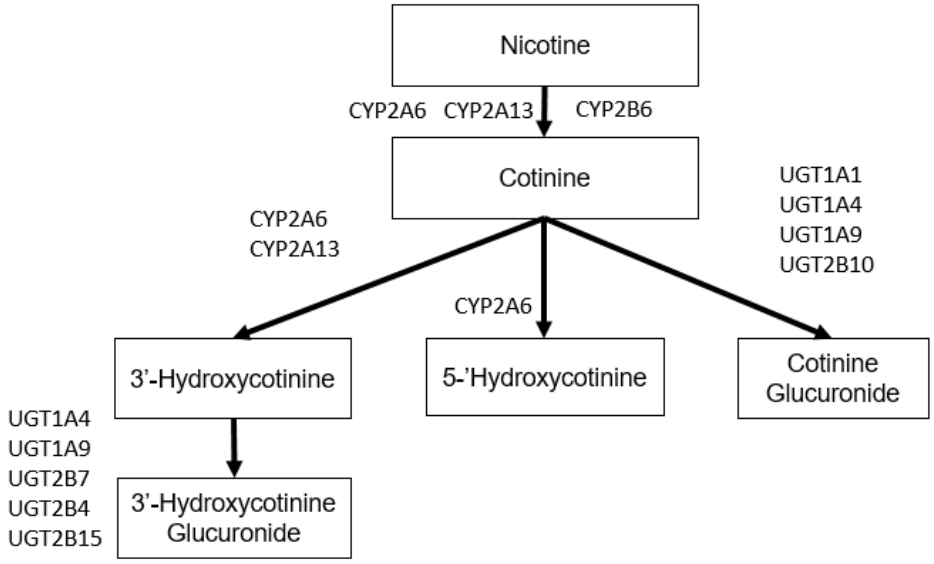


Figure 3.1. Nicotine Metabolism Pathway with Enzymes Responsible for the Pathways.

Twin Studies.

One twin study of nicotine metabolism has been published [180]. This study used blood and urine samples to assess cotinine. Additive genetic influences (A) accounted for nearly 60% of the variance in nicotine metabolism in twins, while unique environmental influences (E) accounted for the remaining 40% with no shared environmental

influences. Further examination of the twin correlations (rMZ = 0.68, rDZ = 0.25) also suggested some non-additive genetic effects, such as dominance effects (as evidenced by the DZ twin correlation being less than half of the rMZ, or 0.34), contributing to nicotine metabolism. This study did not have the power to detect non-additive effects and encouraged further research to study this source of variance.

Single and Multi-Sample Genome-Wide Association Studies of Nicotine Metabolism

One single-sample GWAS was conducted using nicotine metabolism as the outcome variable and reported 1,732 genome-wide significant associations [181]. This study examined nicotine metabolism as measured by cotinine glucuronide. Specifically, the outcome was defined as the ratio of cotinine glucuronide to total cotinine (similar to the NMR) which was associated with 1,241 SNPs at the genome-wide significant level. In an effort to reduce the number of associations, researchers used the genome-wide significant SNPs in a stepwise regression to determine which SNPs accounted for the majority of the variance. Of the 1,241 SNPs associated with cotinine glucuronidation, 15 were associated after running stepwise regression. Four of these SNPs were in or near *UGT2B10* (on chromosome 4) which encodes for the enzyme UDP-glucuronosyltransferase 2B10, which is responsible for the glucuronidation of both nicotine and cotinine. The strongest single SNP association was for rs115765562 (p = 1.6×10^{-155}), an intergenic location on chromosome 4. This SNP is highly correlated with a *UGT2B10* splice site variant, rs116294140, which together with rs6175900 (Asp67Tyr) explains 24.3% of the variation in cotinine glucuronidation [182]. Further results indicated *FAM107B* (Family with Sequence Similarity 107 Member B), *CERS3* (Ceramide

Synthase 3), and *SLC2A14* (Solute Carrier Family 2 Member 14) had intronic variants associated with nicotine metabolism. All other reported results were in intergenic regions.

In addition to cotinine glucuronidation, this study also examined nicotine glucuronidation, which is defined as the ratio of nicotine *N*-glucuronide to total nicotine. Using this definition of the outcome, 492 SNPs showed genome-wide significance. After putting genome-wide significant SNPs into the regression model, two SNPs remained significant after the stepwise regression procedure. Both significant SNPs were intronic variants. The first was located on chromosome 4 in *UGT2B10*, the same gene associated with cotinine glucuronidation. The other significant SNP was intronic to *SHFM1* (Split hand/foot malformation type 1), which encodes for 26S proteasome complex subunit DSS1. It is thought this protein plays a role in the completion of the cell cycle.

One GWAS was conducted using a multi-sample (5 studies, N = 5,185 participants) GWAS design [183]. This study reported 1,267 genome-wide significant hits for NMR, COT, and 3HC. These SNPs were all located on either chromosome 4 or chromosome 19. Those on chromosome 4 were all located within *TMPRSS11E* gene (Transmembrane Serine Protease 11E), a novel finding for nicotine metabolism. Previous studies of this gene have indicated a role in cognition [184]. The findings on chromosome 19 were tightly linked to *CYP2A6* with the most significant two SNPs being located within this gene. Additionally, there were several highly significant SNPs either near *CYP2A6* (rs113288603) or near the *CYP2B6* gene. Both *CYP* genes have previously been reported to be significant with substance metabolism [37,185–189].

Functional Analysis. No DAVID analysis was performed for nicotine metabolism because only two studies examined the phenotype.

3. Quantitative Measures

The goal of measuring smoking quantity is to establish an estimate of the amount of nicotine exposure. Nicotine content in typical conventional cigarettes is generally regulated in the United States. A typical cigarette sold in the US contains approximately 11mg of nicotine [190]. Consequently, one of the most common ways (5 of 5 reported studies) of measuring regular smoking is to measure the number of cigarettes smoked per day (CPD). This value can be transformed into the number of cigarettes consumed during other time frames (e.g., per week, month, or year) (1 of 5 GWASs). Other studies have also examined maximum cigarettes smoked per day when the participant was at maximum smoking levels (1 of 5 studies). Most studies (4/5) chose to bin the CPD measure either as a binary variable based on a threshold (e.g., smoking at least 10 cigarettes per day) or binned into other categories (e.g., 1-10 CPD, 11-20, 21-30, etc.; 1 out of 5 studies).

Twin Studies

Continuous Measures of Cigarettes per Day. To date, 7 adult twin studies have measured CPD. In general, these studies indicate a higher proportion of variance due to genetic factors compared to measures of smoking initiation. A study of Dutch twins ages 12-24 (mean age = 17.7, N = 1,676 twin pairs) reported a substantial estimate of additive genetic influences (A = 0.86, 95% CI = 0.70-0.94) and a much smaller

contribution due to shared environmental effects (C = 0.54, 95% CI 0.25-0.95) for average number of cigarettes smoked per day [191]. Similarly, in a study of Australian twins, researchers used average number of cigarettes smoked per day as the measure of CPD and reported significant genetic (A = 0.40) and shared environmental effects (C = 0.12) [104]. A study of US twins (N = 94 pairs) also used average number of cigarettes smoked as the outcome; however, this study also asked twins to report their average CPD for the year they smoked the heaviest. This study reported significant genetic (A = 0.40) but no shared environmental effects [106]. Further research in a sample of African-American adults with a mean age of 46.9 (SD = 13.9; N = 200 same-sex twin pairs) reported a significant heritability for pack-years of smoking, though this estimate was lower than when measured as continuous CPD (A = 0.43) (Whitfield 2007). These results provide evidence of genetic influences on cigarette quantity using a continuous measure.

Ordinal Measures of Cigarettes per Day. Other studies have chosen various ways of categorizing smoking quantity twins from the Netherlands (N = 3,657 pairs; average age 28.7 years for DZ twins and 24.7 for MZ twins) by categorizing CPD into 5 ordinal bins; the study reported significant genetic (A = 0.51) and shared environmental effects (C =0.30) [136]. Another study of Finnish twins (N = 9,880 pairs; age range 24-88) dichotomized smoking quantity with the threshold of smoking 20 cigarettes per day. This study reported significant genetic effects (A = 0.54) but no shared environmental effects. A final US study examined CPD as a continuous measure from 1,078 twin pairs (aged 18-25) and reported significant genetic (A = 0.50) but not shared environmental

influences [192]. Regardless of the measure of quantity smoked, there was relatively consistent agreement in the estimate of A with a range of roughly 0.40 to 0.54.

Single and Multi-Sample Genome-Wide Association Studies of Quantitative Measures

The earliest GWAS of smoking quantity were conducted in 2014-2015 with low sample sizes (N ranged from ~ 500 individuals to a little over 5,000) [146,193]. Unsurprisingly, these early studies did not report any genome-wide significant associations. These studies examined cigarettes smoked per day, the maximum number of cigarettes smoked per day, and the average number of cigarettes smoked per day; with each measure self-reported by participants. Further, these studies were of the single-sample or multi-sample nature, with the multi-sample study utilizing a discovery and replication sample (both N < 400 participants, massively underpowered).

A more recent single-sample GWAS completed in an older adult sample[194] (age range: 62-81) was also underpowered (N = 2,063) but did report genome-wide significant associations between several variants and cigarettes per day. Three variants (rs4300632, rs11074386, rs11074388) in the gene *CLEC19A* (C-Type Lectin Domain Family 10 Containing 19A) were associated at the genome-wide level. This gene plays a role in carbohydrate bonding (the oxidation of one or more hydroxy groups).

A multi-sample (N = 3 samples) study of middle to older adults (age range: 44-81; N = 13,551) also reported genome-wide significant associations between the number of cigarettes smoked per day with *CHRNA3*, *CHRNA5*, and *CHRNB4* [195]. These genes code for different subunits of the nicotinic acetylcholine receptors (nAChR). However, all the resulting receptors react to nicotine as an acetylcholine agonist,

compelling the receptor to react. Additional results from this study included the aminoglycoside phosphotransferase domain-containing protein 1 (*AGPHD1)*, which aids in the transferring of phosphorous-containing groups and lysine degradation. In addition, iron responsive element binding protein 2 (*IREB2)* which assists in regulating the translation and stability of mRNAs that affect iron homeostasis was significantly associated with smoking quantity.

Another study assessed smoking quantity by asking participants the number of cigarettes smoked and then categorized participants as never smokers, former smokers, nondaily smokers, and daily smokers [196]. This multi-sample (N = 2) study and used data from 12,804 Hispanic participants ages 18-74. Similar to prior GWASs, this study reported genome-wide significant associations between their measure of nicotine use (i.e., self-reported category) and the nAChR genes, specifically *CHRNA3*, *CHRNA5*, and *CHRNB4*. Importantly, this study provides convergent evidence that the nAChR gene cluster is important when examining the molecular genetic influences on quantitative measures of smoking [196].

Consortia-Based Genome-Wide Association Results of Quantitative Measures

One consortium has examined CPD, though the consortia was small by current standards (N = 5,354). This study did not report any genome-wide significant results for the CPD phenotype.

Functional Analysis. DAVID analysis identified 19 pathways with significant association with genome-wide significant SNPs related to quantitative measures of

nicotine (e.g., cigarettes per day). Each of these associated pathways included the genes *CHRNA3*, *CHRN5*, and *CHRNB4*. *PARD3* and *TMC5* were also represented in two pathways each. In general, these pathways were involved in the behavioral response to nicotine as well as acetylcholine receptor action. These actions included ion channel activity (i.e., transmembrane transfer of an ion via the channel that opens when a specific ligand has been bound by the channel complex) as well as ion transport (i.e., directed movement of charged atoms). Genes that are involved in acetylcholine receptor action (e.g., *CHRNA2*) have previously been associated with GWAS results for quantitative measures of smoking such as CPD.

Table 3.2. DAVID-Identified Gene Clusters and Biological Systems for Cigarettes Per Day

Category	Term	Genes	*p*
GOTERM_BP_DIRECT	GO:0035095~behavioral response to nicotine	*CHRNA3, CHRNB4, CHRNA5*	7.78E-04
GOTERM_BP_DIRECT	GO:0006811~ion transport	*CHRNA3, CHRNB4, CHRNA5, TMC5*	0.00125425
INTERPRO	IPR002394:Nicotinic acetylcholine receptor	*CHRNA3, CHRNB4, CHRNA5*	0.00157543
GOTERM_MF_DIRECT	GO:0004889~acetylcholine-activated cation-selective channel activity	*CHRNA3, CHRNB4, CHRNA5*	9.25E-04
GOTERM_MF_DIRECT	GO:0015464~acetylcholine receptor activity	*CHRNA3, CHRNB4, CHRNA5*	9.25E-04
GOTERM_CC_DIRECT	GO:0005892~acetylcholine-gated channel complex	*CHRNA3, CHRNB4, CHRNA5*	0.00274975
GOTERM_MF_DIRECT	GO:0042166~acetylcholine binding	*CHRNA3, CHRNB4, CHRNA5*	9.25E-04
INTERPRO	IPR027361:Nicotinic acetylcholine-gated receptor, transmembrane domain	*CHRNA3, CHRNB4, CHRNA5*	0.00157543
GOTERM_BP_DIRECT	GO:0007271~synaptic transmission, cholinergic	*CHRNA3, CHRNB4, CHRNA5*	0.00476613
GOTERM_MF_DIRECT	GO:0015276~ligand-gated ion channel activity	*CHRNA3, CHRNB4, CHRNA5*	0.00134041
GOTERM_BP_DIRECT	GO:0098655~cation transmembrane transport	*CHRNA3, CHRNB4, CHRNA5*	0.00603842

Table 3.2 (continued). DAVID-Identified Gene Clusters and Biological Systems for Cigarettes Per Day

Category	Term	Genes	*p*
INTERPRO	IPR018000:Neurotransmitter-gated ion-channel, conserved site	*CHRNA3, CHRNB4, CHRNA5*	0.00212964
INTERPRO	IPR006029:Neurotransmitter-gated ion-channel transmembrane domain	*CHRNA3, CHRNB4, CHRNA5*	0.00212964
INTERPRO	IPR006201:Neurotransmitter-gated ion-channel	*CHRNA3, CHRNB4, CHRNA5*	0.00212964
INTERPRO	IPR006202:Neurotransmitter-gated ion-channel ligand-binding	*CHRNA3, CHRNB4, CHRNA5*	0.00212964
UP_KEYWORDS	Ion channel	*CHRNA3, CHRNB4, CHRNA5, TMC5*	0.02406641
UP_KEYWORDS	Ligand-gated ion channel	*CHRNA3, CHRNB4, CHRNA5*	0.02406641
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	*CHRNA3, CHRNB4, CHRNA5, PARD3*	0.0212414
GOTERM_CC_DIRECT	GO:0030054~cell junction	*CHRNA3, CHRNB4, CHRNA5, PARD3*	0.03088171

Note. Category refers to the original database or resource where the term originates. Term refers to the detailed function in an annotation source.

4. Nicotine Dependence

Researchers have long been concerned with classification of nicotine dependence (ND) and have several valid and reliable tools with which to do so. The most common (2 of 5 studies) of these scales is the Fagerström Test for Nicotine Dependence (FTND) [197], also referred to more recently as the Fagerström Test for Cigarette Dependence (FTCD; 1 of 5 studies) [198,199]. Further, researchers have also used the time to first cigarette in the morning (TTFC) as a proxy for ND, as it is assumed that individuals who are more dependent will need a cigarette sooner (typically within five minutes of waking) than individuals who use but aren't dependent (1 of 5 studies). Finally, symptoms of ND (spent a great deal of time getting, using or getting over effects of CIG; used CIG more often or in larger amounts than intended; built up a tolerance so that the same amount of CIGs has less effect than before; CIG use kept you from working, going to school, taking care of children, or engaging in recreational activities; CIG use caused emotional or psychological problems; CIG use caused health problems; wanted or tried to stop or cut down CIG use), as defined by the American Psychiatric Association [200] the DSM-IV-TR, binned into categories based on the sum of symptoms has been used (1 of 5 studies) to assess ND.

Twin Studies of Nicotine Dependence

Fagerström Test for Nicotine Dependence. Three twin studies examining ND via the FTND have most often been used to estimate A, C, and E for ND. A study of Virginia twins (N = 6,805; age range = 20-59) reported significant genetic effects (A = 0.67) but no shared environmental effects [135]. A contemporary study was completed in a

sample of Dutch twins (N = 1,572, mean age = 30.2) and reported similar findings [201]. Specifically, this study found significant genetic effects (A = 0.75) but no influences due to shared environmental effects [201]. A study of Swedish twins (N = 5,040; age range = 22-57) reported lower, but significant effect, of additive genetic influences (A = 0.39, 95% CI = 0.29-0.49) [202]. Consistent with the previous studies, there was no effect of shared environmental influences.

DSM Symptoms. Three twin studies used the DSM definition of ND. Typically, twin studies have examined the number of symptoms rather than using the diagnosis of ND. Twin studies are often an initial study design employed in genetic epidemiology and the DSM is constantly evolving. As such, twin studies may have used different versions of the DSM to assess the symptoms. An older study of Vietnamese twins from the US (N = 9,414; age range: 35-53) utilized the Mental Health Diagnostic Interview Schedule III-revised (DIS-III-R) which is a standardized interview based on the DSM-III-R [138]. Nicotine dependence was assessed as an ordinal variable with 3 bins based on a sum score of symptoms endorsed: 0-2 symptoms (no dependence), 3-4 symptoms (mild dependence), 5-7 symptoms (high/severe dependence). This study reported significant genetic (A = 0.55, 95% CI = 0.40-0.61) and smaller shared environmental factors (C = 0.04; 95% CI = 0.00-0.17).

More recent research transitioned to using the DSM-IV for examining nicotine dependence. A study conducted in 2016 utilized 7,285 Australian twins with a mean age of roughly 30 to examine ND [203]. A sum score of ND items from the DSM-IV (range 0-7) was used. This study reported significant genetic effects (A = 0.57, 95% CI = 0.43-0.71) and nonsignificant effects of the shared environment (C = 0.02, 95% CI = -0.10-0.14).

Another study utilizing the DSM-IV used the diagnosis of ND, which they defined as having 3 or more symptoms (of the 7 total) in the past 12-months. Using a sample of 5,580 Australian twins (age range: 11-18), this study reported significant genetic effects (A = 0.56, 95% CI = 0.40-0.63) but no shared environmental [204] .

Heaviness of Smoking Index. Heaviness of smoking index (HSI) is an index of two items, the time to first cigarette and cigarettes smoked per day [205], which was used by one twin study. The HSI showed a greater influence of genetic effects (A = 0.71) than when ND was measured via the FTND [204]. There was no significant effect of shared environmental factors for the HSI.

Single and Multi-Sample Genome-Wide Association Studies of Nicotine Dependence

*Fagerström Test for Nicotine Dependence***.** GWAS that have studied ND have been multi-sample in nature (four in total), with only one single sample study employed. The first GWAS was published in 2012 [206], using the FTCD (Fagerström Test for Cigarette Dependence) as the measure of ND. Participants (N = 3,365) were selected from samples under the SAGE project, including the Collaborative Genetic Study of Nicotine Dependence (COGEND), the Collaborative Study on the Genetics of Alcoholism (COGA), and the Family Study of Cocaine Dependence (FSCD). The authors report seven genome-wide significant associations all in or near *CHRNB3* on chromosome 8 with an additional three variants suggestive ($p < 10^{-7}$). The most significant association was rs1451240 and was protective against ND (OR = 0.65; 95% CI = 0.56-0.76, $p = 2.44 \times 10^{-8}$).

The FTND was used to quantify ND in a multi-sample (N = 2) study with 7,646 individuals [207]. This study reported 67 genome-wide significant associations (Table S3.3). Nearly all (58 number of 66 variants) of the detected variants were located on chromosome 14. There were four significant associations on chromosome 8. Additionally, two variants with significant associations were located on chromosome 18. The two results on chromosome 18 were intergenic, with no known function. Other results on chromosome 8 were all located within *DLC1*, Deleted in Liver Cancer 1 which has been implicated in hepatocellular carcinomas. The majority of the associations on chromosome 14 were located within *FAM179B* (most significant SNP: rs114962601, $p = 6.53 \times 10^{-10}$), which has been implicated in the function of primary cilia (an organelle in eukaryotic cells which serve as sensory organelles). Three additional genes were implicated in addition to *FAM179B*. *KLHL28*, Kelch Like Family Member 28, which influences protein binding and has been shown to be differentially expressed in neural tissue [208]. *C14orf18*, chromosome 14 open reading frame 18, a gene with currently uncharacterized function. *FANCM*, Fanconi anemia complementation group M, a gene responsible for DNA repair. *PRPF39*, pre-mRNA processing factor 39, which is involved in processing of RNA and mRNA.

***DSM*.** The most current GWAS for ND as measured using DSM criteria was a single sample study of twins from Finland using DSM-IV diagnosis as well as symptom count [194]. This study of 1,715 individuals (mean age: 55) reported 2 genome-wide significant associations. The most significant association was detected for a SNP located on chromosome 18 which is in the *SLC14A2* gene (rs117354958, $p = 3.55 \times 10^{-8}$). The product of this gene is the solute carrier family 14 member 2 protein. It has been

implicated in protein binding as well as urea transport. The other significant SNP was located in *AP2A2*, rs369708413, $p = 6.58 \times 10^{-8}$. The product of this gene, adaptor related protein complex 2 subunit alpha 2 (*AP2A2*) is a protein that assists in other protein binding and transport.

Consortia-Based Genome-Wide Association Results of Nicotine Dependence

One consortia-based based GWAS examined both the FTND and the time to first cigarette (TTFC) in the morning [209]. For the FTND, there was only one significant variant (rs16969968) located in the *CHRNA5* gene, highlighting again the important of the nAChR genes. This study further examined gene sets (defined *a priori*) and reported the *CHRNA3-CHRNB4-CHRNA5* gene set was statistically associated with FTND ($p = 3.96 \times 10^{-19}$).

When examining the TTFC phenotype, similar results emerged. The previous SNP from *CHRNA5* (rs16969968) was also highly associated with TTF (6.21×10^{-9}); however, there were additional variants significantly associated with TTFC. These included SNPs from *SORBS2* (sorbin and SH3 domain containing 2, rs28567706; involved in cellular structure and RNA binding), *AA333164* (rs117029742, a long non-coding RNA with unknown function), and *BG182718* (rs10133756, a long non-coding RNA with unknown function). When examining gene sets, there were two sets that were associated with TTFC. The first was the same *CHRNA3-CHRNB4-CHRNA5* gene set that was associated with the FTND (6.21×10^{-9}). Also, the gene set *CHRNB3-CHRNA6* was significantly associated with TTFC (8.83×10^{-8}), providing further evidence for the role of the nAChR genes.

Functional Analysis. DAVID analyses of genome-wide significant SNPs for nicotine dependence identified 30 significantly associated pathways (Table 3.3). Each significant pathway was represented with several of the nAChR genes (e.g., *CHRNA3*, *CHRNB4*). As such, it is unsurprising that the significantly associated pathways (13 pathways) overwhelmingly were involved in some sort of acetylcholine activity including binding, ion channel activation, or gate control (all Benjamini-Hochberg $p < 1.06 \times 10^{-6}$). Other associated pathways were involved with the behavioral or molecular response to nicotine which is reasonable given that nicotine acts as an agonist of the nAChRs. This leads to excess dopamine in the reward pathways of the brain, culminating in pleasurable feelings, which may help explain the dependence on the drug. Further network enrichments were found in postsynaptic cell membrane function and neural [949527]6,118 individuals from the Million Veterans Project (MVP; age range = 58-64) and also categorized individuals as ex-smokers on the basis of self-reported questions (i.e., current versus former) [122]. This study reported 8 genome-wide significant SNPs. 4 of these SNPs were in intergenic regions (rs11210228, rs34735365, rs77648866, rs112270518), 1 was in the 5' untranslated regions (5'UTR) of *CHRNA2* (rs2565060), and the other 3 were intronic to various genes. These intronic SNPs were located on *DRD2* (rs61902807), *CYP2A6* (rs56113850), and *CHRNA4* (rs6011779), genes which have been implicated in various tobacco use phenotypes.

Table 3.3. DAVID-Identified Gene Clusters and Biological Systems for Nicotine Dependence

Category	Term	Genes	*p*
INTERPRO	IPR002394:Nicotinic acetylcholine receptor	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	5.31E-09
GOTERM_MF_DIRECT	GO:0004889~acetylcholine-activated cation-selective channel activity	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	9.33E-09
GOTERM_MF_DIRECT	GO:0015464~acetylcholine receptor activity	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	9.33E-09
GOTERM_CC_DIRECT	GO:0005892~acetylcholine-gated channel complex	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.74E-08
INTERPRO	IPR027361:Nicotinic acetylcholine-gated receptor, transmembrane domain	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.31E-08
GOTERM_MF_DIRECT	GO:0042166~acetylcholine binding	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.21E-08
GOTERM_BP_DIRECT	GO:0007271~synaptic transmission, cholinergic	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	3.73E-07
GOTERM_MF_DIRECT	GO:0015276~ligand-gated ion channel activity	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	5.15E-08
GOTERM_BP_DIRECT	GO:0098655~cation transmembrane transport	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	7.25E-07
INTERPRO	IPR018000:Neurotransmitter-gated ion-channel, conserved site	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.71E-07
INTERPRO	IPR006029:Neurotransmitter-gated ion-channel transmembrane domain	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.71E-07
INTERPRO	IPR006201:Neurotransmitter-gated ion-channel	*CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6*	1.71E-07

Table 3.3 (continued). DAVID-Identified Gene Clusters and Biological Systems for Nicotine Dependence

Category	Term	Genes	*p*
INTERPRO	IPR006202:Neurotransmitter-gated ion-channel ligand-binding	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	1.71E-07
UP_KEYWORDS	Ligand-gated ion channel	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	1.06E-05
GOTERM_BP_DIRECT	GO:0035095~behavioral response to nicotine	CHRNA3, CHRNB4, CHRNA5, CHRNA4	1.50E-05
UP_KEYWORDS	Cell junction	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, DLC1, CHRNA6, SORBS2, TMEM163, DSC3	1.65E-05
UP_KEYWORDS	Postsynaptic cell membrane	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	1.93E-04
GOTERM_BP_DIRECT	GO:0007274~neuromuscular synaptic transmission	CHRNB3, CHRNA5, CHRNA4, CHRNA6	3.43E-04
UP_KEYWORDS	Synapse	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6, TMEM163	2.81E-04
GOTERM_CC_DIRECT	GO:0045211~postsynaptic membrane	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	5.57E-04
GOTERM_BP_DIRECT	GO:0035094~response to nicotine	CHRNB4, CHRNB3, CHRNA4, CHRNA6	8.01E-04
GOTERM_BP_DIRECT	GO:0007165~signal transduction	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, DLC1, PDE2A, CHRNA6, APOL3, VAV2	8.01E-04
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	9.27E-04
GOTERM_CC_DIRECT	GO:0030054~cell junction	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6, TMEM163	0.00137524
UP_KEYWORDS	Ion channel	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	0.00321989

Table 3.3 (continued). DAVID-Identified Gene Clusters and Biological Systems for Nicotine Dependence

Category	Term	Genes	*p*
KEGG_PATHWAY	hsa04725:Cholinergic synapse	CHRNA3, CHRNB4, CHRNA4, CHRNA6	0.00736798
GOTERM_BP_DIRECT	GO:0006811~ion transport	CHRNA3, CHRNB4, CHRNA5, CHRNA4	0.02273874
UP_KEYWORDS	Cell membrane	SLC14A2, CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6, PDE2A, SORBS2, AP2A2, RPE65, DSC3, MFSD2A	0.01730657
UP_KEYWORDS	Transport	CHRNA3, CHRNB4, SLC14A2, CHRNB3, CHRNA5, CHRNA4, CHRNA6, AP2A2, APOL3, MFSD2A	0.02228539
UP_KEYWORDS	Ion transport	CHRNA3, CHRNB4, CHRNB3, CHRNA5, CHRNA4, CHRNA6	0.02793515

Note. Category refers to the original database or resource where the term originates. Term refers to the detailed function in an annotation source.

5. Smoking Cessation

Smoking cessation (SC) is the final point on the smoking continuum, representing abstinence from tobacco use. As cessation is a process, it is possible for individuals to relapse and return to regular smoking or nicotine dependence. Abstinence is rarely achieved on the first cessation attempt. Cessation is a process and may have several episodes of relapse or returning to regular smoking and nicotine dependence [210].

Twin Studies

One study has examined SC in a twin study design [211]. This study examined the genetic and environmental influences on failed SC (i.e., an individual who attempted to quit, but relapsed) attempts. This study of 4,112 twins from the Vietnam Era Registry estimated significant genetic effects on failed smoking cessation attempts (A = 0.54, 95% CI = 0.40-0.62). Variance due to shared environmental influences did not significantly contribute to SC.

Single and Multi-Sample Genome-Wide Association Studies of Smoking Cessation

Two GWASs have been performed using single samples. The first was from a sample of Bangladeshi adults (N = 5,354) aged between 18 and 75 years old [145]. Measuring SC as a self-reported smoking status (i.e., current or former smoker), researchers reported no genome-wide significant findings. The other single sample GWAS utilized 286,118 individuals from the Million Veterans Project (MVP; age range = 58-64) and also categorized individuals as ex-smokers on the basis of self-reported questions (i.e., current versus former) [122]. This study reported 8 genome-wide significant SNPs. 4 of

these SNPs were in intergenic regions (rs11210228, rs34735365, rs77648866, rs112270518), 1 was in the 5' untranslated regions (5'UTR) of *CHRNA2* (rs2565060), and the other 3 were intronic to various genes. These intronic SNPs were located on *DRD2* (rs61902807), *CYP2A6* (rs56113850), and *CHRNA4* (rs6011779), genes which have been implicated in various tobacco use phenotypes.

Consortia-Based Genome-Wide Association Results of Smoking Cessation

Three separate consortia examined SC, all categorizing former smokers using self-report. The Tobacco and Genetic Consortium was the first to analyze smoking cessation using data from 16 studies (N = 74,053) [156]. They reported one genome-wide significant SNP on *DBH* (rs3025343, $p = 3.56 \times 10^{-8}$). The GWAS and Sequencing Consortium of Alcohol and Nicotine (GSCAN) also performed a GWAS on SC with a larger sample (N = 547,219) from 24 different studies [163]. GSCAN reported 24 genome-wide significant SNPs for SC. Half (12) of these significant SNPs were intergenic. Two SNPs were intronic to *CHRNA4* (rs6011779, rs4809543) with another intronic to *CHRNA5* (rs518425). Additionally, there was one SNP that was intronic to *CYP2A6* (rs56113850). Other genes with genome-wide significant SNPs (one each) were *DBH* (rs1611124), *SOX6* (rs7109376), *SEMA6D* (rs591143), *ISL2* (rs3866543), *PDE1C* (rs7778443), *IRF4* (rs12203592), *KLHDC8B* (rs7617480), and *PPP6C* (rs12378015). A final consortium used data from 61 studies (N = 622,409; N = 121,543 former smokers) to examine smoking cessation [174]. This study reported two genome-wide significant SNPs for smoking cessation on *TOB2* (transducer of ERBB2; rs202664) and *CCDC141* (Coiled-Coil Domain Containing 141; rs150493199). *TOB2* is involved in the regulation

of the cell cycle progression (i.e., from cell formation to cell death), while *CCDC141* is involved with cell adhesion.

Functional Analysis. Functional analysis of five studies conducted with DAVID indicated 41 significantly associated pathways involved in smoking cessation (Table 3.4). Similar to other tobacco use phenotypes, acetylcholine related pathways were significantly associated. Pathways dealing with acetylcholine receptor structure and activity were associated with *CHRNA3*, *CHRNB4*, *CHRNA5*, *CHRNA4*, and *CHRNA6*, including the most associated pathway which was associated with the nicotinic acetylcholine receptor ($p = 6.11 \times 10^{-7}$). A secondary pathway related to chemical synaptic transmission was associated with *CHRNA5*, *CHRNA4*, *CHRNA6*, *LPAR3*, and *DBH*. Thus, two general pathways emerged from the DAIVD analysis. Most pathways were involved in neuronal signaling, either via receptor or ligand gate control and synaptic functions. Another set of pathways was associated with iron and oxygen flow. These pathways were also associated with Cytochrome P450, an enzyme encoded by CYP2A6, previously associated with smoking cessation.

Table 3.4. DAVID-Identified Gene Clusters and Biological Systems for Smoking Cessation

Category	Term	Genes	*p*
INTERPRO	IPR002394:Nicotinic acetylcholine receptor	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	6.11E-07
GOTERM_CC_DIRECT	GO:0005892~acetylcholine-gated channel complex	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.47E-06
INTERPRO	IPR027361:Nicotinic acetylcholine-gated receptor, transmembrane domain	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.06E-06
GOTERM_MF_DIRECT	GO:0004889~acetylcholine-activated cation-selective channel activity	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.24E-06
GOTERM_MF_DIRECT	GO:0015464~acetylcholine receptor activity	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.24E-06
GOTERM_MF_DIRECT	GO:0042166~acetylcholine binding	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.40E-06
GOTERM_BP_DIRECT	GO:0007271~synaptic transmission, cholinergic	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	1.80E-05
GOTERM_BP_DIRECT	GO:0035095~behavioral response to nicotine	*CHRNA3, CHRNB4, CHRNA5, CHRNA4*	1.80E-05
GOTERM_MF_DIRECT	GO:0015276~ligand-gated ion channel activity	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	4.08E-06

Table 3.4 (continued). DAVID-Identified Gene Clusters and Biological Systems for Smoking Cessation

Category	Term	Genes	*p*
INTERPRO	IPR018000:Neurotransmitter-gated ion-channel, conserved site	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	6.30E-06
GOTERM_BP_DIRECT	GO:0098655~cation transmembrane transport	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	3.05E-05
INTERPRO	IPR006029:Neurotransmitter-gated ion-channel transmembrane domain	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	6.30E-06
INTERPRO	IPR006201:Neurotransmitter-gated ion-channel	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	6.30E-06
INTERPRO	IPR006202:Neurotransmitter-gated ion-channel ligand-binding	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	6.30E-06
UP_KEYWORDS	Ligand-gated ion channel	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	2.34E-04
GOTERM_BP_DIRECT	GO:0035094~response to nicotine	*CHRNB4, CREB1, CHRNA4, CHRNA6*	8.14E-04
UP_KEYWORDS	Postsynaptic cell membrane	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	0.00188089
KEGG_PATHWAY	hsa04725:Cholinergic synapse	*CHRNA3, CHRNB4, CREB1, CHRNA4, CHRNA6*	0.00896826
GOTERM_CC_DIRECT	GO:0045211~postsynaptic membrane	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	0.00471956
GOTERM_BP_DIRECT	GO:0007626~locomotory behavior	*CHRNA3, CHRNB4, CHRNA4, DBH*	0.00767406
GOTERM_BP_DIRECT	GO:0007268~chemical synaptic transmission	*CHRNA5, CHRNA4, CHRNA6, LPAR3, DBH*	0.00781503

Table 3.4 (continued). DAVID-Identified Gene Clusters and Biological Systems for Smoking Cessation

Category	Term	Genes	*p*
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6, LPAR3*	0.01462578
GOTERM_BP_DIRECT	GO:0007274~neuromuscular synaptic transmission	*CHRNA5, CHRNA4, CHRNA6*	0.01438218
GOTERM_BP_DIRECT	GO:0006811~ion transport	*CHRNA3, CHRNB4, CHRNA5, CHRNA4*	0.01438218
UP_KEYWORDS	Synapse	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	0.01144954
UP_KEYWORDS	Ion channel	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6*	0.01144954
GOTERM_MF_DIRECT	GO:0005506~iron ion binding	*CYP2A7, CYP2A6, CYP2B6, EGLN2, ACO2*	0.00106543
GOTERM_MF_DIRECT	GO:0016705~oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	*CYP2A7, CYP2A6, CYP2B6, EGLN2*	0.00106543
UP_KEYWORDS	Monooxygenase	*CYP2A7, CYP2A6, CYP2B6, DBH*	0.00377331
GOTERM_MF_DIRECT	GO:0008392~arachidonic acid epoxygenase activity	*CYP2A7, CYP2A6, CYP2B6*	0.00228772
GOTERM_MF_DIRECT	GO:0016712~oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	*CYP2A7, CYP2A6, CYP2B6*	0.00228772

Table 3.4 (continued). DAVID-Identified Gene Clusters and Biological Systems for Smoking Cessation

Category	Term	Genes	*p*
GOTERM_BP_DIRECT	GO:0019373~epoxygenase P450 pathway	*CYP2A7, CYP2A6, CYP2B6*	0.00781503
UP_KEYWORDS	Iron	*CYP2A7, CYP2A6, CYP2B6, EGLN2, ACO2*	0.01092355
GOTERM_MF_DIRECT	GO:0008395~steroid hydroxylase activity	*CYP2A7, CYP2A6, CYP2B6*	0.00623443
INTERPRO	IPR002401:Cytochrome P450, E-class, group I	*CYP2A7, CYP2A6, CYP2B6*	0.02314858
INTERPRO	IPR017972:Cytochrome P450, conserved site	*CYP2A7, CYP2A6, CYP2B6*	0.02715231
INTERPRO	IPR001128:Cytochrome P450	*CYP2A7, CYP2A6, CYP2B6*	0.02929044
UP_KEYWORDS	Oxidoreductase	*CYP2A7, CYP2A6, CYP2B6, EGLN2, DBH*	0.04642428
UP_KEYWORDS	Signal	*CCDC134, CHRNA3, CHRNB4, TNFRSF6B, HSPA5, CHRNA5, LAMB2, CHRNA4, SEMA6D, CHRNA6, PLA2G3, DBH, CYP2A7, CYP2B6, ENPP2, NUCB2*	7.12E-04
UP_SEQ_FEATURE	signal peptide	*CCDC134, CHRNA3, CHRNB4, TNFRSF6B, HSPA5, CHRNA5, LAMB2, CHRNA4, SEMA6D, CHRNA6, PLA2G3, ENPP2, NUCB2*	0.03475165
KEGG_PATHWAY	hsa04080:Neuroactive ligand-receptor interaction	*CHRNA3, CHRNB4, CHRNA5, CHRNA4, CHRNA6, LPAR3*	0.01462578

Note. Category refers to the original database or resource where the terms orient. Term refers to a detailed item in an annotation source.

Conclusion. Several individual genes have been reported to be associated with smoking cessation. However, functional analysis identified pathways relevant for acetylcholine receptor structure and activity were significantly associated with smoking cessation. This is consistent result as all phenotypes with DAVID analysis have identified this gene cluster and action as relevant for smoking phenotypes.

6. Other Tobacco Product Use.

Snus

Snus are an alternative tobacco product that are popular in Europe and are similar to chewing tobacco or dip. In contrast to chewing tobacco, snus are more finely cut with some being in powder form inside pouches or loose for later packaging by the user. Snus expose individuals to nicotine which may lead users to nicotine dependence and other negative health outcomes [212]. A twin study of nicotine dependence, as measured by the FTND, was conducted using 5,040 Swedish male twins [202]. This study reported significant additive genetic (0.32; 95% CI = 0.23-0.41) and unique environmental (0.68, 95% CI = 0.59-0.77) influences. This study also evaluated the twins ND arising from CIG use and found similar estimates of A (0.39) and E (0.61).

Another twin study of Norwegian twins (N = 3,862) asked about two phenotypes related to snu use: initiation and quantity of use [109]. Initiation was assessed as, “Have you ever used snus regularly for at least a month?” while quantity was probed with, “When you used snus the most, how many times per day did you use it?” Using these definitions, researchers reported significant genetic effects for both initiation (0.51) and

quantity used (0.55), as well as significant shared environmental influences (0.29 for initiation and 0.23 for quantity).

Electronic Cigarettes

Electronic cigarettes (ECIG) are just beginning to be evaluated in the genetic epidemiology literature. As of May 2020, studies of ECIGs remain sparse; however, there have been several studies examining differing aspects of ECIG use. A recent twin of ECIG initiation defined as self-reported ever use of ECIGs reported genetic (A = 0.25) and shared environmental (C = 0.42) contributions to ECIG initiation as well as overlap in these factors with CIG initiation (r_g = 0.76, r_c = 0.68) [213].

To date, no GWAS studies have been conducted of ECIG use. However, other studies[58,124] have used GWAS data to generate polygenic risk scores (PRS) and study ECIG use [214]. Both studies examining PRS used self-reported ever use without any additional caveats to classify individuals as initiators or not. In addition, one study examined the age of initiation as a secondary outcome. Both studies created PRSs for conventional cigarette initiation (SI; and CPD in the case of Allegrini) and then applied those scores in a secondary (target) data set to examine if genetic influences on cigarettes could also indicate if those influences also impact ECIG ever use. While no influence was detected for the SI PRS and ECIG ever use, there was a significant finding between CPD PRS and ECIG ever use in the Allegrini study. In contrast, Khouja and colleagues reported a significant effect of SI PRS and ECIG ever use (OR = 1.24), though they did not examine CPD to generate PRS. Taken together, these results provide an unclear answer to the genetic overlap between CIG and ECIG use.

CONCLUSIONS AND RECOMMENDATIONS

This paper has demonstrated how measurement of tobacco phenotypes has led to inconsistent results in GWAS and twin study designs. It was generally expected that the same genetic association results would be identified regardless of measure, particularly in studies of consortia or single studies with sufficiently large sample sizes. However, this did not occur. The high variability in tobacco use measures within a specific behavior (e.g., regular smoking, nicotine dependence) has multiple measures for quantifying a specific form of tobacco use. It is expected that among the many other limitations of different genetic influences than if another measure had been used.

Areas of Results Consistency

The nicotine acetylcholine receptor genes were consistently associated with various facets of nicotine use such as quantitative measures of smoking (e.g., CPD), nicotine dependence, and smoking cessation. Additionally, the *CYP2A6* gene was significantly associated with several nicotine use phenotypes including: nicotine metabolism and smoking cessation. These consistent results indicate that genetic influences are impactful; however, the inconsistent results suggest the genetic architecture is more complex. Aggregating genetic effects may help elucidate how genetics contribute to complex phenotypes.

Consistent findings from GWAS identify biological functions that may help explain their repeated associations with tobacco use. Genes responsible for acetylcholine receptor activity was associated for nearly every tobacco use phenotype. Further, ligand gate ion-channel activity was also consistently associated. These biological pathways

help explain the addictive properties of nicotine. Each leads to increased neurotransmitters (such as dopamine, glutamate, and GABA) in the brain which lead to rewarding sensations.

Genome-wide significant SNPs combined by these biological functions which may lead to more consistent interpretations of results than seeking to replicate the same variant. Genes, rather than SNPs, in a biological pathway are reported which help lead to more consistent interpretation of GWAS results. For instance, while specific SNPs may not be replicated across phenotypes, each tobacco use phenotype was associated with cell signaling and receptor structure or function. This consistency has been demonstrated in this paper via the use of DAVID annotation and the identification of pathways significantly associated, or over-represented, for a similar set of genes. For example, the nicotinic acetylcholine receptor genes were consistently associated with DAVID pathways for most of the conceptual phenotypes examined here. However, it should be noted that DAVID relies on published associations of genetic variants and biological function. It is possible that the results presented here are not comprehensive due to publishing bias.

Further, this functional annotation would allow for multiple phenotypes to be examined in each paper. Looking at individual smoking behaviors in the continuum (i.e., only focusing on smoking initiation or nicotine dependence) is not logical as facets of tobacco use build upon one another (e.g., one cannot be a regular smoker without having initiated smoking) [215]. Many biological pathways were shared across nicotine phenotypes, such as coding for receptors which influence the transmembrane ion channels. The nAChR genes were consistently identified with each nicotine use

phenotype as were biological pathways for gate control and receptor binding. These results suggest that there is biological plausibility for overlap between each nicotine phenotype. Therefore, future GWAS should examine multiple behaviors rather than several single tobacco use behaviors.

Limitations of GWAS in Tobacco Use

In addition to issues related to inconsistencies of tobacco measurement, there are other general limitations in GWAS designs. These studies often suffer from weak statistical power due to small samples (N~ 1,000) [143,216]. Small sample sizes may lead to biased estimates of effect and/or inconsistent results across studies. One solution to increase sample sizes is through consortia and meta-analyses of substance use phenotypes. However, studies contributing to consortia often vary slightly in how they operationalize and measure tobacco use (e.g., studies of smoking cessation may use differing definitions of successful cessation based on the length of time since last smoking). This leads to challenges in the ability to test the same measures across samples. Many consortia currently use secondary data, but as more data is collected, especially involving novel products, it would be advantageous to utilize the same operational definitions of tobacco use stages. To date, GWAS results have produced inconsistent results as would be expected due to outcome measurement heterogeneity. Many studies report genome-wide significant SNPs that have not been replicated in other studies.

GWAS results are reported as SNP ID (e.g., rs number, or chromosome and base pair), followed by the gene in which the SNP is present (if applicable, some

variants may be in intergenic regions), and the measure of association (i.e., beta value or odds ratio) with an associated *p* value and confidence interval. Sometimes researchers will also choose to report the chromosome, arm, and location of the significant SNP (e.g., Chr13p5.5). There may be an inconsistency of results when reporting in this manner as rs numbers may change based on the build used for the reference genome. It is also difficult to aggregate results in a meaningful manner as direct replication of a particular SNP may be difficult. Other methods of aggregating results are needed for determining relevance of GWAS analyses.

Recommendations for Future Studies of Tobacco Use in Genetically-Informative Samples

Future studies should continue to examine differences between major groupings of the population. For instance, twin studies report differing effects of additive genetic influences based on sex. However, only two GWAS[144,162] studies examined molecular effects by sex. Likewise, twin studies have demonstrated different influences of A on smoking phenotypes by race/ethnicity. However, two GWAS [196,207]explicitly examined racial differences. Genetic influences also change as an individual ages; however, only 6 GWAS [122,145,181,195,217,218] explicitly looked at age groups. These GWASs typically chose older individuals (age > 45 years) as participants whereas younger individuals were grouped in with the entire age spectrum (ages 18+). Future studies should continue to stratify GWAS results by environmental influences that have previously identified as associated with tobacco use whenever feasible. Lower levels of income have been associated with greater rates of tobacco use when compared with individuals

at higher income levels. Likewise, other SES variables (such as educational attainment) have been associated with higher tobacco use. Other environmental variables beyond demographic variables should also be examined, including exposure to marketing. Individuals who receive coupons for tobacco products are more likely to use tobacco products. Similarly, individuals who are exposed to marketing promotions are more likely to use tobacco products compared to people who are not [219].

Further, more standardized data collection of tobacco use phenotypes is needed. The movement to standardize phenotypic data collection in genetic studies continues to grow. Different measures of tobacco use have led to inconsistent results in GWAS of tobacco use. Consistently identifying the phenotype with the same operational measures of tobacco use across studies may help increase the statistical power via ensuring the same operational measures are used without needing to harmonize data during analysis. This could also lead to greater phenotypic refinement in appropriate samples, leading to a greater chance of detecting and replicating genetic influences. Reducing the heterogeneity in measurement will allow for consortia to be built faster and with more ease, as well as allowing meta-analyses to occur more efficiently. Multiple measures could also be used to evaluate the phenotypes to ensure that studies capture all variations of the phenotype.

Additional attention should be paid to phenotypes that may be missing from the current measures of tobacco use. GWAS results of smoking cessation generally focused on current versus former smoking status. However, no analyses have been conducted with nicotine withdrawal as the outcome measure. Likewise, there was no analysis for time to failure of cessation (i.e., duration of abstinence during an attempt to

quit smoking). Future GWAS should report on multiple phenotypes in the smoking continuum rather than focusing on very specific behaviors.

ECIG Recommendations. Results from studies of CIG use should be used to guide future studies of ECIG use. There are four recommendations for future studies of genetically informative samples when it comes to measuring tobacco use as administer via ECIGs. First, biological confirmation of smoking status whenever feasible is needed, through expired CO for CIG users and through plasma/urine cotinine levels for both CIG and ECIG users [220,221]. ECIG users may not know if their e-liquid contains nicotine or not [222,223]. Therefore, nicotine exposure would not be captured by self-report. Biological confirmation of recent tobacco use should help distinguish those currently use versus those who have not, leading to greater external validity of results. Future research should examine if salivary COT levels are feasible for large scale epidemiological studies compared to plasma/urine COT. Another important factor to consider, as tobacco use progresses into the ECIG era, is how to handle dual use of CIG and ECIG products. Other research has shown that dual users of CIG and ECIG are a distinct class of tobacco user (see Chapter 5) and future research needs to appropriately model this relationship. GWAS moving forward will need to consider this possibility. While outcome group assignment may seem straight forward, dual users need to be considered a distinct class of user. Future work should delineate exactly how dual users are distinct; for example, how much time must an individual use both products to be considered a dual user? How long after cessation of one product is one still considered

to be a dual user? These questions need to be addressed to ensure accurate results are generated.

Misclassifying individuals as CIG or ECIG users when they are dual users may lead to biased estimates. Future GWAS should take this into account and run multinomial regressions rather than logistic models. While it is possible to account for other tobacco product use as a covariate in the regression, it is not optimal when compared to running multinomial models. Absent running multinomial regression, results could be stratified on tobacco use (CIG-exclusive, ECIG-exclusive, and dual use) which would also present results for each specific tobacco use group.

ECIG research should probe participants to report at what concentration of nicotine a participant's e-liquid is set at to allow for greater accuracy in the estimation of nicotine used by an individual. Though there are after-market modifications to ECIG devices which change the amount of nicotine administered, knowing the concentration will allow researchers to estimate how much nicotine is being used [21].

Finally, results from genetically informed studies should continue to report on functional relevance for tobacco use. Functional relevance may be closely related to clinical relevance which may lead to real world changes by focusing therapies on interrupting the biological pathways from genetic influence to exposure. Continued investigations of functional relevance with polygenic methods may help identify pathways that significantly contribute to tobacco use.

The area of tobacco use is currently evolving as novel tobacco use devices continue to be introduced into the market. Consequently, genetic epidemiological studies across tobacco products are also rapidly evolving. For example, an update to

ECIGs involves heat-not-burn products (e.g., IQOS) which were introduced in the United States in 2019 [224]. Such rapid evolution in product development has implications for the study of nicotine dependence in genetically informative samples. ECIG remain a highly popular product and the measures used in genetic epidemiology studies need to keep pace. In particular, given the variability in CIG initiation results, effectively measuring ECIG initiation should be a high priority. For example, is it enough to classify an individual as an ECIG user if they only use a friend's device or must they own their own device in order to be classified as a user? Similarly, what about individuals who use ECIG devices but use e-liquid that is nicotine-free? These are important issues as different definitions of an ECIG user may yield differing results as happened with CIG samples.

CHAPTER 4: ELECTRONIC CIGARETTE GENOME-WIDE ASSOCIATION AND POLYGENIC SCORES AMONG SELF-IDENTIFIED WHITE PARTICIPANTS: TEST OF OVERLAPPING GENETIC INFLUENCES WITH CONVENTIONAL CIGARETTE INITIATION

INTRODUCTION

The genetic factors that influence conventional cigarette (CIG) initiation may also influence electronic cigarette (ECIG) initiation. Results from Chapter 2 suggest a significant overlap in additive genetic influences between ECIG initiation and CIG initiation in young adults. Chapter 3 summarized several previously reported genetic loci that contribute to CIG initiation. Further, prior studies using genome-wide polygenic scores (GPS) demonstrate the genetic overlap between CIG initiation and ECIG initiation [58,225]. This suggests that similar genes contribute to the liability of CIG as well as ECIG initiation and encourages additional study to detect specific genetic loci associated with ECIG initiation by taking advantage of the genetic overlap shared between CIG and ECIG use.

The Role of Age on CIG and ECIG Initiation

The age of CIG and ECIG initiation for tobacco use is increasing, with more individuals initiating in young adulthood as compared to adolescence [226]. The proportion of individuals who initiated CIGs in young adulthood (age 18-23) more than doubled from 20.6% (2002) to 42.6% in 2018. Similarly, the prevalence of ECIG initiation is increasing among adults. For example, the prevalence of ECIG initiation at age 18 was

estimated at 8.3%. That prevalence grew to 33.8% by age 25 in analysis of the Population Assessment of Tobacco and Health (PATH) study [227]. Additionally, twin studies report that genetic influences on CIG smoking initiation are larger at older ages compared to younger ages (age range = 12 to 18, see Chapter 3) [95,228,229]. While the role of genetic influences on ECIG initiation in adults remains unclear (see Chapters 2 and 3), the epidemiological evidence suggests that it will be important to study adults across the life course rather than only addressing younger users. Further, given the trends related to increasing magnitude of genetic effects on CIG initiation as age increases, such patterns may also apply to ECIG initiation. Therefore, although substantial effort related to the study of CIG and ECIG initiation is focused on young adults (ages 18-24) and adolescents (ages 12-17), study of the etiology underlying CIG and ECIG initiation across adulthood is necessary.

Genome-Wide Association Studies Identify Specific Genetic Loci Contributing to CIG Initiation

Prior genome-wide association studies (GWAS) have identified genetic variants in several loci (i.e., specific locations in the genome) that are associated with CIG initiation (see Chapter 3 for an in-depth discussion). A genetic association is a single test of association which uses genotypic data from a genetic marker to test for statistical associations between a genetic variant (e.g., single nucleotide polymorphism, SNP) at a specific locus and CIG initiation. A GWAS expands this test to thousands of genetic markers that test for associations with CIG initiation with SNPs across located throughout the genome. Therefore, GWAS produces results that identify specific

locations within the genome that are associated with an outcome (i.e., phenotypes such as CIG and ECIG initiation).

To date, 12 GWASs have consistently identified significant associations between SNPs located in the nicotinic acetylcholine receptor genes (e.g., *CHRNA3*, *CHRNB4*), with various tobacco use behaviors, including initiation [122,144,145,147–150,162–164,174,217,230] (Chapter 3). These genes encode subunits of nicotinic acetylcholine neuronal receptors which mediates fast signal transmission at synapses. Neuronal nicotinic acetylcholine receptors reside on all neurons and their influences function throughout the brain [231]. Neuronal nicotinic acetylcholine receptors are activated by acetylcholine and nicotine (and other drugs). When nicotine binds to a nicotinic acetylcholine receptor, it acts as an agonist to potentiate receptor activation. Activation of nicotinic acetylcholine receptors allows sodium (Na^+) into the neuron and results in production of several neurotransmitters, including dopamine, acetylcholine, glutamate, GABA, epinephrine, norepinephrine, and serotonin (Figure 4.1) [232]. Acetylcholine contributes to the reward pathway (along with pathways relating to attention, memory, and arousal [233–235]) leading to rewarding sensations. Addiction is also enhanced through acetylcholine-mediated craving for euphoric feelings produced by this neurotransmitter. Further, consistent evidence has emphasized the role of dopamine, glutamate, and GABA resulting from nicotine exposure which are responsible for information processing, memory, and emotions[232]. Glutamate, which is the brain's primary excitatory neurotransmitter [236], is released in several brain structures such as the ventral tegmental area (VTA) after nicotine exposure and regulates the dopaminergic neurons. Increased firing of dopaminergic neurons in the VTA stimulates dopamine release in the nucleus

accumbens (NAc), which leads to rewarding feelings arising from the mesolimbic system. ECIGs contain nicotine and been reported to change glutamate and dopamine levels in mice [237], which is expected to be mediated by nicotinic acetylcholine receptors and the genetic variants that influence their function.

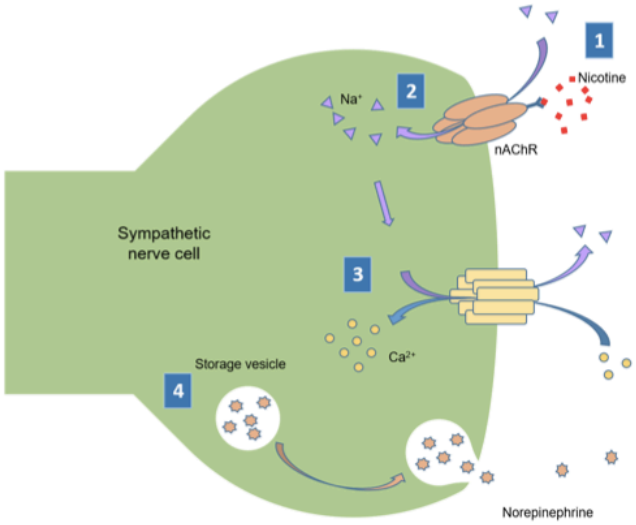


Figure 4.1. Nicotine Acts as an Agonist for Acetylcholine Receptors. Nicotine binds to and stimulates the acetylcholine receptor (1), which allows sodium (Na^+) into the presynaptic neuronal cell (2), which stimulations the calcium ion channel to open (3) releasing Ca^{2+}, potentiating the cell to release neurotransmitters (4) into the synapse. Figure adapted from Price & Martinez, 2019 [238].

CYP2A6 has been associated with several CIG use behaviors including regular smoking, nicotine dependence, and smoking cessation. *CYP2A6* encodes the Cytochrome P450 A6 enzyme. This enzyme is responsible for metabolizing nicotine and cotinine via oxidation [189,239] (see Chapter 3: Nicotine Metabolism). Further, *CYP2A6* is necessary for moving nicotine through the body while the nAChR gene cluster is responsible for rewarding sensations arising from tobacco use. Taken together, prior

GWAS results have identified several important loci within genes that have biological or psychological function associated with CIG use. However, it remains unknown if these same genes play a role in ECIG initiation and use.

Use of Genome-Wide Polygenic Scores Addresses Limitations of GWAS

The concept of "polygenic" factors was first discussed by Sir Ronald Fisher as the influence of many genes making small contributions to a phenotype[240], has been confirmed throughout the GWAS era[142,241], and has led to the recent development and study of genome-wide polygenic scores (GPS). A GPS reflects the average genetic contribution for a phenotype across all measured loci. This approach was developed because large-scale GWAS demonstrated significant associations of hundreds of common variants that contribute small effects to many complex phenotypes [142,240,242–245], including CIG initiation (e.g., GWAS and Sequencing Consortium for Alcohol and Nicotine, GSCAN, [163]). Common variants of small effect are most likely the cause of genetic variation in complex traits [142,242,246] (e.g., outcomes that are due to many genetic and environmental influences). To date, many GWASs have been underpowered due to small sample sizes[247] which limits the ability to detect genetic influences of small effect sizes. Therefore, GWAS may often fail to capture the variation from genes of small effect [248]. The use of GPS to aggregate genetic variants with small effects (i.e., variants with odds ratios less than 1.3) is expected to reflect a more accurate representation of multiple genetic influences on an outcome [249–251].

Studies of GPS use summary statistics (i.e., β values and *p* values) generated from GWAS of large "discovery" samples with sample sizes that have sufficient power to

detect genome-wide significant genetic associations (N > 100,000). Discovery samples generally refer to single sample studies (e.g., UK BioBank) or a consortium of several studies that have appropriate sample sizes to conduct GWAS (e.g., GSCAN). After summary statistics are generated in the discovery sample, they can be applied in smaller "target" samples that would otherwise be underpowered to detect genetic effects (N ~ 1,000) [248,252]. The generation of GPS from a discovery sample and its application of summary statistics to a target sample relies on the assumption that both samples contain participants from the same ancestral group. "Ancestral groups" refer to populations of humans whose biological ancestors come from similar geographic regions (e.g., Europe) and experience common evolutionary selection migration patterns and selection pressures over several generations. Consequently, individuals within a specific ancestral group have largely similar distributions of allele frequencies at most loci throughout the genome. To date, most GWAS studies are conducted in samples with European ancestry groups. Therefore, GPS generation is most often conducted in these populations because they have reliably similar allele distributions across discovery and target samples [253–255].

Genome-wide polygenic scores were first used to examine common genetic variation and its influence on schizophrenia [256]. Using data from the International Schizophrenia Consortium (ISC), polygenic scores were calculated for schizophrenia and applied in independent samples provided by the Molecular Genetics of Schizophrenia (MGS) consortium. The polygenic scores performed adequately, explaining roughly 3% of the variance for schizophrenia in the European sample of MGS. However, the clinical utility of a polygenic score was deemed insufficient to

increase diagnostic accuracy [257]. More recent research has shown improvement in the predictive probability of GPS, though not enough to warrant use in clinical settings [258]. The improvement of the predictive probability of GPS has suggested that with more research, clinical utility could be found [258–262].

CIG GPS Influences on Psychiatric Phenotypes

Significant genetic overlap has been reported between CIG GPS for various CIG phenotypes and other psychiatric outcomes. CIG GPS has been associated with smoking initiation, explaining nearly 5% of the variance in one study [263]. Similarly, a GPS constructed for CPD was associated with CPD [264]. GPS for CIG initiation have also been associated with non-tobacco related phenotypes. GPS for late onset CIG initiation has also been associated with schizophrenia. Individuals who initiated CIG use later in life were at an increase in the odds of reporting a diagnosis of schizophrenia [265]. In addition, GPS for having ever been a regular smoker have been associated with externalizing behaviors, with being a regular smoker associated with an increase in the odds of reporting externalizing behaviors [266]. Further, GPS for age of CIG initiation has been associated with age of regular alcohol drinking, which suggests there may be overlap with substance use [267]. These results suggest that overlap exists between genetic influences for CIG initiation and multiple psychiatric disorders, including other substance use beyond CIGs. It is therefore expected that some genetic influences may be shared between CIG and ECIG initiation.

Associations between CIG GPS and ECIG Initiation

To date, three GPS studies have been conducted on ECIG initiation. Allegrini and colleagues utilized the Tobacco and Alcohol Genetics (TAG) to create GPS for CIG initiation (SI; defined as having smoked 100 or more cigarettes in an individual's lifetime) and cigarettes per day (CPD)[124]. These scores were then applied to a target sample of Netherlands twins who used ECIGs. While the GPS for SI wasn't significantly associated with ECIG use, the GPS calculated from the CPD phenotype was associated. Specifically, the GPS was significantly associated with ECIG initiation among ex-smokers (OR = 1.43) and never smokers (OR = 1.35). These results suggest that there may be some association between genetic influences for CIG and ECIG use; however, the authors acknowledge their study may have been underpowered to detect genetic effects and recommended additional studies.

More recently, GPS for CIG initiation were created from the GWAS and Sequencing Consortium of Alcohol and Nicotine (GSCAN) and used to study their association with ECIG initiation [58]. A significant association was detected between GPS for CIG initiation and ECIG initiation by age 24 (OR = 1.24, 95% CI = 1.14-1.34, $p < 0.001$). Importantly, a separate analysis was conducted to ensure that the GPS for CIG initiation was associated with CIG initiation in their target sample (OR = 1.29, 95% CI = 1.19-1.39, $p < 0.001$), confirming the GPS for CIG initiation was associated with the same phenotype in an independent sample. Taken together, these two analyses show that the genetic influences which are important for CIG initiation are also meaningful for ECIG initiation.

A similar analysis was conducted in the United States utilizing a GPS for regular CIG use and cigarettes per day calculated from GSCAN and applied to ECIG use in a sample of college students [225]. This study reported a significant associated with CIG regular use and ECIG initiation (OR = 1.27, 95% CI = 1.19-1.36, $p < 0.001$) among individuals of European descent. While this study did not model CIG initiation as a GPS, it provides additional preliminary evidence that genetic influences of other CIG phenotypes (quantity of use, regular cigarette use) are also shared with ECIG initiation.

This study builds on the previous results in two ways. First, the molecular genetic variants that may contribute to ECIG initiation remain unknown. Several genome-wide association studies (GWAS) have been conducted on the initiation of CIG use. However, it is unknown if the same variants identified for CIG initiation will also be significant with ECIG initiation. These analyses will first detail the molecular genetic contribution to ECIG lifetime initiation in a sample of US adults, aged 18-93. Second, current results suggest there may be some overlapping genetic influences in a sample of young adults, ages 18-25 (e.g., see Chapter 2) [58,225]. Analyses in this chapter build on the prior results by using data from a community-based sample of unrelated adults (age 18-93) to examine genetic overlap within a larger age range. Taken together these analyses answer two research questions: 1) are there any specific variants that are significantly associated with lifetime ECIG initiation across adulthood and 2) are the molecular genetic influences shared between ECIG and CIG lifetime initiation?

METHODS

Study Description

Data from the Genes for Good (G4G) study were used. The G4G study uses a community-based sample of active Facebook users in the United States aged 18 and older (N = 81,476). In brief, Facebook users participated in the G4G project through a Facebook-specific application (i.e., a third-party add on to the base Facebook page). Participants were taken through a consent process after adding the application to their profile. Participants who used the G4G application on Facebook answered survey questions regarding a variety of behaviors and lifestyles including: sleep, personality, exercise, and drug use (including tobacco, alcohol, marijuana, and other illicit substances). Surveys were divided into two broad categories: 1) health history surveys (baseline surveys focused on behaviors prior to G4G participation) and 2) health tracking surveys (daily behavior tracking from the previous day, such as how many alcoholic drinks they consumed or number of cigarettes smoked). Participants completed surveys at any time and chose the modules they wanted to answer. Participants were recruited via "snowball recruiting"; where individuals are recruited through their peer group (e.g., parents, friends, Facebook groups) via posts that are shared about the participants engagement with G4G (e.g., "I just completed a health tracking survey for Genes for Good!").

A subset of G4G participants also volunteered to give a sample of their DNA via a mailed saliva collection kit (N = 27,469). To be included, participants must have answered the health history survey as well as a minimum of 15 health tracking surveys of their choice. Participants received a free genetic ancestry report and access to their

raw genetic data via a secure (i.e., encrypted) file transfer point (SFTP) over the internet.

As of 18 June 2019, 81,476 individuals had signed up to use the G4G application and completed at least one survey. Of these, 27,469 individuals had been genotyped (33.7%). Data from 20,231 genotyped participants were available at the time of data release. For purposes of this study, individuals were limited to self-identified ancestral (SIA) White participants only (N = 15,881). SIA White participants reduced the influence of possible population stratification on analyses. Population stratification is a source of confounding in GWAS, conceptualized as a phenomenon that arises due to differing selection pressures placed on non-random mating populations. This selection leads to differing allele frequencies between ancestral groups [268]. Different allele frequencies may lead to spurious relationships being detected between genetic markers and phenotypes [269].

Measures of CIG and ECIG Initiation. Two items assessed lifetime CIG and ECIG initiation. CIG lifetime initiation was treated as a dichotomous variable, probed as, "Have you ever tried a cigarette?" ECIG lifetime initiation was treated as a dichotomous variable and measured as, "Do you smoke e-cigarettes?"

Genotyping. Participant DNA was genotyped across approximately 600,000 SNPs using the Illumina Infinium CoreExome-24 v1.0 or v1.1 arrays [163]. Genotyping was completed on nonsynonymous exonic variants (i.e., a mutation that alters the protein coded for by the amino acid sequence), as well as a panel of common genome-wide markers. Additional markers were also genotyped including missense (i.e., mutation of coding for an amino acid), loss-of-function (i.e., mutation leading to less, or

no, function), potential lipid- and myocardial infarction-associated variants, height-associated variants, stop-gain variants (i.e., a mutation that stops transcription prematurely) in 96 genes with loci potentially implicated in type 2 diabetes, blood lipid levels, Alzheimer's disease, nicotine/alcohol metabolism, and other serious but treatable health conditions. Additionally, Neanderthal SNPs from the 1000 Genomes Project, and ancestry informative markers were also genotyped. Genotypes for approximately 30 million additional variants were imputed using Minimac3 [270] using the 1000 Genomes Phase 3 [271] panel as a reference panel. Imputation is an important step because it allows for additional markers to be imputed, increasing the number of genetic markers. These additional markers allow for a more complete interrogation of the genome, above and beyond the directly observed genotypes.

Quality Control. Standard data quality control (QC) procedures were implemented before data analysis (Figure 4.2). This included removing individuals which had excessive missing data (greater than 5% missing) and individuals who were cryptically related (i.e., unknown to either researcher and/or participants that two individuals were biologically related) to one another ($p > 0.05$, 0 participants removed). Individuals with increased heterozygosity rates, defined as more than three standard deviations from the mean were removed (0 individuals. Heterozygosity rate is the proportion of an individual's genome that is heterozygous (i.e., they have one copy of the dominant and recessive allele). Only SIA White participants were used for this study, leaving an analytic sample of 15,881 participants after per-individual QC. Further participants were removed due to missing phenotypic data (85 missing CIG data, 86

missing ECIG data), leaving 15,796 CIG initiators and 15,795 ECIG initiators for analyses.

QC was also done on a per-SNP basis. SNPs were removed if they significantly violated Hardy-Weinberg Equilibrium (HWE; $p < 10^{-8}$, 395 markers removed) as were SNPs with excessive missingness (> 5%, 125,044 markers removed). Violations of HWE suggest that there may be influences (e.g., assortative mating, genetic drift, or founder effects) impacting the frequency of a particular genotype [272]. Two other commonly used QC measures were not utilized: sex checking and INFO score pruning. INFO scores were not included with the data transfer from the G4G study staff. In general, GWAS analyses are limited to high performing SNPs (INFO score >= 0.5) to ensure the highest possible quality of data. Additionally, information was not sent regarding the X chromosome which would be used to check the self-reported sex versus the biological sex of participants. Instead of genetic data regarding the participants, self-reported sex was utilized as a covariate in subsequent GWAS analyses. After completing the per-marker QC, 7,252.506 markers were available for analysis.

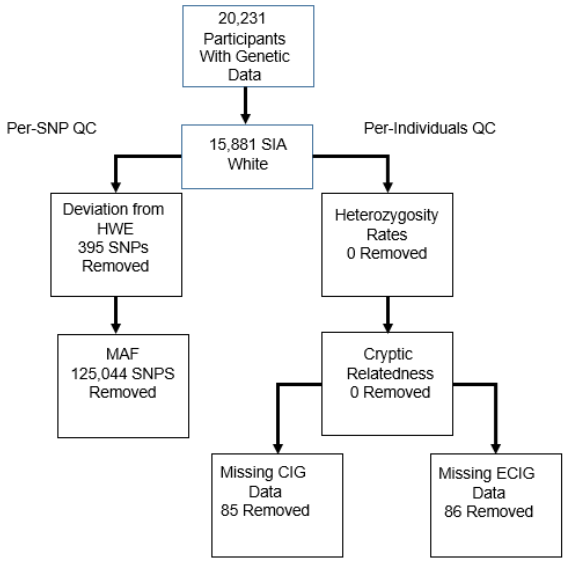


Figure 4.2. Flowchart of Quality Control Procedures and Number of SNPs and Individuals Removed.

Genome-Wide Polygenic Score Calculation. GPS are calculated from a discovery, or training, sample for CIG lifetime initiation. This sample is made up of individuals who are unrelated to the target sample. Markers in this training sample are ranked by their association with the phenotype, generally measured by *p* values [273]. The measure of association (e.g., beta values) are then summed across the markers that are associated at a given level (e.g., all markers with *p* values less than 0.01, 0.05, or 0.10) as shown in the formula below.

$$\boldsymbol{GPS} = \sum_{\boldsymbol{i}}^{\boldsymbol{n}} \boldsymbol{\beta_i}\, \boldsymbol{G_i}$$

where $\boldsymbol{\beta_i}$ is equal to the beta coefficient that is associated with genetic variant $\boldsymbol{G_i}$ for the i^{th} person. Therefore, a GPS is the summed effect (of participants running from *i* to *n*) for all measured genetic variants.

GPS for this study was generated for CIG initiation from summary statistics provided by GSCAN[163]. GPS were generated via PRS-CS, a freely available Python-based software [274]. PRS-CS (Polygenic Risk Score with Continuous Shrinkage) is a relatively new program that calculates GPS under a Bayesian regression framework rather than the frequentist framework presented above. Using this Bayesian framework, the SNP prior probabilities are subject to continuous shrinkage as follows:

$$\beta_j|\varphi_j \sim N(0, \phi\varphi_j), \qquad \varphi_j \sim g$$

where β_j is the effect size of the jth SNP which is contingent on a mixing distribution (g) and the variance is multiplied by ϕ a scaling parameter (10^{-6}, 10^{-4}, 10^{-2}, 1, or autoscaling). Using a known LD reference panel (in this analysis, the European sample from 1000 Genomes), individual level regression models are able to put a posterior probability on each SNP in the sample (assuming the SNPs are overlapping). This continuous shrinkage removes the need for *p* value pruning as is commonplace in non-Bayesian programs (e.g., PRSice-2, PLINK). Continuous shrinkage reduces the priors towards the average effect for the SNP. This shrinkage is updated iteratively constantly shrinking the priors as new information is added. For this study, the amount of variance explained (Nagelkerke R^2) was calculated for both a null model (i.e., a model that uses all terms outside of the GPS in a logistic regression) and a model with the GPS, with the two models then compared [275].

GSCAN summary statistics were calculated on individuals of European ancestry only [276]. The use of summary statistics to calculate GPS works best within a single ancestral group and predicting across ancestry groups (i.e., any non-European ancestry group) may lead to biased estimates [277,278].

Covariates. Several covariates were included in the statistical analyses. These items have previously been associated with CIG and ECIG use in other peer reviewed studies. Previous research has reported a difference between males and females in terms of ECIG use [279,280]. Sex was included as a dichotomous variable (male vs. female). Age has also been associated with ECIG use with younger ages using ECIGS more than older individuals [281]. Age was assessed as a ten-level ordinal variable with age binned into 10 year gaps after age 21 by G4G study staff. Education has been associated with ECIG use with individuals of higher education being less likely to use ECIG compared to individuals with lower education [46]. Education was recoded into a four-level ordinal item with bins reflecting: 1) less than high school (HS), 2) HS Graduate, GED, or Some College, 3) Associate's Degree, and 4) College Graduate or More. The analytic sample was restricted to SIA White individuals, so there was no additional variable to denote race in the GWAS, though the first 7 Principal Components were adjusted for (see below). To account for dual use, opposite tobacco product use (i.e., ECIG users' CIG use) was used as a covariate.

Analytic Strategy

Univariate GWAS was conducted using standard methodologies via PLINK v1.9 [72]. GWAS involves a series of regressions wherein each measured SNP (both genotyped and imputed) is regressed on the outcome of interest.

Ancestry Principal Components Analysis. PLINK v1.9 (--pca) was used to calculate ancestry PCs among G4G SIA White participants to further address the possibility of population stratification, even within SIA White individuals. Prior to running

the PCA, SNPs were restricted to those that were overlapping with the 1KG reference genome (N_{SNP} = 80,104; Figure 4.3. Seven ancestry PCs were retained to account for the majority of the variance of ECIG initiation (Figure 4.4). Seven PCs were retained rather than the first 10 as the scree plot of PCs suggested that after 7, no significant additional proportion of variance was explained (Figure 4.4).

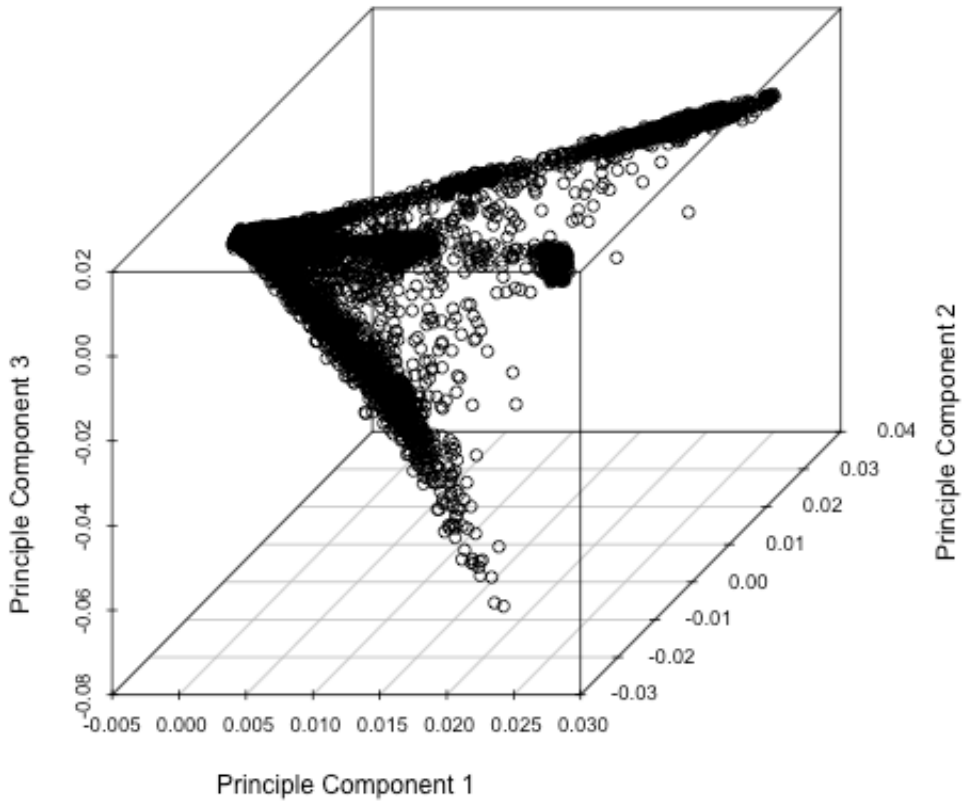


Figure 4.3. Three-Dimensional Plot of the First Three Principal Components for Self-Identified White Participants.

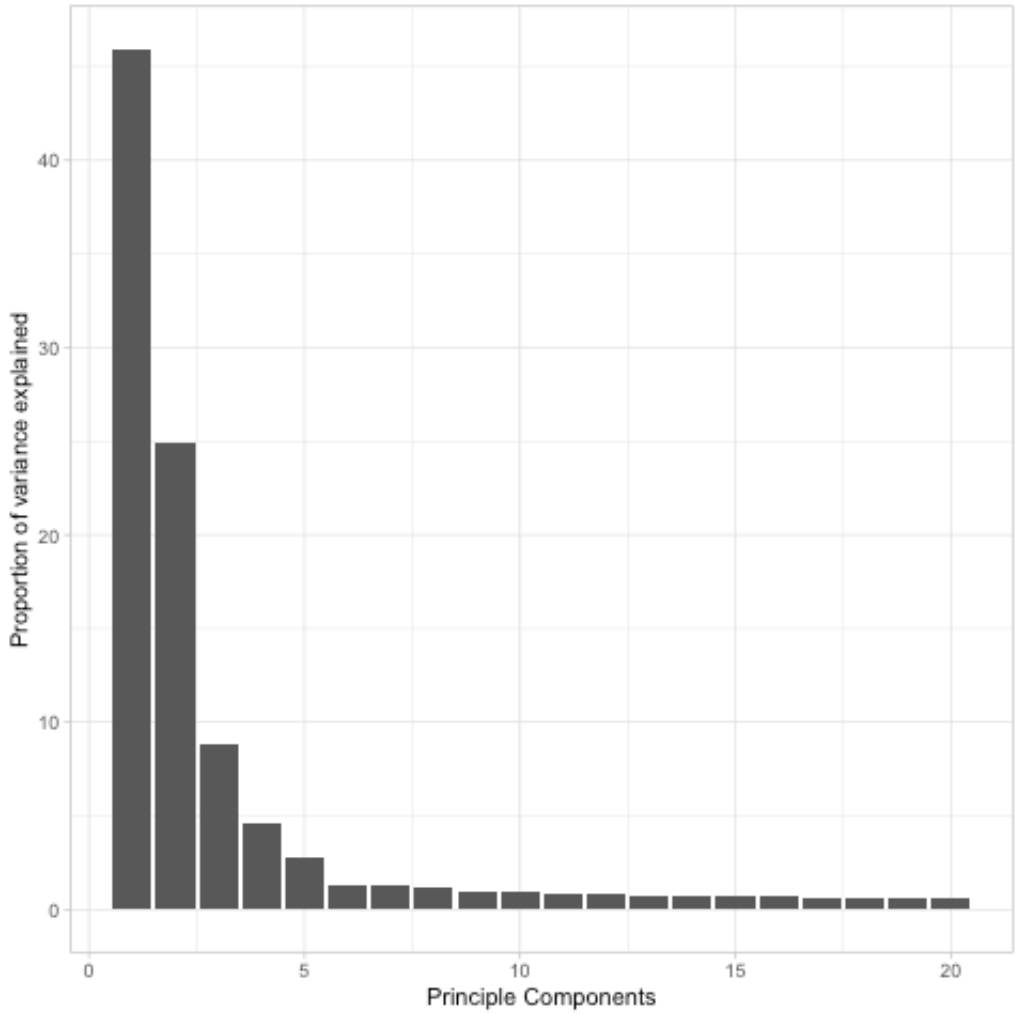


Figure 4.4. Proportion of Variance (y-axis) Explained by the SIA White Principal Components (x-axis).

Predictive Probability of GPS

The area under the curve (AUC) was generated from a receiver operator characteristic (ROC) curve to examine the predictive probability utility of the GPS (i.e., how well does a model with GPS accurately predict an individual's ECIG use) [282,283]. The ROC curve was calculated by comparing a binary outcome (Y, or ECIG initiation) with a continuous predictor (X, or GPS). In a ROC curve, each level of X is evaluated as a candidate cut point which will discriminate an individual's classification on the binary Y (i.e., an ECIG user or not). These predicted classifications are then compared with the observed value

of Y, from which the number of true positives and false positives can be computed. The sensitivity (i.e., the probability that an observation with a positive outcome is correctly classified as positive [sensitivity = True Positives /(True Positive + False Negatives)]) and specificity (i.e., the probability that an observation with a negative outcome is correctly classified as negative [specificity = True Negatives/(True Negatives + False Positives)]) may then be calculated and graphed against one another with the x-axis being the false positive rate (1 – specificity) and the y-axis being the true positive rate (sensitivity). The ROC then calculates the AUC, or the probability of accurate outcome group assignment based on the regression results [250,284].

RESULTS

Sample Representativeness

In general, participants who were genotyped were less likely to have used either CIGs or ECIGs (6.6% of those genotyped were ECIG users while 8.4% of those not genotyped were, 68.2% of genotyped participants were CIG users while 72.3% of non-genotyped participants used CIGs; Table 4.1). Genotyped individuals differed from ungenotyped participants in several respects: they were more likely to fall in the age 22-30 age group (37.6% vs 35.7% respectively).; had a larger proportion of men (29.7% vs 22.8%); were more likely to be college graduates (18.6% vs 14.7%); and more likely to have health insurance (92% vs 90%). All differences were significant at the $p < 0.001$ level. These analyses suggest there are systematic differences in participants who were genotyped compared to those who were not.

Table 4.1. Descriptive Statistics for Genotyped vs Not Genotyped Participants

	Genotyped N (%)	Not Genotyped N (%)
ECIG		
Yes	1,299 (6.6)	2,978 (8.4)
No	18,274 (93.4)	32,545 (91.6)
CIG		
Yes	13,358 (68.2)	25,676 (72.3)
No	6,216 (31.8)	9,849 (27.7)
Sex		
Male	5,837 (29.7)	8,241 (22.8)
Female	13,847 (70.3)	27,908 (77.2)
Age Range		
18-21	1,143 (5.8)	2,279 (6.3)
22-30	7,404 (37.6)	12,834 (35.7)
31-40	5,597 (28.4)	10,110 (28.1)
41-50	2,416 (12.3)	5,050 (14.0)
51-60	1,854 (9.4)	3,516 (9.8)
61-70	986 (5.0)	1,776 (4.9)
70+	284 (1.4)	379 (1.1)
Education		
Less than HS	387 (2.0)	953 (2.6)
HS Grad/GED/Some College	7,642 (38.8)	16,279 (45.0)
Associate's Degree	7,998 (40.6)	13,624 (37.7)
College Graduate or More	3,656 (18.6)	5,310 (14.7)
Insurance Status		
Covered	18,118 (92.0)	32,536 (90.0)
Not Covered	1,430 (7.3)	3,373 (9.3)
I Don't Know	135 (0.7)	257 (0.7)

Note. All χ^2 statistics significant at $p < 0.001$.

Conventional Cigarettes. The majority of participants who were asked "Have you ever tried a cigarette?" responded "Yes" (N = 11,058; 70%; Table 4.2). Lifetime initiators of CIG were more female (72.1%) compared to non-initiators (69%). A greater proportion of initiators were older (34.7% were over 41) compared to non-initiators (21.3%). Initiators also reported less education (82% of initiators had an Associate's Degree or lower) compared to non-initiators (78.6% had an Associate's Degree or

lower. Finally, initiators reported fewer participants without health insurance (7.4%) compared to non-initiators (5.7%). All differences were significant at the $p < 0.001$ level. These results suggest differences between initiators and non-initiators of CIGs across several important variables.

Table 4.2. Descriptive Statistics of CIG Lifetime Initiation

	CIG Initiation	
	Yes	**No**
	N (%)	**N (%)**
Sex		
Male	3,084 (27.9)	1,458 (31.0)
Female	7,974 (72.1)	3,281 (69.0)
Age Range		
18-21	236 (2.1)	490 (10.3)
22-30	3,449 (31.2)	2,133 (45.0)
31-40	3,533 (31.9)	1,110 (23.4)
41-50	1,640 (14.8)	445 (9.4)
51-60	1,278 (11.6)	335 (7.1)
61-70	720 (6.5)	174 (3.7)
70+	292 (1.8)	52 (1.1)
Education		
Less than HS	207 (1.9)	79 (1.7)
HS Grad/GED/Some College	4,391 (39.7)	1,592 (33.6)
Associate Degree	4,474 (40.5)	2,055 (43.4)
College Graduate or More	1,985 (18.0)	1,013 (21.4)
Insurance Status		
Covered	10,187 (92.1)	4,437 (93.6)
Not Covered	823 (7.4)	268 (5.7)
I Don't Know	47 (0.4)	34 (0.7)

Note. All χ^2 statistics significant at $p < 0.001$.

Electronic Cigarette Initiation. ECIG lifetime initiation was measured with, "Do you smoke e-cigarettes?" to which almost 7% of participants (N = 1,050; 6.6%; Table 4.3) reported they had initiated ECIGs. Most ECIG lifetime initiators were between the ages of 22-30 (36.7%) or 31-40 (34.3). Likewise, ECIG lifetime initiators had lower education levels (6.9% were a College Graduate) compared to non-initiators (19.8%

had a College Degree). Finally, a greater proportion of ECIG lifetime initiators did not have insurance coverage (11.8%) compared to participants who did not initiate ECIGs (6.6%). All differences were significant at the $p < 0.001$ level. These results suggest differences between ECIG initiators versus non-initiators across several important variables.

Table 4.3. Descriptive Statistics of ECIG Lifetime Initiation

	ECIG Initiation	
	Yes	**No**
	N (%)	**N (%)**
Sex		
Male	386 (36.8)	4,156 (28.2)
Female	664 (63.2)	10,590 (71.8)
Age Range		
18-21	49 (4.7)	677 (4.6)
22-30	385 (36.7)	5,197 (35.2)
31-40	360 (34.3)	4,282 (29.0)
41-50	138 (13.1)	1,947 (13.2)
51-60	88 (8.4)	1,525 (10.3)
61-70	24 (2.3)	870 (5.9)
70+	6 (0.6)	248 (1.7)
Education		
Less than HS	33 (3.1)	253 (1.7)
HS Grad/GED/Some College	590 (56.2)	5,393 (36.6)
Associate's Degree	355 (33.8)	6,173 (41.9)
College Graduate or More	72 (6.9)	2,926 (19.8)
Insurance Status		
Covered	917 (87.3)	13,706 (93.0)
Not Covered	124 (11.8)	967 (6.6)
I Don't Know	9 (0.1)	72 (0.5)

Note. All χ^2 statistics significant at $p < 0.001$.

Dual Use of ECIG and CIG. ECIG and CIGs are often used concurrently, known as dual use. G4G participants were rarely ECIG-exclusive initiators (0.3%) with most people having initiated CIGs only (63.6%; Table 4.4). Nearly 30% of participants had tried both products. These results suggest that ECIG-exclusive initiation is rare.

Table 4.4. Distribution of Tobacco Lifetime Initiation Among Self-Identified White Participants with Genotypic Data

	N	%
Tobacco Lifetime Initiation		
No Initiation	1,008	6.4
CIG-Exclusive Initiation	10,082	63.6
ECIG-Exclusive Initiation	45	0.3
Dual Use Initiation	4,706	29.7

Main Results

CIG Initiation

There were no genome-wide significant associations (Figure 4.5). However, three SNPs had suggestive associations ($p \leqq 1 \times 10^{-6}$). The first genome-wide suggestive SNP was in an intergenic region on chromosome 18 (18:61021122, OR = 1.24, $p = 2.5 \times 10^{-7}$) between *HMGN1P31* (High Mobility Group Nucleosome Binding Domain 1 Pseudogene 31) and *CDH20* (Cadherin 20). Neither of these genes has been associated with tobacco use in prior literature. The second suggestive SNP was on chromosome 2 (chromosome identification number: base pair location- 2:84368347, OR = 0.88, $p = 9.11 \times 10^{-7}$) near the gene *SUCLG1* (Succinyl-CoA ligase GDP/ADP-forming subunit alpha). No significant associations between this gene and tobacco use have been reported to date. The final genome-wide suggestive SNP was located on chromosome 11 (11:42405437, OR = 0.63, $p = 7.24 \times 10^{-7}$) near the *LINC0240* (Long Non-Coding RNA 2740) gene. This gene has not been associated with tobacco use in prior studies.

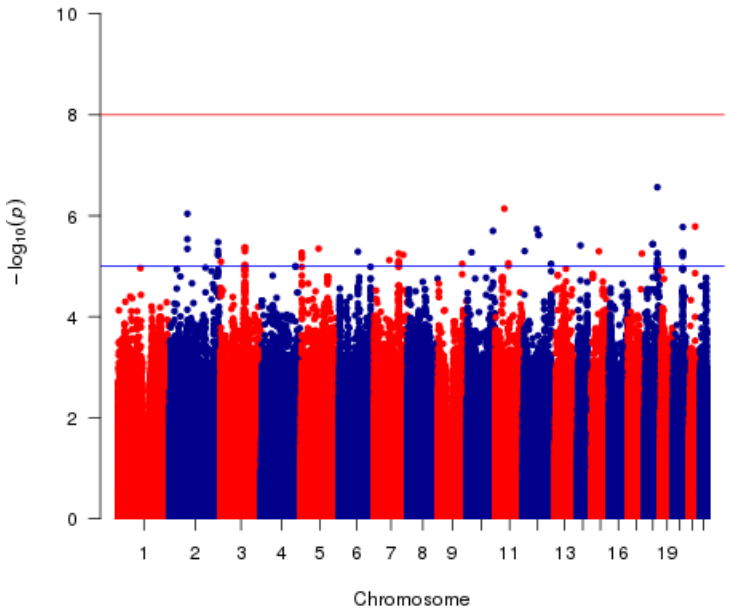


Figure 4.5 Manhattan Plot of CIG Initiation Adjusted for Covariates.

ECIG Initiation

Similar to the CIG results, no SNP reached genome-wide significance in the ECIG GWAS, though there were four SNPs reached the genome-wide suggestive threshold ($p \leqq 1 \times 10^{-6}$; Figure 4.6). The most significant SNP was on chromosome 13 located in the *N4BP2L1* gene (NEDD4 Binding Protein 2 Like 1, 13:32403784, OR = 0.62, $p = 7.49 \times 10^{-7}$). This gene has previously been linked to several cancers, including breast cancer [285]. One SNP located on chromosome 2 (2:115364757) in the *DPP10* (Dipeptidyl Peptidase Like 10) gene was genome-wide suggestive (OR =1.28, $p = 5.19 \times 10^{-7}$). This gene has previously been linked to asthma, a respiratory disease [286,287]. A third SNP on chromosome 15 (15:49010393) within the *SECISBP2L* (SECIS Binding Protein 2 Like)

gene was genome-wide suggestive (OR = 0.44, p = 8.01 x 10^{-7}). This gene has not been associated with tobacco use or a possible health consequence of tobacco use. Finally, a SNP on chromosome 6 (6:33902823) was genome-wide suggestive (OR = 0.79, p = 1.75 x 10^{-7}); however, this SNP is located in an uncharacterized location (*LOC105375026*).

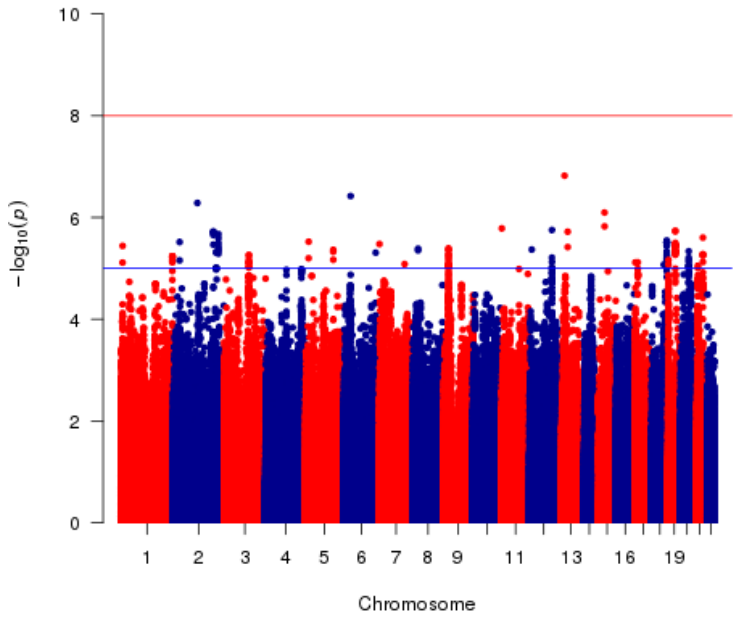


Figure 4.6. Manhattan Plot of ECIG Initiation Adjusted for Covariates.

GPS Results

Raw GPS values had very little variance (Figure 4.7a). Consequently, GPS were transformed into z-scores to have adequate range of variation (Figure 4.7b) as recommended by Choi and colleagues [252,288,289].

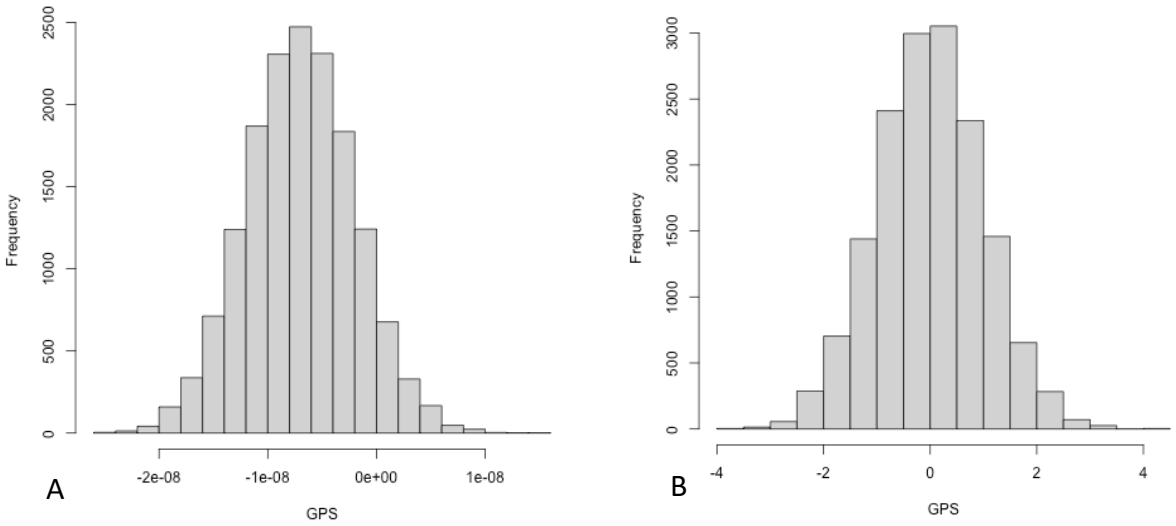


Figure 4.7. Distribution of Raw (Panel A) and Transformed (Panel B) Genome-wide Polygenic Scores.

There was no significant association between GPS and ECIG initiation after controlling for cross tobacco product use, insurance status, education level, income level, gender, and the first 7 ancestry PCs before or after the transformation. A null model was calculated wherein all predictors were added to a logistic regression except for the GPSs (Nagelkerke $R^2 = 0.0595$) and compared against a model with the GPSs and all covariates (Nagelkerke $R^2 = 0.0596$). The difference in Nagelkerke R^2 between the models was 0.0001, suggesting a very small amount of the variance was explained with the addition of the GPS. The AUC analysis showed the model performed fairly well (AUC = 0.75, Figure 4.8). Additionally, there was decent discrimination for ECIG

initiation[290] (AUC >= 0.70 is the threshold for acceptable discrimination). However, the difference in Nagelkerke R^2 suggests that this association is due to the covariates rather than the GPS.

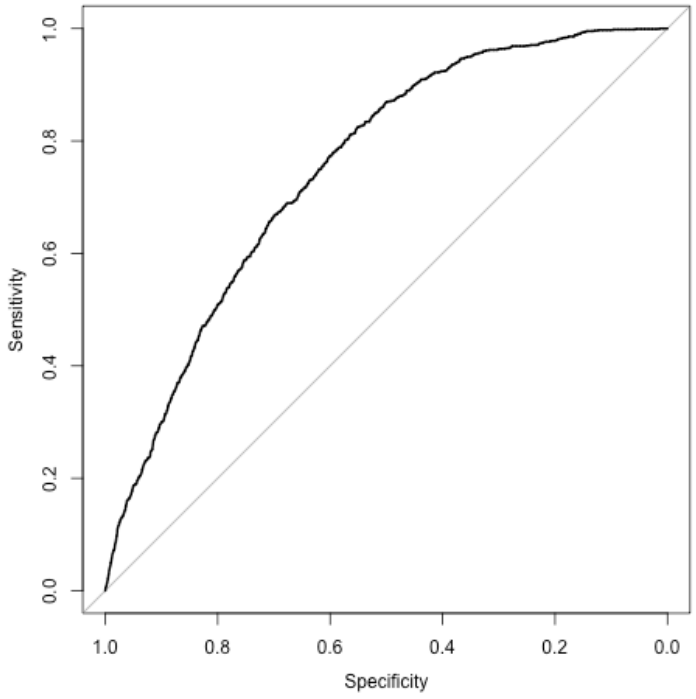


Figure 4.8. Receiver Operator Curve from the Full Model, Including Genome-wide Polygenic Score.

DISCUSSION

This is the first GWAS of ECIG use to date. Genome-wide significant SNPs were not detected for either CIG initiation or ECIG initiation after accounting for covariates. These results are typical of GWAS results in small samples (this sample N = 15,881). Further, GPS calculated from CIG initiation in a large training sample were not significantly associated with ECIG initiation in the test sample after adjusting for covariates.

This study found phenotypic similarities to other published reports. ECIG initiation has been reported to be near 40% in wave 4 of the Population Assessment of Tobacco

and Health (PATH) [291]. Importantly, this study did not differentiate between dual and exclusive use, so it possible the majority of participants were dual users similar to G4G. Similarly, Spit for Science (S4S) reported nearly 41% of participants had initiated ECIG, though dual use was not account for. It is unsurprising that this is a higher proportion than G4G as S4S is focused on college aged participants. Younger participants use ECIGs more frequently than older age ranges [292]. Further, these estimates of ECIG use, when considering dual users, is in line with Khouja and colleagues estimate from ALSPAC (ECIG initiation = 30%) [58]. Further, this study reported 64% of participants had initiated CIG use, similar to the G4G study [58]. Dual use was higher in this sample with 95% of ECIG initiators also engaging in CIG initiation. Prevalence of ECIG-exclusive initiation was reported at 0.3% in G4G, which is similar to ECIG-exclusive initiation in PATH (0.4%) [293].

Genome-Wide Suggestive SNPs for CIG and ECIG Initiation

There were no genome-wide significant SNPs for CIG initiation; however, there were two genome-wide suggestive SNPs. The most significant SNP was an intergenic region of chromosome 18 (18:61021122), with the nearest gene being *CDH20* (Cadherin 20). Previous research has utilized SNPs in this gene to create a genotype score for successful smoking cessation in a clinical trial of nearly 500 smokers examining nicotine replacement therapy [294]. On the other side of this intergenic location, this SNP is bounded by *HMGN1P31* (High Mobility Group Nucleosome Binding Domain 1 Pseudogene 31), a gene with an unknown function and has not been associated with substance use in previous research.

Likewise, no SNP reached genome-wide significance for ECIG initiation after accounting for covariates. However, several genes were genome-wide suggestive. The most intriguing of these suggestive markers was 2:115364757 in the *DPP10* gene. *DPP10* has previously been associated with asthma, a common respiratory disease which is also associated with ECIG initiation [295–297]. This biologically plausible gene should be marked for further investigation, especially with the rise of EVALI (E-cigarette or vaping associated lung injury [298,299].

That no SNP reached genome-wide significance is not unexpected in this study. GWASs require large sample sizes due to the small effect sizes of SNPs of common variation (MAF > 1%). While the G4G is a large sample, it is still underpowered to detect such small effects. Post-hoc power analyses indicated that this sample had about 5% power to detect genetic effects, with nearly 300,000 participants needed to reach 80% power. Further, ECIG-exclusive initiation in this study was estimated at 0.3%. To detail ECIG exclusive initiation in GWAS, using this proportion as a starting point, a sample size of more than 39 million participants would be needed. Future research efforts should continue to build larger data sets, ensuring greater statistical power to detect small effects from common variation. In addition, ensuring that all ancestries are represented in the analyses will increase sample size and increase external validity.

Association between Genome-Wide Polygenic Scores and ECIG Use

This study reported no significant association between the GPS for CIG initiation and ECIG initiation in a community-based sample. To date, three other studies have examined CIG GPS and ECIG initiation with inconsistent results. The initial study of CIG

initiation GPS did not find an association overall between CIG GPS and lifetime ECIG initiation in a small sample (N ~ 4,000) of twins from the Netherlands. When analyses were stratified by past tobacco use, a significant association was reported for former smokers of CIG [124]. A more recent study reported a significant association between lifetime CIG initiation GPS and ECIG initiation in a sample of young adults (age = 24) from the UK [58]. The final study did not examine a lifetime CIG initiation GPS, but did report significant associations with other GPS built from other CIG phenotypes (cigarettes per day, regular cigarette use) in an American college aged sample (18-25) [225]. Therefore, it would be advantageous for future studies to examine multiple facets of the smoking rather than focusing in on one behavior.

While there was no statistical association between the GPS and ECIG initiation, the model still performed adequately. Compared to model with only covariates and phenotypic variables, a model with GPS altered the Nagelkerke R^2 by less than 0.01%. However, as shown in Figure 4.9, the area under the curve is sufficient to be classified as adequately predictive of ECIG use. Most likely this association is drive by the dual use of ECIG and CIG use. The prevalence of ECIG-exclusive use is extremely small (0.3% of the sample) which is in line with other published research (e.g., Population Assessment of Tobacco and Health ECIG-exclusive lifetime initiation is 0.4% (unweighted)) [293]. Additionally, all the phenotypic covariates included in the model have previously been associated with ECIG use, leading to a model that accurately predicts ECIG use without genotypic data.

This study has several limitations. First, the proportion of ECIG-exclusive use was very low (0.3%). This suggests that any genetic variants may be masked by dual

initiation of ECIG and CIGs. This was accounted for in these analyses by using the opposite tobacco product as a covariate in regression models. Models that included genetic risk did not significantly alter the AUC suggesting that there is no additional information gained by using PRS in this sample. Secondly, even though the GPS takes into account variants with small effects, it is possible that these variants were not captured. This would lead to an incomplete picture of the genetic architecture shared between CIG and ECIG initiation. Thirdly, these analyses were limited to SIA White participants. Technological improvements have been made to allow for cross-ancestry estimate of GPS via PRS-CSx [300]. Future studies should continue to examine ancestral groups other than those of European descent

This study also demonstrated several strengths in addition to the limitations. The large sample size for GPS generation met the minimum for statistical power [273]. Further, this study replicated previous null findings of CIG GPS and ECIG lifetime initiation [124]. Lastly, the proportions of tobacco use were similar to other published reports

The *DPP10* gene, a novel gene for ECIG initiation, was identified as genome-wide suggestive. While this study was underpowered to detect genetic effects via GWAS, this gene should be marked for replication in other samples with greater power. Particular attention should be paid to this possible gene due to its previous association with respiratory diseases. These preliminary results report that this gene is associated with increased odds of ECIG use (OR = 1.29). This gene may play a role in the emergence of EVALI cases and should continue to be researched.

CHAPTER 5: THE EFFECT OF COUPON RECEIPT ON THE RELATIONSHIP BETWEEN INCOME AND PAST 12-MONTH ELECTRONIC AND CONVENTIONAL CIGARETTE USE IN ADULTS

INTRODUCTION

Tobacco use is more common among individuals at lower income levels. In 2016, approximately 32% of households with an annual income of less than $20,000 per year used tobacco products for at least some days. In comparison, approximately 12% of households making more than $100,000 per year used tobacco products [301]. 34 million Americans were estimated to live in poverty (e.g., annual household income of $24,339 for a family with two adults and two children) [302] in 2013-2014. Consequently, tobacco use may affect a significant proportion of the American population who are also financially vulnerable.

The Association between Income and Tobacco Use

CIG use creates a greater health burden and financial stress on low-income individuals compared to individuals with higher levels of income [303]. Individuals at lower income levels spend a greater proportion of their income on tobacco products [304]. Further, CIG use as well as ECIG use, are risk factors for several chronic diseases (e.g., cancers, cardiovascular disease), and these conditions impact individuals at lower income levels more severely [305,306]. Low-income populations are also less likely to have access to health insurance or health care compared to individuals at higher levels of income. Individuals who do not have health insurance tend to have worse health outcomes than

individuals with health insurance and generally receive worse quality health care compared to those with health insurance [307]. Low-income populations are particularly vulnerable to the negative consequences of tobacco use as a result of an increased immediate financial burden resulting from regular tobacco expenditures. Previous research has reported significant associations between CIG-exclusive use and dual use such that individuals with lower incomes were more likely to be users (Friedman & Horn, 2019). There are no reported significant associations between ECIG-exclusive use and income (Friedman & Horn, 2019).

Initial ECIG Price Point as an Obstacle to Initiation

Coupons produced by tobacco companies are a cigarette expenditure minimizing strategy (CEMS) that can reduce immediate costs precluding tobacco initiation and use. CEMS reduces the immediate purchase cost of various forms of tobacco use. Engaging in coupon use as a CEMS has been associated with increased use of tobacco products [308]. For example, receipt of coupons was associated with greater odds of CIG initiation [309], smoking relapse [310], and switching to regular smoking from experimental use [311] in adolescents and young adults. This strategy has been more widely used in low-income populations to reduce the cost of CIG use [312,313].

The use of coupons as a CEMS is likely to extend to ECIG use since they require a large initial investment to purchase the device. For example, rechargeable ECIG starter kits typically range from $25-$150 or more while the liquid refill kits cost $50-$75 monthly [314]. Therefore, the cost of ECIG may prevent lower income individuals from accessing this product [315]. Receipt of ECIG coupons and marketing materials was

associated with an increased likelihood of trying ECIGs by reducing the purchase price of these devices [316,317]. Consequently, lower income populations may use CEMS to help defray the initial price point of initiating ECIG use. However, to date, it is unclear the degree to which this strategy is used across all income levels or whether the receipt of coupons moderates this association.

Although low income and use of CEMS have been associated with CIG-exclusive and ECIG-exclusive use, it remains unclear whether receipt of coupons focused on a particular tobacco delivery system is also associated with the dual use of ECIG and CIG in adults. Likewise, it is unclear if income is associated with dual use. It is unclear whether receipt of coupons is associated with specific patterns of ECIG and CIG use (i.e., exclusive product use or dual ECIG and CIG use). This study has two aims: (1) describe the relationship between income and tobacco use, as categorized by product-exclusive use or dual use of ECIG and CIGs, and (2) detail how this relationship varies with receipt of coupons for ECIGs or CIGs. We anticipate that (1) the relationship between income and tobacco use will be similar to previously reported associations (Freidman & Horn, 2019), a significant association will be noted for all modalities of tobacco use (CIG/ECIG-exclusive and dual use) and (2) the association between tobacco use and income will receipt be stronger in low income groups and non-significant in higher income groups.

METHODS

Study Population

Data for this study were drawn from the publicly available files from Wave 3 of the Population Assessment of Tobacco and Health (PATH; N = 28,148) [318]. Described in detail elsewhere [319], PATH is a nationally representative longitudinal study of tobacco use and health. Our analytic sample consisted of adults aged 18-99 who had complete data on tobacco use, income, and coupon receipt. Most participants were White (77.7%) and Female (52.0%, Table 5.1).

Variables

Tobacco Use. Tobacco use, the outcome variable, was recoded into a 4-level variable. If a participant reported not using ECIGs or CIGs in the past 12-months, they were coded as a non-user. If an individual marked they had used CIGs in the past 12-months, but not ECIGs, they were coded as a CIG-exclusive user. The reverse is also true, reporting that one had used ECIGs in the past 12-months but not CIGs returned an ECIG-exclusive user. Finally, if the participant reported using both ECIGs and CIGs in the past 12-months, they were coded as a dual user.

Income. Income was divided into a 5-level ordinal variable, using the following prompt, “Which of the following categories best describes your total household income in the past 12 months?” Responses included: less than $10,000, $10,000-$24,999, $25,000-$49,999, $50,000-$99,999, and $100,000 or more as defined by the PATH study staff.

Receipt of Coupons. Coupon and promotion materials were probed with the following items: “In the past 12 months, received discounts or coupons for any of the following products: Cigarettes?” to which participants could endorse or not endorse and “In the past 12 months, received discounts or coupons for any of the following products: E-cigarettes or other electronic nicotine products (including e-liquid)?” to which individuals could endorse or not endorse. Each of these items was treated as binary with response options of yes or no.

Covariates. Several demographic factors (age, gender, race, education), which have previously been associated with ECIG as well as CIG use [320] were included as covariates. Gender was included as a PATH derived binary variable, representing male and female response options. Education was recoded into a four-level ordinal variable defined as less than high school, high school graduate or GED, some college, and bachelor’s degree or higher. The race was captured as a three-level, PATH-defined nominal variable defined as White, Black, or Other. Age was measured in PATH as a seven-level ordinal variable: 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+.

Statistical Analysis

Descriptive statistics were used to summarize the characteristic of the study participants. *Multinomial regression* was used to assess the association between income and past 12-month tobacco use. ECIG and CIG use do not necessarily occur independently of one another and dual use of these products is common [321–323]. McMillian and colleagues (2015) reported about a third of daily and non-daily smokers or CIGs reported they were also using ECIGs [113]. Consistent with other research, dual

users of these delivery systems may represent a distinct group from delivery system specific groups [324], and as such, this study applied multinomial modeling to take this unique group into account.

Moderation Analysis

To assess whether receiving coupons modifies the relation between income and tobacco use, *preliminary moderation analysis* was performed. To date, it remains unclear how coupons may influence the income and tobacco use association. One possible avenue for investigation is to test whether or not coupons moderate the association. Conceptually, a moderation analysis tests whether the moderating variable (M, coupon receipt) influences the direct relationship between the independent variable (X, income) and dependent variable (Y, tobacco use; Figure 5.1). M can moderate the relationship by either increasing or decreasing the magnitude of the pathway, between variables X and Y as shown in Figure 5.1, via pathway b from M to the relationship of X and Y. Previous research has reported that redeeming coupons was associated with a reduction in the odds of past 30-day abstinence of CIG use [325]. Though researchers did not explicitly model moderation, this may suggest that using coupons may influence the pathway by creating a situation where lower income is not prohibitive of CIG use. Explicitly modeling the moderation by coupons may reveal additional insight into why individuals with lower incomes disproportionately use tobacco products. It is conceivable that using a CEMS, such as coupons, may lead to greater uptake or continued use of a tobacco product. They may also influence an individual to switch administration routes or use both products simultaneously. CEMS seeks to minimize the

amount of money spent on a tobacco product; therefore, a moderating relationship may exist between income and tobacco use coupons.

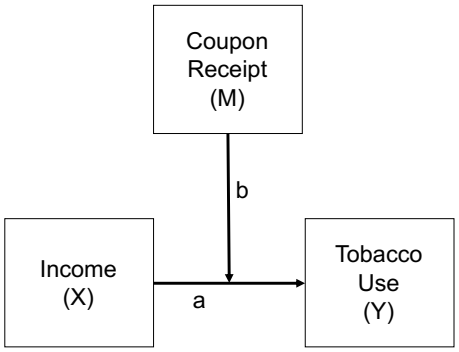


Figure 5.1. Conceptual Model for Moderation Analysis.

Multinomial regression approach allows the separation of the outcome variable into groups of CIG-exclusive users, ECIG-exclusive users, dual users, and non-users and allows for comparisons to be made between unordered groups. Multinomial regression reduces the possibility of misclassification bias by separating the delivery systems used by participants, including those participants who dual used both systems so dual users are not represented as both ECIG and CIG-exclusive users. All models were run in SAS 9.4 (SAS Inc., Cary, NC) and accounted for the complex survey design with PROC SURVEYLOGISTIC. Additionally, these analyses were adjusted for age, race, gender, and education level as these have been previously associated with ECIG and CIG use [326]. Replicate weights were estimated using Fay's variant of balanced repeated replication [327] as detailed in the PATH User Guide. Additionally, preliminary moderation analyses, with coupons as the moderator variable, were conducted to determine how coupon receipt may influence the tobacco use and income relationship. Moderation was tested in SAS by the addition of an interaction term between income and coupon receipt as well as through stratification of results based on coupon receipt.

RESULTS

Descriptive Statistics

Table 5.1. Summary Statistics for PATH Wave 3

Variable	N	Weighted %
Tobacco Use		
Non-User	14,548	72.4
CIG-exclusive	8,171	17.5
ECIG-exclusive	1,373	2.6
Dual User	4,056	7.5
Income		
< $10,000	4,634	11.6
$10,000-$24,999	5,686	19.1
$25,000-$49,999	5,975	22.8
$50,000-$99,999	5,866	26.7
> $100,000	3,917	19.8
ECIG Coupon Receipt		
Yes	1,078	3.4
No	26,962	96.6
CIG Coupon Receipt		
Yes	4,869	12.2
No	23,171	87.8
Gender		
Male	13,788	48.0
Female	14,334	52.0
Race		
White	19,899	77.7
Black	4,494	12.5
Other	2,951	9.89
Education		
< HS	3,714	11.1
HS Grad/GED	8,547	28.2
Some College	9,724	31.7
College grad or higher	6,025	28.0
Age		
18-24	8,453	12.4
25-34	5,824	17.5
35-44	3,972	16.1
45-54	3,804	17.2
55-64	3,389	17.2
65-74	1,891	12.3
75+	813	7.3

Approximately 27.6% of all participants engaged in some form of lifetime tobacco use. Of these, 9,544 (20.1%) used CIG or ECIG in the last 12 months. Most participants engaged in CIG-exclusive use (17.5%). Most participants had an annual household income of $50,000- $99,999 per year (26.7%). Approximately 3.4% of participants received ECIG coupons and 12.2% received CIG coupons (Table 5.1). [328,329]

Table 5.2 Description of Participants Receiving Coupons

	Any Coupon Receipt		
	Yes	No	
Variable	**N (Row %)**	**N (Row %)**	**p-value**
Tobacco Use			< 0.001
Non-User	1,085 (7.5)	13,432 (92.5)	
CIG	2,573 (31.7)	5,551 (68.3)	
ECIG	177 (13.0)	1,189 (87.0)	
Dual User	1,442 (35.8)	2,591 (64.2)	
Income			< 0.001
< $10,000	1,050 (22.8)	3,565 (77.2)	
$10,000 - $24,999	1,301 (23.0)	4,367 (77.0)	
$25,000 - $49,999	1,287 (21.6)	4,678 (78.4)	
$50,000 - $99,999	1,020 (17.4)	4,842 (82.6)	
> $100,000	411 (10.5)	3,501 (89.5)	
Sex			< 0.001
Male	2,360 (17.2)	11,357 (82.8)	
Female	2,912 (20.4)	11,385 (79.6)	
Race			< 0.001
White	4,010 (20.2)	15,832 (79.8)	
Black	727 (16.3)	3,742 (83.7)	
Other	462 (15.7)	2,474 (84.3)	
Age			< 0.001
18-24	981 (11.6)	7,449 (88.4)	
25-34	1,377 (23.7)	4,429 (76.3)	
35-44	948 (24.0)	3,009 (76.0)	
45-54	962 (25.4)	2,821 (74.6)	
55-64	738 (21.9)	2,632 (78.1)	
65-74	225 (12.0)	1,656 (88.0)	
75+	46 (5.7)	765 (94.3)	
Education			< 0.001
< HS	762 (20.6)	2,932 (79.4)	
HS Grad/GED	1,814 (21.3)	6,704 (78.7)	
Some College	2,013 (20.7)	7,692 (79.3)	
College grad or higher	681 (11.3)	5,335 (88.7)	

Individuals with household incomes of $50,000 and higher comprised nearly half of non-tobacco users (46.6%). Participants in the lowest income category (less than $10,000) represented 22.7% of all tobacco users. The second highest income group (those making between $50,000 and $99,999; Table 5.3) had the lowest frequency of tobacco use. Receipt of coupons, either for ECIG or CIG products, was well spread out with nearly equal distributions among the four lowest income categories. Individuals in the highest income category reported the lowest prevalence of coupon receipt (Table 5.3).

Table 5.3. Distribution of Tobacco Use and Coupon Receipt by Income Category

	Less than $10,000		$10,000-$24,999		$25,000-$49,999		$50,000-$99,999		$100,000 or more	
Variable	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Tobacco Use										
Non-Use	1,734	13.0	2,365	17.8	3,016	22.6	3,474	26.1	2,737	20.5
CIG-Exclusive	1,874	24.4	2,068	27.0	1,717	22.4	1,392	18.2	617	8.1
ECIG-Exclusive	192	24.4	265	21.2	299	23.9	287	22.9	209	16.7
Dual Use	834	21.8	988	25.8	943	24.6	713	18.6	354	9.2
Coupon Receipt										
ECIG	144	13.9	224	21.6	260	25.1	258	24.9	151	14.6
CIG	1,012	21.6	1,231	26.3	1,202	25.7	911	19.5	325	6.9

Modeling Results

CIG-Exclusive Use. Relative to non-tobacco users, income was significantly associated with CIG-exclusive use across all income levels after adjusting for covariates. Lower levels of household income were more strongly associated with past 12-month CIG use, with decreasing but still significant ORs for higher income levels relative to those making more than $100,000 ($aOR_{<10k}$ = 4.01, 95% CI = 3.38-4.76; $aOR_{50\text{-}99k}$ = 1.43, 95% CI = 1.24-1.65; Table 5.4). for addition to this significant association, CIG coupon receipt was associated with a roughly 25% decrease in the

odds of CIG use ($aOR_{Coupons}$ = 0.74 95% CI = 0.59-0.92; Table 5.4). The association between income and CIG use also remained significant after also estimating the effect of ECIG coupon receipt (OR= 5.69, 95% CI = 5.08-6.38; Table 5.4). Lower levels of income were more strongly associated with past 12-month CIG use, and attenuated but still significant increases in the odds as the income level rose ($aOR_{<10k}$ = 3.60, 95% CI = 3.02-4.29; $aOR_{50\text{-}99k}$ = 1.30, 95% CI = 1.12-1.51; $aOR_{ECIGCoupons}$ = 5.69, 95% CI = 5.08-6.38; Table 5.4)

ECIG-Exclusive Use. Relative to non-tobacco users, income was not significantly associated with ECIG-exclusive use across all income levels (Table 5.4). This association remained non-significant after adjusting for covariates and ECIG coupon receipt. However, there was a statistically significant association between CIG coupon receipt and ECIG-exclusive past 12-month use ($aOR_{Coupons}$ = 2.32, 95% CI = 1.74-3.10; Table 5.4). Likewise, there was no significant association between income and ECIG use after accounting for covariates and ECIG coupon receipt, though there was a significant association between ECIG past 12-month use and ECIG coupons ($aOR_{ECIGCoupons}$ = 1.40, 95% CI = 1.05-1.88; Table 5.4).

Dual Use. Relative to non-tobacco users, income was significantly associated with dual use across all income levels after adjusting for covariates. Similar to CIG use, the magnitude of the association was the highest as the lowest income level and attenuated, but remained significant, as the income level increased ($aOR_{<10k}$ = 3.65, 95% CI = 2.97-4.48; $aOR_{50\text{-}99k}$ = 1.51, 95% CI = 1.23-1.86; Table 5.4). This association remained significant after adjusting for CIG coupon (OR = 2.62, 95% CI = 2.10-3.28; Table 5.4) or ECIG coupon receipt (OR =7.61, 95% CI = 6.75-8.58; Table 5.4).

Table 5.4. Parameter Estimates for Association between income and Past 12-Month Tobacco Use by ECIG and CIG Coupon Receipt

	CIG User aOR (95% CI)	ECIG User aOR (95% CI)	Dual User aOR (95% CI)
CIG Coupon Receipt			
Income			
> $100,000	Reference	Reference	Reference
< $10,000	**4.01 (3.38-4.76)**	1.00 (0.73-1.26)	**3.65 (2.97-4.48)**
$10,000-$24,999	**3.02 (2.57-3.53)**	1.10 0.82-1.47)	**3.13 (2.61-3.75)**
$25,000-$49,999	**2.00 (1.70-2.26)**	1.03 (0.77-1.37)	**2.34 (1.95-2.82)**
$50,000-$99,999	**1.43 (1.24-1.65)**	0.88 (0.68-1.15)	**1.51 (1.23-1.86)**
Coupon Receipt			
No	Reference	Reference	Reference
Yes	**0.74 (0.59-0.92)**	**2.32 (1.74-3.10)**	**2.62 (2.10-3.28)**
ECIG Coupon Receipt			
Income			
> $100,000	Reference	Reference	Reference
< $10,000	**3.60 (3.02-4.29)**	0.98 (0.72-1.33)	**3.23 (2.62-3.99)**
$10,000-$24,999	**2.73 (2.33-3.21)**	1.10 (0.82-1.47)	**2.77 (2.29-3.34)**
$25,000-$49,999	**1.77 (1.53-2.05)**	1.02 (0.77-1.36)	**2.07 (1.71-2.51)**
$50,000-$99,999	**1.30 (1.12-1.51)**	0.88 (0.68-1.15)	**1.34 (1.08-1.67)**
Coupon Receipt			
No	Reference	Reference	Reference
Yes	**5.69 (5.08-6.38)**	**1.40 (1.05-1.88)**	**7.61 (6.75-8.58)**

Note. The outcome reference group is Non-User; all estimates adjusted for age, race, gender, education level, and complex survey design; bolded values indicate $p < 0.05$

Moderation. No statistically significant moderation of CIG coupon receipt was detected for associations between income and any tobacco product ($F(12, 100) = 1.73$, $p > 0.05$). However, there was evidence of moderation by ECIG coupon receipt on the relationship between income and CIG use ($F(12,100) = 2.73$, $p < 0.001$). Stratified analyses revealed a weak statistically significant relationship (aOR = 2.51, 95% CI = 1.50-4.16) for those who smoked CIG and received ECIG coupons for participants making between $50,000 and $99,999 (Table 5.5).

Table 5.5. Parameter Estimates of Past 12-Month CIG Use by ECIG Coupon Receipt Stratified by Income Level

	< $10,000	**$10,00-$24,999**	**$25,000-$49,999**	**$50,000-$99,999**	**> $100,000**
	β (SE) aOR (95% CI)	**β (SE) aOR (95% CI)**	**β (SE) aOR (95% CI)**	**β (SE) aOR (95% CI)**	**β (SE) aOR (95% CI)**
Receipt					
No	Reference	Reference	Reference	Reference	Reference
Yes	-0.11 (0.34)	0.38 (0.27)	0.05 (0.23)	**0.91 (0.26)**	0.18 (0.26)
	0.89 (0.46-1.75)	1.46 (0.85-2.50)	1.05 (0.67-1.65)	**2.51 (1.50-4.16)**	1.19 (0.71-2.00)

Note. The outcome reference group is Non-User; all estimates adjusted for age, race, gender, education level, and complex survey design; bolded values indicate $p < 0.05$

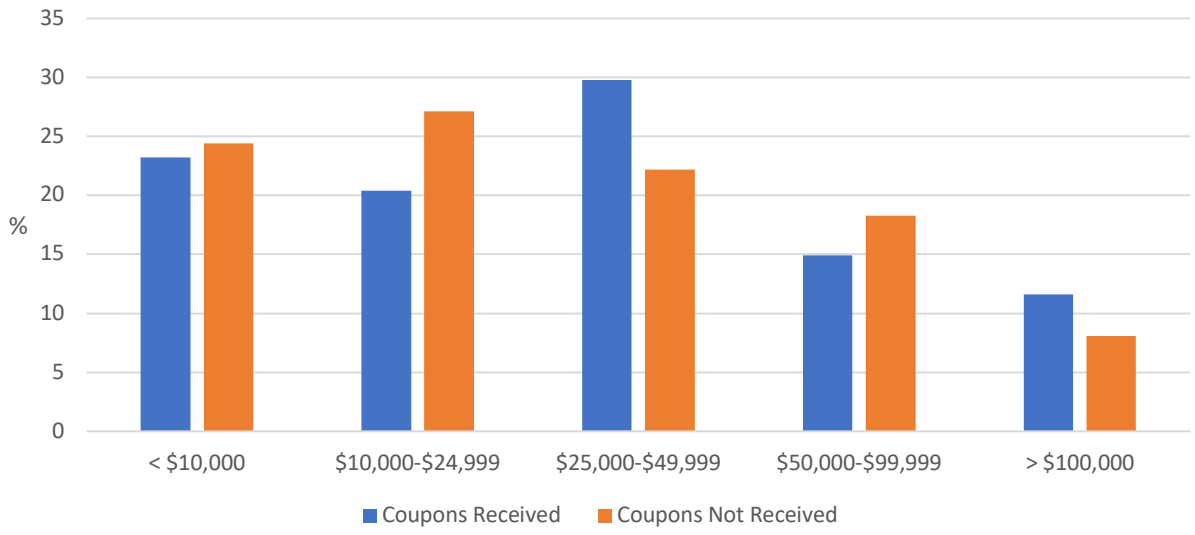


Figure 5.2. Proportion of CIG-Exclusive Users by Income Level and ECIG Coupon Receipt.

DISCUSSION

This study evaluated whether receipt of product-specific coupons influences the association between tobacco use and income in a nationally-representative sample. The results from this study indicate that the patterns of association between income and tobacco use vary across products. Further, receipt of coupons was independently

associated with tobacco use, with receipt of ECIG coupons associated with a reduction in the odds of CIG and dual use.

The prevalence of CIG-exclusive use in this sample (17.5%) was similar to that of other studies (approximately 14%) [330]. Additionally, the estimated prevalence of ECIG exclusive use in Wave 3 PATH was 2.6%, which was similar to those previously reported in other nationally representative samples (1.3% in 2014-2016 NHIS; 4.5% 2016 BRFSS) [52,326]. There was also a higher prevalence of dual use (7%) compared to previous reports (2.7% from Friedman & Horn, 2019), but similar to other nationally representative samples (7.0% 2016 BRFSS; [52]). The difference in the prevalence of ECIG-exclusive use and dual use may be due to the more recent collection for Wave 3 of PATH (2016-2017) compared to previous analyses (National Health Information Survey 2014-2016; Friedman & Horn, 2019) as ECIG-exclusive use has become more common over time in adults 18-25 [113]. Further, ECIG use among existing CIG users (dual use) has also been increasing [331].

Associations between Income and CIG-Exclusive Use and Dual Use

Participants with incomes of less than $10,000 were at greater odds of past 12-month CIG use compared to individuals making more than $100,000. as income increased the magnitude of the association was reduced. These results are similar to previous studies of income and tobacco use [61,113]. Tobacco is thusly used disproportionately by individuals with fewer means to indulge. One method promoted for reducing tobacco use is via raising taxes [332,333]. Therefore, monetary restrictions (e.g., raising taxes) may disproportionately affect more individuals of lower income.

Participants making less than $100,000 per year had a greater odds of past 12-month dual use of ECIG and CIG, compared to participants making more than $100,000 per year. These results are similar to the pattern previously reported for CIG-exclusive use [61,326]. This pattern of association may be due to individuals adding ECIG use to their existing CIG use [334]. As CIG users experiment with ECIGs, it is likely that their patterns of tobacco use should most resemble CIG-exclusive users because they likely continue to engage in CIG use.

There was no significant association between ECIG-exclusive use and income, as has been reported in other studies [326]. ECIG use is generally perceived as a "healthier alternative" to CIG use [335]. Prior literature has also reported individuals with lower income are less likely to engage in positive health behaviors (e.g., smoking cessation) compared to individuals with higher income [336–338]. ECIGs have been marketed products to reduce harmful exposure to the carcinogens in CIG [339] although it is unclear whether these products have been successful for this purpose. Nevertheless, many CIG users perceive ECIG to be safer than CIG [340]. Therefore, among CIG users, ECIG use may be considered a positive health behavior. However, it is possible that lower income smokers do not use ECIGs as harm reduction tools because this belief does not offset the initial higher price of ECIGs. Preferably, individuals would choose not to use tobacco at all; however, using a less harmful product may result in the reduction of lost time and money due to illness attributable to tobacco use. This study provides some initial evidence that reducing the price point of ECIGs via coupons is associated with greater use of ECIGs.

Receipt of Coupons Associated with Varying Patterns Nicotine Use

Receipt of CIG coupons was associated with greater odds of ECIG-exclusive and dual use; however, receipt of CIG coupons was associated with a 26% reduction in the odds of CIG-exclusive use (Table 3). It is possible that providing consumers with coupons means providing them with coupons for all products, including ECIGs, from the manufacturer. In this sample, 62% of participants who received CIG coupons also received ECIG coupons. Thus, individuals may be encouraged to switch to dual or ECIG-exclusive use through the receipt of coupons in manufacturer packs rather than for a specific product. Further research should continue to examine how coupons are received and utilized by consumers.

Receipt of ECIG coupons was associated with greater odds of all forms of tobacco use: ECIG-exclusive use, CIG-exclusive, and dual use. Previous research has reported that individuals who receive coupons from tobacco companies are more likely to report that tobacco companies care about their health and try to make cigarettes as safe as possible [341]. Exposure to tobacco marketing materials and coupon receipt has been associated with an increased willingness of an individual to try ECIGs or CIGs [342], thus increasing the odds of initiation of these tobacco delivery systems. Despite this, the prevalence of CIG use is decreasing [343] while ECIG use is increasing [344]. Further analysis should focus on how coupon receipt influences transitions from product-exclusive use to dual use, though the long-term health benefits or risks of ECIGs have yet to be detailed [345].

The receipt of ECIG coupons was showing an association with greater odds of all tobacco use, additional models were tested to investigate whether ECIG coupons

moderated the relationship between income and tobacco use. Significant statistical evidence of a weak interaction between income and ECIG coupons was detected. Further investigation showed that the moderating effect was driven by individuals making $50,000-$99,999 such that receipt of ECIG coupons was associated with a decrease in CIG use. Participants at lower incomes may have transitioned into a higher income category, but remained in the opt-in group for coupons (i.e., they continued to receive coupons). Individuals may also transition from user to non-user as they move into different income groups, thus rendering a negative effect of coupon receipt. Future studies should examine how transitioning between income groups impacts tobacco use.

Previous research has reported on the possibility that receipt of coupons may contribute to the disparity in smoking by socioeconomic status. Other studies have shown that tobacco companies use direct marketing to target individuals, specifically women of low SES [329,346]. Tobacco companies have used strategies such as coupons to market their products for over 40 years. More recent research has shown that coupon saving is considerably high among adolescents and young adults, creating another barrier to smoking cessation for these populations [347]. Though these analyses do not allow for any conclusions to be drawn about the potential for targeting individuals with low income, future research should examine possible disparities in the receipt of ECIG coupons and determine if ECIG couponing resembles traditional CIG couponing.

Limitations

One major limitation of this study is the inability to distinguish who redeemed and who simply received coupons or promotional items. Previously reported research has

demonstrated that individuals who receive coupons are different from those who do not [348]. Coupon receivers tend to be white, middle aged (25-44), sexual minority females with higher levels of nicotine dependence [348]. Couponing, in the United States, is limited to an "opt-in" situation whereby individuals must agree to receive coupons which may indicate significant differences between those who do and do not receive coupons[346,349]. Future studies should continue to examine those individuals who receive coupons and how these coupons (particularly those for ECs) may influence tobacco use, especially in terms of tobacco use maintenance, tobacco product switching, and tobacco use cessation. Studies should also focus on the receipt of promotional material that takes place where individuals do not need to opt-in. These places may be bars, restaurants, or other private spaces where brand ambassadors are distributing promotional materials (e.g., t-shirts, keychains), samples, or coupons.

These analyses were conducted cross-sectionally, using only wave 3 of PATH. Future studies should probe how coupons influence transitions from product-specific use to dual use or vice versa. The moderation analysis is likely underpowered to detect a significant statistical interaction due to the small sample size since only 3.4% of the sample had received ECIG coupons [350]. As receipt of coupons becomes more common, it may be the case that larger sample sizes will be available to replicate and confirm this preliminary finding. There is also a causal assumption in moderation analysis such that the exposure of interest causes the outcome of interest (pathway a in Figure 5.1); future research should expand on these findings by modeling moderation in a longitudinal analysis. Finally, future studies should continue to study the possible moderation by

using more refined exposure group assignments, such as detailing if the participants actually used the coupons or merely received them.

Strengths

This study also has several strengths along with the above limitations. First, this study used data from a nationally representative sample of non-institutionalized individuals living in the United States. Appropriate statistical modeling allowed for all individuals who completed the questions regarding income, couponing, and tobacco use to be included while also accounting for the complex sampling design of PATH. Therefore, the results from the current study is more generalizable to broader population. Secondly, we used a multinomial logistic regression approach which allowed the outcome groups to be separated by using patterns to avoid misclassification and bias of model estimates. The use of multinomial logistic regression allowed for the groups of users to be modeled simultaneously and allowed for more accurate estimates of current (past 12-month) CIG-exclusive, ECIG-exclusive, and dual use prevalence in the United States. Lastly, these analyses accounted for the *product-specific* couponing and produced interesting results: ECIG coupons had a competing effect on the past 12-month CIG-exclusive use, and may represent an opportunity for individuals to experiment with ECIGs and possible switch products. Future studies need to account for the product being advertised and not just general advertisements or coupons and better understand the influence of the coupons on the potential switching between CIG-exclusive and ECIG or dual use.

Measuring income on an ordinal scale is an additional strength that builds upon previous research. Utilizing the PATH derived income variable allows for a more nuanced view of income rather than the more common use of a percentage of the federal poverty level (FPL). Previous research around income and tobacco use has focused on FPL as the level of measurement for income [113,326,351]. By focusing on the level of income (e.g., less than $10,000 versus greater than $100,000) rather than the percentage of FPL, a clearer picture of the association between income and tobacco use emerges. Regardless of the level of measurement for income, the same pattern is clear: lower levels of income are associated with greater odds of tobacco use relative to individuals with higher income. Specifically, relative to the highest income group, each income group shows an association with tobacco use suggesting that no income group is being targeted for tobacco use though this targeting may be present in targeting of communities rather than individuals [352]. Further research should continue to resolve the association between income and tobacco use by measuring income as a continuous measure as well as modeling individual income within the context of the larger neighborhood environment.

Public Health Implications

Since the Tobacco Master Settlement Agreement of 1998, tobacco companies have reduced, or ceased, their public advertising practices; however, direct mail marketing remains a primary means for tobacco companies to communicate with their consumers [353]. Previous research has shown individuals who receive coupons for tobacco products are less likely to successfully cease tobacco use [325] as well as increase the odds of

progression of smoking to regular or daily smoking from experimental use [354]. These results may suggest that advertising of ECIGs may result in fewer individuals using CIGs exclusively and might foster switching to dual use. For those who are considering ECIG use as a harm reduction tool to CIG use [355], exposure to advertising may offer an opportunity to engage in ECIG use. However, effort should continue to be expended on promoting the notion that there is no safe tobacco product when compared to no product use.

CHAPTER 6: DISCUSSION

This chapter summarizes and synthesizes the results from the research previously detailed in Chapters 2 through 5. It is divided into two sections that address the major knowledge gaps of this dissertation as identified in Chapter 1 (Introduction). The first knowledge gap set out to quantify the relative degree to which genetic and environmental influences impact ECIG lifetime initiation and to what degree are those factors shared with CIG initiation. While there was statistical evidence of genetic overlap there was no detection of genome-wide significant molecular genetic effect in the GWAS. The second knowledge gap was addressed by examining specific genetic and environmental factors associated with CIG and ECIG initiation. Specific genes and biological pathways were found to associated with several CIG phenotypes. Income (a specific environmental influence) was associated with CIG and dual use, but not ECIG-exclusive use, suggesting dual users represent a distinct class of tobacco user.

Knowledge Gap 1 Results: Preliminary Evidence of Genetic and Environmental Overlap for CIG and ECIG Initiation

This dissertation estimated the degree to which there were overlapping genetic and environmental factors shared between CIG and ECIG initiation. Chapter 2 used a twin study of adolescents and young adults and detected significant overlap in additive genetic (r_g = 0.76, p = 1), shared environmental influences (r_c = 0.68, p = 0.32), and unique environmental influences (r_E = 0.87, p = 0.01) between CIG and ECIGs. While these estimates were non-significant, dropping these parameters resulted in worse

model fit, suggesting that these parameters cannot be excluded in bivariate models of ECIG and CIG initiation. Chapter 4 extended twin study results by examining genetic overlap using a genome-wide polygenic score (GPS) for CIG initiation to determine whether there was significant genetic overlap in measured genetic variants with ECIG initiation. Analyses did not identify any significant specific genetic variants that contributed to both phenotypes. However, power analyses showed this study did not have adequate power to detect genetic effects. Nevertheless, prior studies have reported significant association between a GPS for CIG use and ECIG initiation in a sample of young adults (age = 24) from the UK. Additional studies have reported significant associations with GPS for other CIG use behaviors (e.g., regular cigarette use) and ECIG initiation [58,124,225]. Therefore, additional research is needed in appropriate samples to probe the overlapping genetic influences of CIG and ECIG initiation. Additionally, these studies reported prevalence rates of ECIG initiation similar to those identified in Genes for Good (Genes for Good = 30.0%, other samples = 4.7%, Netherlands Twin Registry – 37.3%, S4S-EUR). Further study is recommended in significantly larger samples (N~ 300,000) with appropriate power to verify the genetic effects that are shared between ECIG and CIG initiation [356].

Translating Results from Knowledge Gap 1 for Public Health: Advancing Future Public Health Strategies to Influence ECIG Initiation by Characterizing Genetic and Environmental Overlap with CIG Initiation

Results addressing Knowledge Gap 1 suggest that factors influencing CIG initiation may also impact ECIG initiation. Consequently, it is possible that previous smoking prevention efforts may also impact ECIG use. For example, previous research has shown an association between peer group attitudes and CIG and ECIG use[67,357]. Peers have previously been targeted by public health prevention efforts for CIG use [358–360]. These strategies may be modified to target ECIG use rather than CIG. Therefore, exploring the degree to which strategies that limit CIG initiation may also limit ECIG initiation because similar genetic and environmental influences impact ECIG initiation as CIG.

Knowledge Gap 2 Results: Income and Coupon Receipt

Chapter 5 assessed the association between a specific environmental factor (income) on CIG and ECIG use. These results reflected other published results of income and CIG or ECIG use [46]. Specifically, those with lower income had greater odds of CIG use. However, there was no significant association with ECIG use. Additionally, moderation was tested to determine if the receipt of coupons influenced the association between income and tobacco use. There was a statistically weak, but significant, interaction between ECIG coupons and income and the odds of past 12-month CIG use; receiving ECIG coupons led to a two and half time increase in the odds of using CIGs in

the past 12 months. However, the sample size was quite small which necessitates further research into this possible association with larger samples [350].

Translating Income and Moderation of Coupon Results for Public Health.

Coupon receipt was associated with greater odds of all forms of tobacco use (product-exclusive and dual use). Cross-product coupon receipt was associated with an increase in the odds of being a user of the other product (CIG users who received ECIG coupons OR = 5.69, 95% CI = 5.08-6.38; ECIG users who received CIG coupons OR = 2.32, 95% CI = 1.74-3.10). Individuals who received ECIG coupons were more likely to use CIGs as well as ECIGs. This could be due to several reasons. First, tobacco coupon receipt is an opt-in scenario in the United States. People must willingly sign up to receive coupons and marketing materials from tobacco companies. Second, coupon receipt of any tobacco product may prime an individual to use tobacco regardless of the delivery form [361,362]. This priming effect has been established for CIG use and is used in state-sponsored counter programming to tobacco marketing initiatives [363]. Further research is needed to understand coupon receipt (i.e., what makes individuals opt-in to coupons), if individuals are being target, and the effect coupon receipt has on switching from a product exclusive users to a dual user.

The trend of coupon receipt increasing the odds of tobacco use did not hold true for receipt of CIG coupons and past 12-month CIG use. Receipt of CIG coupons was associated with a roughly 25% reduction in the odds of past 12-month CIG-exclusive use (OR = 0.74, 95% CI = 0.59-0.92). All other coupon receipt was associated with increased odds of tobacco use (Chapter 5, Table 5.4). It is possible that individuals are

being pushed from CIG-exclusive use to dual use as receipt of CIG coupons was associated with a more than two and half fold increase in the odds of dual use. This may happen as consumers are given coupons for all products for a particular manufacturer when they agree to receive marketing (i.e., signing up for Marlboro marketing may also expose the consumer to marketing for JUUL and IQOS, as all are manufactured by Philip Morris). Further research is needed to understand how individuals are receiving and using coupons from tobacco companies.

Translation for Public Health

Public health professionals are interested in preventing the morbidity and premature mortality associated with tobacco use. Detailing this specific environment (knowledge gap 2) has led to two innovations, one for prevention and one for research, that may be utilized by public health professionals. First, individuals who receive coupons may be priming themselves for additional tobacco use [361,362]. Prevention efforts could focus on negating this priming effect by including material for smoking cessation along with the marketing material. Second, smoking cessation studies should consider how individuals transition from ECIG-exclusive user to non-user and the similarities this trajectory may have with CIG use. Additional research should also consider how dual users move toward smoking cessation: how can individuals quickly cease use of both products? Perhaps it is easier to cease use of one product compared to the other, which would inform future cessation efforts as to which product to focus on first. Further research is needed to understand the dynamics of CIG and ECIG use.

Knowledge Gap 2 Results: Genome-Wide Association Study

Chapter 4 used GWAS to determine whether there were specific genetic variants that were associated with ECIGs. No GWAS results reached genome-wide significance, but there were several variants that were genome-wide suggestive. Further, one SNP in these suggestive results was in a biologically plausible pathway linked to respiratory disease. Therefore, further investigation of this variant is encouraged in samples that are adequately powered to detect genetic effects. Additionally, as the prevalence of ECIG use is low, oversampling for ECIG-exclusive users is encouraged to further investigate. Similar to other published GWAS of CIG initiation (see Chapter 3 for further detail), there were no genome-wide significant associations detected in this small sample.

Translating Genetic Association Results for Public Health.

A suggestive SNP in *DPP10* was reported from the univariate GWAS. This gene should be marked for further investigation with ECIG use. This gene has previously been associated with asthma, a respiratory disease[364–366]. EVALI (ECIG or Vaping Associated Lung Injury) has become a health concern for individuals using ECIGs[298]. This gene may contribute to EVALI as *DPP10* is already active within the respiratory system. Further research with adequately powered samples is encouraged, especially as EVALI prevalence increases and the rate of initiation of ECIGs continues to increase.

These results suggest that novel methods of aggregating GWAS results may lead to more actionable findings (Chapter 3). GWAS was aggregated by biological function via DAVID in this dissertation. This aggregation show consistency among

GWAS results that were otherwise somewhat inconsistent. Aggregating in this fashion provides biological insights which may lead to novel treatment development for nicotine use.

Additional Results and Lessons for Future Research and Public Health

The following sections summarize additional insights that were established from this work to advance genetic epidemiological research of ECIGs and tobacco more broadly: 1) inconsistent GWAS results may be due, in part to inconsistent measurement of tobacco use phenotypes and, 2) modeling of tobacco use must consider how dual users impact the results so as to not report biased estimates of effect.

Inconsistent Genetic Epidemiological Studies of CIG Use Encourages Careful Measurement of ECIG Use for Similar Study Designs.

Chapter 3 demonstrated how measures that operationalize a conceptual variable for tobacco use may lead to different results. For example, an adult sample (18 and older) using DSM symptom count reported a significant association with *AP2A2* for ND [194]. In contrast, another adult sample (18 and older) using the FTCD identified an association between ND and a different gene, *CHRNB3*[206]. These results show that the measure may influence the results even though both of these studies were mostly likely underpowered. These results encourage thoughtful consideration of the measures needed to study a phenotype of interest in the study design phase. In addition, using more than one measure for a particularly important phenotype may be necessary for future study aggregation for meta-analysis or for comparisons against other published

results. While measurement may not address all limitation of GWAS, this is an important facet that is correctable.

Future studies of ECIG use should pay attention to the lessons learned from genetic epidemiological studies of CIG use, particularly as measurement of ECIG use is under active development. For instance, a standardized definition of ECIG initiation is strongly encouraged. Individuals may be misclassified as never users if they are required to meet a threshold (e.g., owning a vape, or using one container of e-liquid). Chapter 3 discusses how the inconsistency of results are due, in part, to differing definitions of smoking initiation. For example, *DLC1* was significantly associated with SI when asking individuals to classify themselves as ever versus never smoker [144]. However, when smoking initiation was defined as smoking more than 100 cigarettes in one's lifetime, this gene was not associated with SI[164]. ECIG researchers are advised to not repeat these mistakes to increase the consistency of results. Several measures have been developed to measure ECIG-related nicotine dependence including the Penn State Electronic Cigarette Dependence Index (PS-ECDI) and the e-cigarette Wisconsin Inventory of Smoking Dependence Motives (e-WISDM) among others [367]. As multiple instruments are developed in parallel, there is risk of future inconsistencies for genetically-informed ECIG research. It is impossible to stop the development of multiple measures of ECIG use especially as this area of study is in its infancy. Given this reality, future studies may need to consider the progression of genetic epidemiology studies of CIG use and the related inconsistency of results. Further, researchers must pay careful attention to the operationalization of their variables and ensure they are captured the conceptualized variable adequately if multiple measures are developed.

Measurement, as detailed in Chapter 3, is a key component that requires more thorough thought in the study design phase. Moving forward, it would be best to standardize operational measures of conceptual tobacco behaviors, especially as it applies to ECIG use. ECIGs represent a novel tobacco product that appears to share similar genetic and environmental influences with CIG use. Therefore, ECIG researchers should take heed of the lessons learned from years of CIG research and agree to common items that could be used to assess various facets of ECIG use. There are currently fewer measures of ND arising from ECIG use [368] compared to ND from CIG use. Reducing these measures to a single instrument may be advantageous in future studies of ECIG use, particularly in genetic epidemiology. Which measure researchers choose to assess ND arising from ECIGs with may lead to inconsistent results.

Environmental factors, like genetic factors, have been consistently identified with tobacco use. For instance, having peers who use CIG increases the likelihood that one will use CIGs themselves [357]. Similarly, policies impact the expression of nicotine dependence. The proliferation and ease of access to pharmacological treatments (e.g., the patch, nicotine gum) have led to a decrease in nicotine dependence, though additional avenues should also be examined [369]. These environmental factors present ways for modifying the risk of tobacco use in a low-risk, high-reward fashion (i.e., Changing a policy or ensuring one's child is not associating with those who use could drastically reduce tobacco use). However, the measures of nicotine use may be slightly different, compared to genetic research, as environmental studies may have the time and ability to dig further into tobacco use. Though beyond the scope of this dissertation,

additional time and consideration should be placed into the measurement of environments during the study design phase. Measurement of environmental influences is also an important step in genetically informed studies, although beyond the scope of this paper. Environmental measures should also be given as much attention as how the outcome and genetics are measured.

The Importance of Modeling Tobacco Use Carefully

Chapter 5 showed how dual users are phenotypically similar to CIG-exclusive users as it pertains to income level and tobacco use. Specifically, the magnitude of associations between of income and CIG-exclusive use ($OR_{<10k}$ = 4.01, 95% CI= 3.38-4.76) was similar that of dual users ($OR_{<10k}$ = 3.65, 95% CI = 2.97-4.48). Generally, lower income levels were significantly associated with higher odds of dual and CIG-exclusive use compared to individuals making $100,000 or more per year. In contrast, no statistically significant association between ECIG-exclusive use and income and use was detected ($OR_{<10k}$ = 1.00, 95% CI =0.73-1.26). These results suggest CIG users that dual users are adding ECIG use to their existing behaviors. The patterns of association with income differed by class of CIG/ECIG user: 1) Non-users, 2) CIG-exclusive users, 3) ECIG-exclusive users, and 4) dual users. These results agree with previously published results of CIG and ECIG use and income. Further, ECIG use was not associated with income [46]. Therefore, it is important to include dual use as a category of tobacco user as the pattern of association changes between tobacco use categories. Examining CIG or ECIG use needs to include cross product use in the statistical analysis due to exposure to nicotine via two avenues, which may be used in concert.

This is of particular concern over other tobacco forms due to the high prevalence of dual use among CIG and ECIG users. Dual use was estimated to be 7.5% (weighted) in chapter 5. Therefore, there are many users who engage in both product use. Use of other tobacco products should be used as a covariate in any regression (i.e., ECIG use should be a covariate for models of CIG use) to account for dual use. Multinomial regression should be preferred whenever possible.

Recommendations for Future Work

Future studies should continue to explore the genetic overlap between CIG and ECIG use. While no statistically significant genetic overlap was detected between CIG and ECIG initiation using GPS in Chapter 4, there is still a suggestion of genetic effects for both CIG and ECIG use using a twin study in Chapter 2. However, there remains inconsistencies when comparing measured genetic effects to estimates of heritability from twin studies. This may be due to two reasons. First, the parameter estimates of the magnitude of associations generated from GWAS may be biased. This would lead to differences in estimates of heritability from GWAS compared to those generated from twin studies. In general, these molecular genetic estimates are well short of the heritability estimated by twin studies [370]. There are several reasons as to why this occurs. It may be that the genetic liability is not inherited in an additive fashion but may be the result of non-additive genetic effects (e.g., epistasis or gene-environment interaction). Further, epigenetic processes (the influence of environmental factors that do not involve actual alterations in the DNA, but are involved with differential expression of the genes) rather than genetic effects may influence the heritability of certain

phenotypes [371]. Second, heterogeneity in outcome measurement may lead to inconsistent results across GWAS. Further, this would lead to an inability to replicate GWAS results. It is also possible that the traits are misclassified (i.e., labeling one an ECIG user, when they are actually a dual user) leading to measurement error which would produce inconsistent results [372,373]. Tobacco use is more complex and difficult to assess than other clinical phenotypes that have reported many significant GWAS results (e.g., height). Height is a phenotype that is unmistakable (i.e., height can be accurately ascertained) and unable to be hidden. Tobacco use is more covert compared to height, relies on self-report, and is subject to social desirability bias. Therefore, it may be more difficult to detect and measure though both phenotypes are polygenic and complex [374].

Final Conclusions of Dissertation

There were three conclusions that could be applied to future genetic epidemiologic studies of ECIG initiation and use. First, latent genetic effects were established through a twin study. While this study was small, the results suggest that there are significant genetic influences on ECIG initiation. Therefore, genetic association studies should continue to be investigated for its association with ECIG initiation and use. A second study probed molecular measured genetic effects that were both unique to ECIG and shared with CIG initiation. While this study did not report significant effects, there was a genome-wide suggestive association that was detected in univariate analysis of ECIG. This SNP resides in a gene with biological plausibility related to nicotine function which should be investigated further. These two study designs provided convergent evidence

that genetic influences are important for ECIG initiation. More research is required to understand the precise nature of these genetic effects. ECIGs expose individuals to nicotine creating the possibility of nicotine dependence arising in ECIG users. Nicotine has been linked to many negative health outcomes [375]. ECIG use is growing in popularity [281,376]. Therefore, understanding how individuals initiate use may help prevention efforts for ECIGs, stopping the possible exposure to nicotine and possible development of nicotine dependence.

This dissertation also showed consistency in GWAS results. A DAVID analysis of results from a scoping review of the literature provided consistency of results. These consistencies arose from gene- and biological pathway levels rather than SNPs. The results suggest that aggregating GWAS results would result in more replication of results (i.e., rather than replicating SNPs, replicating genes or biological pathways). Aggregating genetic effects in this manner may guide future research target specific genes and pathways which will lead to creating understanding of the biological and potential prevention and cessation targets.

This dissertation also probed an environmental influence in addition to genetic effects. These results suggest that modeling of CIG and ECIG use needs to account for cross-product use. Future studies of CIG and ECIG also need to account for this dual use, otherwise the study results will be biased.

Therefore, although these results are limited by the sample sizes and prevalence of ECIG initiation, they provide preliminary evidence that genetic influences are associated with ECIG use. Further, specific environmental influences may not impact ECIG users in the same manner as CIG users. Further genetic epidemiologic

investigation is warranted and encouraged from these results. However, future results should attempt to aggregate genetic effects with plausible biological pathways. These results also encourage additional study of ECIG use to increase the health and wellness of society. ECIG use is increasing and continues to expose users to nicotine and the negative health effects associated with this exposure. It is possible other specific environments will be identified that influence CIG and ECIG use in the same manner. Prevention and cessation efforts built around these environments for CIG could be modified and applied to ECIG use.

APPENDIX A. Supplementary Tables for Chapter 3.

Due to the large tables created from Chapter 3, supplementary tables S3.1 to S3.5 are available online at:

https://osf.io/nzgf9/?view_only=23911760b4ad4188aa2e4024b5a9090c

REFERENCES

1. US Department of Health and Human Services. *The Health Consequences of Smoking—50 Years of Progress*. Centers for Disease Control and Prevention (US); 2014. doi:NBK179276
2. Bold KW, Krishnan-Sarin S, Stoney CM. E-cigarette use as a potential cardiovascular disease risk behavior. *Am Psychol*. 2018;73(8):955-967. doi:10.1037/AMP0000231
3. Rossheim ME, Livingston MD, Soule EK, Zeraye HA, Thombs DL. Electronic cigarette explosion and burn injuries, US Emergency Departments 2015-2017. *Tob Control*. 2019;28(4):472-474. doi:10.1136/TOBACCOCONTROL-2018-054518
4. Outbreak of Lung Injury Associated with the Use of E-Cigarette, or Vaping, Products | Electronic Cigarettes | Smoking & Tobacco Use | CDC. Accessed April 4, 2022. https://www.cdc.gov/tobacco/basic_information/e-cigarettes/severe-lung-disease.html
5. Sarigiannis DA, Karakitsios SP, Gotti A, Liakos IL, Katsoyiannis A. Exposure to major volatile organic compounds and carbonyls in European indoor environments and associated health risk. *Environment International*. 2011;37(4):743-765. doi:10.1016/J.ENVINT.2011.01.005
6. Abdel-Shafy HI, Mansour MSM. A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. *Egyptian Journal of Petroleum*. 2016;25(1):107-123. doi:10.1016/J.EJPE.2015.03.011
7. Goniewicz ML, Smith DM, Edwards KC, et al. Comparison of Nicotine and Toxicant Exposure in Users of Electronic Cigarettes and Combustible Cigarettes. *JAMA Network Open*. 2018;1(8):e185937-e185937. doi:10.1001/JAMANETWORKOPEN.2018.5937
8. Wills TA, Soneji SS, Choi K, Jaspers I, Tam EK. E-cigarette use and respiratory disorders: an integrative review of converging evidence from epidemiological and laboratory studies. *Eur Respir J*. 2021;57(1). doi:10.1183/13993003.01815-2019
9. Benowitz NL. Nicotine and Cardiovascular Disease. In: *Effects of Nicotine on Biological Systems*. Birkhäuser Basel; 1991:579-596. doi:10.1007/978-3-0348-7457-1_74
10. Kutlu MG, Parikh V, Gould TJ. Nicotine Addiction and Psychiatric Disorders. *Int Rev Neurobiol*. 2015;124:171. doi:10.1016/BS.IRN.2015.08.004
11. Baker TB, Breslau N, Covey L, Shiffman S. DSM Criteria for Tobacco Use Disorder and Tobacco Withdrawal: A Critique and Proposed Revisions for DSM-5. *Addiction (Abingdon, England)*. 2012;107(2):263. doi:10.1111/J.1360-0443.2011.03657.X
12. Jensen K, Afroze S, Munshi MK, Guerrier M, Glaser SS. Mechanisms for nicotine in the development and progression of gastrointestinal cancers. *Transl Gastrointest Cancer*. 2012;1(1):81-87. doi:10.3978/J.ISSN.2224-4778.2011.12.01
13. Crowley-Weber CL, Dvorakova K, Crowley C, et al. Nicotine increases oxidative stress, activates NF-kappaB and GRP78, induces apoptosis and sensitizes cells to genotoxic/xenobiotic stresses by a multiple stress inducer, deoxycholate:

relevance to colon carcinogenesis. *Chem Biol Interact*. 2003;145(1):53-66. doi:10.1016/S0009-2797(02)00162-X
14. Chu CJ, Yang YC, Wei JX, Zhang L. Association of nicotinic acetylcholine receptor subunit alpha-4 polymorphisms with smoking behaviors in Chinese male smokers. *Chin Med J (Engl)*. 2011;124(11):1634-1638. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med8&NEWS=N&AN=21740768
15. Wassenaar CA, Dong Q, Amos CI, Spitz MR, Tyndale RF. Pilot study of CYP2B6 genetic variation to explore the contribution of nitrosamine activation to lung carcinogenesis. *Int J Mol Sci*. 2013;14(4):8381-8392. doi:10.3390/IJMS14048381
16. Breland A, Soule E, Lopez A, Ramôa C, El-Hellani A, Eissenberg T. Electronic cigarettes: what are they and what do they do? *Ann N Y Acad Sci*. 2017;1394(1):5-30. doi:10.1111/nyas.12977
17. Marques P, Piqueras L, Sanz MJ. An updated overview of e-cigarette impact on human health. *Respir Res*. 2021;22(1). doi:10.1186/S12931-021-01737-5
18. Rehan HS, Maini J, Hungin APS. Vaping versus Smoking: A Quest for Efficacy and Safety of E-cigarette. *Curr Drug Saf*. 2018;13(2):92-101. doi:10.2174/1574886313666180227110556
19. Williams M, Talbot P. Design Features in Multiple Generations of Electronic Cigarette Atomizers. *International Journal of Environmental Research and Public Health*. 2019;16(16). doi:10.3390/IJERPH16162904
20. Noel JK, Rees VW, Connolly GN. Electronic cigarettes: a new “tobacco” industry? *Tob Control*. 2011;20(1):81. doi:10.1136/TC.2010.038562
21. Hiler M, Karaoghlanian N, Talih S, et al. Effects of electronic cigarette heating coil resistance and liquid nicotine concentration on user nicotine delivery, heart rate, subjective effects, puff topography, and liquid consumption. *Exp Clin Psychopharmacol*. 2020;28(5):527-539. doi:10.1037/PHA0000337
22. Yingst J, Foulds J, Hobkirk AL. Dependence and Use Characteristics of Adult JUUL Electronic Cigarette Users. *Subst Use Misuse*. 2021;56(1):61-66. doi:10.1080/10826084.2020.1834582
23. Prochaska JJ, Vogel EA, Benowitz N. Nicotine delivery and cigarette equivalents from vaping a JUULpod. *Tob Control*. Published online 2021. doi:10.1136/TOBACCOCONTROL-2020-056367
24. Vallone DM, Cuccia AF, Briggs J, Xiao H, Schillo BA, Hair EC. Electronic Cigarette and JUUL Use Among Adolescents and Young Adults. *JAMA Pediatrics*. 2020;174(3):277-286. doi:10.1001/JAMAPEDIATRICS.2019.5436
25. Rao P, Liu J, Springer ML. JUUL and combusted cigarettes comparably impair endothelial function. *Tobacco Regulatory Science*. 2020;6(1):30-37. doi:10.18001/TRS.6.1.4
26. JUUL Labs: FDA Investigation Timeline - TobaccoTactics. Accessed April 4, 2022. https://tobaccotactics.org/wiki/juul-labs-fda-investigation-timeline/
27. American Psychiatric Association. *Diagnositc and Statistical Manual of Mental Disorders: DSM-IV-TR*. American Psychiatric Association; 2000.
28. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. American Psychiatric Association; 2013. doi:10.1176/appi.books.9780890425596

29. Prom-Wormley E, Langi G, Clifford J, Real J. Understanding the Role of Genetic and Environmental Influences on the Neurobiology of Nicotine Use. In: Watson R, Zibadi S, eds. *Addictive Substance and Neurological Disease: Alcohol, Tobacco, Caffeine, and Drugs of Abuse in Everyday Lifestyle*. Vol 1. 1st ed. Academic Press; 2017.
30. Kapalka GM. Substances Involved in Neurotransmission. In: *Nutritional and Herbal Therapies for Children and Adolescents*. Elsevier; 2010:71-99. doi:10.1016/B978-0-12-374927-7.00004-2
31. Sara SJ. Locus Coeruleus in time with the making of memories. *Curr Opin Neurobiol*. 2015;35:87-94. doi:10.1016/J.CONB.2015.07.004
32. Maier SF, Watkins LR. Stressor controllability and learned helplessness: The roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. *Neuroscience and Biobehavioral Reviews*. 2005;29:829-841. doi:10.1016/j.neubiorev.2005.03.021
33. McEntee WJ, Crook TH. Glutamate: its role in learning, memory, and the aging brain. *Psychopharmacology (Berl)*. 1993;111(4):391-401. doi:10.1007/BF02253527
34. D'souza MS, Markou A. The "stop" and "go" of nicotine dependence: role of GABA and glutamate. *Cold Spring Harb Perspect Med*. 2013;3(6). doi:10.1101/CSHPERSPECT.A012146
35. Ryan H, Trosclair A, Gfroerer J. Adult current smoking: Differences in definitions and prevalence estimates - NHIS and NSDUH, 2008. *Journal of Environmental and Public Health*. 2012;2012. doi:10.1155/2012/918368
36. National Center for Health Statistics. National Health Interview Survey. *National Center for Health Statistics, Centers for Disease Control and Prevention*. Published online 2009.
37. Muderrisoglu A, Babaoglu E, Korkmaz ET, et al. Effects of Genetic Polymorphisms of Drug Transporter ABCB1 (MDR1) and Cytochrome P450 Enzymes CYP2A6, CYP2B6 on Nicotine Addiction and Smoking Cessation. *Front Genet*. 2020;11. doi:10.3389/FGENE.2020.571997
38. Kotz D, Batra A, Kastaun S. Smoking Cessation Attempts and Common Strategies Employed. *Deutsches Arzteblatt international*. 2020;117(1-2):7-13. doi:10.3238/ARZTEBL.2020.0007
39. Grant BF, Shmulewitz D, Compton WM. Nicotine Use and DSM-IV Nicotine Dependence in the United States, 2001–2002 and 2012–2013. *American Journal of Psychiatry*. 2020;177(11):1082-1090. doi:10.1176/APPI.AJP.2020.19090900/SUPPL_FILE/APPI.AJP.2020.19090900.DS001.PDF
40. Avenevoli S, Merikangas KR. Familial influences on adolescent smoking. *Addiction (Abingdon, England)*. 2003;98 Suppl 1(SUPPL. 1):1-20. doi:10.1046/J.1360-0443.98.S1.2.X
41. Breslau N, Peterson EL. Smoking cessation in young adults: age at initiation of cigarette smoking and other suspected influences. *Am J Public Health*. 1996;86(2):214-220. doi:10.2105/AJPH.86.2.214

42. Buchmann AF, Blomeyer D, Jennen-Steinmetz C, et al. Early smoking onset may promise initial pleasurable sensations and later addiction. *Addiction biology*. 2013;18(6):947-954. doi:10.1111/J.1369-1600.2011.00377.X
43. Kandel DB, Hu MC, Griesler PC, Schaffran C. On the development of nicotine dependence in adolescence. *Drug Alcohol Depend*. 2007;91(1):26-39. doi:10.1016/J.DRUGALCDEP.2007.04.011
44. Mayhew KP, Flay BR, Mott JA. Stages in the development of adolescent smoking. In: *Drug and Alcohol Dependence*. Vol 59. Drug Alcohol Depend; 2000:61-81. doi:10.1016/S0376-8716(99)00165-9
45. Belsky DW, Moffitt TE, Baker TB, et al. Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA Psychiatry*. 2013;70(5):534-542. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med10&NEWS=N&AN=23536134
46. McMillen RC, Gottlieb MA, Whitmore Shaefer RM, Winickoff JP, Klein JD. Trends in electronic cigarette use among U.S. adults: Use is increasing in both smokers and nonsmokers. *Nicotine and Tobacco Research*. 2015;17(10):1195-1202. doi:10.1093/ntr/ntu213
47. Bold KW, Sussman S, O'Malley SS, et al. Measuring E-cigarette dependence: Initial guidance. *Addictive behaviors*. 2018;79:213-218. doi:10.1016/J.ADDBEH.2017.11.015
48. Vogel EA, Cho J, McConnell RS, Barrington-Trimis JL, Leventhal AM. Prevalence of Electronic Cigarette Dependence Among Youth and Its Association With Future Use. *JAMA Netw Open*. 2020;3(2). doi:10.1001/JAMANETWORKOPEN.2019.21513
49. Martínez Ú, Martínez-Loredo V, Simmons VN, et al. How Does Smoking and Nicotine Dependence Change After Onset of Vaping? A Retrospective Analysis of Dual Users. *Nicotine Tob Res*. 2020;22(5):764-770. doi:10.1093/NTR/NTZ043
50. Jankowski M, Krzystanek M, Zejda JE, et al. E-Cigarettes are More Addictive than Traditional Cigarettes-A Study in Highly Educated Young People. *Int J Environ Res Public Health*. 2019;16(13). doi:10.3390/IJERPH16132279
51. Stallings-Smith S, Ballantyne T. Ever Use of E-Cigarettes Among Adults in the United States: A Cross-Sectional Study of Sociodemographic Factors. *Inquiry (United States)*. 2019;56. doi:10.1177/0046958019864479
52. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Published online 2016.
53. Soneji S, Barrington-Trimis JL, Wills TA, et al. Association Between Initial Use of e-Cigarettes and Subsequent Cigarette Smoking Among Adolescents and Young Adults: A Systematic Review and Meta-analysis. *JAMA Pediatr*. 2017;171(8):788-797. doi:10.1001/JAMAPEDIATRICS.2017.1488
54. East K, Hitchman SC, Bakolis I, et al. The Association Between Smoking and Electronic Cigarette Use in a Cohort of Young People. *J Adolesc Health*. 2018;62(5):539-547. doi:10.1016/J.JADOHEALTH.2017.11.301
55. Itier V, Bertrand D. Neuronal nicotinic receptors: from protein structure to function. *FEBS Lett*. 2001;504(3):118-125. doi:10.1016/S0014-5793(01)02702-8

56. Kubota T, Nakajima-Taniguchi C, Fukuda T, et al. CYP2A6 polymorphisms are associated with nicotine dependence and influence withdrawal symptoms in smoking cessation. *The Pharmacogenomics Journal 2006 6:2*. 2006;6(2):115-119. doi:10.1038/sj.tpj.6500348
57. Kremer I, Bachner-Melman R, Reshef A, et al. Association of the serotonin transporter gene with smoking behavior. *Am J Psychiatry*. 2005;162(5):924-930. doi:10.1176/APPI.AJP.162.5.924
58. Khouja JN, Wootton RE, Taylor AE, Smith GD, Munafò MR. Association of genetic liability to smoking initiation with e-cigarette use in young adults: A cohort study. *PLoS Medicine*. 2021;18(3):1-16. doi:10.1371/JOURNAL.PMED.1003555
59. Aldrich MC, Hidalgo B, Widome R, Briss P, Brownson RC, Teutsch SM. The role of epidemiology in evidence-based policy making: A case study of tobacco use in youth. *Annals of Epidemiology*. 2015;25(5):360-365. doi:10.1016/j.annepidem.2014.03.005
60. Leonardi-Bee J, Nderi M, Britton J. Smoking in movies and smoking initiation in adolescents: systematic review and meta-analysis. *Addiction (Abingdon, England)*. 2016;111(10):1750-1763. doi:10.1111/ADD.13418
61. Casetta B, Videla AJ, Bardach A, et al. Association Between Cigarette Smoking Prevalence and Income Level: A Systematic Review and Meta-Analysis. *Nicotine & Tobacco Research*. 2017;19(12):1401-1407. doi:10.1093/ntr/ntw266
62. Vink JM, Willemsen G, Boomsma DI. The association of current smoking behavior with the smoking behavior of parents, siblings, friends and spouses. *Addiction (Abingdon, England)*. 2003;98(7):923-931. doi:10.1046/J.1360-0443.2003.00405.X
63. Kestilä L, Koskinen S, Martelin T, et al. Influence of parental education, childhood adversities, and current living conditions on daily smoking in early adulthood. *Eur J Public Health*. 2006;16(6):617-626. doi:10.1093/EURPUB/CKL054
64. Lovato C, Watts A, Stead LF. Impact of tobacco advertising and promotion on increasing adolescent smoking behaviours. *Cochrane Database of Systematic Reviews*. 2011;(10). doi:10.1002/14651858.cd003439.pub2
65. Paynter J, Edwards R. The impact of tobacco promotion at the point of sale: a systematic review. *Nicotine Tob Res*. 2009;11(1):25-35. doi:10.1093/NTR/NTN002
66. Hwang JH, Park SW. Association between peer cigarette smoking and electronic cigarette smoking among adolescent nonsmokers: A national representative survey. *PLoS ONE*. 2016;11(10):4-6. doi:10.1371/journal.pone.0162557
67. Alexander JP, Williams P, Lee YO. Youth who use e-cigarettes regularly: A qualitative study of behavior, attitudes, and familial norms. *Prev Med Rep*. 2018;13:93-97. doi:10.1016/J.PMEDR.2018.11.011
68. Loukas A, Paddock EM, Li X, Harrell MB, Pasch KE, Perry CL. Electronic Nicotine Delivery Systems Marketing and Initiation Among Youth and Young Adults. *Pediatrics*. 2019;144(3). doi:10.1542/PEDS.2018-3601
69. Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. Socioeconomic Factors, Health Behaviors, and Mortality: Results From a Nationally Representative Prospective Study of US Adults. *JAMA*. 1998;279(21):1703-1708. doi:10.1001/JAMA.279.21.1703

70. Lee J. Effects of health insurance coverage on risky behaviors. *Health Econ*. 2018;27(4):762-777. doi:10.1002/HEC.3634
71. Choi K, Chen JC, Tan ASL, Soneji S, Moran MB. Receipt of tobacco direct mail/email discount coupons and trajectories of cigarette smoking behaviours in a nationally representative longitudinal cohort of US adults. *Tobacco Control*. 2019;28(3):282-288. doi:10.1136/tobaccocontrol-2018-054363
72. Magid HSA, Bradshaw PT, Ling PM, Halpern-Felsher B. Association of Alternative Tobacco Product Initiation With Ownership of Tobacco Promotional Materials Among Adolescents and Young Adults. *JAMA Netw Open*. 2019;2(5). doi:10.1001/JAMANETWORKOPEN.2019.4006
73. Cantrell J, Bennett M, Mowery P, et al. Patterns in first and daily cigarette initiation among youth and young adults from 2002 to 2015. *PLoS ONE*. 2018;13(8):1-20. doi:10.1371/journal.pone.0200827
74. Schneider S, Diehl K. Vaping as a Catalyst for Smoking? An InitialModel on the Initiation of Electronic CigaretteUse and the Transition to Tobacco SmokingAmong Adolescents. *Nicotine and Tobacco Research*. 2016;18(5):647-653. doi:10.1093/ntr/ntv193
75. Chen X, Yu B, Wang Y. Initiation of Electronic Cigarette Use by Age Among Youth in the U.S. *Am J Prev Med*. 2017;53(3):396-399. doi:10.1016/J.AMEPRE.2017.02.011
76. Loukas A, Marti CN, Cooper M, Pasch KE, Perry CL. Exclusive e-cigarette use predicts cigarette initiation among college students. *Addictive behaviors*. 2018;76:343-347. doi:10.1016/J.ADDBEH.2017.08.023
77. Leventhal AM, Strong DR, Kirkpatrick MG, et al. Association of Electronic Cigarette Use With Initiation of Combustible Tobacco Product Smoking in Early Adolescence. *JAMA*. 2015;314(7):700-707. doi:10.1001/JAMA.2015.8950
78. Jones MR, Tellez-Plaza M, Navas-Acien A. Smoking, Menthol Cigarettes and All-Cause, Cancer and Cardiovascular Mortality: Evidence from the National Health and Nutrition Examination Survey (NHANES) and a Meta-Analysis. *PLoS ONE*. 2013;8(10). doi:10.1371/journal.pone.0077941
79. O'Loughlin J, O'Loughlin EK, Wellman RJ, et al. Predictors of Cigarette Smoking Initiation in Early, Middle, and Late Adolescence. *Journal of Adolescent Health*. 2017;61(3):363-370. doi:10.1016/j.jadohealth.2016.12.026
80. Wellman RJ, Sylvestre MP, O'Loughlin EK, et al. Socioeconomic status is associated with the prevalence and co-occurrence of risk factors for cigarette smoking initiation during adolescence. *Int J Public Health*. 2018;63(1):125-136. doi:10.1007/S00038-017-1051-9
81. Sahu M, Prasuna JG. Twin Studies: A Unique Epidemiological Tool. *Indian J Community Med*. 2016;41(3):177-182. doi:10.4103/0970-0218.183593
82. Kaprio J, Koskenvuo M, Langinvainio H. Finnish twins reared apart. IV: Smoking and drinking habits. A preliminary analysis of the effect of heredity and environment. *Acta Genet Med Gemellol (Roma)*. 1984;33(3):425-433. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med2&NEWS=N&AN=6543277

83. Eaves L, Eysenck H. Are twins enough? The analysis of family and adoption data. In: Eysenck HJ, ed. *The Causes and Effects of Smoking*. Maurice Temple Smith; 1980:236-282.
84. Osler M, Holst C, Prescott E, Sørensen TIA. Influence of genes and family environment on adult smoking behavior assessed in an adoption study. *Genet Epidemiol*. 2001;21(3):193-200. doi:10.1002/GEPI.1028
85. Neale M, Cardon L. *Methodology for Genetic Studies of Twins and Familes*. Kluwer Academic Publishers; 1992.
86. Eaves L, Foley D, Silberg J. Has the “Equal Environments” assumption been tested in twin studies? *Twin Res*. 2003;6(6):486-489. doi:10.1375/136905203322686473
87. Richardson K, Norgate S. The equal environments assumption of classical twin studies may not hold. *Br J Educ Psychol*. 2005;75(Pt 3):339-350. doi:10.1348/000709904X24690
88. Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. *Nat Rev Genet*. 2002;3(11):872-882. doi:10.1038/NRG932
89. Kendler KS, Aggen SH, Gillespie N, Czajkowski N, Ystrom E, Reichborn-Kjennerud T. A twin study of cigarette and snus initiation and quantity of use in Norwegian adult twins. *Twin Research and Human Genetics*. 2019;22(2):108-113.
90. Morley KI, Medland SE, Ferreira MA, et al. A possible smoking susceptibility locus on chromosome 11p12: evidence from sex-limitation linkage analyses in a sample of Australian twin families. *Behav Genet*. 2006;36(1):87-99. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=16365831
91. Heath AC, Kirk KM, Meyer JM, Martin NG. Genetic and social determinants of initiation and age at onset of smoking in Australian twins. *Behav Genet*. 1999;29(6):395-407. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=10857245
92. True WR, Heath AC, Scherrer JF, et al. Genetic and environmental contributions to smoking. *Addiction (Abingdon, England)*. 1997;92(10):1277-1287. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=9489045
93. Kendler KS, Myers J, Damaj MI, Chen X. Early smoking onset and risk for subsequent nicotine dependence: a monozygotic co-twin control study. *Am J Psychiatry*. 2013;170(4):408-413. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med10&NEWS=N&AN=23318372
94. Maes HH, Morley K, Neale MC, et al. Cross-Cultural Comparison of Genetic and Cultural Transmission of Smoking Initiation Using an Extended Twin Kinship Model. *Twin Res Hum Genet*. 2018;21(3):179-190. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med15&NEWS=N&AN=29757125
95. Maes HH, Prom-Wormley E, Eaves LJ, et al. A Genetic Epidemiological Mega Analysis of Smoking Initiation in Adolescents. *Nicotine Tob Res*. 2017;19(4):401-409. doi:10.1093/ntr/ntw294

96. Rose R, Broms R, Korhonen T, Dick D, Kaprio J. Genetics of Smoking Behavior. In: Kim Y, ed. *The Handbook of Behavioral Genetics*. Vol 1. 1st ed. Springer; 2009.
97. Tully EC, Iacono WG, McGue M. Changes in genetic and environmental influences on the development of nicotine dependence and major depressive disorder from middle adolescence to early adulthood. *Dev Psychopathol*. 2010;22(4):831-848. doi:10.1017/S0954579410000490
98. Okoli C, Greaves L, Fagyas V. Sex differences in smoking initiation among children and adolescents. *Public Health*. 2013;127(1):3-10. doi:10.1016/J.PUHE.2012.09.015
99. Schmitz LL, Gard AM, Ware EB. Examining sex differences in pleiotropic effects for depression and smoking using polygenic and gene-region aggregation techniques. *Am J Med Genet B Neuropsychiatr Genet*. 2019;180(6):448-468. doi:10.1002/AJMG.B.32748
100. Allen AM, Scheuermann TS, Nollen N, Hatsukami D, Ahluwalia JS. Gender Differences in Smoking Behavior and Dependence Motives Among Daily and Nondaily Smokers. *Nicotine & Tobacco Research*. 2016;18(6):1408. doi:10.1093/NTR/NTV138
101. Thompson AB, Tebes JK, McKee SA. Gender differences in age of smoking initiation and its association with health. *Addiction research & theory*. 2015;23(5):413-420. doi:10.3109/16066359.2015.1022159
102. Li MD, Cheng R, Ma JZ, Swan GE. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction (Abingdon, England)*. 2003;98(1):23-31. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=12492752
103. Heath AC, Martin NG, Lynskey MT, Todorov AA, Madden PAF. Estimating two-stage models for genetic influences on alcohol, tobacco or drug use initiation and dependence vulnerability in twin and family data. *Twin Res*. 2002;5(2):113-124. doi:10.1375/1369052022983
104. Morley KI, Lynskey MT, Madden PAF, Treloar SA, Heath AC, Martin NG. Exploring the inter-relationship of smoking age-at-onset, cigarette consumption and smoking persistence: genes or environment? *Psychol Med*. 2007;37(9):1357-1367. doi:10.1017/S0033291707000748
105. Maes HH, Sullivan PF, Bulik CM, et al. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychol Med*. 2004;34(7):1251-1261. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=15697051
106. Hardie TL, Moss HB, Lynch KG. Genetic correlations between smoking initiation and smoking behaviors in a twin sample. *Addictive behaviors*. 2006;31(11):2030-2037. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=16675152
107. Hamilton AS, Lessov-Schlaggar CN, Cockburn MG, Unger JB, Cozen W, Mack TM. Gender differences in determinants of smoking initiation and persistence in

California twins. *Cancer Epidemiol Biomarkers Prev*. 2006;15(6):1189-1197. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=16775180
108. Broms U, Silventoinen K, Madden PAF, Heath AC, Kaprio J. Genetic architecture of smoking behavior: A study of Finnish adult twins. *Twin Research and Human Genetics*. 2006;9(1):64-72. doi:10.1375/183242706776403046
109. Kendler KS, Aggen SH, Gillespie N, Czajkowski N, Ystrom E, Reichborn-Kjennerud T. A Twin Study of Cigarette and Snus Initiation and Quantity of Use in Norwegian Adult Twins. *Twin Res Hum Genet*. 2019;22(2):108-113. doi:10.1017/THG.2019.9
110. True WR, Xian H, Scherrer JF, et al. Common Genetic Vulnerability for Nicotine and Alcohol Dependence in Men. *Archives of General Psychiatry*. 1999;56(7):655-661. doi:10.1001/ARCHPSYC.56.7.655
111. Koopmans JR, van Doornen LJP, Boomsma DI. Association between Alcohol Use and Smoking in Adolescent and Young Adult Twins: A Bivariate Genetic Analysis. *Alcohol Clin Exp Res*. 1997;21(3):537-546. doi:10.1111/j.1530-0277.1997.tb03800.x
112. Agrawal A, Lynskey MT, Pergadia ML, et al. Early cannabis use and DSM-IV nicotine dependence: a twin study. *Addiction (Abingdon, England)*. 2008;103(11):1896-1904. doi:10.1111/J.1360-0443.2008.02354.X
113. McMillen RC, Gottlieb MA, Whitmore Shaefer RM, Winickoff JP, Klein JD. Trends in electronic cigarette use among U.S. adults: Use is increasing in both smokers and nonsmokers. *Nicotine and Tobacco Research*. 2015;17(10):1195-1202. doi:10.1093/ntr/ntu213
114. Wright S. Correlation and Causation. *Journal of Agricultural Research*. 1921;20:557-585.
115. Wright S. The Method of Path Coefficients. *Annals of Mathematical Statistics*. 1934;5(3):161-215.
116. Medland SE, Neale MC, Eaves LJ, Neale BM. A Note on the Parameterization of Purcell's G × E Model for Ordinal and Binary Data. *Behav Genet*. 2009;39(2):220. doi:10.1007/S10519-008-9247-7
117. R Core Team. R: A language and environment for statistical computing. Published online 2017. https://www.r-project.org/
118. Neale MC, Hunter MD, Pritikin JN, et al. OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika*. 2016;81(2):535-549. doi:10.1007/S11336-014-9435-8
119. Røysamb E, Tambs K. The beauty, logic and limitations of twin studies. *Norsk Epidemiologi*. 2016;26(1-2):35-46. doi:10.5324/NJE.V26I1-2.2014
120. Loukas A, Marti CN, Cooper M, Pasch KE, Perry CL. Exclusive e-cigarette use predicts cigarette initiation among college students. *Addictive behaviors*. 2018;76:343-347. doi:10.1016/J.ADDBEH.2017.08.023
121. Liu M. Meta-analysis of genome-wide association study of 1.2 million people finds novel signals in alcohol and nicotine use. *Behavior Genetics*. 2018;48(6):491. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed19&NEWS=N&AN=624714665

122. Xu K, Li B, McGinnis KA, et al. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nature Communications*. 2020;11(1). doi:10.1038/s41467-020-18489-3
123. Brunzell DH, Stafford AM, Dixon CI. Nicotinic receptor contributions to smoking: insights from human studies and animal models. *Curr Addict Rep*. 2015;2(1):33. doi:10.1007/S40429-015-0042-2
124. Allegrini AG, Verweij KJH, Abdellaoui A, et al. Genetic vulnerability for smoking and cannabis use: Associations with E-cigarette and water pipe use. *Nicotine and Tobacco Research*. 2018;21(6):723-730. doi:10.1093/ntr/nty150
125. Xu X, Bishop EE, Kennedy SM, Simpson SA, Pechacek TF. Annual healthcare spending attributable to cigarette smoking: An update. *American Journal of Preventive Medicine*. 2015;48(3):326-333. doi:10.1016/j.amepre.2014.10.012
126. Cigarette Smoking Among U.S. Adults Hits All-Time Low | CDC Online Newsroom | CDC. Published 2019. Accessed January 2, 2020. https://www.cdc.gov/media/releases/2019/p1114-smoking-low.html
127. Liu Y, Brossard M, Sarnowski C, et al. Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma. *Scientific Reports 2017 7:1*. 2017;7(1):1-10. doi:10.1038/s41598-017-01058-y
128. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*. 2016;5(1). doi:10.1186/s13643-016-0384-4
129. Huang DW, Sherman BT, Zheng X, et al. Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinformatics*. 2009;Chapter 13(SUPPL. 27). doi:10.1002/0471250953.BI1311S27
130. Huang DW, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*. 2007;35(SUPPL.2). doi:10.1093/nar/gkm415
131. Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 2007;8(9). doi:10.1186/GB-2007-8-9-R183
132. Rettew DC, Rebollo-Mesa I, Hudziak JJ, Willemsen G, Boomsma DI. Non-additive and Additive Genetic Effects on Extraversion in 3314 Dutch Adolescent Twins and Their Parents. *Behavior Genetics*. 2008;38(3):223. doi:10.1007/S10519-008-9192-5
133. Keller MC, Coventry WL, Heath AC, Martin NG. Widespread evidence for non-additive genetic variation in Cloninger's and Eysenck's personality dimensions using a twin plus sibling design. *Behav Genet*. 2005;35(6):707-721. doi:10.1007/S10519-005-6041-7
134. Rhee SH, Hewitt JK, Young SE, Corley RP, Crowley TJ, Stallings MC. Genetic and environmental influences on substance initiation, use, and problem use in adolescents. *Arch Gen Psychiatry*. 2003;60(12):1256-1264. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=14662558
135. Maes HH, Sullivan PF, Bulik CM, et al. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence.

Psychological Medicine. 2004;34(7):1251-1261. doi:10.1017/S0033291704002405
136. Vink JM, Beem AL, Posthuma D, et al. Linkage analysis of smoking initiation and quantity in Dutch sibling pairs. *Pharmacogenomics J*. 2004;4(4):274-282. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=15170444
137. Sartor CE, Grant JD, Agrawal A, et al. Genetic and environmental contributions to initiation of cigarette smoking in young African-American and European-American women. *Drug Alcohol Depend*. 2015;157:54-59. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med12&NEWS=N&AN=26482091
138. McCaffery JM, Papandonatos GD, Lyons MJ, Koenen KC, Tsuang MT, Niaura R. Educational attainment, smoking initiation and lifetime nicotine dependence among male Vietnam-era twins. *Psychological Medicine*. 2008;38(9):1287-1297. doi:10.1017/S0033291707001882
139. Heath AC, Cates R, Martin NG, et al. Genetic contribution to risk of smoking initiation: comparisons across birth cohorts and across cultures. *J Subst Abuse*. 1993;5(3):221-246. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med3&NEWS=N&AN=8312729
140. Wills A, Keller M. SNPs and smoking: What can the aggregate of genomewide SNPs tell us about genetic liability to smoking initiation and quantity smoked? *Behavior Genetics*. 2013;43(6):549. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed14&NEWS=N&AN=71325568
141. Kendler KS, Neale MC, Sullivan P, Corey LA, Gardner CO, Prescott CA. A population-based twin study in women of smoking initiation and nicotine dependence. *Psychol Med*. 1999;29(2):299-308. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=10218922
142. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90(1):7-24. doi:10.1016/J.AJHG.2011.11.029
143. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516-1517. doi:10.1126/SCIENCE.273.5281.1516
144. Matoba N, Akiyama M, Ishigaki K, et al. GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nat Hum Behav*. 2019;3(5):471-477. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medc&NEWS=N&AN=31089300
145. Argos M, Tong L, Pierce BL, et al. Genome-wide association study of smoking behaviours among Bangladeshi adults. *J Med Genet*. 2014;51(5):327-333. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med11&NEWS=N&AN=24665060
146. Loukola A, Wedenoja J, Keskitalo-Vuokko K, et al. Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample. *Mol Psychiatry*. 2014;19(5):615-624.

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med11&NEWS=N&AN=23752247
147. Caporaso N, Gu F, Chatterjee N, et al. Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One*. 2009;4(2):e4653. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med7&NEWS=N&AN=19247474
148. Vink JM, Smit AB, de Geus EJC, et al. Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet*. 2009;84(3):367-379. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med7&NEWS=N&AN=19268276
149. Yoon D, Kim YJ, Cui WY, et al. Large-scale genome-wide association study of Asian population reveals genetic factors in FRMD4A and other loci influencing smoking initiation and nicotine dependence. *Hum Genet*. 2012;131(6):1009-1021. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med9&NEWS=N&AN=22006218
150. Siedlinski M, Cho MH, Bakke P, et al. Genome-wide association study of smoking behaviours in patients with COPD. *Thorax*. 2011;66(10):894-902. doi:10.1136/THORAXJNL-2011-200154
151. Cousins RJ, Liuzzi JP, Lichten LA. Mammalian zinc transport, trafficking, and signals. *J Biol Chem*. 2006;281(34):24085-24089. doi:10.1074/JBC.R600011200
152. Ji Y, Yiorkas AM, Frau F, et al. Genome-Wide and Abdominal MRI Data Provide Evidence That a Genetically Determined Favorable Adiposity Phenotype Is Characterized by Lower Ectopic Liver Fat and Lower Risk of Type 2 Diabetes, Heart Disease, and Hypertension. *Diabetes*. 2019;68(1):207-219. doi:10.2337/DB18-0708
153. Cloutier P, Lavallée-Adam M, Faubert D, Blanchette M, Coulombe B. A newly uncovered group of distantly related lysine methyltransferases preferentially interact with molecular chaperones to regulate their activity. *PLoS Genet*. 2013;9(1). doi:10.1371/JOURNAL.PGEN.1003210
154. Sutherland APR, Zhang H, Zhang Y, et al. Zinc finger protein Zbtb20 is essential for postnatal survival and glucose homeostasis. *Mol Cell Biol*. 2009;29(10):2804-2815. doi:10.1128/MCB.01667-08
155. Ferris BG. Epidemiology Standardization Project (American Thoracic Society). *Am Rev Respir Dis*. 1978;118(6 Pt 2):1-120.
156. Furberg H, Kim Y, Dackor J, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*. 2010;42(5):441-447. doi:10.1038/ng.571
157. Mallei A, Baj G, Ieraci A, et al. Expression and Dendritic Trafficking of BDNF-6 Splice Variant are Impaired in Knock-In Mice Carrying Human BDNF Val66Met Polymorphism. *Int J Neuropsychopharmacol*. 2015;18(12):1-10. doi:10.1093/IJNP/PYV069
158. Baj G, Leone E, Chao M v., Tongiorgi E. Spatial segregation of BDNF transcripts enables BDNF to differentially shape distinct dendritic compartments. *Proc Natl Acad Sci U S A*. 2011;108(40):16813-16818. doi:10.1073/PNAS.1014168108
159. Jamal M, der Does W, Elzinga BM, Molendijk ML, Penninx BWJH. Association between smoking, nicotine dependence, and BDNF Val66Met polymorphism with

BDNF concentrations in serum. *Nicotine Tob Res*. 2015;17(3):323-329. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med12&NEWS=N&AN=25183693
160. Chen ZY, Jing D, Bath KG, et al. Genetic variant BDNF (Val66Met) polymorphism alters anxiety-related behavior. *Science*. 2006;314(5796):140-143. doi:10.1126/SCIENCE.1129663
161. di Rosa MC, Zimbone S, Saab MW, Tomasello MF. The Pleiotropic Potential of BDNF beyond Neurons: Implication for a Healthy Mind in a Healthy Body. *Life (Basel)*. 2021;11(11). doi:10.3390/LIFE11111256
162. David SP, Hamidovic A, Chen GK, et al. Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry*. 2012;2:e119. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med9&NEWS=N&AN=22832964
163. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2):237-244. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medc&NEWS=N&AN=30643251
164. Brazel DM, Jiang Y, Hughey JM, et al. Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. *Biol Psychiatry*. 2019;85(11):946-955. doi:10.1016/J.BIOPSYCH.2018.11.024
165. Koppel I, Aid-Pavlidis T, Jaanson K, et al. Tissue-specific and neural activity-regulated expression of human BDNF gene in BAC transgenic mice. *BMC Neuroscience*. 2009;10:68. doi:10.1186/1471-2202-10-68
166. Lu B, Nagappan G, Lu Y. BDNF and synaptic plasticity, cognitive function, and dysfunction. *Handb Exp Pharmacol*. 2014;220:223-250. doi:10.1007/978-3-642-45106-5_9
167. Kowiański P, Lietzau G, Czuba E, Waśkow M, Steliga A, Moryś J. BDNF: A Key Factor with Multipotent Impact on Brain Signaling and Synaptic Plasticity. *Cell Mol Neurobiol*. 2018;38(3):579-593. doi:10.1007/S10571-017-0510-4
168. Suriyaprom K, Tungtrongchitr R, Thawnashom K, Pimainog Y. BDNF Val66Met polymorphism and serum concentrations of BDNF with smoking in Thai males. *Genet Mol Res*. 2013;12(4):4925-4933. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med10&NEWS=N&AN=24301752
169. Merrill RA, Plum LA, Kaiser ME, Clagett-Dame M. A mammalian homolog of unc-53 is regulated by all-trans retinoic acid in neuroblastoma cells and embryos. *Proc Natl Acad Sci U S A*. 2002;99(6):3422-3427. doi:10.1073/PNAS.052017399
170. Maes T, Barceló A, Buesa C. Neuron navigator: a human gene family with homology to unc-53, a cell guidance gene from Caenorhabditis elegans. *Genomics*. 2002;80(1):21-30. doi:10.1006/GENO.2002.6799
171. Muley PD, McNeill EM, Marzinke MA, Knobel KM, Barr MM, Clagett-Dame M. The atRA-responsive gene neuron navigator 2 functions in neurite outgrowth and axonal elongation. *Dev Neurobiol*. 2008;68(13):1441-1453. doi:10.1002/DNEU.20670

172. Neuwald AF, Hirano T. HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Res*. 2000;10(10):1445-1452. doi:10.1101/GR.147400
173. Andrade MA, Bork P. HEAT repeats in the Huntington's disease protein. *Nat Genet*. 1995;11(2):115-116. doi:10.1038/NG1095-115
174. Erzurumluoglu AM, Liu M, Jackson VE, et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Molecular Psychiatry*. Published online 2019. http://www.nature.com/mp/index.html
175. Li M, Jaffe AE, Straub RE, et al. A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nat Med*. 2016;22(6):649-656. doi:10.1038/NM.4096
176. Scott D, Palmer R. The influence of tobacco smoking on adhesion molecule profiles. *Tobacco Induced Diseases*. 2003;1(1):7. doi:10.1186/1617-9625-1-1-7
177. Sobkowiak R, Lesicki A. Absorption, metabolism and excretion of nicotine in humans. *Postepy Biochem*. 2013;59(1):33-44.
178. Avila-Tang E, Al-Delaimy WK, Ashley DL, et al. Assessing secondhand smoke using biological markers. *Tobacco Control*. 2013;22(3):164-171. doi:10.1136/tobaccocontrol-2011-050298
179. Avila-Tang E, Al-Delaimy WK, Ashley DL, et al. Assessing secondhand smoke using biological markers. *Tob Control*. 2013;22(3):164-171. doi:10.1136/TOBACCOCONTROL-2011-050298
180. Swan GE, Benowitz NL, Lessov CN, Jacob IP, Tyndale RF, Wilhelmsen K. Nicotine metabolism: The impact of CYP2A6 on estimates of additive genetic influence. *Pharmacogenetics and Genomics*. 2005;15(2):115-125. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed9&NEWS=N&AN=40562988
181. Patel YM, Stram DO, Wilkens LR, et al. The contribution of common genetic variation to nicotine and cotinine glucuronidation in multiple ethnic/racial populations. *Cancer Epidemiol Biomarkers Prev*. 2015;24(1):119-127. doi:10.1158/1055-9965.EPI-14-0815
182. Chen G, Giambrone NE, Dluzen DF, et al. Glucuronidation genotypes and nicotine metabolic phenotypes: importance of functional UGT2B10 and UGT2B17 polymorphisms. *Cancer Res*. 2010;70(19):7543-7552. doi:10.1158/0008-5472.CAN-09-4582
183. Buchwald J, Chenoweth MJ, Palviainen T, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. *Mol Psychiatry*. Published online 2020. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medp&NEWS=N&AN=32157176
184. Swaminathan S, Huentelman MJ, Corneveaux JJ, et al. Analysis of copy number variation in Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. *PLoS One*. 2012;7(12). doi:10.1371/JOURNAL.PONE.0050640

185. Tyndale RF, Sellers EM. Genetic variation in CYP2A6-mediated nicotine metabolism alters smoking behavior. *Ther Drug Monit*. 2002;24(1):163-171. doi:10.1097/00007691-200202000-00026
186. Ring HZ, Valdes AM, Nishita DM, et al. Gene-gene interactions between CYP2B6 and CYP2A6 in nicotine metabolism. *Pharmacogenet Genomics*. 2007;17(12):1007-1015. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=18004205
187. Binnington MJ, Zhu AZX, Renner CC, et al. CYP2A6 and CYP2B6 genetic variation and its association with nicotine metabolism in South Western Alaska Native people. :429-440. doi:10.1097/FPC.0b013e3283527c1c
188. Caporaso NE, Lerman C, Audrain J, et al. Nicotine metabolism and CYP2D6 phenotype in smokers. *Cancer Epidemiol Biomarkers Prev*. 2001;10(3):261-263. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11303596
189. Pelkonen O, Rautio A, Raunio H, Pasanen M. CYP2A6: a human coumarin 7-hydroxylase. *Toxicology*. 2000;144(1-3):139-147. doi:10.1016/S0300-483X(99)00200-0
190. How Much Nicotine Is in a Cigarette, Cigar, and E-Cigarette? . Accessed April 6, 2022. https://www.healthline.com/health/how-much-nicotine-is-in-a-cigarette
191. Koopmans JR, Slutske WS, Heath AC, Neale MC, Boomsma DI. The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. *Behav Genet*. 1999;29(6):383-393. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=10857244
192. Haberstick BC, Timberlake D, Ehringer MA, et al. Genes, time to first cigarette and nicotine dependence in a general population sample of young adults. *Addiction (Abingdon, England)*. 2007;102(4):655-665. doi:10.1111/J.1360-0443.2007.01746.X
193. Nishizawa D, Kasai S, Hasegawa J, et al. Genome-wide association study identified susceptibility loci associated with nicotine dependence in a Japanese population. *International Journal of Neuropsychopharmacology*. 2012;15:248. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=71592821
194. Hallfors J, Palviainen T, Surakka I, et al. Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway. *Addiction biology*. 2019;24(3):549-561. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=29532581
195. Lutz SM, Frederiksen B, Begum F, et al. Common and Rare Variants Genetic Association Analysis of Cigarettes per Day Among Ever-Smokers in Chronic Obstructive Pulmonary Disease Cases and Controls. *Nicotine Tob Res*. 2019;21(6):714-722. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=29767774

196. Saccone NL, Emery LS, Sofer T, et al. Genome-Wide Association Study of Heavy Smoking and Daily/Nondaily Smoking in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Nicotine Tob Res*. 2018;20(4):448-457. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med15&NEWS=N&AN=28520984
197. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*. 2017;101(1):5-22. doi:10.1016/j.ajhg.2017.06.005
198. Fagerström K. Determinants of tobacco use and renaming the FTND to the Fagerstrom Test for Cigarette Dependence. *Nicotine Tob Res*. 2012;14(1):75-78. doi:10.1093/NTR/NTR137
199. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction*. 1991;86(9):1119-1127. doi:10.1111/j.1360-0443.1991.tb01879.x
200. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*. 2020;11(1):1-3. doi:10.1038/s41467-020-19653-5
201. Vink JM, Willemsen G, Boomsma DI. Heritability of smoking initiation and nicotine dependence. *Behav Genet*. 2005;35(4):397-406. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=15971021
202. Modig K, Silventoinen K, Tynelius P, Kaprio J, Rasmussen F. Genetics of the association between intelligence and nicotine dependence : a study of male Swedish twins. Published online 2011:995-1002. doi:10.1111/j.1360-0443.2011.03384.x
203. Richmond-Rakerd LS, Slutske WS, Lynskey MT, et al. Age at first use and later substance use disorder: Shared genetic and environmental pathways for nicotine, alcohol, and cannabis. *J Abnorm Psychol*. 2016;125(7):946-959. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med13&NEWS=N&AN=27537477
204. Lessov CN, Martin NG, Statham DJ. Defining nicotine dependence for genetic research : Evidence from Australian twins. Published online 2004.
205. Borland R, Yong HH, O'Connor RJ, Hyland A, Thompson ME. The reliability and predictive validity of the Heaviness of Smoking Index and its two components: Findings from the International Tobacco Control Four Country study. *Nicotine & Tobacco Research*. 2010;12(Suppl 1):S45. doi:10.1093/NTR/NTQ038
206. Rice JP, Hartz SM, Agrawal A, et al. CHRNB3 is more strongly associated with Fagerström Test for Cigarette Dependence-based nicotine dependence than cigarettes per day: Phenotype definition changes genome-wide association studies results. *Addiction*. 2012;107(11):2019-2028. doi:10.1111/j.1360-0443.2012.03922.x
207. Gelernter J, Kranzler HR, Sherva R, et al. Genome-wide association study of nicotine dependence in American populations: identification of novel risk loci in both African-Americans and European-Americans. *Biol Psychiatry*. 2015;77(5):493-503. doi:10.1016/J.BIOPSYCH.2014.08.025

208. Fagerberg L, Hallstrom BM, Oksvold P, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 2014;13(2):397-406. doi:10.1074/MCP.M113.035600
209. Chen J, Loukola A, Gillespie NA, et al. Genome-Wide Meta-Analyses of FTND and TTFC Phenotypes. *Nicotine Tob Res*. 2020;22(6):900-909. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=31294817
210. Ockene JK, Emmons K, Mermelstein R, Perkins K. Relapse and Maintenance Issues for Smoking Cessastion. *Health Psychology*. 2000;19(1(Suppl)):17-31. doi:10.1037/0278-6133.19.Suppl1.17
211. Xian H, Scherrer JF, Madden PAF, et al. The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. *Nicotine Tob Res*. 2003;5(2):245-254. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=12745498
212. Siddiqi K, Shah S, Abbas SM, et al. Global burden of disease due to smokeless tobacco consumption in adults: analysis of data from 113 countries. *BMC Med*. 2015;13(1). doi:10.1186/S12916-015-0424-2
213. Fishbein H, Bauer D, Yu Q, et al. Harmonizing Cigar Survey Data across Tobacco Centers of Regulatory Science, Center for Tobacco Products, and Population Assessment of Tobacco and Health Studies: The Cigar Collaborative Research Group. *Nicotine and Tobacco Research*. 2021;23(1):212-218. doi:10.1093/ntr/ntz201
214. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX toolkit: Get the most from your measures. *American Journal of Epidemiology*. 2011;174(3):253-260. doi:10.1093/aje/kwr193
215. Maes HH, Neale MC, Kendler KS, Martin NG, Heath AC, Eaves LJ. Genetic and cultural transmission of smoking initiation: an extended twin kinship model. *Behav Genet*. 2006;36(6):795-808. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=16810566
216. Norton BJ, Strube MJ. Understanding statistical power. *J Orthop Sports Phys Ther*. 2001;31(6):307-315. doi:10.2519/JOSPT.2001.31.6.307
217. Korhonen T, Loukola A, Hallfors J, Salomaa V, Kaprio J. Is Brain-Derived Neurotrophic Factor Associated With Smoking Initiation? Replication Using a Large Finnish Population Sample. *Nicotine Tob Res*. 2020;22(2):293-296. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=30329132
218. Kasai S, Nishizawa D, Hasegawa J, et al. Nociceptin/orphanin FQ receptor gene variation is associated with smoking status in Japanese. *Pharmacogenomics*. 2016;17(13):1441-1451. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med13&NEWS=N&AN=27491383

219. Yu M, Sacco P, Choi HJ, Wintemberg J. Identifying patterns of tobacco use among US middle and high school students: A latent class analysis. *Addictive Behaviors*. 2018;79(June 2017):1-7. doi:10.1016/j.addbeh.2017.11.034
220. Aslan D, Gürbay A, Hayran M, Şengelen M, Paslı D, Hüseyin B. Carbon Monoxide in the Expired Air and Urinary Cotinine Levels of e-Cigarette Users. *Turkish Thoracic Journal*. 2019;20(2):125. doi:10.5152/TURKTHORACJ.2018.18110
221. Benowitz NL, Bernert JT, Foulds J, et al. Biochemical Verification of Tobacco Use and Abstinence: 2019 Update. *Nicotine & Tobacco Research*. 2020;22(7):1086-1097. doi:10.1093/NTR/NTZ132
222. Cooper M, Harrell MB, Perry CL. A Qualitative Approach to Understanding Real-World Electronic Cigarette Use: Implications for Measurement and Regulation. *Preventing Chronic Disease*. 2019;13(1). doi:10.5888/PCD13.150502
223. Morean ME, Bold KW, Kong G, et al. Adolescents' Awareness of the Nicotine Strength and E-cigarette Status of JUUL E-cigarettes. *Drug Alcohol Depend*. 2019;204:107512. doi:10.1016/J.DRUGALCDEP.2019.05.032
224. IQOS in the U.S. Accessed April 7, 2022. https://truthinitiative.org/research-resources/emerging-tobacco-products/iqos-us
225. Cooke ME, Clifford JS, Do EK, et al. Polygenic score for cigarette smoking is associated with ever electronic-cigarette use in a college-aged sample. *Addiction*. 2022;117(4):1071-1078. doi:10.1111/ADD.15716
226. Barrington-Trimis JL, Braymiller JL, Unger JB, et al. Trends in the Age of Cigarette Smoking Initiation Among Young Adults in the US From 2002 to 2018. *JAMA Network Open*. 2020;3(10):e2019022-e2019022. doi:10.1001/JAMANETWORKOPEN.2020.19022
227. Pérez A, Bluestein MA, Kuk AE, Chen B. Age of e-cigarette initiation in USA young adults: Findings from the Population Assessment of Tobacco and Health (PATH) study (2013–2017). *PLOS ONE*. 2021;16(12):e0261243. doi:10.1371/JOURNAL.PONE.0261243
228. Maes HH, Woodard CE, Murrelle L, et al. Tobacco, alcohol and drug use in eight- to sixteen-year-old twins: the Virginia Twin Study of Adolescent Behavioral Development. *J Stud Alcohol*. 1999;60(3):293-305. doi:10.15288/JSA.1999.60.293
229. White VM, Hopper JL, Wearing AJ, Hill DJ. The role of genes in tobacco smoking during adolescence and young adulthood: a multivariate behaviour genetic investigation. *Addiction (Abingdon, England)*. 2003;98(8):1087-1100. doi:10.1046/J.1360-0443.2003.00427.X
230. Furberg H, Ostroff J, Lerman C, Sullivan PF. The public health utility of genome-wide association study results for smoking behavior. *Genome Med*. 2010;2(4):26. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=20423533
231. Zuo L, Tan Y, Li CSR, et al. Associations of rare nicotinic cholinergic receptor gene variants to nicotine and alcohol dependence. *Am J Med Genet B Neuropsychiatr Genet*. 2016;171(8):1057. doi:10.1002/AJMG.B.32476
232. D'Souza MS, Markou A. Neuronal Mechanisms Underlying Development of Nicotine Dependence: Implications for Novel Smoking-Cessation Treatments.

Addiction Science & Clinical Practice. 2011;6(1):4. Accessed April 4, 2022. /pmc/articles/PMC3188825/
233. Ridley RM, Bowes PM, Baker HF, Crow TJ. An involvement of acetylcholine in object discrimination learning and memory in the marmoset. *Neuropsychologia*. 1984;22(3):253-263. doi:10.1016/0028-3932(84)90073-3
234. Himmelheber AM, Sarter M, Bruno JP. Increases in cortical acetylcholine release during sustained attention performance in rats. *Brain Res Cogn Brain Res*. 2000;9(3):313-325. doi:10.1016/S0926-6410(00)00012-4
235. Jones BE. From waking to sleeping: neuronal and chemical substrates. *Trends Pharmacol Sci*. 2005;26(11):578-586. doi:10.1016/J.TIPS.2005.09.009
236. Liechti ME, Markou A. Role of the glutamatergic system in nicotine dependence : implications for the discovery and development of new pharmacological smoking cessation therapies. *CNS Drugs*. 2008;22(9):705-724. doi:10.2165/00023210-200822090-00001
237. Alasmari F, Crotty Alexander LE, Hammad AM, Bojanowski CM, Moshensky A, Sari Y. Effects of Chronic Inhalation of Electronic Cigarette Vapor Containing Nicotine on Neurotransmitters in the Frontal Cortex and Striatum of C57BL/6 Mice. *Frontiers in Pharmacology*. 2019;10(JULY):885. doi:10.3389/FPHAR.2019.00885/BIBTEX
238. Price LR, Martinez J, Antoniewicz L, Mayer B, Holloway A. Cardiovascular, carcinogenic and reproductive effects of nicotine exposure: A narrative review of the scientific literature. *F1000Research 2020 8:1586*. 2020;8:1586. doi:10.12688/f1000research.20062.2
239. Tyndale RF, Sellers EM. Genetic variation in CYP2A6-mediated nicotine metabolism alters smoking behavior. *Ther Drug Monit*. 2002;24(1):163-171. doi:10.1097/00007691-200202000-00026
240. Fisher R. The correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc Edinb*. 1918;53:399-433.
241. Plomin R, DeFries JC, Knopik VS, Neiderhiser JM. Top 10 Replicated Findings from Behavioral Genetics. *Perspect Psychol Sci*. 2016;11(1):3. doi:10.1177/1745691615617439
242. Henriksen MG, Nordgaard J, Jansson LB. Genetics of schizophrenia: Overview of methods, findings and limitations. *Frontiers in Human Neuroscience*. 2017;11:322. doi:10.3389/FNHUM.2017.00322/BIBTEX
243. Gejman P v., Sanders AR, Duan J. The Role of Genetics in the Etiology of Schizophrenia. *Psychiatr Clin North Am*. 2010;33(1):35. doi:10.1016/J.PSC.2009.12.003
244. Sella G, Barton NH. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *https://doi.org/101146/annurev-genom-083115-022316*. 2019;20:461-493. doi:10.1146/ANNUREV-GENOM-083115-022316
245. Gibson G. Rare and Common Variants: Twenty arguments HHS Public Access. *Nat Rev Genet*. 2015;13(2):135-145. doi:10.1038/nrg3118.Rare
246. Robinson MR, Wray NR, Visscher PM. Explaining additional genetic variation in complex traits. *Trends Genet*. 2014;30(4):124. doi:10.1016/J.TIG.2014.02.003

247. Nishino J, Ochi H, Kochi Y, Tsunoda T, Matsui S. Sample size for successful genome-wide association study of major depressive disorder. *Frontiers in Genetics*. 2018;9:227. doi:10.3389/FGENE.2018.00227/BIBTEX
248. Croucha DJM, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc Natl Acad Sci U S A*. 2020;117(32):18924-18933. doi:10.1073/PNAS.2005634117/SUPPL_FILE/PNAS.2005634117.SAPP.PDF
249. Escott-Price V, Shoai M, Pither R, Williams J, Hardy J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiology of Aging*. 2017;49:214.e7-214.e11. doi:10.1016/J.NEUROBIOLAGING.2016.07.018
250. Maher BS. Polygenic Scores in Epidemiology: Risk Prediction, Etiology, and Clinical Utility. *Current Epidemiology Reports 2015 2:4*. 2015;2(4):239-244. doi:10.1007/S40471-015-0055-3
251. Yang S, Zhou X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *The American Journal of Human Genetics*. 2020;106(5):679-693. doi:10.1016/J.AJHG.2020.03.013
252. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759-2772. doi:10.1038/S41596-020-0353-1
253. Gola D, Erdmann J, Läll K, et al. Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease. *Circulation: Genomic and Precision Medicine*. 2020;13:569-575. doi:10.1161/CIRCGEN.120.002932
254. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics 2019 51:4*. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x
255. Osterman MD, Kinzy TG, Bailey JNC. Polygenic Risk Scores. *Current Protocols*. 2021;1(5):e126. doi:10.1002/CPZ1.126
256. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752. doi:10.1038/nature08185
257. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*. 2009;18(18):3525-3531. doi:10.1093/HMG/DDP295
258. Bray MJ, Chen LS, Fox L, et al. Studying the Utility of Using Genetics to Predict Smoking-Related Outcomes in a Population-Based Study and a Selected Cohort. *Nicotine & Tobacco Research*. 2021;23(12):2110-2116. doi:10.1093/NTR/NTAB100
259. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet*. 2019;28(R2):R133-R142. doi:10.1093/HMG/DDZ187
260. Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. Polygenic scores in biomedical research. *Nature Reviews Genetics 2022*. Published online March 30, 2022:1-9. doi:10.1038/s41576-022-00470-z
261. Konuma T, Okada Y. Statistical genetics and polygenic risk score for precision medicine. *Inflamm Regen*. 2021;41(1). doi:10.1186/S41232-021-00172-9

262. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine*. 2020;12(1):1-11. doi:10.1186/S13073-020-00742-5/TABLES/2
263. Pasman JA, Smit K, Vollebergh WAM, et al. Interplay between genetic risk and the parent environment in adolescence and substance use in young adulthood: A TRAILS study. *Development and Psychopathology*. Published online 2021:1-14. doi:10.1017/S095457942100081X
264. Meyers J, Galea S, Aiello A, Uddin M, Wildman D, Koenen K. Examining polygenic risk of cigarette use in the detroit neighborhood health study. *Comprehensive Psychiatry*. 2013;54(8):e28. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed14&NEWS=N&AN=71279881
265. Ohi K, Nishizawa D, Muto Y, et al. Polygenic risk scores for late smoking initiation associated with the risk of schizophrenia. *npj Schizophrenia 2020 6:1*. 2020;6(1):1-7. doi:10.1038/s41537-020-00126-z
266. Hicks BM, Clark DA, Deak JD, et al. Polygenic Score for Smoking is associated with Externalizing Psychopathology and Disinhibited Personality Traits but not Internalizing Psychopathology in Adolescence. *Clin Psychol Sci*. 2021;9(6):1205-1213. doi:10.1177/21677026211002117
267. Vink JM, Hottenga JJ, de Geus EJC, et al. Polygenic risk scores for smoking: predictors for alcohol and cannabis use? *Addiction*. 2014;109(7):1141. doi:10.1111/ADD.12491
268. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003;361(9357):598-604. doi:10.1016/S0140-6736(03)12520-2
269. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nature Genetics 2004 36:4*. 2004;36(4):388-393. doi:10.1038/ng1333
270. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287. doi:10.1038/NG.3656
271. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/NATURE15393
272. Edwards AWF. G. H. Hardy (1908) and Hardy-Weinberg equilibrium. *Genetics*. 2008;179(3):1143-1150. doi:10.1534/GENETICS.104.92940
273. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*. 2013;9(3). doi:10.1371/journal.pgen.1003348
274. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications 2019 10:1*. 2019;10(1):1-10. doi:10.1038/s41467-019-09718-5
275. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika*. 1991;78(3):691-692. doi:10.1093/BIOMET/78.3.691
276. McGuire D, Jiang Y, Liu M, et al. Model-based assessment of replicability for genome-wide association meta-analysis. *Nature Communications 2021 12:1*. 2021;12(1):1-14. doi:10.1038/s41467-021-21226-z
277. Fritsche LG, Ma Y, Zhang D, et al. On cross-ancestry cancer polygenic risk scores. *PLoS Genet*. 2021;17(9). doi:10.1371/JOURNAL.PGEN.1009670

278. Duncan L, Shen H, Gelaye B, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications 2019 10:1*. 2019;10(1):1-9. doi:10.1038/s41467-019-11112-0
279. Wong DN, Fan W. Ethnic and sex differences in E-cigarette use and relation to alcohol use in California adolescents: the California Health Interview Survey. *Public Health*. 2018;157:147-152. doi:10.1016/J.PUHE.2018.01.019
280. Elflein J. Vaping and e-cigarette use by gender U.S. 2018 | Statista. Accessed April 4, 2022. https://www.statista.com/statistics/881837/vaping-and-electronic-cigarette-use-us-by-gender/
281. Dai H, Leventhal AM. Prevalence of e-Cigarette Use Among Adults in the United States, 2014-2018. *JAMA*. 2019;322(18):1824-1827. doi:10.1001/JAMA.2019.15331
282. Oh JJ, Kim E, Woo E, et al. Evaluation of Polygenic Risk Scores for Prediction of Prostate Cancer in Korean Men. *Frontiers in Oncology*. 2020;10:2167. doi:10.3389/FONC.2020.583625/BIBTEX
283. Baker SG. Metrics for Evaluating Polygenic Risk Scores. *JNCI Cancer Spectrum*. 2021;5(1). doi:10.1093/JNCICS/PKAA106
284. Wang Y, Zhu M, Ma H, Shen H. Polygenic risk scores: the future of cancer risk prediction, screening, and precision prevention. *Medical Review*. 2021;1(2):129-149. doi:10.1515/MR-2021-0025
285. Sasahira T, Kurihara M, Nishiguchi Y, Fujiwara R, Kirita T, Kuniyasu H. NEDD 4 binding protein 2-like 1 promotes cancer cell invasion in oral squamous cell carcinoma. *Virchows Arch*. 2016;469(2):163-172. doi:10.1007/S00428-016-1955-4
286. Dahlin A, Sordillo JE, McGeachie M, et al. Genome-wide interaction study reveals age-dependent determinants of responsiveness to inhaled corticosteroids in individuals with asthma. *PLoS One*. 2020;15(3). doi:10.1371/JOURNAL.PONE.0229241
287. Zakarya R, Adcock I, Oliver BG. Epigenetic impacts of maternal tobacco and e-vapour exposure on the offspring lung. *Clin Epigenetics*. 2019;11(1). doi:10.1186/S13148-019-0631-3
288. Lewis CM, Vassos E. Prospects for using risk scores in polygenic medicine. *Genome Med*. 2017;9(1). doi:10.1186/S13073-017-0489-Y
289. Palk AC, Dalvie S, de Vries J, Martin AR, Stein DJ. Potential use of clinical polygenic risk scores in psychiatry-ethical implications and communicating high polygenic risk. *Philosophy, Ethics, and Humanities in Medicine*. 2019;14(4):1-12. doi:10.1186/s13010-019-0073-8
290. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* . 2010;5(9):1315-1316. doi:10.1097/JTO.0B013E3181EC173D
291. Xie Z, Li D. Cross-Sectional Association Between Lifetime Use of Electronic Cigarettes With or Without Marijuana and Self-Reported Past 12-Month Respiratory Symptoms as well as Lifetime Respiratory Diseases in U.S. Adults. *Nicotine Tob Res*. 2020;22(Suppl 1):S70-S75. doi:10.1093/NTR/NTAA194
292. Villarroel MA, Cha AE, Vahratian A. Electronic Cigarette Use Among U.S. Adults, 2018 Key findings Data from the National Health Interview Survey. *NCHS Data*

Brief. 2020;365. Accessed April 4, 2022. https://www.cdc.gov/nchs/products/index.htm.
293. Liu G, Wasserman E, Kong L, Foulds J. A Comparison of Nicotine Dependence among Exclusive E-cigarette and Cigarette Users in the PATH Study. *Prev Med (Baltim)*. 2017;104:86. doi:10.1016/J.YPMED.2017.04.001
294. Rose JE, Behm FM, Drgon T, Johnson C, Uhl GR. Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype score. *Mol Med*. 2010;16(7-8):247-253. doi:10.2119/MOLMED.2009.00159
295. Perez MF, Atuegwu NC, Oncken C, Mead EL, Mortensen EM. Association between Electronic Cigarette Use and Asthma in Never-Smokers. *Ann Am Thorac Soc*. 2019;16(11):1453-1456. doi:10.1513/ANNALSATS.201904-338RL/SUPPL_FILE/DISCLOSURES.PDF
296. Osei AD, Mirbolouk M, Orimoloye OA, et al. The association between e-cigarette use and asthma among never combustible cigarette smokers: Behavioral risk factor surveillance system (BRFSS) 2016 & 2017. *BMC Pulmonary Medicine*. 2019;19(1):1-6. doi:10.1186/S12890-019-0950-3/FIGURES/1
297. Cho JH, Paik SY. Association between Electronic Cigarette Use and Asthma among High School Students in South Korea. *PLOS ONE*. 2016;11(3):e0151022. doi:10.1371/JOURNAL.PONE.0151022
298. Triantafyllou GA, Tiberio PJ, Zou RH, et al. Long-term outcomes of EVALI: a 1-year retrospective study. *The Lancet Respiratory Medicine*. 2021;9(12):e112-e113. doi:10.1016/S2213-2600(21)00415-X/ATTACHMENT/7FE95A05-7615-4038-9E59-A05B6785373E/MMC1.PDF
299. Layden JE, Ghinai I, Pray I, et al. Pulmonary Illness Related to E-Cigarette Use in Illinois and Wisconsin — Final Report. *New England Journal of Medicine*. 2020;382(10):903-916. doi:10.1056/NEJMOA1911614/SUPPL_FILE/NEJMOA1911614_PRELIMINARY-REPORT.PDF
300. Ruan Y, Lin YF, Feng YCA, et al. Improving Polygenic Prediction in Ancestrally Diverse Populations. *medRxiv*. Published online August 25, 2021:2020.12.27.20248738. doi:10.1101/2020.12.27.20248738
301. Economics of tobacco: education, income, and smoking.
302. Jaber RM, Mirbolouk M, Defilippis AP, et al. Electronic cigarette use prevalence, associated factors, and pattern by cigarette smoking status in the United States from NHANES (National health and nutrition examination survey) 2013–2014. *J Am Heart Assoc*. 2018;7(14):2013-2014. doi:10.1161/JAHA.117.008178
303. Guindon GE, Tobin S, Yach D. Trends and affordability of cigarette prices: Ample room for tax increases and related health gains. *Tobacco Control*. 2002;11(1):35-43. doi:10.1136/tc.11.1.35
304. Hiscock R, Bauld L, Amos A, Fidler JA, Munafò M. Socioeconomic status and smoking: A review. *Ann N Y Acad Sci*. 2012;1248(1):107-123. doi:10.1111/j.1749-6632.2011.06202.x
305. Meo SA, Al Asiri SA. Effects of electronic cigarette smoking on human health. *Eur Rev Med Pharmacol Sci*. 2014;18(21):3315-3319. doi:8033 [pii]

306. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the united States: What the patterns tell us. *American Journal of Public Health*. 2010;100(SUPPL. 1). doi:10.2105/AJPH.2009.166082
307. Michael McWilliams J. Health Consequences of Uninsurance among Adults in the United States: Recent Evidence and Implications. *Milbank Quarterly*. 2009;87(2):443-494. doi:10.1111/j.1468-0009.2009.00564.x
308. Choi K, Boyle RG. Changes in cigarette expenditure minimising strategies before and after a cigarette tax increase. *Tobacco Control*. 2018;27(1):99-104. doi:10.1136/tobaccocontrol-2016-053415
309. Rose SW, Glasser AM, Zhou Y, et al. Adolescent tobacco coupon receipt, vulnerability characteristics and subsequent tobacco use: Analysis of PATH Study, Waves 1 and 2. *Tobacco Control*. 2018;27(e1):e50-e56. doi:10.1136/tobaccocontrol-2017-054141
310. Choi K, Chen JC, Tan ASL, Soneji S, Moran MB. Recepit of tobacco direct mail/email coupons and trajectories of cigarette behaviours in a nationally representative longituinal cohort of US adults. *Tobacco Control*. 2019;28(3).
311. Choi K, Soneji S, Tan ASL. Receipt of Tobacco Direct Mail Coupons and Changes in Smoking Status in a Nationally Representative Sample of US Adults. *Nicotine & Tobacco Research*. 2018;20(9). doi:10.1093/NTR/NTX141
312. Dai H, Hao J. Direct Marketing Promotion and Electronic Cigarette Use Among US Adults, National Adult Tobacco Survey, 2013-2014. *Prev Chronic Dis*. 2017;14:E84. doi:10.5888/pcd14.170073
313. Papaleontiou L, Agaku IT, Filippidis FT. Effects of Exposure to Tobacco and Electronic Cigarette Advertisements on Tobacco Use: An Analysis of the 2015 National Youth Tobacco Survey. *Journal of Adolescent Health*. 2020;66(1):64-71. doi:10.1016/j.jadohealth.2019.05.022
314. Whelan C. Electronic Cigarettes: How They Work, Risk, Cost, Benefit & More.
315. English TMA, Smith TB, Song X, Whitman M V. Barriers to electronic cigarette use. *Public Health Nursing*. 2018;35(5):363-368. doi:10.1111/phn.12406
316. Dai H, Hao J. Direct Marketing Promotion and Electronic Cigarette Use Among US Adults, National Adult Tobacco Survey, 2013-2014. *Prev Chronic Dis*. 2017;14:E84. doi:10.5888/pcd14.170073
317. Papaleontiou L, Agaku IT, Filippidis FT. Effects of Exposure to Tobacco and Electronic Cigarette Advertisements on Tobacco Use: An Analysis of the 2015 National Youth Tobacco Survey. *Journal of Adolescent Health*. 2020;66(1):64-71. doi:10.1016/j.jadohealth.2019.05.022
318. United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse and USD of H and HServicesF and DAdministrationC for TP. Population Assessment of Tobacco and Health (PATH) Study [United States] Public-Use Files. Published online 2020. doi:10.3886/ICPSR36498.v11
319. Hyland A, Ambrose BK, Conway KP, et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. *Tobacco Control*. 2017;26(4):371-378. doi:10.1136/tobaccocontrol-2016-052934

320. Friedman AS, Horn SJL. Socioeconomic Disparities in Electronic Cigarette Use and Transitions from Smoking. *Nicotine and Tobacco Research*. 2019;21(10):1363-1370. doi:10.1093/ntr/nty120
321. Bullen C. Electronic Cigarettes for Smoking Cessation. *Current Cardiology Reports*. 2014;16(11):1-8. doi:10.1007/s11886-014-0538-8
322. Weaver SR, Majeed BA, Pechacek TF, Nyman AL, Gregory KR, Eriksen MP. Use of electronic nicotine delivery systems and other tobacco products among USA adults, 2014: results from a national survey. *International Journal of Public Health*. 2016;61(2):177-188. doi:10.1007/s00038-015-0761-0
323. Pepper JK, Brewer NT. Electronic nicotine delivery system (electronic cigarette) awareness, use, reactions and beliefs: A systematic review. *Tobacco Control*. 2014;23(5):375-384. doi:10.1136/tobaccocontrol-2013-051122
324. Lee AS, Hart JL, Walker KL, Keith RJ, Ridner SL. Dual Users and Electronic Cigarette Only Users: Consumption and Characteristics. *International Journal of Healthcare and Medical Sciences*. 2018;4(6):111-116.
325. Choi K, Hennrikus DJ, Forster JL, Moilanen M. Receipt and redemption of cigarette coupons, perceptions of cigarette companies and smoking cessation. *Tobacco Control*. 2013;22(6):418-422. doi:10.1136/tobaccocontrol-2012-050539
326. Friedman AS, Horn SJL. Socioeconomic Disparities in Electronic Cigarette Use and Transitions from Smoking. *Nicotine and Tobacco Research*. 2019;21(10):1363-1370. doi:10.1093/ntr/nty120
327. Judkins DR. Fay's method for variance estimation. *Journal of Official Statistics*. 1990;6(3):223-239.
328. Brock B, Schillo BA, Moilanen M. Tobacco industry marketing: An analysis of direct mail coupons and giveaways. *Tobacco Control*. 2015;24(5):505-508. doi:10.1136/TOBACCOCONTROL-2014-051602
329. Brown-Johnson CG, England LJ, Glantz SA, Ling PM. Tobacco industry marketing to low socioeconomic status women in the U.S.A. *Tob Control*. 2014;23(e2):e139-e146. doi:10.1136/TOBACCOCONTROL-2013-051224
330. Cigarette Smoking Among U.S. Adults Hits All-Time Low | CDC Online Newsroom | CDC. Published 2019. Accessed January 3, 2020. https://www.cdc.gov/media/releases/2019/p1114-smoking-low.html
331. Baig SA, Giovenco DP. Behavioral heterogeneity among cigarette and e-cigarette dual-users and associations with future tobacco use: Findings from the Population Assessment of Tobacco and Health Study. *Addictive Behaviors*. 2020;104. doi:10.1016/j.addbeh.2019.106263
332. van walbeek C, Blecher E, Gilmore A, Ross H. Price and Tax Measures and Illicit Trade in the Framework Convention on Tobacco Control: What We Know and What Research Is Required. *Nicotine & Tobacco Research*. 2013;15(4):767. doi:10.1093/NTR/NTS170
333. Hoek J, Edwards R, Thomson GW, Waa A, Wilson N. Tobacco excise taxes: a health and social justice measure? *Tob Control*. 2021;30(3):258-259. doi:10.1136/TOBACCOCONTROL-2020-055735
334. Baig SA, Giovenco DP. Behavioral heterogeneity among cigarette and e-cigarette dual-users and associations with future tobacco use: Findings from the Population

Assessment of Tobacco and Health Study. *Addictive Behaviors*. 2020;104. doi:10.1016/j.addbeh.2019.106263
335. Drummond MB, Upson D. Electronic cigarettes: Potential harms and benefits. *Ann Am Thorac Soc*. 2014;11(2):236-242. doi:10.1513/AnnalsATS.201311-391FR
336. Campbell DJT, Ronksley PE, Manns BJ, et al. The Association of Income with Health Behavior Change and Disease Monitoring among Patients with Chronic Disease. Rosenbaum JT, ed. *PLoS ONE*. 2014;9(4):e94007. doi:10.1371/journal.pone.0094007
337. Mehta NK, House JS, Elliott MR. Dynamics of health behaviours and socioeconomic differences in mortality in the USA. *Journal of Epidemiology and Community Health*. 2015;69(5):416-422. doi:10.1136/jech-2014-204248
338. Pampel FC, Krueger PM, Denney JT. Socioeconomic disparities in health behaviors. *Annual Review of Sociology*. 2010;36:349-370. doi:10.1146/annurev.soc.012809.102529
339. Farsalinos KE, Le Houezec J. Regulation in the face of uncertainty: The evidence on electronic nicotine delivery systems (e-cigarettes). *Risk Management and Healthcare Policy*. 2015;8:157-167. doi:10.2147/RMHP.S62116
340. Peroskie A, O'Brien E, Poonai K. Perceived relative harm of using e-cigarettes predicts future product switching among US adult cigarette and e-cigarette dual users. *Addiction*. 2019;114(12):2197-2205.
341. Choi K, Hennrikus DJ, Forster JL, Moilanen M. Receipt and redemption of cigarette coupons, perceptions of cigarette companies and smoking cessation. *Tobacco Control*. 2013;22(6):418-422. doi:10.1136/tobaccocontrol-2012-050539
342. Nicksic NE, Snell LM, Rudy AK, Cobb CO, Barnes AJ. Tobacco marketing, e-cigarette susceptibility, and perceptions among adults. *American Journal of Health Behavior*. 2017;41(5):579-590. doi:10.5993/AJHB.41.5.7
343. Cornelius ME, Wang TW, Jamal A, Loretan CG, Neff LJ. Tobacco Product Use Among Adults — United States, 2019. *MMWR Morbidity and Mortality Weekly Report*. 2020;69(46):1736-1742. doi:10.15585/mmwr.mm6946a4
344. Stallings-Smith S, Ballantyne T. Ever Use of E-Cigarettes Among Adults in the United States: A Cross-Sectional Study of Sociodemographic Factors. *Inquiry (United States)*. 2019;56. doi:10.1177/0046958019864479
345. Breland A, Soule E, Lopez A, Ramôa C, El-Hellani A, Eissenberg T. Electronic cigarettes: what are they and what do they do? *Ann N Y Acad Sci*. 2017;1394(1):5-30. doi:10.1111/nyas.12977
346. Brock B, Carlson SC, Moilanen M, Schillo BA. Reaching consumers: How the tobacco industry uses email marketing. *Preventive Medicine Reports*. 2016;4:103-106. doi:10.1016/J.PMEDR.2016.05.020
347. Owotomo O, Maslowsky J, Pasch KE. Historical declines and disparities in cigarette coupon saving among adolescents in the United States, 1997-2013. *Prev Med (Baltim)*. 2017;100:61-66. doi:10.1016/J.YPMED.2017.04.011
348. Osman A, Queen T, Choi K, Goldstein AO. Receipt of direct tobacco mail/email coupons and coupon redemption: Demographic and socioeconomic disparities among adult smokers in the United States. *Preventive Medicine*. 2019;126. doi:10.1016/j.ypmed.2019.105778

349. Lewis M, Ling PM. “Gone are the days of mass-media marketing plans and short term customer relationships”: tobacco industry direct mail and database marketing strategies. *Tob Control*. 2016;25(4):430-436. doi:10.1136/TOBACCOCONTROL-2015-052314
350. Memon MA, Cheah JH, Raymayah T, Ting H, Chuah F, Cham TH. Moderation Analysis: Issues and Guidelines. *Journal of Applied Structural Equation Modeling*. 2019;3(1):1-9.
351. Harlow AF, Stokes A, Brooks DR. Socioeconomic and Racial/Ethnic Differences in E-Cigarette Uptake among Cigarette Smokers: Longitudinal Analysis of the Population Assessment of Tobacco and Health (PATH) Study. *Nicotine and Tobacco Research*. 2019;21(10):1385-1393. doi:10.1093/ntr/nty141
352. Lee JGL, Henriksen L, Rose SW, Moreland-Russell S, Ribisl KM. A systematic review of neighborhood disparities in point-of-sale tobacco marketing. *American Journal of Public Health*. 2015;105(9):e8-e18. doi:10.2105/AJPH.2015.302777
353. Brock B, Carlson SC, Moilanen M, Schillo BA. Reaching consumers: How the tobacco industry uses email marketing. *Preventive Medicine Reports*. 2016;4:103-106. doi:10.1016/j.pmedr.2016.05.020
354. Choi K, Soneji S, Tan ASL. Receipt of Tobacco Direct Mail Coupons and Changes in Smoking Status in a Nationally Representative Sample of US Adults. *Nicotine & Tobacco Research*. 2018;20(9). doi:10.1093/NTR/NTX141
355. Oh AY, Kacker A. Do electronic cigarettes impart a lower potential disease burden than conventional tobacco cigarettes?: Review on E-cigarette vapor versus tobacco smoke. *Laryngoscope*. 2014;124(12):2702-2706. doi:10.1002/lary.24750
356. Hong EP, Park JW. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics*. 2012;10(2):117. doi:10.5808/gi.2012.10.2.117
357. Kobus K. Peers and adolescent smoking. *Addiction (Abingdon, England)*. 2003;98 Suppl 1(SUPPL. 1):37-55. doi:10.1046/J.1360-0443.98.S1.4.X
358. Campbell R, Starkey F, Holliday J, et al. An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *The Lancet*. 2008;371(9624):1595-1602. doi:10.1016/S0140-6736(08)60692-3
359. Dobbie F, Purves R, McKell J, et al. Implementation of a peer-led school based smoking prevention programme: a mixed methods process evaluation. *BMC Public Health*. 2019;19(1). doi:10.1186/S12889-019-7112-7
360. Bilgiç N, Günay T. Evaluation of effectiveness of peer education on smoking behavior among high school students. *Saudi Medical Journal*. 2018;39(1):74. doi:10.15537/SMJ.2018.1.21774
361. Harris JL, Pierce M, Bargh JA. Priming effect of antismoking PSAs on smoking behaviour: a pilot study. *Tob Control*. 2014;23(4):285-290. doi:10.1136/TOBACCOCONTROL-2012-050670
362. Shi Z, Wang AL, Fairchild VP, et al. Addicted to green: priming effect of menthol cigarette packaging on brain response to smoking cues. *Tob Control*. Published online October 1, 2021:tobaccocontrol-2021-056639. doi:10.1136/TOBACCOCONTROL-2021-056639

363. Hersey JC, Niederdeppe J, Ng SW, Mowery P, Farrelly M, Messeri P. How state counter-industry campaigns help prime perceptions of tobacco industry practices to promote reductions in youth smoking. *Tob Control*. 2005;14(6):377-383. doi:10.1136/TC.2004.010785
364. Gao J, Li W, Willis-Owen SA, et al. Polymorphisms of PHF11 and DPP10 are associated with asthma and related traits in a Chinese population. *Respiration*. 2010;79(1):17-24. doi:10.1159/000235545
365. Weiss ST, Raby BA, Rogers A. Asthma genetics and genomics 2009. *Curr Opin Genet Dev*. 2009;19(3):279-282. doi:10.1016/J.GDE.2009.05.001
366. Zhang Y, Poobalasingam T, Yates LL, et al. Manipulation of dipeptidylpeptidase 10 in mouse and human in vivo and in vitro models indicates a protective role in asthma. *Dis Model Mech*. 2018;11(1). doi:10.1242/DMM.031369
367. Piper ME, Baker TB, Benowitz NL, Smith SS, Jorenby DE. E-cigarette Dependence Measures in Dual Users: Reliability and Relations with Dependence Criteria and E-cigarette Cessation. *Nicotine and Tobacco Research*. 2020;22(5):756-763. doi:10.1093/ntr/ntz040
368. Piper ME, Baker TB, Benowitz NL, Smith SS, Jorenby DE. E-cigarette Dependence Measures in Dual Users: Reliability and Relations With Dependence Criteria and E-cigarette Cessation. *Nicotine Tob Res*. 2020;22(5):756-763. doi:10.1093/NTR/NTZ040
369. Lerman C, Niaura R. Applying genetic approaches to the treatment of nicotine dependence. *Oncogene*. 2002;21(48):7412-7420. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=12379882
370. Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008;456(7218):18-21. doi:10.1038/456018a
371. Wray NR, Maier R. Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability. *Current Epidemiology Reports 2014 1:4*. 2014;1(4):220-227. doi:10.1007/S40471-014-0023-3
372. Wray NR, Lee SH, Kendler KS. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *European Journal of Human Genetics 2012 20:6*. 2012;20(6):668-674. doi:10.1038/ejhg.2011.257
373. Yeager DS, Krosnick JA. The validity of self-reported nicotine product use in the 2001-2008 National Health and Nutrition Examination Survey. *Med Care*. 2010;48(12):1128-1132. doi:10.1097/MLR.0B013E3181EF9948
374. Bonnie RJ, Stratton K, Kwan LY. The Effects of Tobacco Use on Health. Published online July 23, 2015. Accessed April 18, 2021. https://www.ncbi.nlm.nih.gov/books/NBK310413/
375. Mirbolouk M, Charkhchi P, Kianoush S, et al. Prevalence and distribution of e-cigarette use among U.S. adults: Behavioral risk factor surveillance system, 2016. *Annals of Internal Medicine*. 2018;169(7):429-438. doi:10.7326/M17-3440

STATISTICAL CODE

```
### ANALYSIS FOR CHAPTER 2
.#####----------------####
##### Equating Sexes  Elelctronic Cigarettes ####
#####----------------####

#rm(list=ls())

#source('http://openmx.psyc.virginia.edu/getOpenMx.R')
setwd('C:/Users/cliffordjs/Desktop/Research Projects/Adolescent and Young Adult Twin
Study/6. Liz Edits')
#rm(list = ls(all = TRUE))
source("C:/Users/cliffordjs/Desktop/Super wicked important R
files/GenEpiHelperFunctions.R")
source("C:/Users/cliffordjs/Desktop/Super wicked important R files/miFunctions.R")
#source('http://openmx.ssri.psu.edu/getOpenMx.R')
#omxGetNPSOL()

require(MASS)
require(OpenMx)
require(psych)
require(polycor)
mxOption( NULL, "Default optimizer","CSOLNP" )


# Call data, NAs = NA

setwd('C:/Users/cliffordjs/Desktop/Research Projects/Adolescent and Young Adult Twin
Study/0. Raw Data')

# was twindata2019.csv

data2<-read.csv("TwinData4.csv", header=T,na.strings=c("9999", "NA"))
names(data2)
table(data2$zyg2)


###-------------------------------------------------###
###           UNIVARIATE FOR EC                     ###
###-------------------------------------------------###


# set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
vars      <- c("ecigEver3")
```

```
#vars      <- c("cccigEver3", "ecigEver3")   #reverse order of variables to see if same
results emerge
nv        <- 1          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
nth <- 1  # Number of thresholds per variable (only for binary data)

# Subset the data to only the things I need
twinDatauni  <- data2[,c(selVars,'zyg2')]
describe(twinDatauni)
summary(twinDatauni)
dim(twinDatauni)

#twinData2<-na.omit(twinData)
#summary(twinData2)
#dim(twinData2)
twinDataBin <-twinDatauni
dim(twinDataBin)
table(twinDataBin$zyg2)

# Factorize Ordinal Variables using the mxFactor option
twinDataBin[,c(1,2)]  <- mxFactor(twinDataBin[,c(1,2)], levels = c(0:nth))

# 1=MZM, 2= MZF, 3=DZM, 4=DZF, 5=ODZ
#Vars      <- c("cccigEver3")
#nv        <- 1          # number of variables
#ntv       <- nv*2       # number of total variables
#selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")

# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #males = T1, females = T2


polychor(mzfData$ecEver3_T1,mzfData$ecEver3_T2, std.err=T)
polychor(mzmData$ecEver3_T1,mzmData$ecEver3_T2, std.err=T)
polychor(dzfData$ecEver3_T1,dzfData$ecEver3_T2, std.err=T)
polychor(dzmData$ecEver3_T1,dzmData$ecEver3_T2, std.err=T)
polychor(dzoData$ecEver3_T1,dzoData$ecEver3_T2, std.err=T)


# Set Starting Values /
svLTh     <- 0.8    # start value for first threshold
```

```
svlTh     <- 1       # start value for increments
#svTh      <- c(0.7,1,0.7,1)
svTh      <- matrix(rep(c(svLTh,(rep(svlTh,nth-1)))),nrow=nth,ncol=ntv)     # start value
for thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=ntv)     # lower bounds
for thresholds

#svTh      <- c(1,1)                  # start value for thresholds
svPa      <- .4                      # start value for path coefficient
svPaD     <- vech(diag(svPa,nv,nv))    # start values for diagonal of covariance matrix
svPe      <- .8                      # start value for path coefficient for e
svPeD     <- vech(diag(svPe,nv,nv))    # start values for diagonal of covariance matrix
lbPa      <- .00001                  # start value for lower bounds
lbPaD     <- diag(lbPa,nv,nv)          # lower bounds for diagonal of covariance matrix
lbPaD[lower.tri(lbPaD)] <- -2        # lower bounds for below diagonal elements
lbPaD[upper.tri(lbPaD)] <- NA          # lower bounds for above diagonal elements


# Set Starting Values
aLabs     <- paste("a",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
cLabs     <- paste("c",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
eLabs     <- paste("e",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
var1thM     <- paste("var1M","_th",1:nth, sep="")
var2thM     <- paste("var2M","_th",1:nth, sep="")
var1thF     <- paste("var1F","_th",1:nth, sep="")
var2thF     <- paste("var2F","_th",1:nth, sep="")

thUB      <- 2

# ------------------------------------------------------------------------------
# PREPARE MODEL

# ACE Model
# Create Algebra for expected Mean Matrices to include differing thresholds for males
and females
meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
threGm     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values= svTh,
                    labels=c(var1thM,var2thM), name="threGm",lbound=-2, ubound=2 )
threGf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                    labels=c(var1thF,var2thF), name="threGf",lbound=-2, ubound=2  )
threGmf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                     labels=c(var1thM,var2thM,var1thF,var2thF), name="threGmf", lbound=-
2, ubound=2  )
# Create Matrices for Path Coefficients
pathA     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=c(0.2),
#replacing svPaD
```

```
                label=aLabs, lbound=lbPaD, name="a" )
pathC    <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=c(0.4),
                label=cLabs, lbound=lbPaD, name="c" )
pathE    <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE,
                values=0.4, label=eLabs, lbound=lbPaD, name="e" )

# Create Algebra for Variance Comptwonts
covA     <- mxAlgebra( expression=a %*% t(a), name="A" )
covC     <- mxAlgebra( expression=c %*% t(c), name="C" )
covE     <- mxAlgebra( expression=e %*% t(e), name="E" )

# Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
covP     <- mxAlgebra( expression= A+C+E, name="V" )
covMZ    <- mxAlgebra( expression= A+C, name="cMZ" )
covDZ    <- mxAlgebra( expression= 0.5%x%A+ C, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)),
name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)),
name="expCovDZ" )

# Create Algebra for Standardization
matI     <- mxMatrix( type="Iden", nrow=nv, ncol=nv, name="I")
invSD    <- mxAlgebra( expression=solve(sqrt(I*V)), name="iSD")

# Calculate genetic and environmental correlations
corA     <- mxAlgebra( expression=solve(sqrt(I*A))%&%A, name ="rA" ) #cov2cor()
corC     <- mxAlgebra( expression=solve(sqrt(I*C))%&%C, name ="rC" )
corE     <- mxAlgebra( expression=solve(sqrt(I*E))%&%E, name ="rE" )

## Calculate Phenotypic Correlation ##
corP     <- mxAlgebra (expression=solve(sqrt(I*V)) %*% V %*% solve(sqrt(I*V)),
name="rP")

## Calculate Standardized Covariances ##
stCovA   <- mxAlgebra (solve(sqrt(I*V)) %*% A %*% solve(sqrt(I*V)), name="stCovA")
stCovC   <- mxAlgebra (solve(sqrt(I*V)) %*% C %*% solve(sqrt(I*V)), name="stCovC")
stCovE   <- mxAlgebra (solve(sqrt(I*V)) %*% E %*% solve(sqrt(I*V)), name="stCovE")


# Constrain Variance of Binary Variables
matUnv   <- mxMatrix( type="Unit", nrow=nv, ncol=1, name="Unv1" )
var1     <- mxConstraint( expression=diag2vec(V)==Unv1, name="Var1" )

# Create Algebra for Variance Components
rowVC    <- rep('VC',nv)
colVC    <- rep(c('A','C','E','SA','SC','SE'),each=nv)
```

```
estVC     <- mxAlgebra( expression=cbind(A,C,E,A/V,C/V,E/V), name="VC",
dimnames=list(rowVC,colVC))

# Create Confidence Interval Objects
ciACE  <- mxCI(c("stCovA","stCovC", "stCovE"))#
"VC[1,seq(1,3*nv,nv),(2,seq(1,3*nv,nv)),(2,seq(2,3*nv,nv)))]" )

# Data objects for Multiple Groups
dataMZf   <- mxData( observed=mzfData, type="raw" )
dataDZf   <- mxData( observed=dzfData, type="raw" )
dataMZm   <- mxData( observed=mzmData, type="raw" )
dataDZm   <- mxData( observed=dzmData, type="raw" )
dataDZo   <- mxData( observed=dzoData, type="raw" )

# Expectation objects for Multiple Groups
expMZf    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf")
expDZf    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf" )
expMZm    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZm    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZo    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGmf" )
funML     <- mxFitFunctionML()

# Combine Groups
parsZf    <- list( pathA, pathC, pathE,
                covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                meanG, threGf, matI, invSD, matUnv )
parsZm    <- list( pathA, pathC, pathE,
                covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                meanG, threGm, matI, invSD, matUnv )
parsZmf   <- list( pathA, pathC, pathE,
                covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                meanG, threGmf, matI, invSD, matUnv )

modelMZf  <- mxModel( parsZf, meanG, covMZ, expCovMZ, dataMZf, expMZf, funML,
name="MZf" )
modelDZf  <- mxModel( parsZf, meanG, covDZ, expCovDZ, dataDZf, expDZf, funML,
name="DZf" )
modelMZm  <- mxModel( parsZm, meanG, covMZ, expCovMZ, dataMZm, expMZm,
funML, name="MZm" )
modelDZm  <- mxModel( parsZm, meanG, covDZ, expCovDZ, dataDZm, expDZm,
funML, name="DZm" )
```

```
modelDZo  <- mxModel( parsZmf, meanG, covDZ, expCovDZ,dataDZo, expDZo,
funML, name="DZo" )
multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
EcUniAceModel  <- mxModel( "EcUniAce", parsZf, parsZm, parsZmf,
                    modelMZf, modelDZf, modelMZm, modelDZm, modelDZo, multi,
                    estVC, ciACE)

EcUniAceFit <-mxRun(EcUniAceModel, intervals = F)
EcUniAceFit <-mxTryHardOrdinal(EcUniAceFit, intervals = F)
EcUniAceFit$algebras

#tableFitStatistics(QualAceFit, EcUniAceFit)


# Confidence Intervals

EcUniAceFit2 <-mxTryHardOrdinal(EcUniAceFit, intervals = T)
summary(EcUniAceFit2, verbose=T)


  # Test of A
  EcUniNoA <- EcUniAceFit2
  EcUniNoAModel<- omxSetParameters(EcUniNoA, labels=c( "a11"), free=FALSE,
values=0 )
  EcUniNoAfit<- mxTryHardOrdinal(EcUniNoAModel, intervals = T)
  tableFitStatistics(EcUniAceFit, EcUniNoAfit)


  # Test of C
  EcUniNoC <- EcUniAceFit2
  EcUniNoCModel<- omxSetParameters(EcUniNoC, labels=c( "c11"), free=FALSE,
values=0 )
  EcUniNoCFit<- mxTryHardOrdinal(EcUniNoCModel, intervals = F)
  tableFitStatistics(EcUniAceFit, EcUniNoCFit)

  EcUniNoC2 <- EcUniNoCFit
  EcUniNoCModel2<- omxSetParameters(EcUniNoC2, labels=c( "c11"), free=FALSE,
values=0 )
  EcUniNoCFit2<- mxTryHardOrdinal(EcUniNoCModel2, intervals = F)
  tableFitStatistics(EcUniAceFit, EcUniNoCFit2)



  # Test of E
  EcUniNoE <- EcUniAceFit2
```

```
  EcUniNoEModel<- omxSetParameters(EcUniNoE, labels=c("a11", "c11"),
free=FALSE, values=0 )
  EcUniNoEModelFit<- mxTryHardOrdinal(EcUniNoEModel, intervals = F)
  tableFitStatistics(EcUniAceFit, EcUniNoEModelFit)

  ECallmodels<-list(EcUniNoAfit,EcUniNoCFit,EcUniNoEModelFit)
  tableFitStatistics(EcUniAceFit, ECallmodels)



  ###-------------------------------------------------###
  ###           Testing model assumptions for EC      ###
  ###-------------------------------------------------###

  # set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
  Vars      <- c("ecigEver3")
  nv        <- 1          # number of variables
  ntv       <- nv*2       # number of total variables
  selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
  nth <- 1  # Number of thresholds per variable (only for ordinal data)


  mzData_1 <- subset(twinDataBin, zyg2 %in% c(1,2), select = selVars)
  dzData_1 <- subset(twinDataBin, zyg2 %in% c(3,4,5), select = selVars)
  # Set Starting Values
  svLTh     <- -1.5    # start value for first threshold
  svITh     <- 1       # start value for increments
  svTh      <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=nv)     # start value
for thresholds
  lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv)     # lower bounds
for thresholds
  svCor     <- .5      # start value for correlations
  lbCor     <- -0.99   # lower bounds for correlations
  ubCor     <- 0.99    # upper bounds for correlations

  labThMZ   <- c(paste("t",1:nth,"MZ1",sep=""),paste("t",1:nth,"MZ2",sep=""))
  labThDZ   <- c(paste("t",1:nth,"DZ1",sep=""),paste("t",1:nth,"DZ2",sep=""))

  # ------------------------------------------------------------------------
  # PREPARE MODEL

  # Saturated Model
  # Algebra for expected Mean & Threshold Matrices in MZ & DZ twins
  meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
```

```
  thinMZ    <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=TRUE, values=svTh,
lbound=lbTh, labels=labThMZ, name="thinMZ" )
  thinDZ    <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=TRUE, values=svTh,
lbound=lbTh, labels=labThDZ, name="thinDZ" )
  inc       <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=FALSE, values=1,
name="inc" )
  threMZ    <- mxAlgebra( expression= inc %*% thinMZ, name="threMZ" )
  threDZ    <- mxAlgebra( expression= inc %*% thinDZ, name="threDZ" )

  # Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
  corMZ     <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels="rMZ", name="corMZ" )
  corDZ     <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels="rDZ", name="corDZ" )

  # Data objects for Multiple Groups
  dataMZ   <-mxData( observed=mzData_1, type="raw" )
  dataDZ   <-mxData( observed=dzData_1, type="raw" )

  # Objective objects for Multiple Groups
  expMZ     <- mxExpectationNormal( covariance="corMZ", means="meanG",
dimnames=selVars, thresholds="threMZ" )
  expDZ     <- mxExpectationNormal( covariance="corDZ", means="meanG",
dimnames=selVars, thresholds="threDZ" )
  funML     <- mxFitFunctionML()

  # Combine Groups
  modelMZ   <- mxModel( "MZ", meanG, corMZ, thinMZ, inc, threMZ, dataMZ, expMZ,
funML )
  modelDZ   <- mxModel( "DZ", meanG, corDZ, thinDZ, inc, threDZ, dataDZ, expDZ,
funML )
  multi     <- mxFitFunctionMultigroup( c("MZ","DZ") )
  ciCor     <- mxCI( c('MZ.corMZ','DZ.corDZ' ))
  ciThre    <- mxCI( c('MZ.threMZ','DZ.threDZ' ))
  twinSatOrdModel   <- mxModel( "EC Cigs", modelMZ, modelDZ, multi, ciCor, ciThre )

  # ------------------------------------------------------------------------------
  # RUN MODEL

  # Run Saturated Model
  twinSatOrdFit     <- mxRun( twinSatOrdModel, intervals=T )
  twinSatOrdSum     <- summary( twinSatOrdFit )
  twinSatOrdSum
  round(twinSatOrdFit$output$estimate,4)

  # Generate Saturated Model Output
```

```
  rMZ      <- twinSatOrdFit$MZ.corMZ$values[2,1]
  rDZ      <- twinSatOrdFit$DZ.corDZ$values[2,1]
  tMZ      <- twinSatOrdFit$MZ.threMZ$result
  tDZ      <- twinSatOrdFit$DZ.threDZ$result

  twinSatOrdOS     <- twinSatOrdSum$observedStatistics
  twinSatOrdDF     <- twinSatOrdSum$degreesOfFreedom
  twinSatOrdNP     <- length(twinSatOrdSum$parameters[[1]])
  twinSatOrdLLL    <- twinSatOrdFit$output$Minus2LogLikelihood
  twinSatOrdAIC    <- twinSatOrdSum$AIC

  mxCompare(twinSatOrdFit)

  # ------------------------------------------------------------------------
  # RUN SUBMODELS

  # Constrain expected Thresholds to be equal across twin order
  eqThresTwinModel    <- mxModel(twinSatOrdFit, name="eqThresTwin" )
  eqThresTwinModel    <- omxSetParameters( eqThresTwinModel,
label=c("t1MZ1","t1MZ2"), free=TRUE, values=svLTh, newlabels='t1MZ' )

  eqThresTwinModel    <- omxSetParameters( eqThresTwinModel,
label=c("t1DZ1","t1DZ2"), free=TRUE, values=svLTh, newlabels='t1DZ' )

  eqThresTwinFit     <- mxRun( eqThresTwinModel, intervals=F )
  eqThresTwinSum     <- summary( eqThresTwinFit )
  eqThresTwinLLL     <- eqThresTwinFit$output$Minus2LogLikelihood
  mxCompare(twinSatOrdFit, eqThresTwinFit)

  # Constrain expected Thres to be equal across twin order and zygosity
  eqThresZygModel    <- mxModel(eqThresTwinModel, name="eqThresZyg" )
  eqThresZygModel    <- omxSetParameters( eqThresZygModel,
label=c("t1MZ","t1DZ"), free=TRUE, values=svLTh, newlabels='t1Z' )


  eqThresZygFit      <- mxRun( eqThresZygModel, intervals=F )
  eqThresZygSum      <- summary( eqThresZygFit )
  eqThresZygLLL      <- eqThresZygFit$output$Minus2LogLikelihood
  mxCompare(eqThresTwinFit, eqThresZygFit)

  # ------------------------------------------------------------------------

  # Print Comparative Fit Statistics
  SatNested <- list(eqThresTwinFit, eqThresZygFit)
  mxCompare(twinSatOrdFit, SatNested)
```

```
tableFitStatistics(twinSatOrdFit, SatNested)



###-------------------------------------------------###
###            Testing Sex Difference EC            ###
###-------------------------------------------------###
# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #fm

# Set Starting Values
svLTh     <- -1.5    # start value for first threshold
svITh     <- 1       # start value for increments
svTh      <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=nv)     # start value
for thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv)     # lower bounds
for thresholds
svCor     <- .5      # start value for correlations
lbCor     <- -0.99   # lower bounds for correlations
ubCor     <- 0.99    # upper bounds for correlations

labThMZ   <- c(paste("t",1:nth,"MZ1",sep=""),paste("t",1:nth,"MZ2",sep=""))
labThDZ   <- c(paste("t",1:nth,"DZ1",sep=""),paste("t",1:nth,"DZ2",sep=""))
mvar1th     <- paste("mvar1","_th",1:nth, sep="")
mvar2th     <- paste("mvar2","_th",1:nth, sep="")
fvar1th     <- paste("fvar1","_th",1:nth, sep="")
fvar2th     <- paste("fvar2","_th",1:nth, sep="")
dzvar1th     <- paste("dzvar1","_th",1:nth, sep="")
dzvar2th     <- paste("dzvar2","_th",1:nth, sep="")
thUB      <- 2

# ------------------------------------------------------------------------------
# PREPARE MODEL

# General non-scalar ACE Model
# Matrices declared to store a, c, and e Path Coefficients
pathAf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="af11", name="af" )
pathCf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="cf11", name="cf" )
pathEf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="ef11", name="ef" )
```

```
  pathAm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="am11", name="am" )
  pathCm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="cm11", name="cm" )
  pathEm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="em11", name="em" )
  pathRa    <- mxMatrix( "Lower", nrow=1, ncol=1, free=TRUE,  values=1, label="ra11",
name="ra", ubound=1, lbound=0 )

  # Matrices generated to hold A, C, and E computed Variance Components
  covAf     <- mxAlgebra( af %*% t(af), name="Af" )
  covCf     <- mxAlgebra( cf %*% t(cf), name="Cf" )
  covEf     <- mxAlgebra( ef %*% t(ef), name="Ef" )
  covAm     <- mxAlgebra( am %*% t(am), name="Am" )
  covCm     <- mxAlgebra( cm %*% t(cm), name="Cm" )
  covEm     <- mxAlgebra( em %*% t(em), name="Em" )

  # Algebra to compute total variances and standard deviations (diagonal only)
  covPf     <- mxAlgebra( Af+Cf+Ef, name="Vf" )
  covPm     <- mxAlgebra( Am+Cm+Em, name="Vm" )

  # Algebras generated to hold Parameter Estimates and Derived Variance Components
  colVarsZf <- c('Af','Cf','Ef','SAf','SCf','SEf')
  estVarsZf <- mxAlgebra( cbind(Af,Cf,Ef,Af/Vf,Cf/Vf,Ef/Vf), name="VarsZf",
dimnames=list(NULL,colVarsZf))
  colVarsZm <- c('Am','Cm','Em','SAm','SCm','SEm')
  estVarsZm <- mxAlgebra( cbind(Am,Cm,Em,Am/Vm,Cm/Vm,Em/Vm),
name="VarsZm", dimnames=list(NULL,colVarsZm))

  # Algebra for expected Mean and Variance/Covariance Matrices in MZ & DZ twins
  meanGf    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanf", name="expMeanGf" )
  meanGm    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanm", name="expMeanGm" )
  meanGfm   <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label=c("meanf","meanm"), name="expMeanGfm" )
  covMZf    <- mxAlgebra( expression= rbind( cbind(Vf, Af+Cf), cbind(Af+Cf, Vf)),
name="expCovMZf" )
  covDZf    <- mxAlgebra( expression= rbind( cbind(Vf, 0.5%x%Af+Cf),
cbind(0.5%x%Af+Cf, Vf)), name="expCovDZf" )
  covMZm    <- mxAlgebra( expression= rbind( cbind(Vm, Am+Cm), cbind(Am+Cm,
Vm)), name="expCovMZm" )
  covDZm    <- mxAlgebra( expression= rbind( cbind(Vm, 0.5%x%Am+Cm),
cbind(0.5%x%Am+Cm, Vm)), name="expCovDZm" )
  CVfm      <- mxAlgebra( expression= ra%x%(af%*%t(am))+cf%*%t(cm), name="CVfm"
)
```

```
  CVmf      <- mxAlgebra( expression= ra%x%(am%*%t(af))+cm%*%t(cf), name="CVmf"
)
  covDZo    <- mxAlgebra( expression= rbind( cbind(Vf, CVfm), cbind(CVmf, Vm)),
name="expCovDZo" )
  Inc     <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=F, values=1, name="Inc" )

  # MALES
  ThreM    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(mvar1th, mvar2th), lbound=-2, ubound=thUB, name="ThreM")
  ExpThreM   <- mxAlgebra( expression= cbind( ( Inc %*% ThreM  ),
                                    ( Inc %*% ThreM  ) ), name="ExpThreM" )
  # FEMALES
  ThreF    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(fvar1th, fvar2th), lbound=-2, ubound=thUB, name="ThreF")
  ExpThreF   <- mxAlgebra( expression= cbind( ( Inc %*% ThreF  ),
                                    ( Inc %*% ThreF  ) ), name="ExpThreF" )
  ## OS

  ThreOS    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(dzvar1th, dzvar2th), lbound=-2, ubound=thUB, name="ThreOS")
  ExpThreOS   <- mxAlgebra( expression= cbind( ( Inc %*% ThreOS  ),
                                     ( Inc %*% ThreOS  ) ), name="ExpThreOS" )


  # Data objects for Multiple Groups
  dataMZf   <- mxData( observed=mzfData, type="raw" )
  dataDZf   <- mxData( observed=dzfData, type="raw" )
  dataMZm   <- mxData( observed=mzmData, type="raw" )
  dataDZm   <- mxData( observed=dzmData, type="raw" )
  dataDZo   <- mxData( observed=dzoData, type="raw" )

  # Expectation objects for Multiple Groups
  expMZf    <- mxExpectationNormal( covariance="expCovMZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
  expDZf    <- mxExpectationNormal( covariance="expCovDZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
  expMZm    <- mxExpectationNormal( covariance="expCovMZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
  expDZm    <- mxExpectationNormal( covariance="expCovDZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
  expDZo    <- mxExpectationNormal( covariance="expCovDZo",
means="expMeanGfm", dimnames=selVars, thresholds = "ExpThreOS" )
  funML     <- mxFitFunctionML()

  # Combine Groups
```

```
  parsZf    <- list( pathAf, pathCf, pathEf, covAf, covCf, covEf, covPf, estVarsZf, ThreF,
ExpThreF, Inc )
  parsZm    <- list( pathAm, pathCm, pathEm, covAm, covCm, covEm, covPm,
estVarsZm, ThreM, ExpThreM, Inc )
  parsZfm   <- list( pathRa, CVfm, CVmf, ExpThreOS, Inc, ThreOS)
  modelMZf  <- mxModel( parsZf, meanGf, covMZf, dataMZf, expMZf, funML,
name="MZf" )
  modelDZf  <- mxModel( parsZf, meanGf, covDZf, dataDZf, expDZf, funML,
name="DZf" )
  modelMZm  <- mxModel( parsZm, meanGm, covMZm, dataMZm, expMZm, funML,
name="MZm" )
  modelDZm  <- mxModel( parsZm, meanGm, covDZm, dataDZm, expDZm, funML,
name="DZm" )
  modelDZo  <- mxModel( parsZf, parsZm, parsZfm, meanGfm, covDZo, dataDZo,
expDZo, funML, name="DZo" )
  multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
  QualAceModel  <- mxModel( "QualACE", modelMZf, modelDZf, modelMZm,
modelDZm, modelDZo, multi )


  QualAceFit <-mxTryHardOrdinal(QualAceModel, intervals = F)
  summary(QualAceFit)

  ## Coerce threshold to be equal


  eqthres   <-mxModel(QualAceFit, name = "Equal Threshold")
  eqthres   <-omxSetParameters( eqthres, label="mvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")
  eqthres   <-omxSetParameters( eqthres, label="fvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")
  eqthres   <-omxSetParameters( eqthres, label="dzvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")

  eqthresfit<-mxTryHardOrdinal(eqthres, intervals = F)
  summary(eqthresfit)
  tableFitStatistics(QualAceFit, eqthresfit)


  ## Coerce males and females to be equal

  eqsex   <-mxModel(QualAceFit, name = "Equal sexes")
  eqsex   <-omxSetParameters( eqsex, label="am11", free=TRUE, values=0.04,
newlabels="a11")
  eqsex   <-omxSetParameters( eqsex, label="af11", free=TRUE, values=0.04,
newlabels="a11")
```

```
  eqsex   <-omxSetParameters( eqsex, label="cm11", free=TRUE, values=0.5,
newlabels="c11")
  eqsex   <-omxSetParameters( eqsex, label="cf11", free=TRUE, values=0.5,
newlabels="c11")
  eqsex   <-omxSetParameters( eqsex, label="em11", free=TRUE, values=0.3,
newlabels="e11")
  eqsex   <-omxSetParameters( eqsex, label="ef11", free=TRUE, values=0.3,
newlabels="e11")

  eqsexfit<-mxTryHardOrdinal(eqsex, intervals = F)
  tableFitStatistics(QualAceFit, eqsexfit)

  ## NO sex

  nosex   <- omxSetParameters(eqsex, labels="ra11", name="No Sex Effects",
free=FALSE, values=0.5 )
  nosexfit<- mxTryHardOrdinal(nosex, intervals = F)
  nested <-list(eqthresfit, eqsexfit, nosexfit)
  tableFitStatistics(QualAceFit, nested)


  ###-----------------------------------------------------###
  ###            UNIVARIATE FOR CC                    ###
  ###-----------------------------------------------------###

  mxOption(NULL, "Default optimizer", "CSOLNP")
  # set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
  vars      <- c("cccigEver3")
  #vars      <- c("cccigEver3", "ecigEver3")   #reverse order of variables to see if same
results emerge
  nv        <- 1          # number of variables
  ntv       <- nv*2       # number of total variables
  selVars   <- paste(vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
  nth <- 1  # Number of thresholds per variable (only for binary data)

  # Subset the data to only the things I need
  twinDatauni  <- data2[,c(selVars,'zyg2')]
  describe(twinDatauni)
  summary(twinDatauni)
  dim(twinDatauni)

  #twinData2<-na.omit(twinData)
  #summary(twinData2)
  #dim(twinData2)
  twinDataBin <-twinDatauni
```

```
dim(twinDataBin)
table(twinDataBin$zyg2)

# Factorize Ordinal Variables using the mxFactor option
twinDataBin[,c(1,2)]  <- mxFactor(twinDataBin[,c(1,2)], levels = c(0:nth))

# 1=MZM, 2= MZF, 3=DZM, 4=DZF, 5=ODZ
#Vars      <- c("cccigEver3")
#nv        <- 1          # number of variables
#ntv       <- nv*2       # number of total variables
#selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")

# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #males = T1, females = T2

polychor(mzfData$cccever3_T1,mzfData$cccever3_T2, std.err=T)
polychor(mzmData$cccever3_T1,mzmData$cccever3_T2, std.err=T)
polychor(dzfData$cccever3_T1,dzfData$cccever3_T2, std.err=T)
polychor(dzmData$cccever3_T1,dzmData$cccever3_T2, std.err=T)
polychor(dzoData$cccever3_T1,dzoData$cccever3_T2, std.err=T)


# Set Starting Values /
svLTh     <- 0.8    # start value for first threshold
svITh     <- 1      # start value for increments
#svTh      <- c(0.7,1,0.7,1)
svTh      <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=ntv)     # start value
for thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=ntv)     # lower bounds
for thresholds

#svTh      <- c(1,1)                # start value for thresholds
svPa      <- .4                    # start value for path coefficient
svPaD     <- vech(diag(svPa,nv,nv))    # start values for diagonal of covariance matrix
svPe      <- .8                    # start value for path coefficient for e
svPeD     <- vech(diag(svPe,nv,nv))    # start values for diagonal of covariance matrix
lbPa      <- .00001                # start value for lower bounds
lbPaD     <- diag(lbPa,nv,nv)          # lower bounds for diagonal of covariance matrix
lbPaD[lower.tri(lbPaD)] <- -2          # lower bounds for below diagonal elements
lbPaD[upper.tri(lbPaD)] <- NA          # lower bounds for above diagonal elements
```

```
  # Set Starting Values
  aLabs     <- paste("a",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
  cLabs     <- paste("c",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
  eLabs     <- paste("e",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
  var1thM     <- paste("var1M","_th",1:nth, sep="")
  var2thM     <- paste("var2M","_th",1:nth, sep="")
  var1thF     <- paste("var1F","_th",1:nth, sep="")
  var2thF     <- paste("var2F","_th",1:nth, sep="")

  thUB      <- 2

  # ----------------------------------------------------------------------
  # PREPARE MODEL

  # ACE Model
  # Create Algebra for expected Mean Matrices to include differing thresholds for males
and females
  meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
  threGm     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values= svTh,
                      labels=c(var1thM,var2thM), name="threGm",lbound=-2, ubound=2 )
  threGf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                      labels=c(var1thF,var2thF), name="threGf",lbound=-2, ubound=2  )
  threGmf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                       labels=c(var1thM,var2thM,var1thF,var2thF), name="threGmf",
lbound=-2, ubound=2  )
  # Create Matrices for Path Coefficients
  pathA     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=c(0.2),
#replacing svPaD
                     label=aLabs, lbound=lbPaD, name="a" )
  pathC     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=c(0.4),
                     label=cLabs, lbound=lbPaD, name="c" )
  pathE     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE,
                     values=0.4, label=eLabs, lbound=lbPaD, name="e" )

  # Create Algebra for Variance Comptwonts
  covA      <- mxAlgebra( expression=a %*% t(a), name="A" )
  covC      <- mxAlgebra( expression=c %*% t(c), name="C" )
  covE      <- mxAlgebra( expression=e %*% t(e), name="E" )

  # Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
  covP      <- mxAlgebra( expression= A+C+E, name="V" )
  covMZ     <- mxAlgebra( expression= A+C, name="cMZ" )
  covDZ     <- mxAlgebra( expression= 0.5%x%A+ C, name="cDZ" )
  expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)),
name="expCovMZ" )
```

```
  expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)),
name="expCovDZ" )

  # Create Algebra for Standardization
  matI      <- mxMatrix( type="Iden", nrow=nv, ncol=nv, name="I")
  invSD     <- mxAlgebra( expression=solve(sqrt(I*V)), name="iSD")

  # Calculate genetic and environmental correlations
  corA      <- mxAlgebra( expression=solve(sqrt(I*A))%&%A, name ="rA" ) #cov2cor()
  corC      <- mxAlgebra( expression=solve(sqrt(I*C))%&%C, name ="rC" )
  corE      <- mxAlgebra( expression=solve(sqrt(I*E))%&%E, name ="rE" )

  ## Calculate Phenotypic Correlation ##
  corP      <- mxAlgebra (expression=solve(sqrt(I*V)) %*% V %*% solve(sqrt(I*V)),
name="rP")

  ## Calculate Standardized Covariances ##
  stCovA    <- mxAlgebra (solve(sqrt(I*V)) %*% A %*% solve(sqrt(I*V)), name="stCovA")
  stCovC    <- mxAlgebra (solve(sqrt(I*V)) %*% C %*% solve(sqrt(I*V)), name="stCovC")
  stCovE    <- mxAlgebra (solve(sqrt(I*V)) %*% E %*% solve(sqrt(I*V)), name="stCovE")


  # Constrain Variance of Binary Variables
  matUnv    <- mxMatrix( type="Unit", nrow=nv, ncol=1, name="Unv1" )
  var1      <- mxConstraint( expression=diag2vec(V)==Unv1, name="Var1" )

  # Create Algebra for Variance Components
  rowVC     <- rep('VC',nv)
  colVC     <- rep(c('A','C','E','SA','SC','SE'),each=nv)
  estVC     <- mxAlgebra( expression=cbind(A,C,E,A/V,C/V,E/V), name="VC",
dimnames=list(rowVC,colVC))

  # Create Confidence Interval Objects
  ciACE  <- mxCI(c("stCovA","stCovC", "stCovE"))#
"VC[1,seq(1,3*nv,nv),(2,seq(1,3*nv,nv)),(2,seq(2,3*nv,nv)))]" )

  # Data objects for Multiple Groups
  dataMZf   <- mxData( observed=mzfData, type="raw" )
  dataDZf   <- mxData( observed=dzfData, type="raw" )
  dataMZm   <- mxData( observed=mzmData, type="raw" )
  dataDZm   <- mxData( observed=dzmData, type="raw" )
  dataDZo   <- mxData( observed=dzoData, type="raw" )

  # Expectation objects for Multiple Groups
  expMZf    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                                    dimnames=selVars, thresholds = "threGf")
```

```
  expDZf    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                                    dimnames=selVars, thresholds = "threGf" )
  expMZm    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                                    dimnames=selVars, thresholds = "threGm" )
  expDZm    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                                    dimnames=selVars, thresholds = "threGm" )
  expDZo    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                                    dimnames=selVars, thresholds = "threGmf" )
  funML     <- mxFitFunctionML()

  # Combine Groups
  parsZf    <- list( pathA, pathC, pathE,
                 covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                 meanG, threGf, matI, invSD, matUnv )
  parsZm    <- list( pathA, pathC, pathE,
                 covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                 meanG, threGm, matI, invSD, matUnv )
  parsZmf   <- list( pathA, pathC, pathE,
                 covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                 meanG, threGmf, matI, invSD, matUnv )

  modelMZf  <- mxModel( parsZf, meanG, covMZ, expCovMZ, dataMZf, expMZf, funML,
name="MZf" )
  modelDZf  <- mxModel( parsZf, meanG, covDZ, expCovDZ, dataDZf, expDZf, funML,
name="DZf" )
  modelMZm  <- mxModel( parsZm, meanG, covMZ, expCovMZ, dataMZm, expMZm,
funML, name="MZm" )
  modelDZm  <- mxModel( parsZm, meanG, covDZ, expCovDZ, dataDZm, expDZm,
funML, name="DZm" )
  modelDZo  <- mxModel( parsZmf, meanG, covDZ, expCovDZ,dataDZo, expDZo,
funML, name="DZo" )
  multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
  CcUniAceModel  <- mxModel( "CcUniAce", parsZf, parsZm, parsZmf,
                       modelMZf, modelDZf, modelMZm, modelDZm, modelDZo, multi,
                       estVC, ciACE)

  CcUniAceFit <-mxRun(CcUniAceModel, intervals = F)
  CcUniAceFit <-mxTryHardOrdinal(CcUniAceFit, intervals = F)
  CcUniAceFit$algebras


  # Confidence Intervals

  CcUniAceFitCIs <- mxTryHardOrdinal(CcUniAceFit, intervals = T)
  summary(CcUniAceFitCIs, verbose=T)
  #CcUniAceFitCIs$algebras
```

```
   # Test of A
  CcUniNoA <- CcUniAceFit
  CcUniNoAModel<- omxSetParameters(CcUniNoA, labels=c( "a11"), free=FALSE,
values=0 )
  CcUniNoAfit<- mxTryHardOrdinal(CcUniNoAModel, intervals = T)
  tableFitStatistics(CcUniAceFit, CcUniNoAfit)


  # Test of C
  CcUniNoC <- CcUniAceFit
  CcUniNoCModel<- omxSetParameters(CcUniNoC, labels=c( "c11"), free=FALSE,
values=0 )
  CcUniNoCFit<- mxTryHardOrdinal(CcUniNoCModel, intervals = F)
  tableFitStatistics(CcUniAceFit, CcUniNoCFit)


  # Test of E Only Model
  CcUniNoE <- CcUniAceFit
  CcUniNoEModel<- omxSetParameters(CcUniNoE, labels=c("a11","c11"), free=FALSE,
values=0 )
  CcUniNoEModelFit<- mxTryHardOrdinal(CcUniNoEModel, intervals = F)
  tableFitStatistics(CcUniAceFit, CcUniNoEModelFit)

  CCallmodels<-list(CcUniNoAfit,CcUniNoCFit,CcUniNoEModelFit)
tableFitStatistics(CcUniAceFit, CCallmodels)




###-------------------------------------------------###
###            Testing model assumptions for CC      ###
###-------------------------------------------------###

# set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
Vars      <- c("cccigEver3")
nv        <- 1          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
nth <- 1  # Number of thresholds per variable (only for ordinal data)

mzData_1 <- subset(twinDataBin, zyg2 %in% c(1,2), select = selVars)
dzData_1 <- subset(twinDataBin, zyg2 %in% c(3,4,5), select = selVars)
# Set Starting Values
svLTh     <- -1.5    # start value for first threshold
```

```
svlTh     <- 1       # start value for increments
svTh      <- matrix(rep(c(svLTh,(rep(svlTh,nth-1)))),nrow=nth,ncol=nv)     # start value for
thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv)     # lower bounds
for thresholds
svCor     <- .5      # start value for correlations
lbCor     <- -0.99   # lower bounds for correlations
ubCor     <- 0.99    # upper bounds for correlations

labThMZ   <- c(paste("t",1:nth,"MZ1",sep=""),paste("t",1:nth,"MZ2",sep=""))
labThDZ   <- c(paste("t",1:nth,"DZ1",sep=""),paste("t",1:nth,"DZ2",sep=""))

# ------------------------------------------------------------------------------
# PREPARE MODEL

# Saturated Model
# Algebra for expected Mean & Threshold Matrices in MZ & DZ twins
meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
thinMZ    <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=TRUE, values=svTh,
lbound=lbTh, labels=labThMZ, name="thinMZ" )
thinDZ    <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=TRUE, values=svTh,
lbound=lbTh, labels=labThDZ, name="thinDZ" )
inc       <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=FALSE, values=1,
name="inc" )
threMZ    <- mxAlgebra( expression= inc %*% thinMZ, name="threMZ" )
threDZ    <- mxAlgebra( expression= inc %*% thinDZ, name="threDZ" )

# Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
corMZ     <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels="rMZ", name="corMZ" )
corDZ     <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels="rDZ", name="corDZ" )

# Data objects for Multiple Groups
dataMZ   <-mxData( observed=mzData_1, type="raw" )
dataDZ   <-mxData( observed=dzData_1, type="raw" )

# Objective objects for Multiple Groups
expMZ     <- mxExpectationNormal( covariance="corMZ", means="meanG",
dimnames=selVars, thresholds="threMZ" )
expDZ     <- mxExpectationNormal( covariance="corDZ", means="meanG",
dimnames=selVars, thresholds="threDZ" )
funML     <- mxFitFunctionML()

# Combine Groups
```

```
modelMZ   <- mxModel( "MZ", meanG, corMZ, thinMZ, inc, threMZ, dataMZ, expMZ,
funML )
modelDZ   <- mxModel( "DZ", meanG, corDZ, thinDZ, inc, threDZ, dataDZ, expDZ,
funML )
multi     <- mxFitFunctionMultigroup( c("MZ","DZ") )
ciCor     <- mxCI( c('MZ.corMZ','DZ.corDZ' ))
ciThre    <- mxCI( c('MZ.threMZ','DZ.threDZ' ))
twinSatOrdModel   <- mxModel( "CC Cigs", modelMZ, modelDZ, multi, ciCor, ciThre )

# ------------------------------------------------------------------------------
# RUN MODEL

# Run Saturated Model
twinSatOrdFit     <- mxRun( twinSatOrdModel, intervals=T )
twinSatOrdSum     <- summary( twinSatOrdFit )
twinSatOrdSum
round(twinSatOrdFit$output$estimate,4)

# Generate Saturated Model Output
rMZ       <- twinSatOrdFit$MZ.corMZ$values[2,1]
rDZ       <- twinSatOrdFit$DZ.corDZ$values[2,1]
tMZ       <- twinSatOrdFit$MZ.threMZ$result
tDZ       <- twinSatOrdFit$DZ.threDZ$result

twinSatOrdOS      <- twinSatOrdSum$observedStatistics
twinSatOrdDF      <- twinSatOrdSum$degreesOfFreedom
twinSatOrdNP      <- length(twinSatOrdSum$parameters[[1]])
twinSatOrdLLL     <- twinSatOrdFit$output$Minus2LogLikelihood
twinSatOrdAIC     <- twinSatOrdSum$AIC

mxCompare(twinSatOrdFit)

# ------------------------------------------------------------------------------
# RUN SUBMODELS

# Constrain expected Thresholds to be equal across twin order
eqThresTwinModel    <- mxModel(twinSatOrdFit, name="eqThresTwin" )
eqThresTwinModel    <- omxSetParameters( eqThresTwinModel,
label=c("t1MZ1","t1MZ2"), free=TRUE, values=svLTh, newlabels='t1MZ' )

eqThresTwinModel    <- omxSetParameters( eqThresTwinModel,
label=c("t1DZ1","t1DZ2"), free=TRUE, values=svLTh, newlabels='t1DZ' )

eqThresTwinFit      <- mxRun( eqThresTwinModel, intervals=F )
eqThresTwinSum      <- summary( eqThresTwinFit )
eqThresTwinLLL      <- eqThresTwinFit$output$Minus2LogLikelihood
```

```
mxCompare(twinSatOrdFit, eqThresTwinFit)

# Constrain expected Thres to be equal across twin order and zygosity
eqThresZygModel     <- mxModel(eqThresTwinModel, name="eqThresZyg" )
eqThresZygModel     <- omxSetParameters( eqThresZygModel, label=c("t1MZ","t1DZ"),
free=TRUE, values=svLTh, newlabels='t1Z' )


eqThresZygFit       <- mxRun( eqThresZygModel, intervals=F )
eqThresZygSum       <- summary( eqThresZygFit )
eqThresZygLLL       <- eqThresZygFit$output$Minus2LogLikelihood
mxCompare(eqThresTwinFit, eqThresZygFit)

# ------------------------------------------------------------------------------

# Print Comparative Fit Statistics
SatNested <- list(eqThresTwinFit, eqThresZygFit)
mxCompare(twinSatOrdFit, SatNested)

tableFitStatistics(twinSatOrdFit, SatNested)

#####--------------------####
##### Equating Sexes  CC ####
#####--------------------####

# 1=MZM, 2= MZF, 3=DZM, 4=DZF, 5=ODZ

Vars      <- c("cccigEver3")
nv        <- 1          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
# Subset the data to only the things I need
twinDatauni  <- data2[,c(selVars,'zyg2')]
describe(twinDatauni)
summary(twinDatauni)
dim(twinDatauni)

#twinData2<-na.omit(twinData)
#summary(twinData2)
#dim(twinData2)
twinDataBin <-twinDatauni
dim(twinDataBin)
table(twinDataBin$zyg2)

# Factorize Ordinal Variables using the mxFactor option
twinDataBin[,c(1,2)]  <- mxFactor(twinDataBin[,c(1,2)], levels = c(0:nth))
```

```
# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #fm

# Set Starting Values
svLTh     <- -1.5    # start value for first threshold
svITh     <- 1       # start value for increments
svTh      <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=nv)     # start value for
thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv)     # lower bounds
for thresholds
svCor     <- .5      # start value for correlations
lbCor     <- -0.99   # lower bounds for correlations
ubCor     <- 0.99    # upper bounds for correlations

labThMZ   <- c(paste("t",1:nth,"MZ1",sep=""),paste("t",1:nth,"MZ2",sep=""))
labThDZ   <- c(paste("t",1:nth,"DZ1",sep=""),paste("t",1:nth,"DZ2",sep=""))
var1th     <- paste("var1","_th",1:nth, sep="")
var2th     <- paste("var2","_th",1:nth, sep="")
thUB      <- 2

# ------------------------------------------------------------------------------
# PREPARE MODEL

# General non-scalar ACE Model
# Matrices declared to store a, c, and e Path Coefficients
pathAf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label="af11", name="af" )
pathCf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label="cf11", name="cf" )
pathEf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label="ef11", name="ef" )
pathAm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="am11", name="am" )
pathCm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="cm11", name="cm" )
pathEm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label="em11", name="em" )
pathRa    <- mxMatrix( "Lower", nrow=1, ncol=1, free=TRUE,  values=1, label="ra11",
name="ra", ubound=1, lbound=0 )
```

```
# Matrices generated to hold A, C, and E computed Variance Components
covAf     <- mxAlgebra( af %*% t(af), name="Af" )
covCf     <- mxAlgebra( cf %*% t(cf), name="Cf" )
covEf     <- mxAlgebra( ef %*% t(ef), name="Ef" )
covAm     <- mxAlgebra( am %*% t(am), name="Am" )
covCm     <- mxAlgebra( cm %*% t(cm), name="Cm" )
covEm     <- mxAlgebra( em %*% t(em), name="Em" )

# Algebra to compute total variances and standard deviations (diagonal only)
covPf     <- mxAlgebra( Af+Cf+Ef, name="Vf" )
covPm     <- mxAlgebra( Am+Cm+Em, name="Vm" )

# Algebras generated to hold Parameter Estimates and Derived Variance Components
colVarsZf <- c('Af','Cf','Ef','SAf','SCf','SEf')
estVarsZf <- mxAlgebra( cbind(Af,Cf,Ef,Af/Vf,Cf/Vf,Ef/Vf), name="VarsZf",
dimnames=list(NULL,colVarsZf))
colVarsZm <- c('Am','Cm','Em','SAm','SCm','SEm')
estVarsZm <- mxAlgebra( cbind(Am,Cm,Em,Am/Vm,Cm/Vm,Em/Vm), name="VarsZm",
dimnames=list(NULL,colVarsZm))

# Algebra for expected Mean and Variance/Covariance Matrices in MZ & DZ twins
meanGf    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanf", name="expMeanGf" )
meanGm    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanm", name="expMeanGm" )
meanGfm   <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label=c("meanf","meanm"), name="expMeanGfm" )
covMZf    <- mxAlgebra( expression= rbind( cbind(Vf, Af+Cf), cbind(Af+Cf, Vf)),
name="expCovMZf" )
covDZf    <- mxAlgebra( expression= rbind( cbind(Vf, 0.5%x%Af+Cf),
cbind(0.5%x%Af+Cf, Vf)), name="expCovDZf" )
covMZm    <- mxAlgebra( expression= rbind( cbind(Vm, Am+Cm), cbind(Am+Cm, Vm)),
name="expCovMZm" )
covDZm    <- mxAlgebra( expression= rbind( cbind(Vm, 0.5%x%Am+Cm),
cbind(0.5%x%Am+Cm, Vm)), name="expCovDZm" )
CVfm      <- mxAlgebra( expression= ra%x%(af%*%t(am))+cf%*%t(cm), name="CVfm" )
CVmf      <- mxAlgebra( expression= ra%x%(am%*%t(af))+cm%*%t(cf), name="CVmf" )
covDZo    <- mxAlgebra( expression= rbind( cbind(Vf, CVfm), cbind(CVmf, Vm)),
name="expCovDZo" )
Inc     <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=F, values=1, name="Inc" )

# MALES
ThreM    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(mvar1th, mvar2th), lbound=-2, ubound=thUB, name="ThreM")
ExpThreM   <- mxAlgebra( expression= cbind( ( Inc %*% ThreM  ),
```

```
                                ( Inc %*% ThreM  ) ), name="ExpThreM" )
# FEMALES
ThreF    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(fvar1th, fvar2th), lbound=-2, ubound=thUB, name="ThreF")
ExpThreF   <- mxAlgebra( expression= cbind( ( Inc %*% ThreF  ),
                                ( Inc %*% ThreF  ) ), name="ExpThreF" )
## OS

ThreOS    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(dzvar1th, dzvar2th), lbound=-2, ubound=thUB, name="ThreOS")
ExpThreOS   <- mxAlgebra( expression= cbind( ( Inc %*% ThreOS  ),
                                ( Inc %*% ThreOS  ) ), name="ExpThreOS" )


# Data objects for Multiple Groups
dataMZf   <- mxData( observed=mzfData, type="raw" )
dataDZf   <- mxData( observed=dzfData, type="raw" )
dataMZm   <- mxData( observed=mzmData, type="raw" )
dataDZm   <- mxData( observed=dzmData, type="raw" )
dataDZo   <- mxData( observed=dzoData, type="raw" )

# Expectation objects for Multiple Groups
expMZf    <- mxExpectationNormal( covariance="expCovMZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
expDZf    <- mxExpectationNormal( covariance="expCovDZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
expMZm    <- mxExpectationNormal( covariance="expCovMZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
expDZm    <- mxExpectationNormal( covariance="expCovDZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
expDZo    <- mxExpectationNormal( covariance="expCovDZo", means="expMeanGfm",
dimnames=selVars, thresholds = "ExpThreOS" )
funML     <- mxFitFunctionML()

# Combine Groups
parsZf    <- list( pathAf, pathCf, pathEf, covAf, covCf, covEf, covPf, estVarsZf, ThreF,
ExpThreF, Inc )
parsZm    <- list( pathAm, pathCm, pathEm, covAm, covCm, covEm, covPm,
estVarsZm, ThreM, ExpThreM, Inc )
parsZfm   <- list( pathRa, CVfm, CVmf, ExpThreOS, Inc, ThreOS)
modelMZf  <- mxModel( parsZf, meanGf, covMZf, dataMZf, expMZf, funML,
name="MZf" )
modelDZf  <- mxModel( parsZf, meanGf, covDZf, dataDZf, expDZf, funML, name="DZf"
)
modelMZm  <- mxModel( parsZm, meanGm, covMZm, dataMZm, expMZm, funML,
name="MZm" )
```

```
modelDZm  <- mxModel( parsZm, meanGm, covDZm, dataDZm, expDZm, funML,
name="DZm" )
modelDZo  <- mxModel( parsZf, parsZm, parsZfm, meanGfm, covDZo, dataDZo,
expDZo, funML, name="DZo" )
multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
QualAceModel  <- mxModel( "QualACE", modelMZf, modelDZf, modelMZm,
modelDZm, modelDZo, multi )


QualAceFit <-mxTryHardOrdinal(QualAceModel, intervals = F)
summary(QualAceFit)

## Coerce threshold to be equal


eqthres   <-mxModel(QualAceFit, name = "Equal Threshold")
eqthres   <-omxSetParameters( eqthres, label="mvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")
eqthres   <-omxSetParameters( eqthres, label="fvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")
eqthres   <-omxSetParameters( eqthres, label="dzvar1_th1", free=TRUE, values=0.5,
newlabels="var1_th1")

eqthresfit<-mxTryHardOrdinal(eqthres, intervals = F)
summary(eqthresfit)
tableFitStatistics(QualAceFit, eqthresfit)

## Coerce males and females to be equal

eqsex   <-mxModel(QualAceFit, name = "Equal sexes")
eqsex   <-omxSetParameters( eqsex, label="am11", free=TRUE, values=0.04,
newlabels="a11")
eqsex   <-omxSetParameters( eqsex, label="af11", free=TRUE, values=0.04,
newlabels="a11")
eqsex   <-omxSetParameters( eqsex, label="cm11", free=TRUE, values=0.5,
newlabels="c11")
eqsex   <-omxSetParameters( eqsex, label="cf11", free=TRUE, values=0.5,
newlabels="c11")
eqsex   <-omxSetParameters( eqsex, label="em11", free=TRUE, values=0.3,
newlabels="e11")
eqsex   <-omxSetParameters( eqsex, label="ef11", free=TRUE, values=0.3,
newlabels="e11")

eqsexfit<-mxTryHardOrdinal(eqsex, intervals = F)
tableFitStatistics(QualAceFit, eqsexfit)
```

```
## NO sex

nosex   <- omxSetParameters(eqsex, labels="ra11", free=FALSE, values=0.5,
name="No sex Effects" )
nosexfit<- mxTryHardOrdinal(nosex, intervals = F)
nested <-list(eqthresfit,eqsexfit, nosexfit)
tableFitStatistics(QualAceFit, nested)




###-----------------------------------------------------###
###            BIVARIATE ANALYSIS                   ###
###-----------------------------------------------------###

# set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
vars      <- c("ecigEver3", "cccigEver3")
#vars      <- c("cccigEver3", "ecigEver3")   #reverse order of variables to see if same
results emerge
nv        <- 2          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
nth <- 1  # Number of thresholds per variable (only for binary data)

# Subset the data to only the things I need
twinData  <- data2[,c(selVars,'zyg2')]
describe(twinData)
summary(twinData)
dim(twinData)

#twinData2<-na.omit(twinData)
#summary(twinData2)
#dim(twinData2)
twinDataBin <-twinData
dim(twinDataBin)
table(twinDataBin$zyg2)

# Factorize Ordinal Variables using the mxFactor option
twinDataBin[,c(1,3)]  <- mxFactor(twinDataBin[,c(1,3)], levels = c(0:nth))
twinDataBin[,c(2,4)]  <- mxFactor(twinDataBin[,c(2,4)], levels = c(0:nth))


# Twin correlations

mzdat <- subset(twinDataBin, zyg2==c(1) | zyg2 ==2, selVars)
```

```
dzdat <- subset(twinDataBin, zyg2==c(3) | zyg2 ==4 | zyg2==5, selVars)

polychor(mzdat$ecEver3_T1, mzdat$ecEver3_T2,std.err = T)
.65+(1.96*.12)
.65-(1.96*.12)
polychor(mzdat$cccever3_T1, mzdat$cccever3_T2,std.err = T)
.62+(1.96*.12)
.62-(1.96*.12)

polychor(dzdat$ecEver3_T1, dzdat$ecEver3_T2,std.err = T)
.55+(1.96*.11)
.55-(1.96*.11)
polychor(dzdat$cccever3_T1, dzdat$cccever3_T2,std.err = T)
.52+(1.96*.11)
.52-(1.96*.11)

# 1=MZM, 2= MZF, 3=DZM, 4=DZF, 5=ODZ
#Vars      <- c("cccigEver3")
#nv        <- 1          # number of variables
#ntv       <- nv*2       # number of total variables
#selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")

# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #males = T1, females = T2

# Set Starting Values /
svLTh     <- 0.8    # start value for first threshold
svlTh     <- 1       # start value for increments
#svTh      <- c(0.7,1,0.7,1)
svTh      <- matrix(rep(c(svLTh,(rep(svlTh,nth-1)))),nrow=nth,ncol=ntv)     # start value
for thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=ntv)     # lower bounds
for thresholds

#svTh      <- c(1,1)                  # start value for thresholds
svPa      <- .4                      # start value for path coefficient
svPaD     <- vech(diag(svPa,nv,nv))    # start values for diagonal of covariance matrix
svPe      <- .8                      # start value for path coefficient for e
svPeD     <- vech(diag(svPe,nv,nv))    # start values for diagonal of covariance matrix
lbPa      <- .00001                  # start value for lower bounds
lbPaD     <- diag(lbPa,nv,nv)          # lower bounds for diagonal of covariance matrix
lbPaD[lower.tri(lbPaD)] <- 0         # lower bounds for below diagonal elements
```

```
lbPaD[upper.tri(lbPaD)] <- NA          # lower bounds for above diagonal elements


# Set Starting Values
aLabs     <- paste("a",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
cLabs     <- paste("c",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
eLabs     <- paste("e",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
var1thM     <- paste("var1M","_th",1:nth, sep="")
var2thM     <- paste("var2M","_th",1:nth, sep="")
var1thF     <- paste("var1F","_th",1:nth, sep="")
var2thF     <- paste("var2F","_th",1:nth, sep="")

thUB      <- 2

# ------------------------------------------------------------------------------
# PREPARE MODEL

# ACE Model
# Create Algebra for expected Mean Matrices to include differing thresholds for males
and females
meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
threGm     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values= svTh,
                    labels=c(var1thM,var2thM), name="threGm",lbound=-2, ubound=2 )
threGf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                    labels=c(var1thF,var2thF), name="threGf",lbound=-2, ubound=2  )
threGmf     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                     labels=c(var1thM,var2thM,var1thF,var2thF), name="threGmf", lbound=-
2, ubound=2  )
# Create Matrices for Path Coefficients
pathA     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE,
values=c(0.7,0.5,0), #replacing svPaD
                    label=aLabs, lbound=lbPaD, name="a" )
pathC     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE,
values=c(0.8,0.6,0.8),
                    label=cLabs, lbound=lbPaD, name="c" )
pathE     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE,
                    values=svPeD, label=eLabs, lbound=lbPaD, name="e" )

# Create Algebra for Variance Comptwonts
covA      <- mxAlgebra( expression=a %*% t(a), name="A" )
covC      <- mxAlgebra( expression=c %*% t(c), name="C" )
covE      <- mxAlgebra( expression=e %*% t(e), name="E" )

# Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
covP      <- mxAlgebra( expression= A+C+E, name="V" )
covMZ     <- mxAlgebra( expression= A+C, name="cMZ" )
```

```
covDZ     <- mxAlgebra( expression= 0.5%x%A+ C, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)),
name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)),
name="expCovDZ" )

# Create Algebra for Standardization
matI      <- mxMatrix( type="Iden", nrow=nv, ncol=nv, name="I")
invSD     <- mxAlgebra( expression=solve(sqrt(I*V)), name="iSD")

# Calculate genetic and environmental correlations
corA      <- mxAlgebra( expression=solve(sqrt(I*A))%&%A, name ="rA" ) #cov2cor()
corC      <- mxAlgebra( expression=solve(sqrt(I*C))%&%C, name ="rC" )
corE      <- mxAlgebra( expression=solve(sqrt(I*E))%&%E, name ="rE" )

## Calculate Phenotypic Correlation ##
corP      <- mxAlgebra (expression=solve(sqrt(I*V)) %*% V %*% solve(sqrt(I*V)),
name="rP")

## Calculate Standardized Covariances ##
stCovA    <- mxAlgebra (solve(sqrt(I*V)) %*% A %*% solve(sqrt(I*V)), name="stCovA")
stCovC    <- mxAlgebra (solve(sqrt(I*V)) %*% C %*% solve(sqrt(I*V)), name="stCovC")
stCovE    <- mxAlgebra (solve(sqrt(I*V)) %*% E %*% solve(sqrt(I*V)), name="stCovE")


# Constrain Variance of Binary Variables
matUnv    <- mxMatrix( type="Unit", nrow=nv, ncol=1, name="Unv1" )
var1      <- mxConstraint( expression=diag2vec(V)==Unv1, name="Var1" )

# Create Algebra for Variance Components
rowVC     <- rep('VC',nv)
colVC     <- rep(c('A','C','E','SA','SC','SE'),each=nv)
estVC     <- mxAlgebra( expression=cbind(A,C,E,A/V,C/V,E/V), name="VC",
dimnames=list(rowVC,colVC))

# Create Confidence Interval Objects
ciACE  <- mxCI(c("rA", "rC", "rE", "stCovA","stCovC", "stCovE"))#
"VC[1,seq(1,3*nv,nv),(2,seq(1,3*nv,nv)),(2,seq(2,3*nv,nv)))]" )

# Data objects for Multiple Groups
dataMZf   <- mxData( observed=mzfData, type="raw" )
dataDZf   <- mxData( observed=dzfData, type="raw" )
dataMZm   <- mxData( observed=mzmData, type="raw" )
dataDZm   <- mxData( observed=dzmData, type="raw" )
dataDZo   <- mxData( observed=dzoData, type="raw" )
```

```
# Expectation objects for Multiple Groups
expMZf    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf")
expDZf    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf" )
expMZm    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZm    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZo    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGmf" )
funML     <- mxFitFunctionML()

# Combine Groups
parsZf    <- list( pathA, pathC, pathE,
               covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
               meanG, threGf, matI, invSD, matUnv )
parsZm    <- list( pathA, pathC, pathE,
               covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
               meanG, threGm, matI, invSD, matUnv )
parsZmf   <- list( pathA, pathC, pathE,
               covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
               meanG, threGmf, matI, invSD, matUnv )

modelMZf  <- mxModel( parsZf, meanG, covMZ, expCovMZ, dataMZf, expMZf, funML,
name="MZf" )
modelDZf  <- mxModel( parsZf, meanG, covDZ, expCovDZ, dataDZf, expDZf, funML,
name="DZf" )
modelMZm  <- mxModel( parsZm, meanG, covMZ, expCovMZ, dataMZm, expMZm,
funML, name="MZm" )
modelDZm  <- mxModel( parsZm, meanG, covDZ, expCovDZ, dataDZm, expDZm,
funML, name="DZm" )
modelDZo  <- mxModel( parsZmf, meanG, covDZ, expCovDZ,dataDZo, expDZo,
funML, name="DZo" )
multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
BivBinAceModel  <- mxModel( "BivBinACE", parsZf, parsZm, parsZmf,
                    modelMZf, modelDZf, modelMZm, modelDZm, modelDZo, multi,
                    estVC, ciACE)

BivBinAceFit <-mxRun(BivBinAceModel, intervals = F)
BivBinAceFit <-mxTryHardOrdinal(BivBinAceFit, intervals = F)
BivBinAceFit2 <- mxTryHardOrdinal(BivBinAceFit, intervals = F)
BivBinAceFit3 <- mxTryHardOrdinal(BivBinAceFit2, intervals = F)
BivBinAceFit4 <- mxTryHardOrdinal(BivBinAceFit3, intervals = F)
BivBinAceFit5 <- mxTryHardOrdinal(BivBinAceFit4, intervals = F)
BivBinAceFit6 <- mxTryHardOrdinal(BivBinAceFit, intervals = F)
```

```
summary(BivBinAceFit6, verbose = T)
BivBinAceFit6$algebras

#BivBinAceFit7 <- mxBootstrap(BivBinAceFit6)
#summary(BivBinAceFit7)
#BivBinAceFit7b <- mxBootstrap(BivBinAceFit7, replications = 1000)
#summary(BivBinAceFit7b)
#BivBinAceFit7b$algebras

#BivBinAceFit7c <- mxBootstrap(BivBinAceFit7b, replications = 1500)
#summary(BivBinAceFit7c)
#BivBinAceFit7c$algebras

#BivBinAceFit7d <- mxBootstrap(BivBinAceFit7c, replications = 2000)
#summary(BivBinAceFit7d)
#BivBinAceFit7d$algebras
# Confidence Interval calculation below

# Test of covA
BivBinAceModel8 <- BivBinAceFit6
BivBinAceModel8<- omxSetParameters(BivBinAceModel8, labels=c( "a21"),
free=FALSE, values=0 )
BivBinAceFit8<- mxTryHardOrdinal(BivBinAceModel8, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit8)

#BivBinAceFit8a <- omxRunCI(BivBinAceFit8)
summary(BivBinAceFit8, verbose=F)
#BivBinAceFit8$algebras


# Test of covC
BivBinAceModel9 <- BivBinAceFit6
BivBinAceModel9<- omxSetParameters(BivBinAceModel9, labels=c( "c21"),
free=FALSE, values=0 )
BivBinAceFit9<- mxRun(BivBinAceModel9, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit9)


# Test of covE
BivBinAceModel10 <- BivBinAceFit6
BivBinAceModel10<- omxSetParameters(BivBinAceModel10, labels=c("e21"),
free=FALSE, values=0 )
BivBinAceFit10<- mxTryHardOrdinal(BivBinAceModel10, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit10)

# Test of rP
```

```
BivBinAceModel11 <- BivBinAceFit6
BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="a21",
free=FALSE, values=0 )
BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="c21",
free=FALSE, values=0 )
#BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="e21",
free=FALSE, values=0 )
BivBinAceFit11<- mxTryHardOrdinal(BivBinAceModel11, intervals = F)

tableFitStatistics(BivBinAceFit6, c(BivBinAceFit8, BivBinAceFit9, BivBinAceFit10,
BivBinAceFit11))

# Test of covA/A
#BivBinAceModel12 <- BivBinAceFit7a
#BivBinAceModel12<- omxSetParameters(BivBinAceModel12, labels="a22",
free=FALSE, values=0 )
#BivBinAceFit12<- mxTryHardOrdinal(BivBinAceModel12, intervals = F)


# Confidence Intervals
mxOption(NULL, "Default optimizer", "CSOLNP")
BivBinAceFit7a <- omxRunCI(BivBinAceFit7d)
summary(BivBinAceFit7a, verbose=T)
BivBinAceFit7a$algebras


# Test of CE model

BivBinAceModel12 <- BivBinAceFit6
BivBinAceModel12<- omxSetParameters(BivBinAceModel12, labels=c( "a11","a21",
"a22"), free=FALSE, values=0, name = "CE model" )
BivBinAceFit12<- mxTryHardOrdinal(BivBinAceModel12, intervals = F)
summary(BivBinAceFit12, verbose=T)
tableFitStatistics(BivBinAceFit6, BivBinAceFit12)

# Test of AE Model
BivBinAceModel13 <- BivBinAceFit6
BivBinAceModel13<- omxSetParameters(BivBinAceModel13, labels=c( "c11","c21",
"c22"), free=FALSE, values=0, name = "AE model" )
BivBinAceFit13<- mxTryHardOrdinal(BivBinAceModel13, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit13)

# Test of E Model
BivBinAceModel16 <- BivBinAceFit6
```

```
BivBinAceModel16<- omxSetParameters(BivBinAceModel16,
labels=c("a11","a21","a22", "c11","c21", "c22"), free=FALSE, values=0, name = "E
model" )
BivBinAceFit16<- mxTryHardOrdinal(BivBinAceModel16, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit16)

bivnested <-list(BivBinAceFit12, BivBinAceFit13,BivBinAceFit16)
tableFitStatistics(BivBinAceFit6, bivnested)



# Test of E crosspaths Model
BivBinAceModel17 <- BivBinAceFit6
BivBinAceModel17<- omxSetParameters(BivBinAceModel17, labels=c("a21","c21"),
free=FALSE, values=0, name = "E model" )
BivBinAceFit17<- mxTryHardOrdinal(BivBinAceModel17, intervals = T)
summary(BivBinAceFit17, verbose=T)
tableFitStatistics(BivBinAceFit6, BivBinAceFit17)

BivBinAceModel18 <- BivBinAceFit17
BivBinAceFit18<- mxTryHardOrdinal(BivBinAceModel18, intervals = T)
summary(BivBinAceFit18, verbose=T)

BivBinAceModel19 <- BivBinAceFit18
BivBinAceFit19<- mxTryHardOrdinal(BivBinAceModel19, intervals = T)
summary(BivBinAceFit19, verbose=T)


# Using CE model, testing of C21

BivBinAceModel14 <- BivBinAceModel12
BivBinAceModel14<- omxSetParameters(BivBinAceModel14, labels=c( "c21"),
free=FALSE, values=0, name = "CE model No C21" )
BivBinAceFit14<- mxTryHardOrdinal(BivBinAceModel14, intervals = F)
tableFitStatistics(BivBinAceFit12, BivBinAceFit14)

# Using CE model, testing of E21

BivBinAceModel15 <- BivBinAceModel14
BivBinAceModel15<- omxSetParameters(BivBinAceModel15, labels=c( "e21"),
free=FALSE, values=0, name = "CE model No Cross Paths" )
BivBinAceFit15<- mxTryHardOrdinal(BivBinAceModel15, intervals = F)
tableFitStatistics(BivBinAceFit14, BivBinAceFit15)
tableFitStatistics(BivBinAceFit12, BivBinAceFit15)
```

```
###-----------------------------------------###
###            Correlated Factors Model     ###
###-----------------------------------------###

mxOption( NULL, "Default optimizer","CSOLNP" )

# Create Functions to assign labels
laLower   <- function(la,nv) { paste(la,rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="_")
}
laSdiag   <- function(la,nv) { paste(la,rev(nv+1-sequence(1:(nv-1))),rep(1:(nv-1),(nv-
1):1),sep="_") }
laFull    <- function(la,nv) { paste(la,1:nv,rep(1:nv,each=nv),sep="_") }
laDiag    <- function(la,nv) { paste(la,1:nv,1:nv,sep="_") }
laSymm    <- function(la,nv) { paste(la,rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="_")
}

# Create Algebra for expected Mean Matrices to include differing thresholds for males
and females
meanG     <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
threGm    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values= svTh,
                  labels=c(var1thM,var2thM), name="threGm",lbound=-2, ubound=2 )
threGf    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                  labels=c(var1thF,var2thF), name="threGf",lbound=-2, ubound=2  )
threGmf   <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svTh,
                   labels=c(var1thM,var2thM,var1thF,var2thF), name="threGmf", lbound=-
2, ubound=2  )

# Matrices a, c, and e to store a, c, and e path coefficients
pathA   <- mxMatrix( type="Diag", nrow=nv, ncol=nv, free=TRUE, values=.5,
label=laDiag("a",nv), lbound=.0001, name="a" )
pathC   <- mxMatrix( type="Diag", nrow=nv, ncol=nv, free=TRUE, values=.5,
label=laDiag("c",nv), lbound=.0001, name="c" )
pathE   <- mxMatrix( type="Diag", nrow=nv, ncol=nv, free=TRUE, values=.5,
label=laDiag("e",nv), lbound=.0001, name="e" )


pathRa   <- mxMatrix( type="Stand", nrow=nv, ncol=nv, free=TRUE, values=.4,
label=laSdiag("ra",nv), lbound=-1, ubound=1, name="Ra" )
pathRc   <- mxMatrix( type="Stand", nrow=nv, ncol=nv, free=TRUE, values=.4,
label=laSdiag("rc",nv), lbound=-1, ubound=1, name="Rc" )
pathRe  <- mxMatrix( type="Stand", nrow=nv, ncol=nv, free=TRUE, values=.4,
label=laSdiag("re",nv), lbound=-1, ubound=1, name="Re" )

# Matrices A, C, and E compute variance components

covA    <- mxAlgebra( expression=a %*% (Ra) %*% t(a), name="A" )
```

```
covC     <- mxAlgebra( expression=c %*% (Rc) %*% t(c), name="C" )
covE     <- mxAlgebra( expression=e %*% (Re) %*% t(e), name="E" )

# Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
covP      <- mxAlgebra( expression= A+C+E, name="V" )
covMZ     <- mxAlgebra( expression= A+C, name="cMZ" )
covDZ     <- mxAlgebra( expression= 0.5%x%A+ C, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)),
name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)),
name="expCovDZ" )

matI      <- mxMatrix( type="Iden", nrow=nv, ncol=nv, name="I" )
invSDm    <- mxAlgebra( expression=solve(sqrt(I*V)), name="iSD" )

## Calculate Standardized Covariances ##
stCovA    <- mxAlgebra (solve(sqrt(I*V)) %*% A %*% solve(sqrt(I*V)), name="stCovA")
stCovC    <- mxAlgebra (solve(sqrt(I*V)) %*% C %*% solve(sqrt(I*V)), name="stCovC")
stCovE    <- mxAlgebra (solve(sqrt(I*V)) %*% E %*% solve(sqrt(I*V)), name="stCovE")


# Data objects for Multiple Groups
dataMZf   <- mxData( observed=mzfData, type="raw" )
dataDZf   <- mxData( observed=dzfData, type="raw" )
dataMZm   <- mxData( observed=mzmData, type="raw" )
dataDZm   <- mxData( observed=dzmData, type="raw" )
dataDZo   <- mxData( observed=dzoData, type="raw" )

# Expectation objects for Multiple Groups
expMZf    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf")
expDZf    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGf" )
expMZm    <- mxExpectationNormal( covariance="expCovMZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZm    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGm" )
expDZo    <- mxExpectationNormal( covariance="expCovDZ", means="meanG",
                          dimnames=selVars, thresholds = "threGmf" )
funML     <- mxFitFunctionML()


# Algebras generated to hold Parameter Estimates and Derived Variance Components
colVarsZ <- paste(selVars,rep(c('A','C','E','SA','SC','SE'),each=nv),sep="")
estVarsZ <- mxAlgebra( cbind(A,C,E,A/V,C/V,E/V), name="VarsZ",
dimnames=list(NULL,colVarsZ))
```

```
# Combine Groups
makeModel  <- function(name) {
  parsZf    <- list( pathA, pathC, pathE, pathRa, pathRc, pathRe, covA, covC, covE,
covP, estVarsZ )
  parsZm    <- list( pathA, pathC, pathE, pathRa, pathRc, pathRe, covA, covC, covE,
covP, estVarsZ )
  modelMZf  <- mxModel( parsZf ,meanG, covMZ, dataMZf, expCovMZ, funML,
name="MZf" )
  modelDZf  <- mxModel( parsZf, meanG, covDZ, dataDZf, expCovDZ, funML,
name="DZf" )
  modelMZm  <- mxModel( parsZm, meanG, covMZ, dataMZm, expCovMZ, funML,
name="MZm" )
  modelDZm  <- mxModel( parsZm, meanG, covDZ, dataDZf, expCovDZ, funML,
name="DZm" )
  modelDZo  <- mxModel( parsZmf, meanG, covDZ, expCovDZ,dataDZo, expCovDZ,
funML, name="DZo" )
  minus2ll  <- mxAlgebra( MZf.objective+ DZf.objective+ MZm.objective+ DZm.objective,
name="m2LL" )
  name      <- mxModel( name, parsZf, parsZm, modelMZf, modelDZf, modelMZm,
modelDZm, minus2ll)

}
corrfactormodel <- makeModel("Correlated Factor Model")



# Combine Groups
parsZf    <- list( pathA, pathC, pathE, pathRa, pathRc, pathRe,
                   covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                   meanG, threGf, matI, invSD, matUnv )
parsZm    <- list( pathA, pathC, pathE, pathRa, pathRc, pathRe,
                   covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                   meanG, threGm, matI, invSD, matUnv )
parsZmf   <- list( pathA, pathC, pathE, pathRa, pathRc, pathRe,
                   covA, covC, covE, covP, corA, corC, corE, corP, stCovA, stCovC, stCovE,
                   meanG, threGmf, matI, invSD, matUnv )

modelMZf  <- mxModel( parsZf, meanG, covMZ, expCovMZ, dataMZf, expMZf, funML,
name="MZf" )
modelDZf  <- mxModel( parsZf, meanG, covDZ, expCovDZ, dataDZf, expDZf, funML,
name="DZf" )
modelMZm  <- mxModel( parsZm, meanG, covMZ, expCovMZ, dataMZm, expMZm,
funML, name="MZm" )
modelDZm  <- mxModel( parsZm, meanG, covDZ, expCovDZ, dataDZm, expDZm,
funML, name="DZm" )
```

```
modelDZo  <- mxModel( parsZmf, meanG, covDZ, expCovDZ,dataDZo, expDZo,
funML, name="DZo" )
multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
corrfactormodel  <- mxModel( "Correlated Factors", parsZf, parsZm, parsZmf,
                    modelMZf, modelDZf, modelMZm, modelDZm, modelDZo, multi,
                    estVC, ciACE)

# ------------------------------------------------------------------------------
# RUN MODEL

#
corrfactorfit   <- mxRun(corrfactormodel)
summary(corrfactorfit)
corrfactorfit$algebras

corrfactorfit2<-mxTryHardOrdinal(corrfactorfit)
summary(corrfactorfit2)
corrfactorfit2$algebras

corrfactornoc <-corrfactorfit
corrfactornoc <-omxSetParameters(corrfactornoc, labels = "rc_2_1", free=F, values=0)
corrfactornocfit <-mxTryHardOrdinal(corrfactornoc, intervals = F)
tableFitStatistics(corrfactorfit2, corrfactornocfit)

corrfactornoa <-corrfactorfit
corrfactornoa <-omxSetParameters(corrfactornoa, labels = "ra_2_1", free=F, values=0)
corrfactornoafit <-mxTryHardOrdinal(corrfactornoa, intervals = F)
tableFitStatistics(corrfactorfit2, corrfactornoafit)


# Confidence Intervals

corrfactorfit3 <- omxRunCI(corrfactorfit2)
summary(corrfactorfit3, verbose=T)
corrfactorfit3$algebras



corrfactorfit4 <- omxRunCI(corrfactorfit3)
summary(corrfactorfit4, verbose=T)



###-------------------------------------------------------###
###                Bivariate Model Assumptions          ###
###-------------------------------------------------------###
```

```
# ------------------------------------------------------------------------------------------------------------
----

# PREPARE DATA

# Select Variables for Analysis

Vars      <- c('ecigEver3', 'cccigEver3')
nv        <- 2          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")

# Specify Thresholds for Ordinal Variables
## nth: number of thresholds; fcat: first category; lcat: last category; ncat: number of
categories;
nth1      <- 1 ;  fcat1     <- 0 ;  lcat1     <- fcat1+nth1 ;  ncat1     <- nth1+1
nth       <- max(nth1)

# Specify Arguments for Threshold Matrices
## lth: lowest threshold; ith: increment;
lth1      <- 0 ;  ith1      <-  0 ;
lth2      <- 0 ;  ith2      <-  0 ;
thFree    <- c(rep(T,nth1),rep(F,nth-nth1))

#thValues  <- matrix(rep(c(lth1,(rep(ith1,nth-1)),lth2,(rep(ith2,nth-1)),lth3,(rep(ith3,nth-
1))),nv),nrow=nth,ncol=nv)
thValues  <- matrix(c(lth1,(rep(ith1,nth-1)),lth2,(rep(ith2,nth-1))),nrow=nth,ncol=nv)
thLBound  <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv)
#thLBound  <- matrix(c(-3,0, -3,0, -3,0),nrow=nth,ncol=nv)


# Select Data for Analysis
twinData  <- data2[,c(selVars,'zyg2')]
#twinData  <- FTall[,c(selVars,'zygroup5')]
describe(twinData)
twinDataBin <- twinData

# Factorize Ordinal Variables
twinDataBin[,c(1,nv+1)]  <- mxFactor(twinDataBin[,c(1,nv+1)], levels = c(0:nth1))
twinDataBin[,c(2,nv+2)]  <- mxFactor(twinDataBin[,c(2,nv+2)], levels = c(0:nth1))
#twinDataBin[,c(3,nv+3)]  <- mxFactor(twinDataBin[,c(3,nv+3)], levels = c(0:nth3))

# Create Datasets by Zygosity- 5 group
dataBinMZm  <- subset(twinDataBin, zyg2==1, selVars)
dataBinMZf  <- subset(twinDataBin, zyg2==2, selVars)
```

```
dataBinDZm  <- subset(twinDataBin, zyg2==3, selVars)
dataBinDZf  <- subset(twinDataBin, zyg2==4, selVars)
dataBinDZo  <- subset(twinDataBin, zyg2==5, selVars)

# Set Starting Values
svLTh <- -1.5 # start value for first threshold
svITh <- 1 # start value for increments
svTh <- matrix(rep(c(svLTh,(rep(svITh,nth-1)))),nrow=nth,ncol=nv) # start value for
thresholds
lbTh <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=nv) # lower bounds for
thresholds
svCor <- .5 # start value for correlations
lbCor <- -0.99 # lower bound for correlations
ubCor <- 0.99 # upper bound for correlations

# Create Labels
labThMZM <- labTh("MZM",selVars,nth)
labThDZM <- labTh("DZM",selVars,nth)
labThMZF <- labTh("MZF",selVars,nth)
labThDZF <- labTh("DZF",selVars,nth)
labThDZO <- labTh("DZO",selVars,nth)

labThZ <- labTh("Z",selVars,nth)
labCrMZM <- labSdiag("corMZM",ntv)
labCrDZM <- labSdiag("corDZM",ntv)
labCrMZF <- labSdiag("corMZF",ntv)
labCrDZF <- labSdiag("corDZF",ntv)
labCrDZO <- labSdiag("corDZO",ntv)
labCrZ <- labSdiag("corZ",ntv)

# ------------------------------------------------------------------------------------------------------------
----
# PREPARE MODEL
# Create Algebra for expected Mean & Threshold Matrices
meanG <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="meanG" )
thinMZM <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=thFree, values=svTh,
lbound=lbTh, labels=labThMZM, name="thinMZM" )
thinDZM <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=thFree, values=svTh,
lbound=lbTh, labels=labThDZM, name="thinDZM" )
thinMZF <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=thFree, values=svTh,
lbound=lbTh, labels=labThMZF, name="thinMZF" )
thinDZF <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=thFree, values=svTh,
lbound=lbTh, labels=labThDZF, name="thinDZF" )
thinDZO <- mxMatrix( type="Full", nrow=nth, ncol=ntv, free=thFree, values=svTh,
lbound=lbTh, labels=labThDZO, name="thinDZO" )
```

```
inc <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=FALSE, values=1, name="inc"
)
threMZM <- mxAlgebra( expression= inc %*% thinMZM, name="threMZM" )
threDZM <- mxAlgebra( expression= inc %*% thinDZM, name="threDZM" )
threMZF <- mxAlgebra( expression= inc %*% thinMZF, name="threMZF" )
threDZF <- mxAlgebra( expression= inc %*% thinDZF, name="threDZF" )
threDZO <- mxAlgebra( expression= inc %*% thinDZO, name="threDZO" )

# Create Algebra for expected Correlation Matrices
corMZM <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels=labCrMZM,
            name="corMZM" )
corDZM <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels=labCrDZM,
            name="corDZM" )
corMZF <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels=labCrMZF,
            name="corMZF" )
corDZF <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels=labCrDZF,
            name="corDZF" )
corDZO <- mxMatrix( type="Stand", nrow=ntv, ncol=ntv, free=TRUE, values=svCor,
lbound=lbCor, ubound=ubCor, labels=labCrDZO,
            name="corDZO" )

# Create Data Objects for Multiple Groups
dataMZM <- mxData( observed=dataBinMZm, type="raw" )
dataDZM <- mxData( observed=dataBinDZm, type="raw" )
dataMZF <- mxData( observed=dataBinMZf, type="raw" )
dataDZF <- mxData( observed=dataBinDZf, type="raw" )
dataDZO <- mxData( observed=dataBinDZo, type="raw" )

# Create Expectation Objects for Multiple Groups
# Note- Means are set to zero and the thresholds change.  So,everyone can have the
same meanG but the thresholds vary.

expMZM <- mxExpectationNormal( covariance="corMZM", means="meanG",
dimnames=selVars, thresholds="threMZM" )
expDZM <- mxExpectationNormal( covariance="corDZM", means="meanG",
dimnames=selVars, thresholds="threDZM" )
expMZF <- mxExpectationNormal( covariance="corMZF", means="meanG",
dimnames=selVars, thresholds="threMZF" )
expDZF <- mxExpectationNormal( covariance="corDZF", means="meanG",
dimnames=selVars, thresholds="threDZF" )
expDZO <- mxExpectationNormal( covariance="corDZO", means="meanG",
dimnames=selVars, thresholds="threDZO" )
```

```
funML <- mxFitFunctionML()

# Create Model Objects for Multiple Groups
modelMZM <- mxModel( meanG, corMZM, thinMZM, inc, threMZM, dataMZM, expMZM,
funML, name="MZM" )
modelDZM <- mxModel( meanG, corDZM, thinDZM, inc, threDZM, dataDZM, expDZM,
funML, name="DZM" )
modelMZF <- mxModel( meanG, corMZF, thinMZF, inc, threMZF, dataMZF, expMZF,
funML, name="MZF" )
modelDZF <- mxModel( meanG, corDZF, thinDZF, inc, threDZF, dataDZF, expDZF,
funML, name="DZF" )
modelDZO <- mxModel( meanG, corDZO, thinDZO, inc, threDZO, dataDZO, expDZO,
funML, name="DZO" )

multi <- mxFitFunctionMultigroup( c("MZM", "DZM", "MZF", "DZF", "DZO") )

# Create Confidence Interval Objects
ciCor <- mxCI( c('MZM.corMZM','DZM.corDZM', 'MZF.corMZF','DZF.corDZF',
'DZO.corDZO'  ))
#ciThre <- mxCI( c('MZ.threMZ','DZ.threDZ' ))

# Build Saturated Model with Confidence Intervals
modelSAT <- mxModel( "BivSAT", modelMZM, modelDZM, modelMZF, modelDZF,
modelDZO, multi, ciCor )
modelSAT <- mxOption(modelSAT, "mvnRelEps", 1e-3)
# ----------------------------------------------------------------------------------------------------------
----
# RUN MODEL
# Run Saturated Model
mxOption( NULL, "Default optimizer","CSOLNP" )

fitSAT <- mxRun( modelSAT, intervals=FALSE)
fitSAT <- mxTryHardOrdinal(fitSAT, intervals=FALSE, scale=0.5 )
sumSAT <- summary( fitSAT )
# Print Goodness-of-fit Statistics & Parameter Estimates
fitGofs(fitSAT)
fitEsts(fitSAT)
mxGetExpected( fitSAT, c("thresholds","covariance"))
options("max.print"=1100)
summary(fitSAT, verbose =TRUE)
# ----------------------------------------------------------------------------------------------------------
----
# RUN SUBMODELS
# Constrain expected Thresholds
# to be equal across Twin Order
```

```
modelETO <- mxModel( fitSAT, name="threeETOo" )

## CIGS
  modelETO <- omxSetParameters( modelETO,
label=c("t1MZMcccigEver3_T2","t1MZMcccigEver3_T1"), free=T, values=svLTh,
newlabels='t1MZMcccigEver3', strict=F )
  modelETO <- omxSetParameters( modelETO,
label=c("t1DZMcccigEver3_T2","t1DZMcccigEver3_T1"), free=T, values=svLTh,
newlabels="t1DZMcccigEver3" , strict=F)
  modelETO <- omxSetParameters( modelETO,
label=c("t1MZFcccigEver3_T2","t1MZFcccigEver3_T1"), free=T, values=svLTh,
newlabels="t1MZFcccigEver3" , strict=F)
  modelETO <- omxSetParameters( modelETO,
label=c("t1DZFcccigEver3_T2","t1DZFcccigEver3_T1"), free=T, values=svLTh,
newlabels="t1DZFcccigEver3", strict=F )
  modelETO <- omxSetParameters( modelETO,
label=c("t1DZOcccigEver3_T2","t1DZOcccigEver3_T1"), free=T, values=svLTh,
newlabels="t1DZOcccigEver3", strict=F )

## ECIGS
  modelETO <- omxSetParameters( modelETO, label=c("t1MZMecigEver3_T1",
"t1MZMecigEver3_T2"), free=T, values=svlTh, newlabels="t1MZMecigEver3", strict=F )
  modelETO <- omxSetParameters( modelETO, label=c("t1DZMecigEver3_T1",
"t1DZMecigEver3_T2"), free=T, values=svlTh, newlabels="t1DZMecigEver3", strict=F )
  modelETO <- omxSetParameters( modelETO, label=c("t1MZFecigEver3_T1",
"t1MZFecigEver3_T2"), free=T, values=svlTh, newlabels="t1MZFecigEver3", strict=F )
  modelETO <- omxSetParameters( modelETO, label=c("t1DZFecigEver3_T1",
"t1DZFecigEver3_T2"), free=T, values=svlTh, newlabels="t1DZFecigEver3", strict=F )
  modelETO <- omxSetParameters( modelETO, label=c("t1DZOecigEver3_T1",
"t1DZOecigEver3_T2"), free=T, values=svlTh, newlabels="t1DZOecigEver3", strict=F )

modelETO <-  mxOption(modelETO, "mvnRelEps", 1e-3)
fitETO <- mxTryHardOrdinal( modelETO, intervals=F)
fitGofs(fitETO); fitEsts(fitETO)

# Constrain expected Thresholds to be equal across Twin Order and Zygosity
modelETZ <- mxModel( fitETO, name="twoETZo" )
modelETZ<- omxSetParameters(modelETZ, label=c("t1MZMcccigEver3",
"t1MZFcccigEver3", "t1DZFcccigEver3","t1DZMcccigEver3", "t1DZOcccigEver3"), free =
T, values=svlTh, newlabels="t1cccigEver3", strict=F)
modelETZ<- omxSetParameters(modelETZ, label=c("t1MZMecigEver3",
"t1MZFecigEver3", "t1DZFecigEver3","t1DZMecigEver3", "t1DZOecigEver3"), free = T,
values=svlTh, newlabels="t1ecigEver3", strict=F)
fitETZ <- mxTryHardOrdinal( modelETZ, intervals=F )
fitGofs(fitETZ); fitEsts(fitETZ)
```

```
# Print Comparative Fit Statistics
satNested <- list(fitETO,fitETZ)
tableFitStatistics(fitSAT, satNested)


#####--------------------####
##### Equating Sexes     ####
#####--------------------####


# set the number of variables per twin (nv) and total variables per twin pair (ntv) for
automation
vars      <- c("ecigEver3", "cccigEver3")
#vars      <- c("cccigEver3", "ecigEver3")   #reverse order of variables to see if same
results emerge
nv        <- 2          # number of variables
ntv       <- nv*2       # number of total variables
selVars   <- paste(vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")
nth <- 1  # Number of thresholds per variable (only for binary data)

# Subset the data to only the things I need
twinData  <- data2[,c(selVars,'zyg2')]
describe(twinData)
summary(twinData)
dim(twinData)

#twinData2<-na.omit(twinData)
#summary(twinData2)
#dim(twinData2)
twinDataBin <-twinData
dim(twinDataBin)
table(twinDataBin$zyg2)

# Factorize Ordinal Variables using the mxFactor option
twinDataBin[,c(1,3)]  <- mxFactor(twinDataBin[,c(1,3)], levels = c(0:nth))
twinDataBin[,c(2,4)]  <- mxFactor(twinDataBin[,c(2,4)], levels = c(0:nth))


# Twin correlations

mzdat <- subset(twinDataBin, zyg2==c(1) | zyg2 ==2, selVars)
dzdat <- subset(twinDataBin, zyg2==c(3) | zyg2 ==4 | zyg2==5, selVars)

# 1=MZM, 2= MZF, 3=DZM, 4=DZF, 5=ODZ
#Vars      <- c("ecigEver3", "cccigEver3")
#nv        <- 2          # number of variables
```

```
#ntv       <- nv*2       # number of total variables
#selVars   <- paste(Vars,c(rep("_T1",nv),rep("_T2",nv)),sep="")

# Select Data for Analysis
mzfData    <- subset(twinDataBin, zyg2==2, selVars)
dzfData    <- subset(twinDataBin, zyg2==4, selVars)
mzmData    <- subset(twinDataBin, zyg2==1, selVars)
dzmData    <- subset(twinDataBin, zyg2==3, selVars)
dzoData    <- subset(twinDataBin, zyg2==5, selVars) #males = T1, females = T2

# Set Starting Values /
svLTh     <- 0.8    # start value for first threshold
svlTh     <- 1       # start value for increments
#svTh      <- c(0.7,1,0.7,1)
svTh      <- matrix(rep(c(svLTh,(rep(svlTh,nth-1)))),nrow=nth,ncol=ntv)     # start value
for thresholds
lbTh      <- matrix(rep(c(-3,(rep(0.001,nth-1))),nv),nrow=nth,ncol=ntv)     # lower bounds
for thresholds

#svTh      <- c(1,1)                  # start value for thresholds
svPa      <- .4                      # start value for path coefficient
svPaD     <- vech(diag(svPa,nv,nv))    # start values for diagonal of covariance matrix
svPe      <- .8                      # start value for path coefficient for e
svPeD     <- vech(diag(svPe,nv,nv))    # start values for diagonal of covariance matrix
lbPa      <- .00001                  # start value for lower bounds
lbPaD     <- diag(lbPa,nv,nv)          # lower bounds for diagonal of covariance matrix
lbPaD[lower.tri(lbPaD)] <- 0         # lower bounds for below diagonal elements
lbPaD[upper.tri(lbPaD)] <- NA          # lower bounds for above diagonal elements


# Set Starting Values
aLabs     <- paste("a",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
cLabs     <- paste("c",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
eLabs     <- paste("e",rev(nv+1-sequence(1:nv)),rep(1:nv,nv:1),sep="")
mvar1th     <- paste("mvar1","_th",1:nth, sep="")
mvar2th     <- paste("mvar2","_th",1:nth, sep="")
fvar1th     <- paste("fvar1","_th",1:nth, sep="")
fvar2th     <- paste("fvar2","_th",1:nth, sep="")
#dzvar1th    <-paste("var1DZ", "_th", 1:nth, sep="")
#dzvar2th    <-paste("var2DZ", "_th", 1:nth, sep="")


thUB      <- 2

# ------------------------------------------------------------------------------
# PREPARE MODEL
```

```
# General non-scalar ACE Model
# Matrices declared to store a, c, and e Path Coefficients
pathAf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label=c("af11", "af21", "af22"), name="af" )
pathCf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label=c("cf11", "cf21", "cf22"), name="cf" )
pathEf    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound =
-0.99, values=.6, label=c("ef11", "ef21", "ef22"), name="ef" )
pathAm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label=c("am11", "am21", "am22"), name="am" )
pathCm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label=c("cm11", "cm21", "cm22"), name="cm" )
pathEm    <- mxMatrix( "Lower", nrow=nv, ncol=nv, free=TRUE, ubound = 0.99, lbound
= -0.99, values=.6, label=c("em11", "em21", "em22"), name="em" )
pathRa    <- mxMatrix( "Lower", nrow=1, ncol=1, free=TRUE,  values=1, label="ra11",
name="ra", ubound=1, lbound=0 )

# Matrices generated to hold A, C, and E computed Variance Components
covAf     <- mxAlgebra( af %*% t(af), name="Af" )
covCf     <- mxAlgebra( cf %*% t(cf), name="Cf" )
covEf     <- mxAlgebra( ef %*% t(ef), name="Ef" )
covAm     <- mxAlgebra( am %*% t(am), name="Am" )
covCm     <- mxAlgebra( cm %*% t(cm), name="Cm" )
covEm     <- mxAlgebra( em %*% t(em), name="Em" )

# Algebra to compute total variances and standard deviations (diagonal only)
covPf     <- mxAlgebra( Af+Cf+Ef, name="Vf" )
covPm     <- mxAlgebra( Am+Cm+Em, name="Vm" )

# Algebras generated to hold Parameter Estimates and Derived Variance Components
colVarsZf <- c('Af','Cf','Ef','SAf','SCf','SEf')
estVarsZf <- mxAlgebra( cbind(Af,Cf,Ef,Af/Vf,Cf/Vf,Ef/Vf), name="VarsZf",
dimnames=list(NULL,colVarsZf))
colVarsZm <- c('Am','Cm','Em','SAm','SCm','SEm')
estVarsZm <- mxAlgebra( cbind(Am,Cm,Em,Am/Vm,Cm/Vm,Em/Vm), name="VarsZm",
dimnames=list(NULL,colVarsZm))

# Algebra for expected Mean and Variance/Covariance Matrices in MZ & DZ twins
#meanGf    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanf", name="expMeanGf" )
#meanGm    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label="meanm", name="expMeanGm" )
#meanGfm   <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=20,
label=c("meanf","meanm"), name="expMeanGfm" )
```

```
meanGf    <- mxMatrix( type="Zero", nrow=1, ncol=ntv, label="meanf",
name="expMeanGf" )
meanGm    <- mxMatrix( type="Zero", nrow=1, ncol=ntv, label="meanm",
name="expMeanGm" )
meanGfm   <- mxMatrix( type="Zero", nrow=1, ncol=ntv, label=c("meanf","meanm"),
name="expMeanGfm" )
covMZf    <- mxAlgebra( expression= rbind( cbind(Vf, Af+Cf), cbind(Af+Cf, Vf)),
name="expCovMZf" )
covDZf    <- mxAlgebra( expression= rbind( cbind(Vf, 0.5%x%Af+Cf),
cbind(0.5%x%Af+Cf, Vf)), name="expCovDZf" )
covMZm    <- mxAlgebra( expression= rbind( cbind(Vm, Am+Cm), cbind(Am+Cm, Vm)),
name="expCovMZm" )
covDZm    <- mxAlgebra( expression= rbind( cbind(Vm, 0.5%x%Am+Cm),
cbind(0.5%x%Am+Cm, Vm)), name="expCovDZm" )
CVfm      <- mxAlgebra( expression= ra%x%(af%*%t(am))+cf%*%t(cm), name="CVfm" )
CVmf      <- mxAlgebra( expression= ra%x%(am%*%t(af))+cm%*%t(cf), name="CVmf" )
covDZo    <- mxAlgebra( expression= rbind( cbind(Vf, CVfm), cbind(CVmf, Vm)),
name="expCovDZo" )
Inc     <- mxMatrix( type="Lower", nrow=nth, ncol=nth, free=F, values=1, name="Inc" )

# MALES
ThreM    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(mvar1th, mvar2th), lbound=-2, ubound=thUB, name="ThreM")
ExpThreM   <- mxAlgebra( expression= cbind( ( Inc %*% ThreM  ),
                                ( Inc %*% ThreM  ) ), name="ExpThreM" )
# FEMALES
ThreF    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(fvar1th, fvar2th), lbound=-2, ubound=thUB, name="ThreF")
ExpThreF   <- mxAlgebra( expression= cbind( ( Inc %*% ThreF  ),
                                ( Inc %*% ThreF  ) ), name="ExpThreF" )
## OS

ThreOS    <-mxMatrix( type="Full", nrow=nth, ncol=nv, free=c(T, T), values=,
labels=cbind(mvar1th ,fvar2th), lbound=-2, ubound=thUB, name="ThreOS")
ExpThreOS   <- mxAlgebra( expression= cbind( ( Inc %*% ThreOS  ),
                                 ( Inc %*% ThreOS  ) ), name="ExpThreOS" )


# Data objects for Multiple Groups
dataMZf   <- mxData( observed=mzfData, type="raw" )
dataDZf   <- mxData( observed=dzfData, type="raw" )
dataMZm   <- mxData( observed=mzmData, type="raw" )
dataDZm   <- mxData( observed=dzmData, type="raw" )
dataDZo   <- mxData( observed=dzoData, type="raw" )

# Expectation objects for Multiple Groups
```

```
expMZf    <- mxExpectationNormal( covariance="expCovMZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
expDZf    <- mxExpectationNormal( covariance="expCovDZf", means="expMeanGf",
dimnames=selVars, thresholds = "ExpThreF" )
expMZm    <- mxExpectationNormal( covariance="expCovMZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
expDZm    <- mxExpectationNormal( covariance="expCovDZm",
means="expMeanGm", dimnames=selVars, thresholds = "ExpThreM" )
expDZo    <- mxExpectationNormal( covariance="expCovDZo", means="expMeanGfm",
dimnames=selVars, thresholds = "ExpThreOS" )
funML     <- mxFitFunctionML()

# Combine Groups
parsZf    <- list( pathAf, pathCf, pathEf, covAf, covCf, covEf, covPf, estVarsZf, ThreF,
ExpThreF, Inc )
parsZm    <- list( pathAm, pathCm, pathEm, covAm, covCm, covEm, covPm,
estVarsZm, ThreM, ExpThreM, Inc )
parsZfm   <- list( pathRa, CVfm, CVmf, ExpThreOS, Inc, ThreOS)
modelMZf  <- mxModel( parsZf, meanGf, covMZf, dataMZf, expMZf, funML,
name="MZf" )
modelDZf  <- mxModel( parsZf, meanGf, covDZf, dataDZf, expDZf, funML, name="DZf"
)
modelMZm  <- mxModel( parsZm, meanGm, covMZm, dataMZm, expMZm, funML,
name="MZm" )
modelDZm  <- mxModel( parsZm, meanGm, covDZm, dataDZm, expDZm, funML,
name="DZm" )
modelDZo  <- mxModel( parsZf, parsZm, parsZfm, meanGfm, covDZo, dataDZo,
expDZo, funML, name="DZo" )
multi     <- mxFitFunctionMultigroup( c("MZf","DZf","MZm","DZm","DZo") )
QualAceModel  <- mxModel( "QualACE", modelMZf, modelDZf, modelMZm,
modelDZm, modelDZo, multi )


QualAceFit <-mxTryHardOrdinal(QualAceModel, intervals = F)
summary(QualAceFit)

## Coerce threshold to be equal across variables, see around line 500


eqthres   <-mxModel(QualAceFit, name = "Equal Thresholds")
#eqthres   <-omxSetParameters( eqthres, label=c("mvar1_th1", "mvar2_th1"),
free=TRUE, values=0.5, newlabels="var_th1")
#eqthres   <-omxSetParameters( eqthres, label=c("fvar1_th1", "fvar2_th1"), free=TRUE,
values=0.5, newlabels="var_th1")
eqthres   <-omxSetParameters( eqthres, label=c("mvar1_th1", "fvar1_th1"), free=TRUE,
values=0.5, newlabels="var1_th")
```

```
eqthres   <-omxSetParameters( eqthres, label=c("mvar2_th1", "fvar2_th1"), free=TRUE,
values=0.5, newlabels="var2_th")


eqthresfit<-mxTryHardOrdinal(eqthres, intervals=F)

satNested <- list(eqthresfit)
tableFitStatistics(QualAceFit, satNested)

## Coerce males and females to be equal

eqsex   <-mxModel(eqthresfit, name = "Equal sexes")
eqsex   <-omxSetParameters( eqsex, label=c("am11", "am21", "am22"), free=TRUE,
values=0.04, newlabels=c("a11", "a21", "a22")   )
eqsex   <-omxSetParameters( eqsex, label=c("af11", "af21", "af22"), free=TRUE,
values=0.04, newlabels=c("a11", "a21", "a22")  )
eqsex   <-omxSetParameters( eqsex, label=c("cm11", "cm21", "cm22"), free=TRUE,
values=0.5, newlabels=c("c11", "c21", "c22"))
eqsex   <-omxSetParameters( eqsex, label=c("cf11", "cf21", "cf22"), free=TRUE,
values=0.5, newlabels=c("c11", "c21", "c22"))
eqsex   <-omxSetParameters( eqsex, label=c("em11", "em21", "em22"), free=TRUE,
values=0.3, newlabels=c("e11", "e21", "e22") )
eqsex   <-omxSetParameters( eqsex, label=c("ef11", "ef21", "ef22"), free=TRUE,
values=0.3, newlabels=c("e11", "e21", "e22") )

eqsexfit<-mxTryHardOrdinal(eqsex, intervals = F)
tableFitStatistics(QualAceFit, eqsexfit)

## NO sex

nosex   <- omxSetParameters(eqsex, labels="ra11", free=FALSE, values=0.5,
name="No sex Effects" )
#nosex   <- omxSetParameters(eqsex, labels="meanf", free=T, values=0,
newlabels="meanm", name="No sex Effects" )
nosexfit<- mxTryHardOrdinal(nosex, intervals = F)
nested <-list(eqthresfit,eqsexfit, nosexfit)
tableFitStatistics(QualAceFit, nested)
parameterSpecifications(nosexfit)



# Check BivBinACEFIT is equal to nosex model

testmodel<-omxSetParameters(BivBinAceFit6, labels=c("var1M_th1", "var1F_th1"), free
=T, values = 0.3, newlabels="var1_th", name="Testing Model")
```

```
testmodel<-omxSetParameters(testmodel, labels=c("var2M_th1", "var2F_th1"), free =T,
values = 0.2, newlabels="var2_th")

testfit<-mxTryHardOrdinal(testmodel, intervals=F)
tableFitStatistics(nosexfit, testfit)

# Test of covA
BivBinAceModel8 <- BivBinAceFit6
BivBinAceModel8<- omxSetParameters(BivBinAceModel8, labels=c( "a21"),
free=FALSE, values=0 , name = "Test of CovA")
BivBinAceFit8<- mxTryHardOrdinal(BivBinAceModel8, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit8)

#BivBinAceFit8a <- omxRunCI(BivBinAceFit8)
#summary(BivBinAceFit8, verbose=F)
#BivBinAceFit8$algebras


# Test of covC
BivBinAceModel9 <- BivBinAceFit6
BivBinAceModel9<- omxSetParameters(BivBinAceModel9, labels=c( "c21"),
free=FALSE, values=0, name = "Test of CovC" )
BivBinAceFit9<- mxTryHardOrdinal(BivBinAceModel9, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit9)


# Test of covE
BivBinAceModel10 <- BivBinAceFit6
BivBinAceModel10<- omxSetParameters(BivBinAceModel10, labels=c("e21"),
free=FALSE, values=0, name="Test of CovE" )
BivBinAceFit10<- mxTryHardOrdinal(BivBinAceModel10, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit10)

# Test of rP
BivBinAceModel11 <- BivBinAceFit6
BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="a21",
free=FALSE, values=0, name = "Test of rP" )
BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="c21",
free=FALSE, values=0 )
#BivBinAceModel11<- omxSetParameters(BivBinAceModel11, labels="e21",
free=FALSE, values=0 )
BivBinAceFit11<- mxTryHardOrdinal(BivBinAceModel11, intervals = F)

tableFitStatistics(BivBinAceFit6, c(BivBinAceFit8, BivBinAceFit9, BivBinAceFit10,
BivBinAceFit11))
```

```
# Test of CE model

BivBinAceModel12 <- BivBinAceFit6
BivBinAceModel12<- omxSetParameters(BivBinAceModel12, labels=c( "a11","a21",
"a22"), free=FALSE, values=0, name = "CE model" )
BivBinAceFit12<- mxTryHardOrdinal(BivBinAceModel12, intervals = F)
summary(BivBinAceFit12, verbose=T)
tableFitStatistics(BivBinAceFit6, BivBinAceFit12)

# Test of AE Model
BivBinAceModel13 <- BivBinAceFit6
BivBinAceModel13<- omxSetParameters(BivBinAceModel13, labels=c( "c11","c21",
"c22"), free=FALSE, values=0, name = "AE model" )
BivBinAceFit13<- mxTryHardOrdinal(BivBinAceModel13, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit13)

# Test of E Model
BivBinAceModel16 <- BivBinAceFit6
BivBinAceModel16<- omxSetParameters(BivBinAceModel16,
labels=c("a11","a21","a22", "c11","c21", "c22"), free=FALSE, values=0, name = "E
model" )
BivBinAceFit16<- mxTryHardOrdinal(BivBinAceModel16, intervals = F)
tableFitStatistics(BivBinAceFit6, BivBinAceFit16)

bivnested <-list(BivBinAceFit12, BivBinAceFit13,BivBinAceFit16)
tableFitStatistics(BivBinAceFit6, bivnested)



# Test of E crosspaths Model
BivBinAceModel17 <- BivBinAceFit6
BivBinAceModel17<- omxSetParameters(BivBinAceModel17, labels=c("a21","c21"),
free=FALSE, values=0, name = "E model" )
BivBinAceFit17<- mxTryHardOrdinal(BivBinAceModel17, intervals = T)
summary(BivBinAceFit17, verbose=T)
tableFitStatistics(BivBinAceFit6, BivBinAceFit17)

BivBinAceModel18 <- BivBinAceFit17
BivBinAceFit18<- mxTryHardOrdinal(BivBinAceModel18, intervals = T)
summary(BivBinAceFit18, verbose=T)

BivBinAceModel19 <- BivBinAceFit18
BivBinAceFit19<- mxTryHardOrdinal(BivBinAceModel19, intervals = T)
summary(BivBinAceFit19, verbose=T)
```

```
# Using CE model, testing of C21

BivBinAceModel14 <- BivBinAceModel12
BivBinAceModel14<- omxSetParameters(BivBinAceModel14, labels=c( "c21"),
free=FALSE, values=0, name = "CE model No C21" )
BivBinAceFit14<- mxTryHardOrdinal(BivBinAceModel14, intervals = F)
tableFitStatistics(BivBinAceFit12, BivBinAceFit14)

# Using CE model, testing of E21

BivBinAceModel15 <- BivBinAceModel14
BivBinAceModel15<- omxSetParameters(BivBinAceModel15, labels=c( "e21"),
free=FALSE, values=0, name = "CE model No Cross Paths" )
BivBinAceFit15<- mxTryHardOrdinal(BivBinAceModel15, intervals = F)
tableFitStatistics(BivBinAceFit14, BivBinAceFit15)
tableFitStatistics(BivBinAceFit12, BivBinAceFit15)
```

```
## ANALYSIS FOR CHAPTER 4

## Title: Genes for Good phenotypic data cleaning and analysis
## Author: James Clifford (cliffordjs@vcu.edu)


require(car)
require(polycor)
require(gmodels)
setwd("~/Desktop/G4G_Data")
getwd()

samp <- read.table("masked-id-pass-samples.txt", header = F)
samp[,1]  <- as.character(samp[,1])
dim(samp)

## Read in the main outcome data
## Tobacco use
tobdata<-read.csv("~/Desktop/G4G_Data/G4G_tobacco.csv", header=T, na.strings = "")
dim(tobdata)
names(tobdata)

tobdata<-subset(tobdata, user_id %in% samp[,1])
dim(tobdata)


table(tobdata[,9], useNA = 'always')
table(tobdata[,22], useNA = 'always')

newtobdat<-subset(tobdata, user_id %in% samp[,1])
dim(newtobdat)
table(newtobdat[,6])
table(newtobdat[,22])

table(newtobdat[,6], newtobdat[,22])
ans<-polychor(newtobdat[,6], newtobdat[,22])
pchisq(ans$chisq, ans$df, lower.tail=T)

## Get the EC data into a format with just ID and EC ever use
ecdata <-tobdata[,c(1,22)]
names(ecdata)

## Add in column for family id (user ID) as this column is required for PLINK
ecdata2<-cbind(famid=ecdata$user_id, ecdata)
head(ecdata2)
```

```
## Recode from string variables into numeric
## 1 = No, 2 = Yes. This is consistent with PLINK pheno file notation
ecdata2$e_cigs<-recode(ecdata2$e_cigs, "'No' = 1; 'Yes'= 2")
head(ecdata2)
table(ecdata2[,3])

## Write EC phenotype file to space delimited text
#write.table(ecdata2, file = "G4G_ecuse.txt", sep = " ", col.names=F, row.names=F,
quote=F)


## Repeating previous steps but with CC ever use
ccdata <-tobdata[,c(1,6)]
head(ccdata)

## Add in column for family id (user ID) as this column is required for PLINK
ccdata2<-cbind(famid=ccdata$user_id, ccdata)
head(ccdata2)

## Recode from string variables into numeric
## 1 = No, 2 = Yes. This is consistent with PLINK pheno file notation
ccdata2$ever_tried_cig<-recode(ccdata2$ever_tried_cig, "'No' = 1; 'Yes'= 2")
head(ccdata2)
table(ccdata2[,3])

## Write CC phenotype file to space delimited text
#write.table(ccdata2, file = "G4G_ccuse2.txt", sep = " ", col.names=F, row.names=F,
quote = F)


## Create new R object for phenotypic analyses

phenodat<-tobdata[,c(1,6,22,28,29,30)]
names(phenodat)


# Recode tobacco varaibles
phenodat$ever_tried_cig<-recode(phenodat$ever_tried_cig, "'No' = 1; 'Yes'= 2")
phenodat$e_cigs<-recode(phenodat$e_cigs, "'No' = 1; 'Yes'= 2")

table(phenodat[,2])
table(phenodat[,3])

## Parents smoke
table(phenodat[,4])
```

```
phenodat$parents_smoke<-recode(phenodat$parents_smoke, "'Neither
smokes/smoked'=1; 'Yes, both'=2;'Yes, one of them' =2")
table(phenodat[,4])

## Friends smoke
table(phenodat[,5])
phenodat$friends_smoke<-recode(phenodat$friends_smoke, "'None of them smoke'=1;
'Yes, a few' = 2; 'Yes, most of them'=2")
table(phenodat[,5])

## Friends smoke under 18
table(phenodat[,6])
phenodat$friends_smoke_under_18<-recode(phenodat$friends_smoke_under_18,
"'None of them smoked'=1; 'Yes, a few' = 2; 'Yes, most of them'=2")
table(phenodat[,6])



## Read in the demographic data, note there are two separate demo files
demodata<-read.csv("~/Desktop/G4G_Data/G4G_demos_a.csv", header=T, na.strings
= "")
dim(demodata)
names(demodata)
demodatasmall<-demodata[,c(1,3,4,28,38,39,41)]
dim(demodatasmall)
names(demodatasmall)

# recode Health insurance, 1= Yes, 2 = No, 3 = I don't know
demodatasmall$insure_r <-recode (demodatasmall[,7], "'I do not know' = 3; 'No'=2;
'Yes'=1")
table(demodatasmall$insure_r )

# recode gender, male = 1, female =2
demodatasmall$gender_r <-recode(demodatasmall[,2], "'female'=2; 'male'=1")
table(demodatasmall$gender_r)

# Recode education
demodatasmall$education_r<-recode(demodatasmall[,5], "'No high school diploma or
GED'=1;'Some college but no degree'=2;'High school graduate or GED'=2;
                        'Associates degree'=3; 'Bachelors degree (such as  BA, AB, BS,
or BBA)'=3;'Masters degree or higher (such as MA, MS, MBA, PhD, MD, and so on)'=4
                        ")

table(demodatasmall$education_r)
```

```
demodata2<-read.csv("~/Desktop/G4G_Data/G4G_demos2.csv", header=T, na.strings
= "")
dim(demodata2)
names(demodata2)
demodata2small<-demodata2[,c(1,3,4,19, 14,15,13)]
dim(demodata2small)
names(demodata2small)


# recode Health insurance, 1= Yes, 2 = No, 3 = I don't know
demodata2small$insure_r <-recode (demodata2small[,7], "'I do not know' = 3; 'No'=2;
'Yes'=1")
table(demodata2small$insure_r )

# recode gender, male = 1, female =2
demodata2small$gender_r <-recode(demodata2small[,2], "'female'=2; 'male'=1")
table(demodata2small$gender_r)

# Recode education, < HS, some college, college degree, master or higher

demodata2small$education_r<-recode(demodata2small[,5], "'No high school diploma or
GED'=1;'Some college but no degree'=2;'High school graduate or GED'=2;
                        'Associate degree'=3; 'Bachelor degree (like a  BA, AB, BS, or
BBA)'=3;'Master degree or higher (such as MA, MS, MBA, PhD, MD, and so on)'=4
                        ")

table(demodata2small$education_r)



totaldemo<-rbind(demodatasmall, demodata2small)
dim(totaldemo)
names(totaldemo)

## recode age range
totaldemo$age_range_r<-recode(totaldemo[,3], "'18-21' =1; '21-30'=2;'30-40'=3;'40-
50'=4;
                      '50-60'=5;'60-70'=6; '70+'=7")
table(totaldemo[,11])

### Filter out individuals without Genetic data

demofilt<-subset(totaldemo, user_id %in% samp[,1])
dim(demofilt)

# Remove phenotypic duplicates
```

```
finaldemo<-subset(demofilt, !duplicated(demofilt[,1]))
table(finaldemo[,2], useNA = 'always')
dim(finaldemo)

# filter pheno data to Whte only data
racedemo<-totaldemo[totaldemo$race == "White" | totaldemo$race=="White or
European",]
dim(racedemo)


whiteIDs<-(racedemo[,1])
head(whiteIDs)
whiteIDs<-as.data.frame(whiteIDs)
head(whiteIDs)

whitepheno<-subset(finaldemo, user_id %in% whiteIDs[,1])
dim(whitepheno)
# 15,927 individuals

# Find out dual users in Whites
whitetob<-subset(phenodat, user_id %in% whiteIDs[,1])
table(whitetob$e_cigs, whitetob$ever_tried_cig)

# 1008 never users
# 45 ECIG exclusive
# 4706 dual users
# 10082 CIG exclusive

# Find genotyped participants' tobacco use

tobuse<-subset(phenodat, user_id %in%finaldemo[,1])
dim(tobuse)
names(tobuse)
table(tobuse$e_cigs, tobuse$ever_tried_cig)


## Add in PCs
setwd("/Users/jamesclifford/Desktop/G4G/PCA")

pca<-read.table("plink_pca_test_white.eigenvec")

# remove extra column
pca <-pca[,-1]

# set names
names(pca)[1] <- "IID"
```

```
names(pca)[2:ncol(pca)] <- paste0("PC", 1:(ncol(pca)-1))
head(pca)
dim(pca)
names(pca)

## remove individuals who aren't present
finalgen<-subset(pca, IID %in% finaldemo[,1])
dim(finalgen)
# 15881

## Take only first 7 PCs
finalgen<-finalgen[,c(1:8)]
names(finalgen)
# Combine PCs with phenotypic data
names(whitepheno)[1]<-"IID"

whitefinal<-merge(whitepheno, finalgen, by ="IID")

dim(whitefinal)
# 15,881 individuals
names(whitefinal)

## Write Covariate file with only white participants


# write.table(finalcovars, "G4G_white_covars.txt", quote=F, row.names=F,
col.names=F)
# test<-read.table("G4G_white_covars.txt")
# head(test)


## Write covariate file with ECIG use

names(whitetob)[1] <- "IID"
ecigtest<-whitetob[,c(1,3)]

whiteecigfinal<-merge(whitefinal, ecigtest, by = "IID")
dim(whiteecigfinal)
names(whiteecigfinal)
# 15798

whiteecigfinal<-whiteecigfinal[,c(1,1,8:19)]

# write.table(whiteecigfinal, "G4G_white_ECIG_covars.txt", quote=F, row.names=F,
col.names=F)
# test2<-read.table("G4G_white_ECIG_covars.txt")
```

```
# head(test2)

## Write covar file with CIG use

cigtest<-whitetob[,c(1,2)]

whitecigfinal<-merge(whitefinal, cigtest, by = "IID")
dim(whitecigfinal)
names(whitecigfinal)
# 15798

whitecigfinal<-whitecigfinal[,c(1,1,8:19)]

# write.table(whitecigfinal, "G4G_white_CIG_covars.txt", quote=F, row.names=F,
col.names=F)
# test3<-read.table("G4G_white_CIG_covars.txt")
# head(test3)

## Add in ECIG and CIG data

tobdata2<-merge(ecdata2, ccdata2, by = "user_id")
dim(tobdata2)
# 20105

names(tobdata2)
tobdata2<-subset(tobdata2, select = c("user_id", "e_cigs", "ever_tried_cig"))
head(tobdata2)

tobdata3<-subset(tobdata2, user_id %in%whitefinal[,1])
dim(tobdata3)
#15798

names(tobdata3)[1]<-"IID"
## Merge phenotypic/PC data with tobacco data

finaldata<-merge(tobdata3, whitefinal, by = "IID")
dim(finaldata)
# 15798
names(finaldata)


### Create Table 1

require(table1)
finaldata$age_range_r <- factor(finaldata$age_range_r, levels = 1:7,
                                labels=c("18-21", "22-30", "31-40", "41-50", "51-60",
```

```
                                    "61-70", "70+"))
finaldata$education_r <-factor(finaldata$education_r, levels=1:4, labels=c("Less than
HS",
                                                                    "HS Grad/GED/Some College",
"Associates Degree", "College Graduate or More"))

finaldata$insure_r    <-factor(finaldata$insure_r, levels=1:3, labels=c("Covered", "Not
Covered", "I Dont Know"))

finaldata$gender_r    <-factor(finaldata$gender_r, levels = 1:2, labels=c("Male",
"Female"))

finaldata$ever_tried_cig<-factor(finaldata$ever_tried_cig, levels = 1:2, labels=c("No",
"Yes"))
finaldata$e_cigs<-factor(finaldata$e_cigs, levels = 1:2, labels=c("No", "Yes"))


label(finaldata$age_range_r)      <- "Age Range"
label(finaldata$education_r)      <- "Education Level"
label(finaldata$insure_r)         <- "Insurance Status"
label(finaldata$gender_r)         <- "Sex"

table1(~ gender_r+age_range_r+ education_r+ insure_r|ever_tried_cig, data=finaldata,
overall=F)
table1(~ gender_r+age_range_r+ education_r+ insure_r|e_cigs, data=finaldata,
overall=F)
table(finaldata$e_cigs)
table(finaldata$ever_tried_cig)


## Bivariate

### CIG

CrossTable(finaldata$gender_r, finaldata$ever_tried_cig, chisq=T)

## p < 0.0001

CrossTable(finaldata$age_range_r,  finaldata$ever_tried_cig,chisq=T)

## p < 0.0001

CrossTable(finaldata$education_r,  finaldata$ever_tried_cig,chisq=T)

## p < 0.0001
```

```
CrossTable(finaldata$insure_r,  finaldata$ever_tried_cig,chisq=T)

## p < 0.0001


### ECIG

CrossTable(finaldata$gender_r,  finaldata$e_cigs,chisq=T)

## p < 0.0001

CrossTable(finaldata$age_range_r, finaldata$e_cigs, chisq=T)

## p < 0.0001

CrossTable(finaldata$education_r, finaldata$e_cigs, chisq=T)

## p < 0.0001

CrossTable(finaldata$insure_r, finaldata$e_cigs, chisq=T)

## p < 0.0001

### CIG x ECIG

finaldata$ever_tried_cig<-factor(finaldata$ever_tried_cig, levels = 1:2, labels=c("No",
"Yes"))
finaldata$e_cigs<-factor(finaldata$e_cigs, levels = 1:2, labels=c("No", "Yes"))


CrossTable(finaldata$e_cigs, finaldata$ever_tried_cig, chisq=T)

(4694*1005)/(45*100052)
## Crude OR = 1.048; those who smoke cigarettes are ~4% more likely to use ECIGs
than
## non-smokers

## create genotype or not and rerun bivaraite with that

totaldemo$geno<-ifelse(totaldemo$user_id %in% samp[,1], 2, 1)

# Remove phenotypic duplicates
totaldemo<-subset(totaldemo, !duplicated(totaldemo[,1]))
#table(totaldemo[,2], useNA = 'always')
dim(totaldemo)
table(totaldemo$geno)
```

```
## Label for easy readin'
totaldemo$age_range_r <- factor(totaldemo$age_range_r, levels = 1:7,
                          labels=c("18-21", "22-30", "31-40", "41-50", "51-60",
                                "61-70", "70+"))
totaldemo$education_r <-factor(totaldemo$education_r, levels=1:4, labels=c("Less than
HS",
                                                        "HS Grad/GED/Some College",
"Associates Degree", "College Graduate or More"))

totaldemo$insure_r    <-factor(totaldemo$insure_r, levels=1:3, labels=c("Covered", "Not
Covered", "I Dont Know"))

totaldemo$gender_r    <-factor(totaldemo$gender_r, levels = 1:2, labels=c("Male",
"Female"))

totaldemo$geno        <-factor(totaldemo$geno, levels=1:2, labels = c("Not Genotyped",
"Genotyped"))


CrossTable(totaldemo$gender_r, totaldemo$geno, chisq=T)

## p < 0.0001

CrossTable( totaldemo$age_range_r,totaldemo$geno, chisq=T)

## p < 0.0001

CrossTable(totaldemo$education_r, totaldemo$geno, chisq=T)

## p < 0.0001

CrossTable(totaldemo$insure_r,totaldemo$geno,  chisq=T)

## p < 0.0001


tobdata_geno<-read.csv("~/Desktop/G4G_Data/G4G_tobacco.csv", header=T,
na.strings = "")


## Get the EC data into a format with just ID and EC ever use
genoecdata <-tobdata_geno[,c(1,22)]
names(genoecdata)

## Recode from string variables into numeric
```

```
## 1 = No, 2 = Yes. This is consistent with PLINK pheno file notation
genoecdata$e_cigs<-recode(genoecdata$e_cigs, "'No' = 1; 'Yes'= 2")
head(genoecdata)
table(genoecdata[,2])


## Repeating previous steps but with CC ever use
genoccdata <-tobdata_geno[,c(1,6)]
head(genoccdata)

## Recode from string variables into numeric
## 1 = No, 2 = Yes. This is consistent with PLINK pheno file notation
genoccdata$ever_tried_cig<-recode(genoccdata$ever_tried_cig, "'No' = 1; 'Yes'= 2")
head(genoccdata)
table(genoccdata[,2])

newtobgen<-merge(genoccdata, genoecdata, by = "user_id")
dim(newtobgen)

genodat<-merge(newtobgen, totaldemo, by = "user_id")
dim(genodat)
# 55,104

genodat$ever_tried_cig<-factor(genodat$ever_tried_cig, levels = 1:2, labels=c("No",
"Yes"))
genodat$e_cigs<-factor(genodat$e_cigs, levels = 1:2, labels=c("No", "Yes"))


CrossTable(genodat$e_cigs,genodat$geno,  chisq=T)

CrossTable( genodat$ever_tried_cig,genodat$geno, chisq=T)



## Overlap of ECIG and CIG use
names(genodat)
dim(genodat)
table(genodat$ever_tried_cig, genodat$e_cigs)
CrossTable(genodat$ever_tried_cig, genodat$e_cigs, chisq=T)



### COMMAND LINE CODES AND BASH SCRIPTS

## Plink QC for Genes for Good (G4G)
## Using 1000 Genomes (1KG) as reference panel
```

Note that though this is written in R, there are very few R commands
Author: James Clifford (cliffordjs@vcu.edu) with special thanks/acknowledgments
to Dr Roseann Peterson

Remove duplicate SNPs

/vcu_gpfs2/home/cliffordjs/bin/./plink --bfile
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged --list-duplicate-vars --suppress-first
--make-bed --out /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged_pruned

Merge reference and data

/vcu_gpfs2/home/cliffordjs/bin/./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg --
bmerge /vcu_gpfs2/home/cliffordjs/g4g_geno --allow-no-sex --geno 0.05 --make-bed --
out /vcu_gpfs2/home/GfG/refgenome/1kg_merged

2504 individuals in 1KG
20231 in G4G
1.28.21 this is throwing several errors: multiple variants for rs #, multiple
chromosomes for variant,
variant '.'?
variants with multiple positions: rs9442277, rs571228985
rs6658405
SNPs with multiple chromosomes: rs2789523, rs554199249, .
Tried flipping on .missnp file, no good

How many SNPs overlap 1KG and G4G

Genotyping rate
/vcu_gpfs2/home/cliffordjs/bin/./plink --bfile
/vcu_gpfs2/home/GfG/refgenome/1kg_merged
-- freq --out merge_genoRate

Total genotyping rate is

Light QC on overlapping SNPs
limit 1KG to overlapping SNPs

/vcu_gpfs2/home/cliffordjs/bin/./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg
--extract /vcu_gpfs2/home/GfG/refgenome/1kg_merged --allow-no-sex --make-bed --out
/vcu_gpfs2/home/GfG/refgenome/1kg_g4gsnps

Light QC --mind 0.05 --geno 0.5 --maf 0.01 --hwe 0.0000000005 (5xe-10)

```
/vcu_gpfs2/home/cliffordjs/bin/./plink --bfile
/vcu_gpfs2/home/GfG/refgenome/1kg_g4gsnps
--mind 0.05 --geno 0.05 --maf 0.01 --hwe 0.0000000005 --allow-no-sex --make-bed
--out /vcu_gpfs2/home/GfG/refgenome/1kg_g4gsnps_qc


## PCA
setwd("/vcu_gpfs2/home/cliffordjs")

data1<-read.table("chr22_illumina.txt")
data2<-read.table("chr22_g4g.txt")

test<-merge(data1, data2)
head(test)
dim(test)
# 5130 SNPs

data1<-read.table("chr21_illumina.txt")
#4407 SNPs

data2<-read.table("chr21_g4g.txt")
# 7027 SNPs

test<-merge(data1, data2)
head(test)
dim(test)

# 4345 SNPs

data1<-read.table("illumina_snp_ids.txt")
# 547667

data2<-read.table("G4G_snpids2.txt")

# 313085

test<-merge(data1, data2)
head(test)
dim(test)

# 308,985 SNPs

## Looking at G4G and 1KG

require(data.table)
require(dplyr)
```

```
data1<-fread("1kg_snpids.txt", header=F)
#data1<-as.data.frame(data1)
# 84,358,431

#data2<-read.table("G4G_snpids2.txt")
data2<-fread("G4G_snpids2.txt", header=F)

# 313085 SNPs

test<-inner_join(data1, data2)
head(test)
dim(test)

# 312,304 SNPs


## Create list of new snp ids for 1 kg
# awk '{print $2,$1":"$4":"$6":"$5}' 1kg.bim >> 1kg_ids2.txt

# Remove Duplicates
#
# --allow-extra-chr
# --bfile /vcu_gpfs2/home/GfG/refgenome/1kg
# --exclude /vcu_gpfs2/home/GfG/refgenome/1kg_dups_id.txt
# --make-bed
# --out /vcu_gpfs2/home/GfG/refgenome/1kg_dupsremoved


# update SNP namesta

# --allow-extra-chr
# --bfile /vcu_gpfs2/home/GfG/refgenome/1kg_dupsremoved
# --make-bed
# --out 1kg_updated_ids
# --update-name /vcu_gpfs2/home/GfG/refgenome/1kg_ids.txt


# Run in plink

#./plink --bfile /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged3 --extract
# /vcu_gpfs2/home/cliffordjs/overlapping_gfg_1kg.txt --make-bed --out
1KG_G4G_merge
# --allow-no-sex


# Genotyping rate
```

```
# ./plink --bfile /vcu_gpfs2/home/GfG/cleangenetic/Filtered/1KG_G4G_merge --freq --
out
# /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merge_genoRate



# Limit 1KG to overlapping SNPs
# ./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg_update_ids  \
# --extract /vcu_gpfs2/home/GfG/cleangenetic/Filtered/1KG_G4G_merge.bim --allow-
no-sex --make-bed
# --out /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps --allow-extra-chr

# 229,934 variants and 2504 people pass filters and QC.

# Light QC --mind 0.05 --geno 0.05 --maf 0.01 --hwe 0.0000000005 (5.0xe-10)
# ./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps --mind 0.05 --geno 0.05
--maf 0.01 --hwe 0.0000000005
# --keep-allele-order --chr 1-22 \
# --allow-no-sex --make-bed --out /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps_qc

# 27,618 removed due to HWE
# 3,978 removed due to MAF

# 198,338 variants and 2504 people pass filters and QC.

# Prune SNps
# ./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps_qc
# --indep-pairwise 1500 150 0.2 --allow-no-sex --keep-allele-order
# --out /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps_qc_prune02

# after pruning: 118,224 of 198,338 variants removed


### Quick check G4G quality overlap SNPs
# Light QC --mind 0.05 --geno 0.05 --maf 0.01 --hwe 0.0000000005 (5.0xe-10)
# ./plink --bfile /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged3 --extract
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/1KG_G4G_merge.bim --mind 0.05 --geno
0.05 --maf 0.01 --hwe 0.0000000005 --keep-allele-order --chr 1-22 \
# --allow-no-sex --make-bed --out
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/G4G_1kgSNPs_qc

# 0 removed to missing genotype data
# 0 removed due to HWE
# 8 removed due to MAF
```

```
# 229,924 variants and 20231 people pass filters and QC

## Create 1KGP-G4G bed file on pruned 0.2
#./plink --bfile /vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps --extract
/vcu_gpfs2/home/GfG/refgenome/1kg_G4Gsnps_qc_prune02.prune.in
# --allow-no-sex --keep-allele-order --make-bed --out
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/G4G_1kgSNPs_qc_pruned

# 80,114 variants and 2504 people pass filters and QC.


### Create Pedind file
setwd('/Users/jamesclifford/Desktop/G4G/PCA/')
# add row number to make sure order has not changed
#awk '{print $0 "\t" NR}' G4G_1kgSNPs_qc_pruned_newids.fam >
G4G_1kgSNPs_qc_pruned_fam_NR.txt
#scp
cliffordjs@fenn.vcu.edu:/vcu_gpfs2/home/GfG/refgenome/G4G_1kgSNPs_qc_pruned_f
am_NR.txt /Users/jamesclifford/Desktop/G4G/PCA
fam<-read.table('/Users/jamesclifford/Desktop/G4G/PCA/Old
files/G4G_1kgSNPs_qc_pruned_fam_NR.txt', header=F)
colnames(fam)<-c("FID", "IID", "V3" ,"V4" ,"V5" ,"V6" ,"Index")
head(fam)
dim(fam)
# Add G4G IDs
#awk '{print $1 "\t" $2 "\t" "G4G"}' G4G_1kgSNPs_qc.fam > G4G_ID.txt
#scp cliffordjs@fenn.vcu.edu:/vcu_gpfs2/home/GfG/cleangenetic/Filtered/G4G_ID.txt
/Users/jamesclifford/Desktop/G4G/PCA
g4g<-read.table('/Users/jamesclifford/Desktop/G4G/PCA/Old files/G4G_ID.txt', header =
F)
colnames(g4g)<-c("FID", "IID", "pop")
head(g4g)
dim(g4g)

# 1KGP IDs
kgp<-read.table('/Users/jamesclifford/Desktop/G4G/PCA/Old files/1KGP_pop.txt',
header = T)
colnames(kgp)
#"IID" "fam" "pop"
kgpID<-subset(kgp, select=c(fam, IID, pop))
colnames(kgpID)<-c("FID","IID", "pop")

# merge here to remove extra samples from 1kgp_pop.txt

table(kgpID$pop)
# 3500 ppl
```

```
dim(fam)

require(dplyr)

test<-inner_join(fam, kgpID, by =c("IID", "FID"))
dim(test)
# returns 824 people, should be 2504


test<-merge(fam, kgpID, by=c("IID", "FID"),all.x=T, no.dups = T)
dim(test)
table(test$pop, useNA="always")
#1680 NAs?


testIDS<-subset(test, select=c(FID, IID, pop))
dim(testIDS)

## new ID file from website
## https://github.com/WeiYang-BAI/Impu-Reference-Panel-
Reconstruction/blob/master/1000GP_Phase3.sample

kgp2<-read.table('1kg_pops.txt', header=T)
dim(kgp2)
names(kgp2)

kgp2IDS<-subset(kgp2, select=c("ID", "ID", "POP"))
colnames(kgp2IDS)<-c("FID","IID", "pop")
dim(kgp2IDS)
names(kgp2IDS)
table(kgp2IDS$pop)

test<-merge(fam, kgp2IDS, by=c("IID", "FID"),all.x=T, no.dups = T)
dim(test)
head(test)
table(test$pop, useNA="always")

testIDS<-subset(test, select=c(FID, IID, pop))
dim(testIDS)
table(testIDS$pop, useNA = 'always')
# Merge 1KGP pop data
pops<-rbind(testIDS,g4g)
dim(pops)
names(pops)
table(pops$pop, useNA = 'always')
```

```
# Do a merge and keep all x
#fam_pops<-merge(fam, pops, all.x=T, by=c("IID"))
fam_pops<-merge(fam, pops, all.y=T, by=c("IID"))
table(fam_pops$pop)
table(is.na(fam_pops$pop))
colnames(fam_pops)
pedindI<-subset(fam_pops, select=c(FID.y, IID, V3, V4, V5, pop, Index))
#write.table(pedindI,
file="/Users/jamesclifford/Desktop/G4G/PCA/1KGP_G4G_pop_index.txt", row.names =
T, col.names = T, quote=F)
# Confirmed Index
pedind<-subset(fam_pops, select=c(FID.y, IID, V3, V4, V5, pop))
#write.table(pedind,
file="/Users/jamesclifford/Desktop/G4G/PCA/1KGP_G4G_qc_pruned.pedind",
row.names = F, col.names = F, quote=F)

## GWAS
CIG

PLINK v1.90b6.18 64-bit (16 Jun 2020)
Options in effect:
  --allow-no-sex
  --bfile /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged3
  --covar /vcu_gpfs2/home/cliffordjs/G4G_white_ECIG_covars.txt
  --keep /vcu_gpfs2/home/cliffordjs/G4G_white_ids.txt
  --logistic hide-covar
  --out
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/Adjust_Assoc/ECIG_adjusted_hidecovar_
White_CC_phenocovars_pc
s
  --pheno /vcu_gpfs2/home/GfG/cleanphenotypic/G4G_ccuse2.txt


ECIG

PLINK v1.90b6.18 64-bit (16 Jun 2020)
Options in effect:
  --allow-no-sex
  --bfile /vcu_gpfs2/home/GfG/cleangenetic/Filtered/merged3
  --covar /vcu_gpfs2/home/cliffordjs/G4G_white_CIG_covars.txt
  --keep /vcu_gpfs2/home/cliffordjs/G4G_white_ids.txt
  --logistic hide-covar
  --out
/vcu_gpfs2/home/GfG/cleangenetic/Filtered/Adjust_Assoc/CIG_adjusted_hidecovar_Wh
ite_ECIG_phenocovars_p
cs
```

```
  --pheno /vcu_gpfs2/home/GfG/cleanphenotypic/G4G_ecuse.txt



## Manhattan


require(data.table)
require(qqman)

setwd("/vcu_gpfs2/home/GfG/cleangenetic/Filtered/Adjust_Assoc/")


ccdatadj<-
fread("ECIG_adjusted_hidecovar_White_CC_phenocovars_pcs.assoc.logistic", header
= T)
head(ccdatadj)

tiff("cc_adj_forECIG_withwhitepcs.tiff")
manhattan(ccdatadj, chr = "CHR", bp = "BP", snp = "SNP", p = "P", main = "Manhat
tan Plot for CIG-Adjusted Analyses, White Only PCs",ylim=c(0,11),
        col = c("red", "blue4"), suggestiveline=-log10(1e-05), genomewideline
        = -log10(1e-08), annotatePval = -log10(1e-07),
        chrlabs = c(1:22))
dev.off()

# Manhattan plot without annotation

tiff("cc_noannotation_adj_forECIG_withwhitepcs_3302022.tiff")
manhattan(ccdatadj, chr = "CHR", bp = "BP", snp = "SNP", p = "P", main =
"",ylim=c(0,11),
        col = c("red", "blue4"), suggestiveline=-log10(1e-05), genomewideline
        = -log10(1e-08), annotateTop=F,
        chrlabs = c(1:22))
dev.off()


## Find Suggestive/Significant Variants
ccSNPS<-subset(ccdatadj, P <= 1e-6)
ccSNPS
#   CHR         SNP       BP A1 TEST NMISS    OR  STAT        P
#   10 10:122876485 122876485  A  ADD 15796 1.771 5.005 5.574e-07  Gene:
FAM24B-CUZD1
#   18  18:61024122  61024122  T  ADD 15796 1.234 5.158 2.500e-07  Intergenic:
HMGN1P31, CDH20
```

```
# ECIG Adjusted results

#   2  2:84368347 84368347  T  ADD 15795 0.8810 -4.910 9.114e-07 #SUCLG1
#  11 11:42405437 42405437  A  ADD 15795 0.6264 -4.955 7.243e-07 # LINC02740
#  18 18:61024122 61024122  T  ADD 15795 1.2370  5.142 2.715e-07 Intergenic:
HMGN1P31, CDH20


ecdatadj<-
fread("CIG_adjusted_hidecovar_White_ECIG_phenocovars_pcs.assoc.logistic", header
= T)
head(ecdatadj)

tiff("ec_adj_forCIG_withwhitepcs.tiff")

manhattan(ecdatadj, chr = "CHR", bp = "BP", snp = "SNP", p = "P", main = "Manhat
tan Plot for ECIG-Adjusted Analyses", ylim = c(0,11),
        col = c("red", "blue4"), suggestiveline=-log10(1e-05), genomewideline
        = -log10(1e-08), annotatePval = -log10(1e-08),
        chrlabs = c(1:22))
dev.off()

## No annotation manhattan
tiff("ec_noannotation_adj_forCIG_withwhitepcs3302022.tiff")

manhattan(ecdatadj, chr = "CHR", bp = "BP", snp = "SNP", p = "P", main = "", ylim =
c(0,11),
        col = c("red", "blue4"), suggestiveline=-log10(1e-05), genomewideline
        = -log10(1e-08), annotateTop= F,
        chrlabs = c(1:22))
dev.off()



## Find Suggestive/Significant Variants
ecSNPs<-subset(ecdatadj, P <= 1e-6)
ecSNPs

dim(ecSNPs)

# CHR        SNP        BP A1 TEST NMISS     OR   STAT        P
# 2 2:216131813 216131813  A  ADD 15795 0.6215 -4.929 8.276e-07  Gene: XRCC5
# 2 2:216131851 216131851  G  ADD 15795 0.6172 -5.038 4.706e-07  Gene: XRCC5
# 2 2:216132049 216132049  A  ADD 15795 0.6195 -4.942 7.736e-07 Gene: XRCC5
# 2 2:216133672 216133672  C  ADD 15795 0.6237 -4.899 9.657e-07  Gene: XRCC5
```

```
# 6  6:33902823  33902823  T  ADD 15795 0.7852 -5.224 1.752e-07  Gene:
LOC105375026
# 8  8:25281329  25281329  G  ADD 15795 0.6349 -4.995 5.895e-07  Gene: DOCK5
# 13 13:32403784  32403784  T  ADD 15795 0.6188 -5.379 7.494e-07  Gene:N4BP2L1
# 13 13:47437096  47437096  C  ADD 15795 0.5112 -4.918 8.761e-07  Intergenic:
GNGSP5, RN7SL700P


# CIG ADJUSTED
# CHR         SNP         BP A1 TEST NMISS     OR   STAT         P
# 2 2:115364757 115364757  C  ADD 15795 1.2840  5.019 5.192e-07 Gene: DPP10
# 6  6:33902823  33902823  T  ADD 15795 0.7881 -5.079 3.801e-07 Gene:
LOC105375026
# 13 13:32403784  32403784  T  ADD 15795 0.6207 -5.252 1.508e-07 Gene:N4BP2L1
# 15 15:49010393  49010393  T  ADD 15795 0.4387 -4.935 8.009e-07 Gene:
SECISBP2L


## PCA PLOTS
#scp cliffordjs@fenn.vcu.edu:/vcu_gpfs2/home/cliffordjs/bin/*white.eigenval .
#scp cliffordjs@fenn.vcu.edu:/vcu_gpfs2/home/cliffordjs/bin/*white.eigenvec .

setwd("/Users/jamesclifford/Desktop/G4G/PCA")
require(tidyverse)
pca <- read_table2("plink_pca_test_white.eigenvec", col_names = FALSE)
eigenval <- scan("plink_pca_test_white.eigenval")

# remove extra column
pca <-pca[,-1]

# set names
names(pca)[1] <- "ind"
names(pca)[2:ncol(pca)] <- paste0("PC", 1:(ncol(pca)-1))

pca <- as_tibble(data.frame(pca))

# first convert to percentage variance explained
pve <- data.frame(PC = 1:20, pve = eigenval/sum(eigenval)*100)


# make plot of variance explained
png("PLINK_var_exp_white_3Aug2021.tiff")
a <- ggplot(pve, aes(PC, pve)) + geom_bar(stat = "identity")
a + ylab("Percentage variance explained") + theme_light()
dev.off()
```

```
pca2<-data.frame(pca)
dim(pca2)
firstten<-pca2[,1:11]
dim(firstten)
head(firstten)

# Examine PC1 vs PC2
png("PLINK_pca_white_3Aug21.tiff")
plot(firstten$PC1,firstten$PC2, main = "First 2 PCs with PLINK method, White Only",
xlab="First PC",
    ylab = "Second PC")
dev.off()


require(scatterplot3d)
# Examine PC1, PC2, and PC3

png("PLINK_pca_3d_white_3Aug21.tiff")
scatterplot3d(firstten$PC1
            , firstten$PC2, firstten$PC3, main = "3D Plot of First 3 PCs via PLINK, White
Only")
dev.off()



### PRS
### CREATE PRS


Rscript /vcu_gpfs2/home/cliffordjs/bin/PRSice.R \
 --prsice /vcu_gpfs2/home/cliffordjs/bin/PRSice_linux \
 --base /vcu_gpfs2/home/GfG/sumstats/noheaderfinalSIsumstats.txt \
 --target /vcu_gpfs2/home/GfG/cleangentic/Filtered/SIA_White \
 --binary-target T \
 --pheno /vcu_gpfs2/home/GfG/cleanphenotypic/G4G_ecuse.txt \
 --cov /vcu_gpfs2/home/GfG/cleanphenotypic/White_covars.txt \
 --chr-id c:l \
 --base-maf MAF:0.01 \
 --stat BETA \
 --beta \

setwd("~/Desktop/PRS-Reg")

require(data.table)
require(tidyverse)
```

```
require(fmsb)

p.threshold <- c(0.001,0.05,0.1,0.2,0.3,0.4,0.5)
# Read in the phenotype file
phenotype <- fread("G4G_ecuse.txt", header=F)
ccuse<-fread("G4G_ccuse2.txt", header = F)

# Read in the covariates
covariate <- fread("White_G4G_covars_pcs.txt", header=F)
head(covariate)

# make column names for p and c

colnames(phenotype)<- c("IID", "FID", "EC")
covariate<-covariate[,c(1:6, 8:14)] # Remove the extra 3 PCs
head(covariate)
colnames(covariate)<- c("IID", "FID", "insure", "gender", "education", "age_range",
"PC1","PC2","PC3","PC4","PC5","PC6","PC7")
head(covariate)
colnames(ccuse)<- c("IID", "FID", "CC")
head(ccuse)


# Now merge the files
pheno <- merge(phenotype, covariate,  by=c("IID", "FID"))
table(pheno$EC, useNA = 'always')
phenocc<-merge (pheno, ccuse, by =c("IID", "FID"))

phenonomiss<-na.omit(phenocc)

# Recode ECs to 1 and 0, no missing data
phenonomiss$EC <-ifelse(phenonomiss$EC == 1, 1, 0)

# Recode CCs to 1 and 0, no missing data
phenonomiss$CC <-ifelse(phenonomiss$CC == 1, 1, 0)

# We can then calculate the null model (model without PRS) using a logistic regression

prs <- fread("SI_CIG_PRScs.profile", header=T)

pheno.prs<-merge(phenonomiss, prs[,c("IID","FID", "SCORE")], by=c("IID", "FID"))
head(pheno.prs)

## Histogram of PRS Scores
#tiff("Unnormalized_GPS.tiff")
#hist(pheno.prs$SCORE, xlab= "GPS", main ="")
```

```
#dev.off()

# Table of covars
table(pheno.prs$insure)
table(pheno.prs$gender)
table(pheno.prs$education)
table(pheno.prs$age_range)
table(pheno.prs$EC)
pheno.prs$EC<-factor(pheno.prs$EC, levels = c(0:1))

pheno.prs$CC<-factor(pheno.prs$CC, levels = c(0:1) )
table(pheno.prs$CC)

pheno.prs$gender<-factor(pheno.prs$gender, levels = c(1:2))
table(pheno.prs$gender)
pheno.prs$insure<-factor(pheno.prs$insure, levels =c(1:3))
table(pheno.prs$insure)

pheno.prs$education<-factor(pheno.prs$education, levels = c(1:4))
table(pheno.prs$education)

pheno.prs$age_range<-factor(pheno.prs$age_range, levels=c(1:7))
table(pheno.prs$age_range)

null.model <-
glm(EC~CC+insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC5+PC6+
PC7, family=binomial, data=pheno.prs)

# And the R2 of the null model is
null.r2 <- NagelkerkeR2(null.model)

model <-
glm(EC~SCORE+CC+insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC
5+PC6+PC7, family=binomial, data=pheno.prs)

# model R2 is obtained as
model.r2 <- NagelkerkeR2(model)

# R2 of PRS is simply calculated as the model R2 minus the null R2
prs.r2 <- model.r2$R2-null.r2$R2

# We can also obtain the coeffcient and p-value of association of PRS as follow

prs.coef <- summary(model)$coeff
prs.beta <- as.numeric(prs.coef[1])
prs.se <- as.numeric(prs.coef[2])
```

```
prs.p <- as.numeric(prs.coef[4])

# We can then store the results

prs.result <- rbind(data.frame(R2=prs.r2, P=prs.p, BETA=prs.beta,SE=prs.se))

#write.file(prs.result, "/vcu_gpfs2/home/cliffordjs/PRS_out.txt", quote=F, row.names=F)

### AUC

predprob <- predict(model, type = "response")

library(pROC)
rocCurve <- roc(EC ~ predprob, data = pheno.prs)
tiff("roccurve3302022.tiff")
plot(rocCurve)
dev.off()

auc(rocCurve)

#write.file(rocCurve,"/vcu_gpfs2/home/cliffordjs/AUC_out.txt", quote=F, row.names=F)


## normalized GPS

pheno.prs$scaled<-scale(pheno.prs$SCORE)

## Histogram of PRS Scores
tiff("Normalized_GPS.tiff")
hist(pheno.prs$scaled, xlab= "GPS", main ="")
dev.off()

model2 <-
glm(EC~scaled+insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC5+PC
6+PC7, family=binomial, data=pheno.prs)
summary(model2)


prs.coef <- summary(model2)$coeff
prs.beta <- as.numeric(prs.coef[1])
prs.se <- as.numeric(prs.coef[2])
prs.p <- as.numeric(prs.coef[4])
# model R2 is obtained as
model.r2 <- NagelkerkeR2(model)

# R2 of PRS is simply calculated as the model R2 minus the null R2
```

```
prs.r2 <- model.r2$R2-null.r2$R2

exp(cbind(coef(model), confint(model)))

prs.result <- rbind(data.frame(R2=prs.r2, P=prs.p, BETA=prs.beta,SE=prs.se))


predprob <- predict(model, type = "response")

## Create ROC and AUC figure

library(pROC)
rocCurve <- roc(EC ~ predprob, data = pheno.prs)
tiff("roccurve.tiff")
plot(rocCurve)
dev.off()

auc(rocCurve)




###-----------------------------------------###
###           CIG USE AND PRS               ###
###-----------------------------------------###

ccuse<-fread("G4G_ccuse2.txt", header = F)

# Read in the covariates (here, it is sex)
covariate <- fread("White_G4G_covars_pcs.txt", header=F)
head(covariate)

# make column names for p and c

colnames(ccuse)<- c("IID", "FID", "CC")
head(ccuse)
covariate<-covariate[,c(1:6, 8:14)] # Remove the extra 3 PCs
head(covariate)
colnames(covariate)<- c("IID", "FID", "insure", "gender", "education", "age_range",
"PC1","PC2","PC3","PC4","PC5","PC6","PC7")
head(covariate)

# Now merge the files
ccpheno <- merge(ccuse, covariate, by=c("IID", "FID"))
table(ccpheno$CC, useNA = 'always')
ccphenonomiss<-na.omit(ccpheno)
```

```
# Recode CIGs to 1 and 0, no missing data
ccphenonomiss$CC <-ifelse(ccphenonomiss$CC == 1, 1, 0)

# We can then calculate the null model (model without PRS) using a logistic regression

prs <- fread("SI_CIG_PRScs.profile", header=T)

ccpheno.prs<-merge(ccphenonomiss, prs[,c("IID","FID", "SCORE")], by=c("IID", "FID"))
head(ccpheno.prs)
table(ccpheno.prs$insure)
table(ccpheno.prs$gender)
table(ccpheno.prs$education)
table(ccpheno.prs$age_range)
table(ccpheno.prs$CC)
ccpheno.prs$CC<-factor(ccpheno.prs$CC, levels = c(0:1))

ccpheno.prs$gender<-factor(ccpheno.prs$gender, levels = c(1:2))
table(ccpheno.prs$gender)
ccpheno.prs$insure<-factor(ccpheno.prs$insure, levels =c(1:3))
table(pheno.prs$insure)

ccpheno.prs$education<-factor(ccpheno.prs$education, levels = c(1:4))
table(ccpheno.prs$education)

ccpheno.prs$age_range<-factor(ccpheno.prs$age_range, levels=c(1:7))
table(ccpheno.prs$age_range)

null.model <-
glm(CC~insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC5+PC6+PC7,
family=binomial, data=ccpheno.prs)

# And the R2 of the null model is
null.r2 <- NagelkerkeR2(null.model)

model <-
glm(CC~SCORE+insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC5+P
C6+PC7, family=binomial, data=ccpheno.prs)

# model R2 is obtained as
model.r2 <- NagelkerkeR2(model)

# R2 of PRS is simply calculated as the model R2 minus the null R2
prs.r2 <- model.r2$R2-null.r2$R2

# We can also obtain the coeffcient and p-value of association of PRS as follow
```

```
prs.coef <- summary(model)$coeff["scaled",]
prs.beta <- as.numeric(prs.coef[1])
prs.se <- as.numeric(prs.coef[2])
prs.p <- as.numeric(prs.coef[4])

# We can then store the results

prs.result <- rbind(data.frame(R2=prs.r2, P=prs.p, BETA=prs.beta,SE=prs.se))

ccpheno.prs$scaled<-scale(ccpheno.prs$SCORE)

model2 <-
glm(CC~scaled+insure+gender+education+age_range+PC1+PC2+PC3+PC4+PC5+PC
6+PC7, family=binomial, data=ccpheno.prs)
summary(model2)


prs.coef <- summary(model2)$coeff["scaled",]
prs.beta <- as.numeric(prs.coef[1])
prs.se <- as.numeric(prs.coef[2])
prs.p <- as.numeric(prs.coef[4])
# model R2 is obtained as
model.r2 <- NagelkerkeR2(model)

# R2 of PRS is simply calculated as the model R2 minus the null R2
prs.r2 <- model.r2$R2-null.r2$R2

exp(cbind(coef(model2), confint(model2)))


## Power calcualtion

require(genpwr)
pw<-genpwr.calc(calc="ss", model = "logistic", Case.Rate = 0.066, OR = 1.02,
True.Model = "Additive",
            Test.Model = "Additive", Alpha= 0.05, Power = 0.8, MAF =0.1)

pw

pw2<-genpwr.calc(calc="Power", model = "logistic", Case.Rate = 0.066, OR = 1.02,
True.Model = "Additive",
             Test.Model = "Additive", Alpha= 0.05, N=15881, MAF =0.1)

pw2
```

CHAPTER 5 - SAS

```
* Do an import wizard of SPSS data file;
* Do an import wizard of SPSS weights;

proc contents data = weights;
run;

data path3;
merge path3 weights;
by PERSONID;
run;

data path3;
set path3;
use=.;
ec = 0;
cc = 0;
if R03_AV1002_12M = 1 or R03_AV1004 = 1 then ec = 1; * recodes any past 12-month
use into yes/no;
if R03_AC1002_12M = 1 or R03_AC1004 =1 then cc = 1; * recodes any past 12-month
use into yes/no;
if ec = 1 and cc =1 then use = 3;
if ec = 0 and cc = 1 then use = 2;
if ec = 1 and cc = 0 then use = 1;
if ec = 0 and cc = 0 then use = 0;
run;

 * NOte for use, 1 = EC user, 2 = CC user, 3 = Dual user, 0 = non-user;

 proc format;
 value use
 1 = 'EC user'
 2 = 'CC user'
 3 = 'Dual user'
 0 = 'Non-user';
 run;


 proc format;
 value single
 1 = 'User'
 0 = 'Non-user';
 run;
```

```
proc freq data = path3;
table R03_AC1002_12M R03_AC1004;
run;

proc freq data = path3;
table R03R_A_AM0030 * use;
format use use.;
run;

proc freq data = path3;
table R03R_A_AM0030*R03_AX0708_02;
run;

proc freq data = path3;
table R03R_A_AM0030*R03_AX0708_01;
run;

proc freq data =path3;
table use  ;
format use use. ;
run;

proc freq data= path3;
table R03R_A_AM0030;
run;

proc freq data = path3;
table ec*cc ec2*cc2;
run;
proc freq data = path3;
table use2;
format use2 use.;
run;


proc freq data = path3;
table R03_AV1002_12M * R03_AC1002_12M;
run;

proc freq data = path3;
table R03R_A_SEX R03R_A_AGECAT7 R03R_A_RACECAT3 R03R_A_AM0018;
run;

* create new data set with truncated education, less than hs, hs, some college, college
degree;
```

```
data path4;
set path3;
education = .;
if R03R_A_AM0018 = 1 then education = 1;
if R03R_A_AM0018 = 2 or R03R_A_AM0018 = 3 then education =2;
if R03R_A_AM0018 = 4 then education = 3;
if R03R_A_AM0018 = 5 or R03R_A_AM0018 = 6 then education = 4;
run;

proc format;
value edu
1 = 'less than HS'
2 = 'hs'
3 = 'some college'
4 = 'bs or higher'
;
run;

proc surveyfreq data=path4 varmethod=BRR (fay=0.3);
table use;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;



proc surveyfreq data=path4 varmethod=BRR (fay=0.3);
table cc ec;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format ec single. cc single.;
run;
quit;


proc surveyfreq data=path4 varmethod=BRR (fay=0.3);
table R03R_A_AM0030 R03_AX0708_01 R03_AX0708_02;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;
```

```
* Multinomial regression of income and tobacco use;

proc surveylogistic data=path4 varmethod=BRR (fay=0.3);
class use (ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')R03R_A_SEX
(ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;


* Multinomial regression code, note that link = glogit;
* EC coupons;

proc surveylogistic data=path3 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_02 (ref='2 = Not Marked')/ param=ref;
model use = R03R_A_AM0030 R03_AX0708_02/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;



* adjusted for age, sex, race, education;

proc surveylogistic data=path3 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_02 (ref='2 = Not Marked')
R03R_A_SEX (ref = '1 = Male') R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old')
R03R_A_RACECAT3 (ref = '1 = White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03_AX0708_02 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
```

```
run;
quit;

** Moderation model for EC coupons;

proc surveylogistic data=path4 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_02 (ref='2 = Not Marked')
R03R_A_SEX (ref = '1 = Male') R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old')
R03R_A_RACECAT3 (ref = '1 = White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03_AX0708_02 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education R03R_A_AM0030*R03_AX0708_02/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;

** Stratified model for EC coupons;

proc sort data=path4 out=path5;
by R03_AX0708_02;
run;

proc freq data=path5;
table R03_AX0708_02;
run;


proc surveylogistic data=path5 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03R_A_SEX (ref = '1 = Male') R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old')
R03R_A_RACECAT3 (ref = '1 = White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education /link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
domain R03_AX0708_02; * this allows for multiple results from regression, not where or
by;
run;
quit;
```

```
* CC Coupons;

proc surveylogistic data=path3 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_01 (ref='2 = Not Marked')/ param=ref;
model use= R03R_A_AM0030 R03_AX0708_01/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;

proc surveylogistic data=path4 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_01 (ref='2 = Not Marked')
R03R_A_SEX (ref = '1 = Male') R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old')
R03R_A_RACECAT3 (ref = '1 = White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03_AX0708_01 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;

** Moderation CIG coupons;

proc surveylogistic data=path4 varmethod=BRR (fay=0.3);
class use(ref='Non-user') R03R_A_AM0030(ref='5 = $100,000 or more')
R03_AX0708_01 (ref='2 = Not Marked')
R03R_A_SEX (ref = '1 = Male') R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old')
R03R_A_RACECAT3 (ref = '1 = White alone')
education (ref = '4')/ param=ref;
model use = R03R_A_AM0030 R03_AX0708_01 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education R03R_A_AM0030*R03_AX0708_01/link=glogit;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
format use use.;
run;
quit;
```

```
** Logistic regressions;
** EC Coupons;

proc surveylogistic data = path3 varmethod=BRR (fay=0.3);
class  R03R_A_AM0030(ref='5 = $100,000 or more')/ param=ref;
model ec (event='User') = R03R_A_AM0030/link=logit;
format ec single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;

* adjusted analyses;
proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class  R03R_A_AM0030(ref='5 = $100,000 or more')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model ec (event='User') = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education/link=logit;
format ec single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;


proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class R03R_A_AM0030(ref='5 = $100,000 or more') R03_AX0708_02 (ref='2 = Not
Marked')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model ec (event='User')= R03R_A_AM0030 R03_AX0708_02 R03R_A_SEX
R03R_A_AGECAT7 R03R_A_RACECAT3 education/link=logit;
format ec single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;



** CC Coupons;

proc surveylogistic data = path3 varmethod=BRR (fay=0.3);
```

```
class  R03R_A_AM0030(ref='5 = $100,000 or more')/ param=ref;
model cc (event='User') = R03R_A_AM0030/link=logit;
format cc single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;

proc surveylogistic data = path3 varmethod=BRR (fay=0.3);
class R03R_A_AM0030(ref='5 = $100,000 or more') R03_AX0708_01 (ref='2 = Not
Marked')/ param=ref;
model cc (event='User')= R03R_A_AM0030 R03_AX0708_01/link=logit;
format cc single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;


* adjusted analyses;
proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class  R03R_A_AM0030(ref='5 = $100,000 or more')R03R_A_AM0030(ref='5 =
$100,000 or more')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model cc (event='User') = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education/link=logit;
format cc single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;

proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class R03R_A_AM0030(ref='5 = $100,000 or more') R03_AX0708_02 (ref='2 = Not
Marked')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model cc (event='User')= R03R_A_AM0030 R03_AX0708_02 R03R_A_SEX
R03R_A_AGECAT7 R03R_A_RACECAT3 education/link=logit;
format cc single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
```

```
quit;

proc surveyfreq data=path4 varmethod=BRR (fay=0.3);
table R03R_A_SEX R03R_A_AGECAT7 R03R_A_RACECAT3 education;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;

* Interaction model;

proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class R03R_A_AM0030(ref='5 = $100,000 or more') R03_AX0708_02 (ref='2 = Not
Marked')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4')/ param=ref;
model cc (event='User')= R03R_A_AM0030 R03_AX0708_02 R03R_A_SEX
R03R_A_AGECAT7 R03R_A_RACECAT3 education
R03R_A_AM0030*R03_AX0708_02/link=logit;
format cc single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;


**** Covariate MODEL;
** CC outcome first, adding EC use as a covariate;
* adjusted analyses;

proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class  R03R_A_AM0030(ref='5 = $100,000 or more')R03R_A_AM0030(ref='5 =
$100,000 or more')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4') ec (ref='1')/ param=ref;
model cc (event='User') = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education ec/link=logit;
format cc single. ec single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;

** EC outcome second, adding CC use as a covariate;
```

```
* adjusted analyses;

proc surveylogistic data = path4 varmethod=BRR (fay=0.3);
class  R03R_A_AM0030(ref='5 = $100,000 or more')R03R_A_AM0030(ref='5 =
$100,000 or more')R03R_A_SEX (ref = '1 = Male')
R03R_A_AGECAT7 (ref = '1 = 18 to 24 years old') R03R_A_RACECAT3 (ref = '1 =
White alone')
education (ref = '4') cc (ref='User')/ param=ref;
model ec (event='User') = R03R_A_AM0030 R03R_A_SEX R03R_A_AGECAT7
R03R_A_RACECAT3 education cc/link=logit;
format cc single. ec single.;
weight R03_A_SWGT;
repweights R03_A_SWGT1 - R03_A_SWGT100;
run;
quit;
```

VITA

James Samuel Clifford was born August 4, 1983 in Cincinnati Ohio. He graduated from Robert E. Lee High School in Springfield, VA and went on to earn a Bachelor of Science in Psychology from Virginia Tech in 2005 and a Master of Science in Experimental Psychology from Radford University in 2007. He has held research assistantships at the University of Kentucky and Brown University. He also has been a faculty member at East Coast Polytechnic Institute University and an adjunct faculty member in the School of Social Work at Virginia Commonwealth University.

ACADEMIC APPOINTMENT HISTORY

2021 – Present

Adjunct Instructor
School of Social Work
Virginia Commonwealth University, Richmond, VA
- SLWK 611: Social Work Research for Advanced-Standing Students
- SLWK 706: Research in Clinical Social Work Practice I
- SLWK 707: Research in Clinical Social Work Practice II

2015 – 2017

Psychology/Mathematics Faculty
Department of Arts and Sciences
ECPI University, Richmond, VA
Courses taught:
- PSY 105: Introduction to Psychology
- PSY 220: Positive Psychology
- MTH 140: Introduction to Statistics
- CAP 480: Arts & Sciences Capstone
- FOR 110: Essentials for Success

2010 – 2011

Research Assistant
Center for Alcohol and Addiction Studies
Division of Biology and Medicine
Brown University, Providence, RI
Researched personalized medicine based on genotype
Mentor: George A. Kenna, Ph.D., R.Ph.

2007-2008

Research Assistant
Center for Drug and Alcohol Research
Department of Behavioral Sciences
University of Kentucky, Lexington, KY

Researched sex differences on pain perception under opiate influences

2006-2007 Graduate Teaching Fellow
Department of Psychology
Radford University, Radford, VA
Courses taught:
PSY 121, Introduction to Psychology

PUBLICATIONS

Peer Reviewed Articles

Clifford, J.S., Lu, J., Do, E.K., Burkhardt, B. Prom-Wormley, E.C. (2021). The association of health literacy and tobacco use: Results from a nationally representative survey. Accepted by the *Journal of Community Health.* Online Ahead of Print. PMID: 34357496.

Cooke, M.E., **Clifford, J.S.**, Do, E.K., Gilman, J., Maes, H.H., Peterson, R.E., Prom-Wormley, E.C., Spit for Science Working Group, Evins, A.E., & Schuster, R.M. (2021). Conventional cigarette polygenic score is associated with E-cigarette use in youth. Accepted by *Addiction*. Online Ahead of Print. PMID: **34636095.**

Usidame, B., Gibson, E.M., Diallo, A., Blondino, C., **Clifford, J.**, Richmond Health and Wellness Community Advisory Board, Zanjani, F., Sargeant, L., Price, E., Slattum, P., Parson, P., & Prom-Wormley, E. (2021). Understanding the preference for receiving mental health and substance use support in African Americans 50 and older. *Journal of Prevention and Intervention in the Community*. PMID: 34053408.

Blondino, C.T., Lu, J., **Clifford, J.S.**, Prom-Wormley, E.C. (2021). The Association between Internalizing and Externalizing Severity with Current Use of Cigarettes, E-cigarettes, and Alcohol in Adults: Wave 1 of the Population Assessment of Tobacco and Health (PATH) Study. *Addictive Behaviors*, *119.* PMID 33901812.

Verhulst, B., Pritikin, J., **Clifford, J.**, & Prom-Wormley, E. (2021). Using genetic marginal effects to study gene-environment interactions with GWAS data. *Behavior Genetics*, *51*, 358-373. PMID: 33899139

Prom-Wormley, E.C.*, **Clifford, J.S.*,** Cooke, M.E., Cecilione, J., Maes, H.H., Do, E.K., & Roberson-Nay, R. (2021). The genetic and environmental contributions to electronic cigarette initiation in a genetically informative sample of young adults. *Nicotine and Tobacco Research*, *23*(5), 856-860. PMID: **33017842.**
* Denotes co-first author

Daly, N., Parsons, M. Johnson, A., Blondino, C., **Clifford, J.S.**, Prom-Wormley, E.C. (2020).
Association between caregiver depression and child after-school program enrollment. *Journal of*
Family Social Work. DOI: https://doi.org/10.1080/10522158.2020.1824954

Blondino, C.T., Gormley, M.A., Taylor, D.S.D.H., Lowery, E., **Clifford, J.**, Burkart, B., Graves, W.C., Lu, J., & Prom-Wormley, E.C. (2020). The association of co-occurring substance use and the effectiveness of opiate treatment programs by intervention type: A systematic review. *Epidemiologic Reviews.* PMID: 32944731.

Gormley, M., Blondino C., Taylor, D., Lowery, E., Graves, W., **Clifford, J.S.,** Prom-Wormley, E.C., & Lu, J. (2020). Assessment of polysubstance use during opioid use disorder treatment in the United States: A systematic review. *Epidemiologic Reviews*. PMID: 33063108.

Do, E.K., Nicksic, N.E., **Clifford, J.S.**, Hayes, A., Fuemmeler, B.F. (2020). Perceived harms of and exposure to tobacco product use, and cigarette, e-cigarette, and dual use among reproductive-aged women from PATH (Waves 1 & 2). *Women & Health*, *60*(9), PMID: 32654622.

Prom-Wormley, E.C., **Clifford, J.S.,** Bourdon, J., Barr, P., Blondino, C., Ball, K. M., Montgomery, J., Davis, J. K., Real, J.E., Edwards, A., Thiselton, D., Wilson, D., Creighton, G.C., & Newbille, C. (2019). Developing community-based strategies with family health history education: Assessing the association between community resident family history and interest in health education. *Social Science and Medicine*. doi: 10.1016/j.socscimed.2019.02.011. PMID: 30862375.

Kenna, G. A., Zywiak, W. H., Swift, R. M., McGeary, J. E., **Clifford, J. S.**, Shoaff, J. R., . . . Leggio, L. (2014). Ondansetron and sertraline may interact with 5-HTTLPR and DRD4 polymorphisms to reduce drinking in non-treatment seeking alcohol-dependent women: Exploratory findings. *Alcohol, 48*(6), 515-522. doi: 10.1016/j.alcohol.2014.04.005. PMID: 25212749.

Kenna, G.A., Zywiak, W.H., Swift, R.M., McGeary, J.E., **Clifford, J.S.**, Shoaff, J.R., Vuittonet, C., Fricchione, S., Brickley, M., Beaucage, K., Haass-Koffler, C.L., & Leggio, L. (2014). Ondansetron reduces naturalistic drinking in non-treatment seeking alcohol dependent individuals with LL *5'-HTTLPR* alleles: A laboratory study. *Alcoholism: Clinical and Experimental Research*, *38*(6), 1567-1574. PMID: 24773166.

Dick, D.M., Nasim, A., Edwards, A., Salvatore, J., Adkins, A., Meyers, J., Yan, J., Cooke, M., **Clifford, J.**,Goyal, N. Halberstadt, L., Ailstock, K., Neale, Z., Opalesky, J., Hancock, L., Donovan, K., Kendler, K.S. (2014). Spit for science: Launching a longitudinal study of genetic and environmental influences on substance use and

emotional health at a large US university. *Frontiers in Behavioral and Psychiatric Genetics*, *5*, 1-12. PMID: 24639683.

Kenna, G.A., Roder-Hanna, N., Leggio, L., Zywiak, W.H., **Clifford, J.**, Edwards, S., Kenna, J.A., Shoaff, J., Swift, R.M. (2012). The association of the 5-HTT gene linked promoter region (5-HTTLPR) polymorphism to psychiatric disorder: A review of psychopathology and pharmacotherapy. *Pharmacogenetics and Personalized Medicine*, 2012(5), 19-35. PMID: 23226060.

Book Chapters

Prom-Wormley, E., Langi, G., **Clifford, J.**, & Real, J. (2016). Understanding the roles of genetic and environmental influences on the neurobiology of nicotine use. In R.R. Watson, *Addictive*
Substance and Neurological Disease. Elsevier: San Diego, CA.

Accepted Papers

Liu, A., Cox, C., Sankoh, M., **Clifford, J.**, Blondino, C.T., Richarson-Lauve, J., Bae, C., Turner, C., Miles, C., Young, K., Gillison-Chew, S., Prom-Wormley, E. (Under Review). The Association Between Loneliness and Increased Mental Health Problems and Substance Use during the COVID-19 Pandemic in Richmond, Virginia. Submitted to the *Virginia Journal of Public Health*.

PRESENTATIONS

Wilson, T.L., **Clifford, J.S**., Blondino, C.T., Prom-Wormley, E.C. (2020, May). *Examining the Association between Race and Mental Health on Lifetime Frequency of E-nicotine use in U.S. adults.* 2020 Virginia Public Health Association Virtual Poster Presentation, Virtual Presentation. https://vapha.org/Virtual-Poster-Presentation-2020/116047050#photo

Mulroy, N.M., **Clifford, J. S.**, Prom-Wormley, E.C. (2020, February). *Association between Marital Status and Smoking Abstinence*. Poster presented at the 27th annual meeting of the Society for Research on Nicotine and Tobacco.

Brown C.K., **Clifford J.S.**, Blondino C.T., & Prom-Wormley E.C. (2020, April). *Utilization of care coordination and routine location of healthcare services in the East End community of Richmond, VA*. 2020 International/Inner City/Rural Preceptorship Program Capstone Reception, Richmond, VA.

Clifford, J.S., Wilson, T., Blondino, C.T., Prom-Wormley, E.C. (2020, March). *The effect of electronic and conventional cigarette coupon receipt on the relationship between income level and past 12-month use in adults in PATH.* Poster presented at

the 26th annual meeting of the Society for Research on Nicotine and Tobacco, New Orleans, LA.

Usidame, B., **Clifford, J.**, Blondino, C., Ball, K., & Prom-Wormley, E. (2020, March). *The association between e-cigarette use and mental health symptoms in American adults.* Poster presented at the 26th annual meeting of the Society for Research on Nicotine and Tobacco, New Orleans, LA.

Taylor, D., Gormley, M., Blondino, C., Lowery, E., **Clifford, J.**, Burkart, B., Devanaboyina, M., Graves, W., Chapman, D., Lu, J., & Prom-Wormley, E. (2019, November). *An examination of polysubstance use in pregnant women during opioid use disorder treatment: A systematic review of studies conducted in the United States.* Poster presented at the annual meeting of the American Public Health Association, Philadelphia, PA.

Gormley, M., Blondino C., Taylor, D., Lowery, E., Graves, W., **Clifford, J.S.,** Prom-Wormley, E.C., & Lu, J. (2019, June). *Assessment of polysubstance use during opioid use disorder treatment in the United States: A systematic review.* Poster presented at the annual meeting of the Society for Epidemiologic Research, Minneapolis, MN.

Blondino, C.T., Gormley, M., Taylor, D.H., Lowery, E., **Clifford, J**., Burkart, B., Graves, W., Lu, J., & Prom-Wormley, E.C. (2019, June). *The impact of polysubstance use on the effectiveness of opioid use disorder therapy by treatment type: A systematic review.* Poster presented at the annual meeting of the Society for Epidemiologic Research, Minneapolis, MN.

Do, E.K., Nicksic, N.E., **Clifford, J.S.**, Fuemmeler, B.F. (2019, June). *Perceived harms of and exposure to tobacco product use and current cigarette, e-cigarette, and dual use among reproductive-aged women from the PATH study (Waves 1 &2)*. Paper presented at the Massey Cancer Center Cancer Research Retreat, Richmond, VA.

Clifford, J.S., Lu, J, Prom-Wormley, E.C. (2019, February). *The association of health information literacy and the use of conventional and electronic cigarettes*. Poster presented at the annual
meeting of the Society for Research on Nicotine and Tobacco, San Francisco, CA.

Clifford, J.S., Ball, K.M., Blondino, C., Prom-Wormley, E.C. (2018, April). *The associations between conventional and electronic cigarette use with tobacco messaging*. Poster presented at the annual conference of the Virginia Public Health Association, Lynchburg, VA.

Ball, K.M., **Clifford, J.**, Blondino, C. Do, E., Maes, H., Prom-Wormley, E.C. (2018, March). *Understanding the associations between opinions towards tobacco and youth current poly-tobacco use.* Poster presented at the 6th triennial conference of Virginia Conference on Youth Tobacco Use, Richmond, VA.

Clifford, J.S., Cecilione, J., Cooke, M. E., Roberson-Nay, R., Prom-Wormley, E.C. (2018, February). *The genetic and environmental contributions to electronic and conventional cigarette use in young adults*. Paper presented at the annual meeting of the Society for Research on Nicotine and Tobacco, Baltimore, MD.

Prom-Wormley, E.C., **Clifford, J.S.**, Cecilione, J., Cooke, M.E., Roberson-Nay, R. (2018, January). *The genetic and environmental contributions to electronic and conventional cigarette use in young adults: A bivariate analysis*. Poster presented at the annual NIDA Genetics Consortium Meeting, Rockville, MD.

Kenna, G.A., Swift, R.M., Zywiak, W., McGeary, J., **Clifford, J.S.**, Shoaff, J., Fricchione, S., Brickley, S.,Beaucage, M., Haas-Koffler, C., Leggio, L. (2014, October). *5-HTTLPR, DRD4 alleles, sertraline and ondansetron interact to reduce drinking non-treatment seeking alcohol dependent women.* Paper to be presented at the International Society of Psychiatric Genetics, Copenhagen, Denmark.

Clifford, J.S., Adkins, A.E., Dick, D.M., Kendler, K.S., Gillespie, N.A. (2014, June). *Multivariate GWAS of alcohol phenotypes in a college-aged sample.* Poster at the 44th annual meeting of the Behavioral Genetics Association Meeting, Charlottesville, VA.

Kenna, G.A., Swift, R.M, Zywiak, W., McGeary,J**., Clifford, J.S.**, Shoaff, J., Fricchione, S., Brickley.M,, Beaaucage, K., Vuittonet, C., Haass-Koffler, C., Leggio,L (2013, December). *A trial matching and mismatching Ondansetron and Sertraline to 5-HTTLPR alleles in non-treatment seeking alcohol dependent individuals*. Paper presented at the American College of Neuropsychopharmacology, Hollywood FL.

Fricchione, S., Brickley, M., Zywiak, W., McGeary, J., Beaucage, K., **Clifford, J**., Shoaff, J., Haass-Koffler, C., Swift, R.M., Leggio, L. Kenna, G.A**.** (2013, October). *A study matching serotonergic drugs to LL vs SS/SL 5'-HTTLPR alleles in non-treatment seeking alcohol dependent individuals*. Poster presented at the 52nd Annual New England Psychological Association, Bridgeport CT.

Brickley, M., Fricchione, S., Leggio, L., Zywiak, W., McGeary,J., Swift, R.M., **Clifford, J.S.**, Shoaff, J., Beaucage,K., Haass-Koffler, C. Kenna, G.A. (2013, October). *Interaction of Ondasetron and Sertraline based on 5-HTTLPR and D4 alleles on drinking in non-treatment seeking alcohol dependent women.* Poster presented at the 52nd Annual New England Psychological Association, Bridgeport CT.

Kenna, G.A., Zyqwiak, W.H., McGeary, J.E., Swift, R.M., **Clifford, J.S.**, Shoaff, J., Brickley, M., Vuittonet, C., Edwards, S., Tavares, T., Fricchione, S., McGeary, C., Beaucage, K., Haass-Koffler, C., & Leggio, L. (2013, September). *Personalized treatment matching of serotonergices based on 5'-HTTLPR and DRD4 in non-treatment seeking alcoholics.* Paper presented at the 14th Congress of the European Society for Biomedical Research in Alcoholism, Warsaw, Poland.

Clifford, J.S., Dick, D.M., Aggen, S.H., Gardner, C.O., Kendler, K.S., & Gillespie, N.A. (2013). *Peer group deviance and alcohol use: Causal models*. Poster presented at the 36th annual meeting of the Research Society on Alcoholism, Orlando, Florida.

Clifford, J.S., Dick, D.M., Aggen, S.H., Gardner, C.O., Kendler, K.S., & Gillespie, N.A. (2013). *Peer group deviance and cannabis use: Causal models.* Poster presented at the 61st Annual Symposium on Motivation, Lincoln, Nebraska.

Kenna, G.A., Leggio, L., Zywiak, W.H., McGeary, J.E., Swift, R.M., **Clifford, J.S.**, Shoaff, J., Edwards, S, Tavares, T., Fricchione, S., McGeary, C., Beaucage, K. (2012, September). *Serotonigenetic matching and mis-matching ondansetron and sertraline based on 5'-HTTLPR alleles in non-treatment seeking alcoholics.* Presented at the 2012 International Society for Biomedical Research on Alcoholism Congress in Sapporo, Japan.

Kenna, G.A., Leggio, L., Zywiak, W., McGeary,J., Swift, R.M., **Clifford, J.S**., Shoaff, J., Edwards, S., Tavares, T., Fricchione, S., McGeary,C.(2012, June). *Interaction of serotonigentic pharmacotherapies based on 5-HTTLPR and D4 alleles on drinking in non-treatment seeking alcohol dependent women.* Poster presented at the 35th annual meeting of the Research Society on Alcoholism, San Francisco, California.

Shoaff, J.R., Leggio, L., Zywiak, W., McGeary,J., Swift, R.M., **Clifford, J.S**., Edwards, S., Tavares, T., Fricchione, S., McGeary, C., Kenna, G.A. (2012, June). *Effect of sertraline on alcohol craving and consumption in the late luteal phase of non-treatment seeking alcohol dependent women.*
Poster presented at the 35th annual meeting of the Research Society on Alcoholism, San Francisco, California.

Clifford, J.S., Tavares, T., Shoaff, J., Fricchione, S., Edwards, S., Leggio, L., McGeary, J.E., Swift, R.M., Zywiak, W.H., Kenna, G.A. (2011, October). *Sertraline and ondansetron as treatment for alcohol dependence in 5-HTTLPR genotyped nontreatment seeking subjects.* Poster presented at the 51st annual meeting of the New England Psychological Association, Fairfield, Connecticut.

Tavares, T., **Clifford, J.S.**, Edwards, S., Fricchione, S.R., Leggio, L., Kenna, G.A., Zywiak, W., McGeary J., Swift, R.M. (2011, Oct.). *The relationship between stress-related growth hormone, alcohol craving, and alcohol consumption in response to GABA-B receptor agonist Baclofen in a human laboratory pilot study*. Poster presented at the 51st annual meeting of the New England
Psychological Association, Fairfield, Connecticut.

Edwards S.M., **Clifford J.**, Fricchione S.R., Tavares T., Tidey J., Kenna G.A., Zywiak W.H., Swift R.M., Leggio L. (2011, October). *Are smokers and drinkers happy? A relationship between subjective happiness and drinking and smoking levels*. Poster presented at the 51st annual meeting of the New England Psychological association, Fairfield, Connecticut.

Fricchione, S., Edwards, S., Tavares, T., **Clifford, J.**, Ferrulli, A., Miceli, A., Addolorato, G., Kenna, G., Swift, R., Leggio, L. (2011, October). *Plasma ghrelin levels correlate with alcohol drinking and craving in alcohol-dependent subjects.* Poster presented at the 51st annual meeting of the New England Psychological Association, Fairfield, Connecticut.

Shoaff, J., Tavares, T., **Clifford, J.S.**, Swift, R.M., Leggio, L., Kenna, G., McGeary, J., Zywiak, W., Grenga, A., Kim, J. (2011, October). *Stimulant effects of alcohol and alcohol craving: Is there a relationship?* Poster presented at the 51st annual meeting of the New England Psychological Association, Fairfield, Connecticut.

Clifford, J.S., Edwards, S., Leggio, L., Swift, R.M., Kenna, G.A. (2011, June). *Shortening the Alcohol Use Disorder Identification Test (AUDIT) based on self-report alcohol use.* Poster presented at the 34th annual Research Society on Alcoholism conference, Atlanta, Georgia.

Edwards, S., **Clifford, J.**, Kenna, G.A., Zywiak, W.H., Swift, R., Leggio, L. (2011, June). *Biobehavioral and psychoneuroendocrine mechanisms of Baclofen and alcohol drinking – preliminary findings.* Poster presented at the 34th annual Research Society on Alcoholism conference, Atlanta, Georgia.

Clifford, J.S., Tavares, T., Leggio, L., Swift, R., McGeary, J.E., Zywiak, W.H., Kenna, G.A. (2011, June). *A within-groups design of nontreatment seeking 5-HTTLPR genotyped alcohol-dependent subjects receiving ondansetron and sertraline: Future directions.* Poster presented at the 41st annual meeting of the Behavioral Genetics Association, Newport, Rhode Island.

Tavares, T., Swift, R.M., Kenna , G.A., Leggio, L., McGeary, J., Zywiak,W., **Clifford, J.S.**, Grenga, A., Kim J. (2011, June). *Genetic Modulators of Alcohol Stimulation and Craving in Humans.* Poster presented at the 41st annual meeting of the Behavioral GeneticsAssociation, Newport, Rhode Island.

Kenna, G.A., Leggio, L., Zywiak, W., Edwards, S., **Clifford, J.**, Swift, R. (2011, May). *Relationship between stress-related hormones, alcohol craving and drinking in a human laboratory paradigm: A pilot study.* poster presented at the Alcoholism and Stress Conference, Volterra Italy.

Clifford, J.S., Tavares, T., Edwards, S., Leggio, L., Swift, R., Kenna, G.A. (2011, April). *Consumption scores on the AUDIT and self-report alcohol use: A quicker method.* Poster presented at the Public Health Research Day at Brown University, Providence, Rhode Island
.
Tavares, T., Swift, R.M., Kenna, G., Leggio, L., McGeary, J., Zywiak, W., **Clifford, J.S.**, Grenga, A., Kim, J. (2011,April). *Stimulant Effects of Alcohol, Alcohol Craving, and Genetic Variants: Is there a Relationship?* Poster presented at the Public Health Research Day at Brown University, Providence, Rhode Island.

Edwards, S., **Clifford, J**., Tavares, T., Kenna, G.A., Zywiak W.H., Swift, R.M., Leggio, L. (2011, April). *Baclofen and Alcohol Drinking: A Pilot Study on the Biobehavioral Mechanisms.* Poster presented at the Public Health Research Day Conference at Brown University, Providence, Rhode Island.

Clifford, J.S., Edwards, S., Leggio, L., Swift, R., Kenna, G.A. (2011, March). *Validation of the Alcohol UseDisorders Identification Test and self-report alcohol use.* Poster presented at the 2011 meeting of the Eastern Psychological Association, Cambridge, Massachusetts.

Edwards, S., **Clifford, J**., Kenna, G.A., Zywiak W.H., Swift, R.M., Leggio, L. (2011, March). *A Pilot Study on the Biobehavioral Mechanisms of Baclofen and Alcohol Drinking.* Poster presented at the 2011 meeting of the Eastern Psychological Association, Cambridge, Massachusetts.

Clifford, J.S., Edwards, S., Tavares, T., Leggio, L., Swift, R., Kenna, G.A. (2011, March). *Self-report alcohol use and consumption scores of the AUDIT: A good predictor?* Poster presented at the 14th annual research symposium on Mental Health Sciences, The Alpert Medical School of Brown University, Providence, RI.

Tavares, T., Swift, R.M., Kenna, G., Leggio, L., McGeary, J., Zywiak, W., **Clifford, J.S.**, Genga, A., Kim, J., Ciminelli, N. (2011, March). *Relationship between the stimulant effects of alcohol and alcohol craving.* Poster presented at the 14th annual research symposium on Mental Health Sciences, The Alpert Medical School of Brown University, Providence, Rhode Island.

Leggio, L., Ferrulli, A., Cardone, S., Edwards, S., **Clifford, J.**, Kenna, G.A., Swift, R.M., Addolorato, G.(2010, October). *New Developments in the Pharmacotherapy of Alcohol Dependence.* Poster presented at the 36th International Medical Advisory Group Conference (IMAG), Frascati, Italy.

Clifford, J.S. & Aspelmeier, J.E. (2007, April). *Terror management theory and self-esteem boosts.* Paper presented at the annual Graduate/Undergraduate Research Forum at Radford University, Radford, Virginia.

Caincross, D., **Clifford, J.S.**, Aspelmeier, J.E. (2007, April). *Attachment and disgust sensitivity.* Poster presented at the annual Graduate/Undergraduate Research Forum at Radford University, Radford, Virginia.

Clifford, J.S., Woleslagle, A., Graninger, L.L., Caron, J., Hylton, K.R., Kerr, J.L., Barnhardt, J., Harris, M., Mackey, E.M. (2005, April). *Sensation seeking and risky driving on a simulated driving task*. Paper presented at the semi-annual meeting of the Virginia Psychological Association, Williamsburg, Virginia.

Hickman, J.S., **Clifford, J.S.**, Hintz, L.M., Pavlak, S.L., Caron, J., Policay, A., Kerr, J.L., Nash, T., Barnhardt, J., Camden, M. (2005, April). *Self-management for safety on a simulated driving task; objective feedback versus self-monitoring.* Paper presented at the semi-annual meeting of the Virginia Psychological Association, Williamsburg, Virginia.

Williamson, K.A., Padgett, M.M., Progen, M., **Clifford, J.S.**, & Valentino, S.E. (2004, December). *Assessing the effect of designated drivers on passengers' levels of intoxication.* Poster presented at the 7th annual meeting of the Maryland Association for Behavior Analysis, Baltimore, Maryland.

Ehrhart, I.J., **Clifford, J.S.**, Geller, E.S., Rayne, S.R., Dula, C.S. (2004, May). *The effect of using a courtesy communication code on driving behaviors.* Paper presented at the 30th annual convention of the Association of Behavior Analysis, Boston, Massachusetts.

Clifford, J.S., Kimbel, H.L., Robichaux, C.B., Smith, R. (2004, April). *What's in your CD case? Exploratory research on lyrical content of music and driving behaviors.* Paper presented at the semi-annual meeting of the Virginia Psychological Association, Roanoke Virginia.

Kimbel, H.L., Robichaux, C.B., **Clifford, J.S.**, Mascio, C., Andrews, R. (2004, April). *Conversational aspects of driving: Measuring the use of cell phones and upper level processing on driver performance.* Paper presented at the semi-annual meeting of the Virginia Psychological Association, Roanoke Virginia.

Clifford, J.S., Dula, C.S., Geller, E.S. (2003, October). *Have You Been Flashed Lately? The effect of a courtesy communication code on driving behaviors.* Paper presented at the semi-annual meeting of the Virginia Psychological Association, Charlottesville, Virginia.

Wiegand, D.M., Clarke, S.W., Burtner, M.L., Turner, P.J., **Clifford, J.S.** (2003, April). *Factors related to intoxication levels of pre-game tailgaters at an NCAA Division 1 football game.* Paper presented at the semi-annual meeting of the Virginia Psychological Association, Tyson's Corner, Virginia.