Genetic Influences on Injection Drug Use

by
Shao-Cheng Wang

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

December, 2015

Abstract

BACKGROUND: The goal of this dissertation was to create a risk phenotype which could be used to represent the intensity of injection drug use and to link this phenotype to genetics. In paper 1, a review of the literature suggests that a) injection drug users are at high risk for HIV and are at the highest risk for dependence among drug users, b) substance use disorders are chronic and recurrent, and c) genetic factors influence opioid, and other drug, addiction. In paper 2, I devised a risk phenotype, injection-years. In paper 3, I conducted genome wide analysis using injection years as the risk phenotype.

METHODS: Data for paper 2 and paper 3 came from the AIDS Linked to the Intravenous Experience (ALIVE) cohort. Paper 2 deals with the problem of missing data in a longitudinal data set, using three different imputation models and examining the sensitivity. Using the imputed ALIVE GWAS cohort I created the risk phenotype, termed "injection years". In paper 3, I conducted three analyses: genome wide association analysis, polygenic risk score analysis, and pathway analysis to explore the association between injection years and genetics.

RESULTS: Our results concerning injection years are consistent with Genberg's findings, which used the same ALIVE cohort. This result suggested that injection years could be used as a tool to measure the intensity of injection behavior. In genome wide analysis, no significant single SNP or gene sets were found. The findings suggest that injection years are influenced by polygenes.

CONCLUSION: Using the appropriate model, multiple imputation can provide reliable information for longitudinal data. Efforts to identify injection drug users who are most likely to be “persistent users” and to identify related genes could help decrease the public health burden and improve personal health.

Thesis Committee

***Committee Members*:**

Brion Maher (Advisor), Associate Professor, Department of Mental Health

Gregory Kirk (Committee Chair), Associate Professor, Department of Epidemiology

Peter Zandi, Associate Professor, Department of Mental Health

Kelly Dunn, Assistant Professor, Department of Psychiatry and Behavioral Sciences

***Alternates*:**

Priya Duggal, Associate Professor, Department of Epidemiology

Rashelle Musci, Assistant Professor, Department of Mental Health

Fernando Goes, Assistant Professor, Department of Psychiatry and Behavioral Sciences

Acknowledgements

To write this acknowledgement, I recalled the fabulous training and excellent mentorships and I am extremely grateful to so many people for their support, guidance, and inspiration for the past four years. This dissertation would not have been possible without your help and I want to recognize each of you individually.

Brion: I remember the first time you gave me the ALIVE GWAS data and taught me how to conduct genome wide association analysis; that moment is really the beginning of my academic career. Without your support and guidance, I could not have overcome stress and frustration to learn the beauty of genetic statistics, even to pass the preliminary oral examination and to complete this dissertation. You are a much more wonderful mentor and advisor than my expectation, and I would like to pay it forward and be a good mentor like you.

Peter: During my first year, every student is required to take your course called "Introduction to Behavioral and Psychiatric Genetics". This course illuminated my knowledge of the psychiatric genetics in the past, the present, and the future. In my third and fourth year, I was the teaching assistant of this course and I learned from you how to schedule a course and to present in class, particularly using non-native language. You led me to the field of psychiatric genetics and I am grateful for your support.

Kirk, Priya, Kelly, and Rashelle: I am deeply grateful to all other my committee members. Dr. Gregory Kirk, who serves as one of the principal investigators in the ALIVE study, without whom I would never have had the opportunity to work with the

ALIVE cohort, Dr. Priya Duggal, who served as the instructor of "Principles of Genetic Epidemiology" and also a study collaborator with the ALIVE study, who provided a methodological support and advice, Dr. Kelly Dunn, your perspective on opioid treatment was invaluable, and Dr. Rashelle Musci, who served on my defense committees and provided advice on latent trajectory models.

Classmates and friends at Hopkins: It is really difficult to study at Hopkins, number one school in public health in the U.S. for twenty years, especially when you are a foreigner. Having a supportive cohort of colleagues like you has made learning and living in Baltimore less difficult. I am particularly grateful to Chris Kaufmann and Kate Musliner. They were my private English consultants and shared their experience with me in academic writing. I hope that we can continue to collaborate in the future.

Mom, Dad, My wife and my daughter: Studying abroad is really burdensome and exhausting, but your love and support helped me get through the good and bad times. My wife brings out the best in me. When I didn't believe in myself, you always told me that I could do it and comforted me with a delicious meal. My daughter's smile encouraged me to carry through to the end; though you always woke me up with your magic, summoning cry in the middle of the night. I am so grateful that my parents supported me to pursue this fantastic PhD program, both emotionally and financially.

Finally, I would like to acknowledge sponsors who supported me to pursue PhD degree. This includes Taiwanese Government Scholarship by the Ministry of Education in Taiwan, a doctoral student scholarship and annual travel fund from the Department of Mental Health.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Substance use disorders

Substance use disorders (SUD) are chronic and recurrent; they affect personal health and wellbeing and cause significant losses in a nation's economic productivity. (Ersche, Clark, London, Robbins, & Sahakian, 2006; Hansen, Oster, Edelsberg, Woody, & Sullivan, 2011) According to DSM-V, substance use disorder is characterized by four groups of symptoms: impaired control, social impairment, risky use, and pharmacological reactions. (American Psychiatric Association, 2013) The risk of developing a SUD may be partly due to heritable factors through underlying behavioral risk dimensions like impulsivity and risk taking. (Kreek, Nielsen, Butelman, & LaForge, 2005)

Individuals can be addicted to any number of psychoactive substances, including legal drugs like nicotine and alcohol. Illicit drugs and their accompanying problems are found the world over. In 2009, an estimated 149 to 271 million people worldwide admitted to having used at least one illicit drug: of that group 15 to 39 million had used an opioid, amphetamine, or cocaine, and 11 to 21 million had used an injected drug. (Degenhardt & Hall, 2012) Opioid overdose and opioid dependence are potentially lethal; moreover, the injection of opioids, cocaine, or amphetamines is a substantial risk factor for transmission of HIV, hepatitis C, and hepatitis B. (Degenhardt & Hall, 2012) The 2009 survey excluded several types of illicit drugs: 3,4-methylenedioxy- N-methylamphetamine (MDMA, or ecstasy), hallucinogens, and inhalants, thus the actual number of illicit drug users is likely even higher. According to the results from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), the 12-

month prevalence in the United States of substance use disorder was 9.35% and of any drug use disorder, 2.00%. The prevalence of opioid use disorder, amphetamine use disorder and cocaine use disorder was 0.35, 0.16, and 0.27, respectively. (Grant et al., 2004) A study by Monitoring the Future found that the number of users of injected heroin rose from 0.3% in 2009 to 0.7% in 2010. (Johnston, O'Malley, Bachman, & Schulenberg, 2013) In addition to the impact on the individual drug user, there is a larger cost to society. In Western countries, alcohol abuse and drug addiction consume approximately 3.5% of gross domestic product. (Pouletty, 2002)

1.2 The human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS)

Acquired immunodeficiency syndrome (AIDS) was first reported in 1981 in California and New York among a small group of homosexual men. (Center for Disease Control, 1981; Friedman-Kien et al., 1981) A classic symptom of AIDS is a complete loss of CD+4 (cluster of differentiation 4) T cells which leads to immune deficiency and an increased susceptibility to opportunistic infections and Kaposi's sarcoma. In 1983 a retrovirus, now called the human immunodeficiency virus (HIV), was implicated as the cause of AIDS, and the transmission pathway was identified as bodily fluids: blood, semen, vaginal secretions and breast milk. Diagnostic tests were developed to identify the infection, which helped prevent its transmission; prevention programs emphasizing risk reduction, condom distribution, and needle exchange are now commonplace. Antiretroviral therapy (ART) is used to treat HIV, of which there are six types: a) nucleoside/nucleotide reverse transcriptase inhibitors, b) nonnucleoside reverse

transcriptase inhibitors, c) protease inhibitors, d) fusion inhibitors, e) CCR5 antagonists, and f) integrase inhibitors. (Panel on Antiretroviral Guidelines for Adults Adolescents, 2009) Currently, highly active antiretroviral therapy (HAART), which combines at least three drugs from two classes of antiretroviral agents, has proven highly effective in slowing the progress of HIV/AIDS. However, the AIDS virus is highly mutable and drug resistance remains a possibility. Prevention remains the most effective means of reducing HIV transmission rates.

HIV was first described in June 1981. Within a decade, the three main transmission routes were identified as via blood, sexual, and perinatal. The virus enters the bloodstream through transfusions with contaminated blood or blood products, needle sharing among injection drug users, and injections with unsterilized needles. Sexual transmission routes of the virus include homosexual contact between men and heterosexual contact from men to women or women to men. Perinatal transmission can occur intrauterine and peripartum. (Friedland & Klein, 1987) Because of the high prevalence of HIV/AIDS among injection drug users, many studies have focused on efforts to stop or slow its spread within this specific population.

1.3 The AIDS Linked to the Intravenous Experience (ALIVE) cohort

The AIDS Linked to the Intravenous Experience (ALIVE) is a longitudinal prospective community-based study of injection drug users in Baltimore. The ALIVE was established in 1988. It is one of the longest-running community-based cohorts of injection drug users. In its early years, the primary goal of the ALIVE study was to

understand the natural history of HIV in the population of IV drug users. Since 1998, the ALIVE study has expanded its focus on participants' access to treatment for HIV and other non-AIDS outcomes including hepatitis C as well as the impact of this treatment. The annual rate for loss-follow-up is five percent; the death rate is two to three percent. The ALIVE study provides an excellent longitudinal sample for research on injection drug users. The ALIVE GWAS subsample, a set of 1197 subjects genotyped on the Affymetrix 6.0 GWAS chip, will be used for this proposed study. (Galai, Safaeian, Vlahov, Bolotin, & Celentano, 2003; Vlahov, Anthony, Muñoz, & Margolick, 1991)

1.4 Personal hazard, social cost, and public health significance

Different substances are associated with different diseases. Smokers experience high rates of lung cancer. Alcoholics suffer from cirrhosis of the liver. Illicit drugs--psychoactive drugs--are associated with a variety of psychological problems, chief among them the high risk of addiction. The signs and symptoms most commonly associated with drug addiction are euphoria, tolerance and withdrawal. (Wise & Bozarth, 1984) One class of highly addictive drug is the opioids. Opioid addiction inflicts damage at all levels of society. Opioids are frequently injected in order to bring on a faster euphoric rush. When drugs are injected, users run the risk of acquiring infectious diseases. (Vlahov et al., 1998) Another consequence of continuous opioid and amphetamine use is functional impairment, in particular cognitive impairment and poor judgment. (Ersche et al., 2006) Lapses in judgment result in risky or dangerous behavior to oneself or others. Furthermore, the cost of the drugs combined with the constant need to assuage the

craving lead addicts to engage in high-risk, often illegal activities that result in run-ins with the police and involvement in the judicial system. (Rounsaville & Kleber, 1985; Rounsaville, Tierney, Crits-Christoph, Weissman, & Kleber, 1982)

Just as society must pay the price for its addiction to alcohol and tobacco, there is also an enormous societal cost to opioid addiction in the United States. (Birnbaum et al., 2011) First, there is the cost to the health care system. Medical spending on individuals with addiction (who not only have comorbid psychiatric disorders but may also have diseases like HIV/AIDS) costs the U.S. nearly $1 billion a year. (Hansen et al., 2011) Second, there is the cost to the American economy: the U.S. loses $42 billion a year in workplace productivity because of this problem. (Hansen et al., 2011) Finally, there is the cost on the criminal justice system; annually, this costs the U.S. $8.2 billion. (Hansen et al., 2011)

Substance use disorders are chronic and recurrent. Early intervention may be the best way to treat patients with complicated clinical conditions. Injection drug users have high rates of HIV; it is therefore critical to identify those who are at higher risk of developing HIV risk behaviors for the purposes of early intervention. It is hoped that the findings from this study may help with clinical and policy efforts to understand the mechanisms of injection patterns, to provide prediction tools, and to connect brain biology to behavior patterns. Further, it is hoped that these findings may contribute to efforts at secondary prevention defined here as the reduction of the impact of a disease or injury that has already occurred. These strategies will be discussed in the following chapters.

1.5 Complex genetic etiology

Genetic epidemiology is a rapidly expanding field of research; it uses a powerful set of tools that helps find answers to questions of etiology. (Burton, Tobin, & Hopper, 2005) In recent decades, the results of classical genetic research using family studies, twin studies, and adoption studies support the premise that psychiatric disorders are not only environmental but also at least partly heritable. These include schizophrenia, bipolar I disorder, substance use disorders and others. (Hopper, Bishop, & Easton, 2005; Shih, Belmonte, & Zandi, 2004) Unlike psychiatric disorders such as anxiety, genetics plays a much more important role in substance use disorders. (Hettema, Prescott, Myers, Neale, & Kendler, 2005) The relative impact of genes and environment varies depending on stage of drug use. As drug use transitions from initiation, to continued use and then dependence, the relative impact of genes increases. (Tsuang et al., 1999; Wray, 2007) Furthermore, it is now easier and faster to perform large computational analyses such as genome wide association studies because of huge advances in computer technology. (Ott, 1974) The discovery of the genetic underpinnings of psychiatric disorders has led to a surge in research using molecular analyses to identify and locate associated genes. Linkage studies, genetic association studies, genome wide association studies and, more recently whole genome sequencing, have become the norm. (Ozaki et al., 2002; Risch & Merikangas, 1996) The achievements of genetic epidemiology are widely apparent; nevertheless, challenges remain. This paper will focus on opioid addiction, one type of substance use disorder, and describe it in detail.

1.6 Guiding principals

1.6.1 Longitudinal data analysis

Substance use disorders are chronic and recurrent. Impulse control, which is a predisposition to and/or a consequence of substance use disorders, is also a core behavioral domain underlying many HIV risk behaviors. Unlike some HIV risk behaviors such as engaging in unprotected sex, substance use disorders are more proximally associated with brain biology. In order to study HIV risk behaviors among injection drug users, we will use the data in a longitudinal manner rather than reducing the data to a cross-section or single measure (e.g., lifetime dependence). The reason is that a longitudinal data provides more valuable information about the long-term trajectories of the HIV risk behaviors of injection drug users. This information may yield a phenotype that reflects severity and intensity in a way that is more strongly related to the underlying biology of the trait. Using longitudinal ALIVE injection data, we borrow the concept of "pack-years" to create our phenotype, injection years, (Caporaso et al., 2009) and analyze the composite HIV risk phenotypes of the trajectories of the participants. (Smith et al., 2015) The lives and relationships of injection drug users tend toward instability; hence, there exists virtually no perfect attendance record for any single participant in this study. For this thesis, we accessed the ALIVE cohort, a longitudinal community-based study, and we use statistical approaches to impute the missing data. Furthermore, the analysis results using this imputed ALIVE sub-cohort serve as the basis for the genome wide association analysis.

1.6.2 Sub-population and sub-phenotypes analysis

The ALIVE cohort is a longitudinal community-based study of injection drug users, primarily African Americans, in Baltimore. This urban minority population is at high risk of illicit drug use and dependence (Reuter, Hsu, Petronis, & Wish, 1998) and there is scant research on the association between their genetic makeup and HIV risk behaviors, including drug injection. We believe that our results may enhance prevention and treatment strategies and aid in identifying subpopulations at high risk of persistent drug use. In addition, it may be possible to identify a composite of HIV risk behaviors that are more amenable to prevention strategies. In addition, in studying this targeted population we avoid a very common pitfall of genome wide association analyses: the use of samples from different geographical regions with individuals of different genetic backgrounds.

Genome wide association studies (GWAS) are a major advance in the field of genetics yet there are limitations. First, GWAS cannot be applied to a select number of disorders which have very low genotypic risks or are influenced by only several rare SNPs. Second, current diagnostic categories might not reflect the heritability of a particular disorder. Third, gene-gene or gene-environment interactions are also factors. (Psychiatric GWAS Consortium Coordinating Committee, 2009) Borrowing the concept of "smoking-years", we create several sub-phenotypes including injection years and the trajectories of composite HIV risk phenotypes. We believe that these phenotypes may be partly genetic in origin. It is hoped that these results will elucidate our understanding of modifiable and fixed components of HIV risk behaviors in this population, and lead to improved treatment and prevention.

1.6.3 Genomics

The genetic underpinnings of substance use disorders are frequently studied. Classic genetic approaches such as twin, family and adoption studies show that there are significant genetic influences on drug addiction. Heritability is the proportion of observed differences on a phenotypic trait among individuals of a population that are due to genetic differences, and the heritability of addiction is estimated at 0.4-0.6. (Kendler, Prescott, Myers, & Neale, 2003; Tsuang et al., 1998) In recent years, huge advances in computer technology have made molecular genetic approaches possible. Linkage studies, genetic association studies, and genome wide association studies can identify and locate associated genes. There have now been a number of studies on addiction behaviors using molecular approaches. (Cornelis et al., 2011; Crabb, Edenberg, Bosron, & Li, 1989; Sulem et al., 2011; Tobacco and Genetics Consortium, 2010; P. Xie et al., 2011)

Classical genetic approaches have shown that addiction is heritable; molecular genetic approaches suggest that specific addiction-related behaviors are associated with specific genes. Herein, we create a phenotype from the ALIVE dataset: injection years. We believe that injection years and the trajectories of HIV risk behaviors are genetically influenced and we therefore use them as phenotypes in genome wide association analysis. Using the results from the genome wide association analysis, we conduct a pathway analysis and a polygenic risk score analysis. The pathway analysis yields information on the biological mechanisms of injection behavior patterns. The polygenic risk score analysis predicts injection behaviors. In brief, we analyze the association between behavioral phenotypes and genetic markers across the entire genome. Using identified genetic markers, the results may link behavioral phenotypes to the biological mechanisms

of addiction, either through the genome-wide significant association of single SNPs in relevant genes or through the overrepresentation of SNPs in specific biological pathways among the set of SNPs below some p-value.

1.7 Chapter introductions

This dissertation is divided into five chapters. In the first chapter, I provide background information on addiction, HIV risk behaviors, genetics and emphasize public health significance. The ALIVE cohort and genetic etiology are introduced briefly in this chapter. There were three key elements of this dissertation: imputing before analyzing longitudinal data, creating a unique behavior pattern as a risk phenotype in a specific population, and using that behavior phenotype in genome wide analysis.

The second chapter is a review of the literature. The findings from the literature review support our proposition that injection drug users are at high risk of HIV infection, and injection behavior is highly associated with addiction, which is partly heritable. These findings suggest that creating a tool to measure the intensity of injection behavior is urgent from both a public health and personal health perspective, and using genome wide analysis on injection years may provide useful information.

Building on the findings in the second chapter, the third chapter describes the risk phenotype, injection years. I describe the importance of longitudinal data, compared to cross sectional data, and the most significant challenge, vast amounts of missing data. In order to solve this problem, I used three different models to impute the entire dataset and conducted sensitivity exams for these three models. The findings pointed to using the universal imputation, a model which jointly imputes the entire panel of data, to create the injection years phenotype.

The fourth chapter is an extension of the third chapter. Injection years, which was created from the imputed data, was used as a risk phenotype in genome wide association analysis, polygenic risk score analysis, and pathway analysis. By using different genomic analyses, I explored the association between injection years and single SNP as well as a group of SNPs with p-values below an arbitrary threshold, and a gene set of SNPs with a shared biological pathway.

The fifth chapter is a discussion and summary of the results. The findings of this dissertation support our proposition that multiple imputation can address the issue of large amounts of missing data in a longitudinal dataset and using injection years to evaluate the intensity of injection behavior is possible. The genetic findings suggest that injection years is influenced by polygenes. In conclusion, I believe that linking genetic studies to observable behaviors, particularly behavioral trends in a longitudinal dataset, can provide valuable information about how genetic factors influence individuals' behavior through specific biological mechanisms.

Chapter 2: A review of the evidence suggesting specific behavior patterns such as the use of injection drugs are related to genetic factors

2.1 Abstract

Injection drug use is a common administration route of substance use disorder. The use of drugs like heroin increases the risk of HIV. Studies have found the use of alcohol, tobacco, marijuana, opioids, and stimulants are prevalent among people with HIV. In particular, heavy drink is very common among people in care for HIV infection. Studies from a number of different countries have found that thirty-one to eighty-four percent of HIV+ individuals are smokers. The use of stimulants like amphetamines and cocaine is widespread among HIV positive injection drug users.

Using Diagnostic and Statistical Manual of Mental Disorders 4th edition, substance-related disorders include substance use disorders or substance-induced disorders. It lists several classes of substances: alcohol, amphetamines, cocaine, nicotine, and opioids. The symptoms and signs including tolerance, withdrawal, increasing uncontrolled intake, spending more time and money on substances, and impaired social, occupational, or recreational function are observed as part of the diagnosis of substance dependence. This complex of symptoms and signs can be categorized into four dimensions: impaired control, social impairment, risky use, and physiological reactions caused by the substance.

Since the 1960's with the first twin studies of alcoholism it has been posited that there is a genetic component to addiction. Scientists are now using molecular approaches in genetic epidemiology to identify the location of addiction-associated genes.

In this thesis, a new phenotype has been created: injection years. We imputed each participant's biannual injection behavior to address the issue of missing information and then took the sum of his or her injection behavior to create an "injection years" phenotype, considering it as a continuous variable. We then performed a descriptive data analysis. Borrowing the concept of smoking pack-years, we believe that the injection years trait is associated with genetic factors. Furthermore, assuming that multiple genes with small effects contribute to injection years, we performed a polygenic risk score analysis and pathway analysis to attempt to ascertain which biological pathways are related to this variable.

2.2 A review of the literature on substance use disorders and HIV

The use of illegal and injectable drugs like heroin increases the risk of HIV. There are three main transmission routes: contaminated blood, sexual contact, and perinatal exposure. Contaminated blood comes from transfusions of blood and blood products, needle sharing among injection drug users, and injection with unsterilized needles. (Friedland & Klein, 1987) Injection drug users are at high risk of becoming HIV-infected regardless of age early detection and advances in treatment make it possible for HIV positive individuals to live longer. The population of HIV-positive patients, in particular those with substance use disorders, is aging. This is a rather serious public health issue and has led to a number of studies on the subject. Edelman et al. reviewed articles related to substance use in older HIV-infected patients and found that the situation of substance use in the older population remains severe. (Edelman, Tetrault, & Fiellin, 2014) Pilowsky

et al. reviewed the articles related to risky behaviors among older HIV-infected patients and he found that substance use increased the risk for HIV risk behaviors and the prevalence of substance use disorder is increasing among the elders in the past decades; thus, older Americans may be at increased risk for HIV infection. (Pilowsky & Wu, 2015) The number of Americans at 65 years or older with substance use disorder is rising because the population is aging and life expectancy is increasing. (Lofwall, Brooner, Bigelow, Kindbom, & Strain, 2005; Rosen, Smith, & Reynolds, 2008; Wu & Blazer, 2014) Although injection drug users are at high risk of HIV regardless of age, this specific population in the United States is also aging. Armstrong et al. estimated that in the United States, the mean age of injection drug users increased from 26 to 42 between 1979 and 2002. (Armstrong, 2007) The longer injection drug users live, the higher the prevalence of HIV infection in that population.

The use of alcohol, tobacco, marijuana, opioids, and stimulants is prevalent among those infected with HIV. There are four types of unhealthy alcohol use: risky drinking, problem drinking, harmful use or alcohol abuse, and alcoholism or alcohol dependence. (Saitz, 2005) Heavy drinking (equal to or more than 5 drinks over 1 to 4 days) is common among people in care for HIV infection. (Burnam et al., 2001; Galvan et al., 2002) Among the HIV positive population, the prevalence of alcohol use disorders is estimated to range from 29% to 60%. (Petry, 1999) Studies of HIV positive people from many different countries show that thirty-one to eighty-four percent of HIV positive individuals are smokers. (Brennan, 2012) HIV positive participants report that in spite of the negative health effects and the cost of cigarettes, they felt more relaxed and better able to manage anxiety, anger and depression by using cigarettes. (Shuter, Bernstein, &

Moadel, 2012) In the same study, alarmingly, 27% of the participants believed that smoking would help *increase* white blood cell counts. (Shuter et al., 2012) Marijuana is also used to self-medicate physical and emotional problems. Prentiss et al. found that HIV positive individuals reported that using marijuana improved their mood and appetite and decreased anxiety, nausea and pain; (Prentiss, Power, Balmas, Tzuang, & Israelski, 2004) they also found that using marijuana improved adherence to medication treatments by decreasing feelings of nausea. (de Jong, Prentiss, McFarland, Machekano, & Israelski, 2005) Injection drug users are at high risk for HIV, with heroin the most commonly injected drug. The use of stimulants like amphetamines and cocaine is widespread among HIV positive injection drug users specifically among homosexual men. In Skeer et al.'s study, 20.7% of subjects reported using methamphetamine and crystal methamphetamine, and 17% reported the use of cocaine. (Skeer et al., 2012) From 2005 to 2010, Mimiaga et al. followed a group of HIV positive individuals in the United States and found that nine percent reported amphetamine use and another nine percent reported crack–cocaine use. (Mimiaga et al., 2013)

2.3 A literature review of genetic epidemiology and substance use disorders

2.3.1 Substance use disorders

DSM-IV classifies substance-related disorders in two ways: substance use disorders and substance-induced disorders. There are eleven classes of substances: alcohol, amphetamines, caffeine, cannabis, cocaine, hallucinogens, inhalants, nicotine, opioids, phencyclidine, and sedatives, hypnotics, or anxiolytics. Substance use disorders

are further identified as either substance abuse disorder or substance dependence disorder. Substance-induced disorders are categorized as substance intoxication, substance withdrawal, and substance-induced mental disorders. The criteria for diagnosing substance dependence include tolerance, withdrawal, increasing of uncontrolled intake, spending more time and money on substance, and impaired social, occupational, or recreational function. The criteria which must be met to make a diagnosis of substance abuse include primarily recurrent substance use resulting in legal problems, physical hazard, and failure to fulfill work, social, or home roles. In contrast, the criteria necessary for defining substance dependence relate to physical changes, behavioral changes, and significant loss of function. (American Psychiatric Association, 2000)

In 2013, DSM-V revised the criteria for defining and diagnosing substance-related disorders. The most significant change is that substance abuse disorder and substance dependence disorder are seen as one category: substance use disorder. In DSM-V, "substance use disorder" includes most of the criteria for "substance dependence disorder" and "substance abuse disorder" from DSM-IV; DSM-V no longer mentions the criterion of "recurrent substance use resulting in legal problems" while adding the criterion of "craving to use substance." Furthermore, DSM-V's criteria for making a diagnosis of substance use disorder can also be used to indicate its current severity. A diagnosis of "mild" indicates the presence of two or three symptoms, "moderate" means the presence of four or five symptoms, and "severe" is the presence of six or more symptoms. (American Psychiatric Association, 2013)

This change between DSM-IV and DSM-V in the view of substance-related disorders reflects the shift from a categorical view to a dimensional approach. A

categorical view is used by clinicians to meet the needs for reporting to health care planners and insurance companies, whereas a dimensional approach conceptualizes a quantitative disorder that is more useful for the purpose of research. (Saunders, Schuckit, Sirovatka, & Regier, 2008)

2.3.2 Symptoms and signs of substance use

Substance use disorder is a complex of symptoms and signs that involve the domains of cognition, behavior, and physiology. The persistent use of a substance can induce changes in brain circuits and cause specific behavior patterns such as craving. According to DSM-V, substance use disorder has four dimensions: impaired control, social impairment, risky use, and physiological reactions caused by substances. Impaired control includes taking larger amounts than intended, multiple unsuccessful efforts to decrease or discontinue use, spending a great deal of time on the substance, and craving. Social impairment refers to family, occupational, or other social problems. Using substances in a physically hazardous situation and/or using substances despite the knowledge of its consequent physical or psychological problems define the category of risky use.

Physiological effects of a substance include tolerance and withdrawal symptoms. Impaired control and risky use are the symptoms and signs resulting from behavioral changes which occurred with persistent use. Social impairment refers to the consequences of persistent use. (American Psychiatric Association, 2013)

2.3.3 Approaches to the study of genetic epidemiology

Since the 1950s, technological advances have changed the field of genetic epidemiology. Genetic epidemiology has become a fundamental clinical practice; family

history is a basic diagnostic tool for many inherited diseases. (Thompson, Orvaschel, Prusoff, & Kidd, 1982) In 1955, Morton et al. demonstrated how the logarithm of odds scores could be used to detect linkage. (Newton E Morton, 1955) In 1974, Ott and colleagues began using a computer program to conduct linkage analysis. (Ott, 1974) In 1996, an association map was developed for complex diseases. (Risch & Merikangas, 1996) In 2002, Ozaki and his colleagues published the first genome wide association study (GWAS) using SNP markers. (Ozaki et al., 2002) In 2007, the Psychiatric GWAS Consortium began to conduct meta-analyses of genome-wide association studies for five psychiatric disorders: autism, attention-deficit hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia.(Sullivan, 2010) In 2009, the International Schizophrenia Consortium developed an innovative approach to evaluating the risk of schizophrenia and bipolar disorder: the polygenic score. (Purcell et al., 2009)

2.3.3.1 Classic and molecular genetic approaches

The purpose of genetic epidemiology is to understand how genetic factors influence health and disease in families and in populations and to interpret the interactions between genetic and environmental factors. Morton defined genetic epidemiology as "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations". (Newton Ennis Morton, 1982) Genetic epidemiology attempts to answer two questions: Is the disease influenced by a genetic component? If the answer is yes, what genes are involved and where are they? Several different types of studies have been designed to answer these questions, twin studies, adoption studies, and family aggregation studies answer the first question, “Is disease X influenced by a genetic component, and if so,

what are the relative contributions of genes and environment?" Segregation studies help to find patterns of inheritance of disease. Linkage studies help to determine which part of which chromosome the disease is associated with. Finally, association studies identify which allele of which gene the disease is associated with. (Williams, Carson, Passmore, Silvestri, & Craig, 2011)

2.3.4 The genetic epidemiology of substance use disorders

As stated in chapter one, genetic epidemiology is a rapidly expanding field of research. (Burton et al., 2005) Both classical genetic research using family studies, twin studies, adoption studies, and molecular genetic research using linkage studies, genetic association studies, and genome wide association studies provide valuable information related to substance use disorders. This section provides a review of the latest genetic research on disorders involving tobacco, alcohol, and opiates. The subsequent section will discuss the possible genetic links between opioid use disorders and specific behavior patterns.

Kreek et al.'s review of the research on genes and addiction offered a three-domain model that included genetics, diverse environmental factors, and drug-induced effects. (Figure 1) (Kreek, Nielsen, & LaForge, 2004) In 1960, Kaij et al. conducted the first study of alcoholism in twins and in 1966, Partanen et al. conducted a similar twin study which further explored the associations between intelligence, personality, and alcohol consumption. These were the earliest studies proposing that specific addictions were heritable, or influenced by genes. (Kaij & Rosenthal, 1961; Partanen, Bruun, & Markkanen, 1966) An adoption study by Cloninger et al. concluded that genes influence alcohol abuse since adopted away probands had a greater resemblance to their biological

relatives than their adoptive family. (Cloninger, Bohman, & Sigvardsson, 1981) Furthermore, Cloninger et al.'s adoption study provided a classic approach to disentangle the influence of genetics from that of environmental factors. (Cloninger et al., 1981) In 1988, Merikangas et al. reported an eight-fold increase in the odds of drug disorders among the relatives of probands with drug disorders, with the greatest odds ratio observed for addiction to the same substance. (Merikangas et al., 1998) Tsuang's twin study posited that both environment and genes influence a person's susceptibility to drug abuse; Tsuang also found that all commonly abused drugs--opiates, marijuana, sedatives, psychedelics, and stimulants--had an overall genetic variance from 0.3 to 0.5. Heroin had the greatest overall genetic variance, 0.54, and a shared genetic variance, 0.2, with other drugs. However, most of these drugs only have a low variance for specific genetic factors and only heroin had the greatest specific genetic variance at 0.4, indicating there could be unique genetic factors affecting opioid abuse. (Tsuang et al., 1998) Kendler and colleagues, in their seminal twin study on substance use disorders, found that lifetime drug use of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates had a range of additive genetic variance, or heritability, of 0.3 to 0.5. (Kendler, Jacobson, Prescott, & Neale, 2003) From twin studies to adoption studies, from alcohol to other substances, these classic genetic studies provide solid evidence that genetics plays an important role in substance use disorders. Subsequently, molecular genetic studies are exploring heritability on a deeper level. In the past twenty years, with huge advances in computer technology and genomic array technology, molecular genetic approaches are identifying or locating specific associated genes.

2.3.4.1 Heritability: nicotine addiction

Many studies have been conducted on the use of alcohol and tobacco. A large number of twin studies have reported significant heritability for tobacco addiction in different populations, regardless of sex or age. (Carmelli, Swan, Robinette, & Fabsitz, 1992; A. Heath, Kirk, Meyer, & Martin, 1999; A. C. Heath, Madden, Slutske, & Martin, 1995; Kendler, Thornton, & Pedersen, 2000; Koopmans, Slutske, Heath, Neale, & Boomsma, 1999) Mcgue et al., in a twin study, found that the heritability of tobacco use and nicotine dependence was 40% to 60%. (McGue, Elkins, & Iacono, 2000)) A family study by Cheng et al. identified a major segregating factor for "ever-smoking", or lifetime cigarette smoking. (Cheng, Swan, & Carmelli, 2000) With data from families in the Collaborative Study on the Genetics of Alcoholism (COGA), some linkage studies have reported that several specific chromosome sections are associated with smoking behaviors. Bergen et al. reported some linkage between smoking behaviors and chromosomes 6, 9 and 19. Additionally, linkage existed between several candidate gene regions and smoking pack-year history. (A. W. Bergen, Korczak, Weissbecker, & Goldstein, 1999) Duggirala et al. also found linkage between smoking pack-year history and a genetic region on chromosome 5q, suggesting that a variant or variants near this marker (D5S1354) on chromosome 5q might be the primary determinant of genetic variation in smoking. (Duggirala, Almasy, & Blangero, 1999) This site is also close to the locus for dopamine receptor D1 which is associated with smoking. (Comings et al., 1997)

Thorgeirsson et al. conducted a GWAS for nicotine dependence and smoking behavior. They found that nicotine dependence and cigarettes per day are associated with rs1051730 which is associated with the nicotinic acetylcholine receptor gene cluster on

chromosome 15q24. (Thorgeirsson et al., 2008) The Tobacco and Genetics Consortium conducted GWAS meta-analyses to examine four smoking phenotypes: smoking initiation, age of smoking initiation, smoking quantity or number of cigarettes smoked per day (CPD) and smoking cessation. (Tobacco and Genetics Consortium, 2010) They found that three loci are associated with CPD and the top most strongly associated SNP is rs1051730, which is in linkage disequilibrium with the nicotinic receptor gene CHRNA3.

2.3.4.2 Heritability: alcohol addiction

Alcohol addiction is partly heritable. Two twin studies estimate the heritability of alcohol addiction at 48% to 58% for males and 51% to 59% for females. (Kendler, Neale, Heath, Kessler, & Eaves, 1994; Prescott & Kendler, 1999) Due to the nearly ubiquitous exposure to the substance, alcohol offers a unique way of understanding the probability of/susceptibility to addiction since all "controls", or non-dependent individuals, are likely to have been exposed. In 1989, the COGA began to identify genes implicated in alcoholism. (Edenberg, 2002) Investigators in the COGA study refined the definition of alcoholism beyond that in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IIIR).Using a family-based linkage study design, they found that drug and/or alcohol dependence are associated with a handful of identified "regions of interest" on human chromosomes 1, 2, 3, 4, 7, and 11. (Foroud et al., 2000; Long et al., 1998; T. Reich et al., 1998; Stallings et al., 2003) In the same way that certain genes make people more susceptible to addiction, other genes seem to make people *less* susceptible. Clinical observations of Asians with different phenotypes in metabolizing alcohol were the earliest studies on the genetics of vulnerability to addiction. Many Asians of the same ethnic descent exhibit a distinct facial flushing and experience severe hangover effects

after consuming alcohol. As a result, many Asians in this population have an aversion to alcohol. Consequently, the biochemical and genetic basis for the facial flushing phenotype in Asians was determined to be due to deficiencies in enzymes responsible for the breakdown and metabolism of alcohol. Several genes that are associated with a susceptibility to alcohol abuse are involved in the metabolism of alcohol. Of great interest is the gene for alcohol dehydrogenase (ADH), the enzyme which initiates the biochemical process of converting ethanol to acetaldehyde. ADH genes, which are linked to alcohol dependence, have been located on chromosomes 4q. (Long et al., 1998; Saccone et al., 2000; Van Eerdewegh et al., 1998) Another gene that has been linked to alcohol dependence is aldehyde dehydrogenase (ALDH) in liver mitochondria, a tetrameric enzyme that converts acetaldehyde to acetyl-CoA. In one study, individuals with inactive ALDH alleles accumulated high levels of acetaldehyde after consuming alcohol, which resulted in facial flush and severe hangover. (Mizoi et al., 1983) ALDH genes protect people of Asian lineage from developing problems with alcohol. (Harada, Agarwal, Goedde, Tagaki, & Ishikawa, 1982; Muramatsu et al., 1995) In brief, this is because the ADH genes are associated with faster metabolism of ethanol, causing acetaldehyde production and accumulation. The ALDH genes are associated with acetaldehyde metabolism, causing its accumulation. The interaction of ADH and ALDH genes contributes to painful hangover and thus appears to lessen the likelihood of developing alcohol-related problems. (Bosron, Ehrig, & Li, 1993; Crabb, Dipple, & Thomasson, 1993)

2.3.5 The highest heritability: opioid addiction

Tsuang et al.'s twin study found that heroin had the greatest overall genetic variance, 0.54, and a shared genetic variance, 0.2, with all drugs. (Tsuang et al., 1998) Wilen and colleagues' study of families in which parents were opioid-or alcohol-dependent reported that children of addicts experienced higher rates of psychopathology including mood, anxiety, and substance use disorders. (Wilens et al., 2002) An adoption study by Cadoret et al found that heritable biological and environmental factors correlated with substance use. These results are consistent with twin studies. (Cadoret, Troughton, O'Gorman, & Heywood, 1986)

Scientists are now using molecular approaches in genetic epidemiology to identify the location of addiction-associated genes. Linkage studies are family-based studies that link to specific regions of the genome rather than a particular gene; they target phenotypes such as physical traits or specific diseases. (Teare & Barrett, 2005) Until twenty years ago, linkage studies could detect only a limited number of candidate genes because the technology was not sufficiently advanced; now, with much more powerful computers, genome-wide linkage studies are possible. Gelernter et al conducted a genome-wide linkage scan for opioid dependence and found that chromosomes 2 and 17 are associated with opioid dependence. (Gelernter et al., 2006) Lachman and colleagues conducted another genome-wide linkage study of opioid dependence and found that a specific region of chromosome 14q is associated with opioid dependence. (Lachman et al., 2007) Genome-wide association studies examine in an hypothesis-free approach that single SNPs throughout the genome are associated with a specific phenotype by comparing the affected individuals to non-affected individuals.(Psychiatric GWAS

Consortium Coordinating Committee, 2009; W. Y. Wang, Barratt, Clayton, & Todd, 2005) Wetherill et al. conducted a genome-wide association study examining the association between candidate genes and substance dependence and found that one single-nucleotide polymorphism (SNP), rs2952621 in the uncharacterized gene LOC151121 on chromosome 2 and another SNP, rs2567261 in ARHGAP28 (Rho GTPase-activating protein 28), are associated with substance dependence. (Wetherill et al., 2015) In a separate genome-wide associations study Gelernter and colleagues found that SNPs from multiple loci-KCNG2*rs62103177 which involved potassium signaling pathways were associated with opioid dependence. (Gelernter et al., 2014)

Some behavior patterns such as impulsivity, risk taking and stress response, which are due to specific personality and physiological traits may make some people more prone to addictive disorders. These patterns may be partially influenced by genetic variation. Moreover, differences in personality and physiological traits may affect different *stages* of addiction. These stages of addiction are chronologically defined as initiation of drug use, regular drug use, abuse/dependence and relapsed use. (Kreek et al., 2005) Clearly, many genes have been found to be associated with addiction. The focus of this work is on genes associated with heroin addiction. These genes can be classified into two gene systems: the dopaminergic system and the mu opioid receptor systems. (Kreek et al., 2004) Several single-nucleotide polymorphisms in the dopaminergic system are associated with heroin addiction. They include rs4680, rs1800497, rs1800955, rs1611115, rs1079597, rs747302, rs1800498, and rs936462. (Hou & Li, 2009; Vereczkei et al., 2013; X. Xie et al., 2013)) The dopamine D4 receptor gene was also found to be associated with novelty-seeking, which is further associated with risk-taking. (Lusher, Chandler, &

Ball, 2001; Schinka, Letsch, & Crawford, 2002) Candidate genes OPRM1, rs1799971, rs7997012, and rs540825 in the mu opioid receptors system are associated with opioid dependence. (A. Bergen et al., 1996; Garriock et al., 2010; Haerian & Haerian, 2013) Table 1 lists characteristics of genes related to heroin/opiate dependence, including protein product, system, location on chromosome, and associated SNP.

Taken together, given the evidence for highest heritability in opioid and cocaine dependence plus lone successes in alcohol and tobacco are for use variation among users (metabolic) I conclude that studying use variation in opioid using subjects would likely yield success.

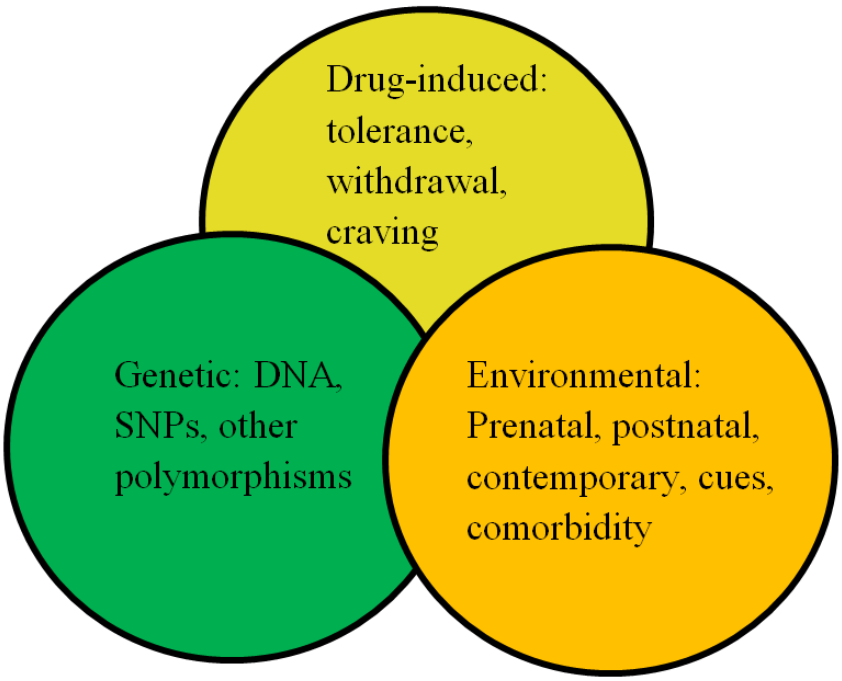
2.4 Discussion: the link between the genetics of opioid addiction and specific HIV risk behaviors

The goal of this paper was to review the literature related to substance use disorders, HIV risk behaviors, and genetics. Previous findings supported that substance use disorders are chronic. Injection drug use, one of HIV risk behaviors, is very common among population with substance use disorders, so they are at high risk of HIV infection. To study the trajectories of injected drug use is definitely important for HIV prevention. Furthermore, the most frequently injection use drugs are heroin and cocaine which had the greatest heritability. (Goldman, Oroszi, & Ducci, 2005) The successes in finding genetic factors of alcohol and tobacco use variation would argue that studying use variation in opioid using subjects would likely yield success.

In the following chapters, a new phenotype --injection years-- has been created. I imputed each participant's injection years to address the issue of missing information and

then took the sum of his or her injection years, considering it a continuous variable. I then performed a descriptive data analysis. Borrowing the concept of smoking pack-years, I believe that injection years will provide a more meaningful phenotype that is more ikely to be associated with genetic factors. (A. W. Bergen et al., 1999) Furthermore, assuming that multiple genes with small effects contribute to injection years, I performed a polygenic risk score analysis and pathway analysis in an attempt to ascertain which biological pathways are related to this variable. **Figure 2** illustrates our hypothesis and the goals of this research project.

Figure 1: Factors contributing to vulnerability to develop a specific addiction

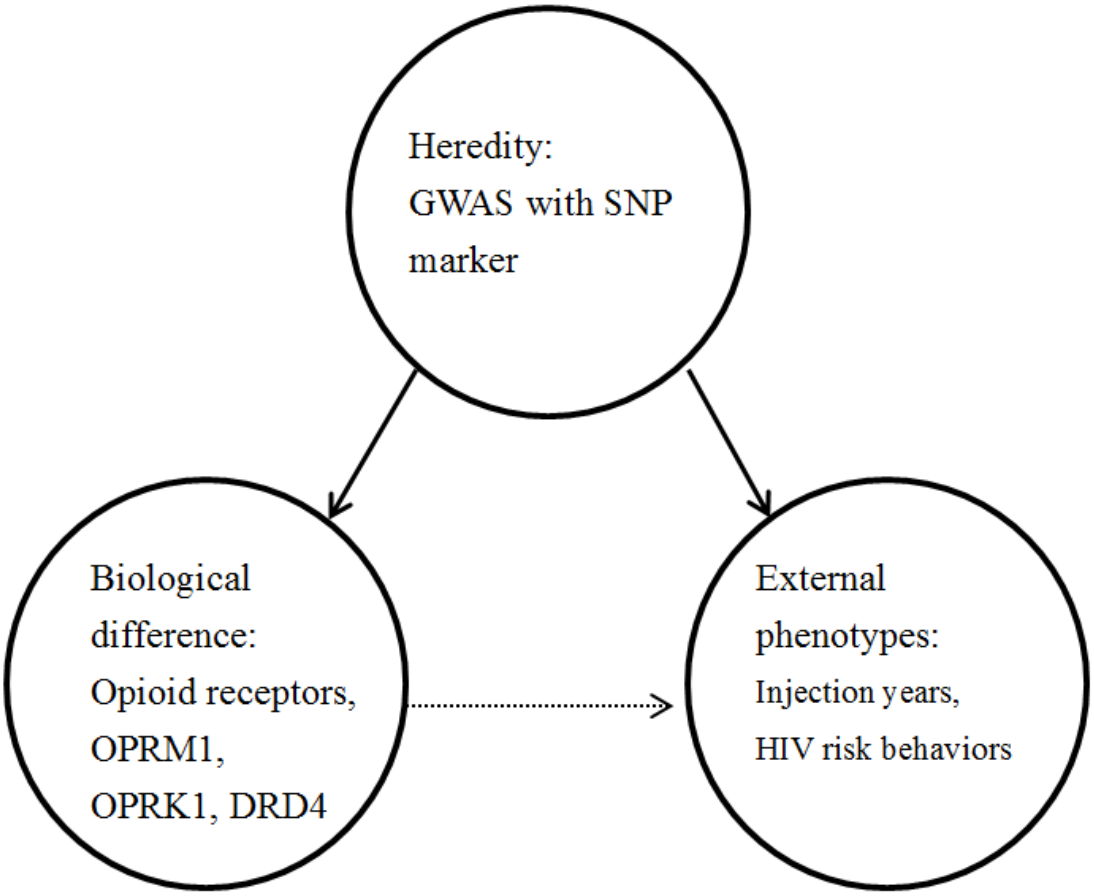


Adapted by Kreek, M. J., D. A. Nielsen, et al. (2004). "Genes associated with addiction." Neuromolecular medicine 5(1): 85-108.

Table 1: Characteristics of genes related to heroin/opiate dependence and addiction

Gene	Protein	System/Function	Chromosomal location	Associated SNP
OPRM1	μ opioid receptor	Opioid	6q24-25	rs1799971
OPRK1	κ opioid receptor	Opioid	8q11.2	rs963549 rs1051660
DRD4	Dopamine receptor D4	Dopaminergic	11q15.5	rs1800955, rs747302, rs936462
TPH2	Tryptophan hydroxylase 2	Serotonergic	12q.21.1	rs4290270 rs7963720
HTR1B	Serotonin receptor 1B	Serotonergic	6q13	rs130058 rs11568817
SLC6A4	Serotonin transporter	Serotonergic	17q11.1-q12	
COMT	Catechol-O-methyl transferase	Catecholaminergic	22q11.2	rs4680
CYP2D6	Cytochrome CYP450	Drug metabolism	22q13.1	

Figure 2: Heredity directly contributes to biological differences among individuals.[+]



[+] Sufficient research exists which suggests that there is a relationship between opioid dependence and opioid receptors; GWAS can connect heredity to the external phenotypes of IDU trajectory, injection years and HIV risk behavior. Our hypothesis is to determine if there is link between biology and external phenotypes. The solid line represents findings supported by the most up-to-date research. The dotted line represents our speculation.

Chapter 3: Injection years used to represent the longitudinal trends of injection drug use

3.1 Abstract

BACKGROUND: Substance use disorders are chronic and recurrent. Functional impairment and unstable lifestyles make it extremely difficult to collect comprehensive longitudinal data on IV drug users. Heroin users typically use needles putting them at high risk of HIV infection. Therefore, the study of a measured injected drug use trend from longitudinal data could provide useful information that will ultimately inform HIV prevention.

METHODS: With data from the AIDS Linked to the Intravenous Experience (ALIVE) Study, this study uses a subgroup with GWAS data (N=1197). For each individual in the subgroup there are twenty-five years of longitudinal data and a single-nucleotide polymorphism (SNP). We devised a variable termed "injection-years": the length of time an individual has been injecting drugs. A hot-deck multiple imputation procedure was used to address the problem of missing data. Models simultaneously imputing all data (Universal), using two prior visits as predictors (two predictor), and one prior visit (one predictor) as a predictor were used for imputation, and sensitivity tests were done for each model.

RESULTS: The ALIVE GWAS subset includes 1,197 subjects, 98% African-American, 287 (24%) females and 882 (76%) males with an average of 19 visits per person. There are 103,599 potential biannual total visits for all participants in the ALIVE GWAS subset, counting from the subject's initial injection drug use to their most recent visit, but two-

thirds of the data is missing. We imputed injection data across the ALIVE GWAS subset and the mean injection year phenotype is 23 years from the universal model, 19 years in the two-predictor, and 18 years in the one-predictor. Results of the sensitivity exam led us to choose the universal model for the analysis. The mean injection years is 24 (it is 23 above), a result consistent with previous work in the ALIVE cohort (Genberg et al).

LIMITATIONS: There is a large amount of missing data in the ALIVE GWAS subset but we recovered much of this data using imputation. Use of data beyond the single item indicating injection use, for example of injection intensity, may provide different results.

CONCLUSION: We imputed with three models. Sensitivity test results supported the universal model, suggesting several causal factors influence injection behavior, including sex and age. After imputing using the universal model, we examined the mean of the sum of injection years across the sample and found it consistent with previous work which inferred injection careers using latent growth models.

KEY WORDS: substance use disorder, addiction, injection, the ALVIE cohort, longitudinal data, imputation

3.2 Introduction

Substance use disorders are chronic and recurrent. People with substance abuse disorders experience changes in behavior that impact their ability to function, and these changes may continue even after someone has quit using. Even worse, long-term addiction impairs intellectual function. Alcoholics, for example, are at high risk of Wernicke-Korsakoff Syndrome. (Thomson, 2000; Victor, Adams, & Collins, 1971) Intellectual functional impairment--specifically, cognitive impairment and poor judgement--can be found in addicts who regularly use opioid and amphetamines. (Ersche

et al., 2006) This functional impairment often translates into their not being able to hold down a steady job or to maintain healthy relationships; it is virtually impossible to follow up on them for lengthy periods of time. Collecting comprehensive longitudinal information from substance-dependent subjects is a challenging proposition; therefore, it is helpful to impute existing longitudinal data.

There are several routes of drug administration: oral, sniffing or intranasal, and injection. Injection produces the fastest effects, often within minutes. Heroin is the most commonly injected drug; heroin addicts inject one to three times a day. Needles are often shared to save money. For these reasons, the administration of injection drugs is popular among regular drug users. In almost all cases, a person who is injecting drugs has a substance use disorder.

HIV enters the bloodstream through transfusions of contaminated blood or blood products, through needle sharing among injection drug users, and through injections with unsterilized needles. (Friedland & Klein, 1987) Injection drug users are therefore at high risk of HIV infection and investigations such as the one performed here can be of benefit to this specific population, particularly because there have been few studies of this at-risk group.

In studying injection drug users, we need a quantifiable way to measure the intensity of their habit. This quantifiable measure must meet two requirements. First, it can be used to track changes in the frequency of injected drug use in our longitudinal dataset. Second, it may be associated with brain biology; in other words, it may link to specific genetic markers. In a study of smokers, the concept of "smoking pack-year" helped researchers to link pack-year history to a specific genetic region on chromosome

5q. (A. W. Bergen et al., 1999; Caporaso et al., 2009; Duggirala et al., 1999) Borrowing this concept, we create a phenotype, injection years. However, it is necessary to deal with a large amount of missing data before starting the analysis.

Genberg et al., using latent growth curve modeling and longitudinal ALIVE injection data, found that I.V. drug users can be categorized into five different groups with unique trajectories: persistent injection, frequent relapse, early cessation, delayed cessation, and late cessation. (Genberg et al., 2011) In this study, using the same data we impute longitudinal injection data in the interest of creating a single injection years phenotype which we then compare to the inferred injection trajectories previously derived for this data set.

3.3 Materials and Methods

3.3.1 The ALIVE cohort

With data from the AIDS Linked to the Intravenous Experience (ALIVE) Study, (Vlahov et al., 1991) this study uses a subgroup with GWAS data (N=1197). Information about each individual in the subgroup includes up to twenty-five years of longitudinal data and genome-wide single-nucleotide polymorphism (SNP) data. The ALIVE study is a longitudinal prospective community-based study of I.V. drug users established in 1988. In its first decade, the ALIVE study attempted to understand the natural history of HIV within this population. Since 1998, ALIVE has focused on participants' access to, treatment for HIV and other non-AIDS outcomes including hepatitis C and the impact of those treatments (Galai et al., 2003; Vlahov et al., 1991)

The initial enrollment in the ALIVE study was 2,938 participants eighteen years and older who had used injection drugs within the previous ten years. Seven hundred

were HIV seropositive. The first participants were 88% African-American, 81% male, with 77% reporting active injection. 1,733 recruits were added to the study in 1994-1995, 1998, 2000, and 2005-2008.

The ALIVE participants are divided into two cohorts. The ALIVE-I study follows a cohort of HIV-positive individuals and a sample of HIV-negative individuals; the ALIVE-II study follows a cohort of HIV- negative individuals. (Galai et al., 2003; Vlahov et al., 1991)

3.3.2 Injection years and multiple imputation

In the COGA study, Bergen et al. created a novel variable, smoking pack-year, which was defined as the number of packs smoked per day for one year. This linkage study found that this external phenotype is associated with a gene region on chromosome 5q. (A. W. Bergen et al., 1999) Caporaso et al. conducted a genome wide association study and linked a similar variable, cigarettes smoked per day, to several genetic markers. (Caporaso et al., 2009) Borrowing their concept, we sought to create a variable, injection years, which we define as the number of years a person has been using injected drugs. We posit that there is a strong possibility that this variable is a better reflection of the biological mechanisms underlying substance dependence and more likely to be associated with specific genetic markers. However, before we could proceed we had to develop a strategy to deal with the large amount of missing data in the ALIVE GWAS subset. Wang et al. demonstrated a hot-deck multiple imputation procedure in longitudinal data. We borrowed this idea in addressing this critical issue. (C. N. Wang, Little, Nan, & Harlow, 2011)

The ALIVE GWAS subgroup is our study sample (N=1197). The dataset comprises a panel of binary variables representing positive (‘1’) or negative (‘0’) reports of injection use twice annually. Each row contains a single subjects injection history coded from age 20 to age 60, with preprocessing to include known but not directly assessed measures. For example, all data points prior to the age of first use were coded to ‘0’ and the data point at the age of first use was coded to ‘1’. For the remaining data, we used all existing reports from each participant between the ages of twenty and sixty and treated all other non-available reports as missing data. Variables were aligned by age, thus each column represents injection in the sample at a specific 6 month age time-point. I also included two covariates as predictors in each model: gender and age of first injection. I compared three imputation models since longitudinal imputation models that account for temporality and are useful for binary data are not widely available. The three approaches are termed: universal, two-predictor and one-predictor. In the universal model, we used all existing reports as predictors. In the two-predictor model, we used two types of existing reports, six months prior and six months later. In the one-predictor model, we used only the six months prior report. We used multivariate imputation by chained equations to fill in missing data. (Buuren & Groothuis-Oudshoorn, 2011; C. N. Wang et al., 2011) Multiple imputation by chained equations was performed by using the program ice in Stata 12. (Royston, 2011; StataCorp, 2011) Multiple imputation by chained equations (or MICE) refers to an approach where, given a set of data with missing values across multiple variables, one variable is imputed at a time as the predicted value from a regression (logistic in this case) of the other variables. After imputation of the first variable, the second variable is imputed using the observed and imputed values of all

other variables, and so on. To estimate uncertainty or variance in the imputation estimate the approach can be repeated multiple times. For imputation in the ALIVE sample, number of iterations was five. We pooled all the imputations and calculated the average value from these imputed results. We then took the sum of the injection years for each participant to create the injection years variable. Sensitivity tests were done by random deletion of observed (ie, present in the actual data) raw data-points and re-imputing it to test accuracy. The number of re-imputations was ten for each model. We checked the rates of agreement between the observed data and the re-imputed data for these three models.

3.3.3 Relationships between injection years and HIV trajectories

Genberg et al. conducted an analysis of latent classes of drug use trajectory using a subsample of 1,716 individuals who had made at least eight follow up visits. (Genberg et al., 2011) Semi-parametric latent class growth mixture modelling was used to determine the number of latent drug use trajectories. (Nagin, 2005) In their final model, five distinct patterns were identified: persistent injection, frequent relapse, early cessation, delayed cessation, and late cessation. Three trajectories were associated with cessation. The early cessation group showed a sharp decline during the first five years, while the delayed cessation and late cessation groups displayed less steeply declining injection levels after approximately ten and fifteen years, respectively. **Figure 3** shows the results of the latest publication from the ALIVE study exploring the trajectories of injection drug use.

Using the same ALIVE cohort, we created the measure, “injection years”, to assess the intensity of injection drug use. Like Genberg et al.’s studies, this study focuses on injection drug use and analyzes the ALIVE longitudinal dataset. The goal is to determine whether there is a correlation between injection years and the trajectories of injection drug use.

3.3.3.1 Latent class growth model

Latent growth modeling approaches, including latent class growth analysis and growth mixture modeling, are most commonly used in the social sciences when tracking longitudinal trajectories of individuals. (Jung & Wickrama, 2008) MacCallum et al reviewed the applications of latent growth curve models in psychological research that used a data set with longitudinal and repeated measure variables, concluding that it is a powerful tool with great benefit in psychological research. (MacCallum & Austin, 2000) These models have been used to understand inter-individual differences in intra-individual change over time. (Nesselroade, 1991) The source dataset, the ALIVE study, is a prospective longitudinal study with repeated measures. There are precedents for applying latent growth models in this study.

Latent growth modeling approaches have previously been applied to longitudinal drug use data in other cohorts. Hill et al suggested that patterns of alcohol use can be classified into four trajectory groups: “early highs”, “increasers”, “late onsetters”, and “non-bingers”. (Hill, White, Chung, Hawkins, & Catalano, 2000) Jackson et al. found four trajectory classes, using growth mixture modeling with five indices of alcohol consumption. (Jackson & Sher, 2005) The Hill and Jackson studies suggested that a heterogeneity of growth trajectories exists within the larger population and furthermore,

that the conventional growth models, which assume that an entire population only has a single growth trajectory are oversimplified. Because this study focuses on a similar type of dependency (the injection behavior of addicts) these same theoretical frameworks can be used to divide the entire population into several subgroups.

Unlike the conventional growth model, growth mixture models make it possible to estimate the subpopulations within the entire population and to provide separate growth models for each latent class with its own estimates of variances and covariate influences. (Muthén & Asparouhov, 2008) Latent class growth analysis is a type of growth mixture model. The assumption of latent class growth analysis is that the variance and covariance estimates within each class are fixed to zero, assuming all individual growth trajectories within a class are homogenous. This approach serves as a preparation for identifying the distinct classes before developing general mixture models. (Kreuter & Muthén, 2007)

3.4 Results

The ALIVE GWAS subset includes 1,197 participants drawn from the ALIVE cohort in 1998. The median age at first visit was 34. There were 287 (24%) females and 882 (76%) males. They visited semiannually with 19 average visits per person. This translates into about 9 years of information on each subject. The population was predominantly African American (98%). About half of the participants were HIV positive when they entered the ALIVE cohort. 1,029 participants were injection drug users when they entered the study, while only 502 were still using injection drugs as of their last visit. When they first entered the study, the number of participants who were using more than once a day, less than once a day and not using were 449, 415, and 159, respectively. After

13 years in the program, 231 of the 496 injection drug users were still using more than once a day. **Table 2** shows the characteristics of the GWAS subset in the ALIVE cohort.

In the ALIVE GWAS subset, the mean age of first injection was 21, and the histogram is right-skewed. More than half used for the first time between ages 14 and 26. The mean age of participants entering the cohort was 34. The years between first use and entering the program were regarded as missing data, although one discreet point was extracted: age at first injection. **Figure 4** shows the distribution of ages at first injection and baseline age upon entry.

To obtain our target variable, injection years, we needed to have information regarding injection status after age of first injection. This presented a problem since so much data was not available. There were 103,599 visits recorded for all the participants in the ALIVE GWAS subset, but this represented just one-third of all potential data from the age of first use forward: 66 percent was missing. The largest gaps of information occurred in injection status between the ages of 20 and 30: 75 percent of the data was not recordable. In other words, the younger the participant, the greater the likelihood of unretrievable information. **Table 3** shows the distribution of missing data in the ALIVE GWAS subset.

We imputed the ALIVE GWAS subset with three models: universal, two-predictor, and one-predictor. Respectively, the mean of the summary of injection years was 23, 19, and 18 years in the universal, two-predictor, and one-predictor model. The mean of the summary of injection years in the universal model was much greater than that of the other two models. In the universal model, we used all other existing data to impute the missing data. Considering there were more “Use” reports between the ages of

30 and 40 and 40 and 50, we imputed more "Use" reports than with the other two models in the age range 20 to 30. **Figure 5** shows the distribution of the imputed injection years of 1,279 individuals from the ALIVE GWAS subset.

Five percent of the fields from the injection data were randomly selected and were treated as missing data. We imputed these data ten times with each of the three models. We compared the imputed injection data with the original injection data and calculated agreement percentage and Kappa for each model. The findings suggested that the highest value for kappa occurred in the universal imputation model. **Table 4** shows the actual value and imputed result agreement in these three models.

After conducting the sensitivity exam for these three imputation models, we chose the universal model for the subsequent analysis. The individuals who were enrolled in both the ALIVE GWAS group and Genberg's study were divided into five groups. (Genberg et al., 2011) The mean injection years were 17, 22, 22, 24, and 29 for early, late, delay, relapse, and persistent groups, respectively. As predicted, the persistent group had the highest mean injection years, and the early quit group had the lowest. For the overlapping group, the mean injection years was 24, a result consistent with that of Genberg et al. (Genberg et al., 2011) **Table 4** shows the distribution summary of injection years among the different trajectory groups of injection drug users.

3.5 Discussion

Substance use disorders have a huge impact on public health and personal health. Illicit psychoactive drugs are associated with a variety of psychological and physical problems. From a public health perspective, in the U.S. there is an enormous societal cost to the use of drugs like opioids. (Birnbaum et al., 2011) The costs to the health care and

criminal justice systems, combined with losses in workplace productivity, add up to $50 billion *annually*. (Birnbaum et al., 2011; Hansen et al., 2011) From a personal health perspective, injection drug use leads to a whole host of physical problems and one of the most serious is the high risk of acquiring HIV. (Friedland & Klein, 1987) There have been few studies of this specific population; we therefore hope this study can be of benefit to this group and to those working on all fronts to help them.

Substance use disorders are chronic and recurrent. Changes in behavior patterns are common among individuals with substance abuse disorders and may continue even after a person has quit using. For example, alcoholics may be irritable and impulsive even after many years of sobriety. Observing the changes in long-term behavior patterns can perhaps help us to understand more about the outcomes of addiction. As stated previously, it is almost impossible to track and collect data from a highly unstable population such as substance dependent individuals for any length of time, much less the amounts of time required for this study. The longer the study, the higher the amounts of missing data. In the ALIVE GWAS subset, investigators tracked their subjects for approximately ten years; thus, there are vast amounts of missing data. In studying injection drug users, we borrowed the concept of "smoking pack-year", which helped researchers to link pack-year history to a specific gene region on chromosome 5q. (A. W. Bergen et al., 1999; Caporaso et al., 2009; Duggirala et al., 1999) We created a phenotype, injection years, to track changes in the status of injected drug use in our longitudinal dataset. We found that it is necessary to account for the large proportion of missing data before starting the analysis. For this purpose, we imputed missing longitudinal data in this study.

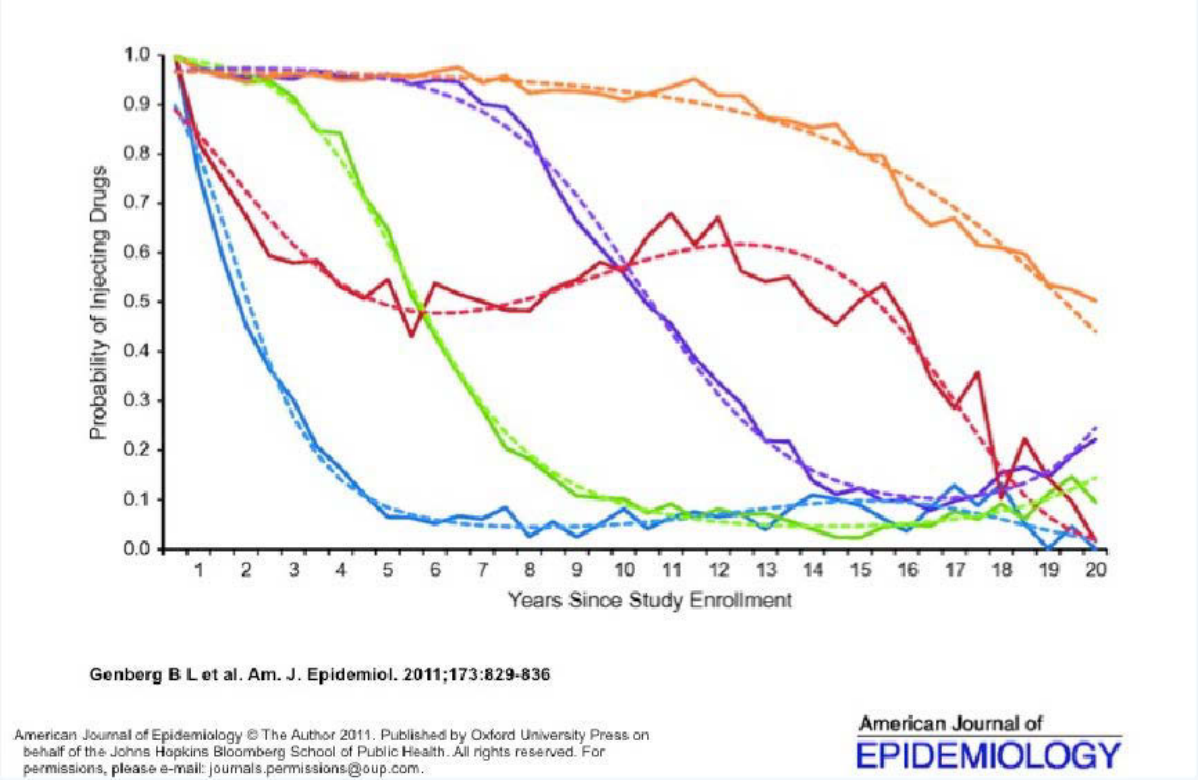
In this study, we imputed the ALIVE GWAS subset with three models: universal, two-predictor, and one-predictor models. In the ALIVE GWAS subset, there is more data missing inform the earlier phases of injection use than in the middle and late phases. In the universal model, we used all the existing reports as predictors with age of first injection and sex, so the imputed reports in the early phase were influenced by the existing reports in the middle and late phases; we supposed that several causal factors continuously influence injection behavior in all phases. In the two-predictor model, we used the two kinds of existing reports, six months prior and six months later, as predictors with age of first injection and sex, so the imputed reports in all phases were only influenced by those existing reports which were closest in time. We supposed that causal factors which influence injection behavior change in different phases but may remain the same in the identical phase. In the one-predictor model, we used only the existing report, six months prior plus age of first injection and sex, as a predictor, so the imputed reports were only influenced by the existing reports at the identical time point. The results of the sensitivity test support the universal model. Using the universal model, the mean of the sum of imputed injection years is consistent with the results of previous work in the ALIVE Cohort. (Genberg et al., 2011)

This study had a number of limitations. The first was the amount of missing data in the ALIVE GWAS subset. Overall, sixty-six percent of the possible data regarding the injection career is missing, if we include the time period between a subjects first injection and entry into the Cohort and assume that a visit was possible twice annually. The 20 to 30 year age range actually has a 75 percent rate of missing data under these parameters. The accuracy of imputation is associated with the amount of existing data. These amounts

of missing data would definitely decrease the accuracy of imputation. Second, the definition of injection years is the number of the years the person has been injecting drugs; we only focused on the length of injection time and ignored the dosage and the injection counts. Although we think our study target, the natural trend of injection behavior, is mostly influenced by the injection years, we cannot exclude the possibility that our study target is also influenced by these other factors.

Using the universal imputation model, we got the mean of the sum of injection years which was similar to previous work in the ALIVE Cohort. (Genberg et al., 2011) The findings of this study support the use of multiple imputation in a longitudinal cohort such as the ALIVE GWAS subset. Clearly, missing data presents a major obstacle for conducting longitudinal studies; multiple imputation using specific models may provide a solution to this problem. We highly recommend using multiple imputation in other longitudinal studies and to examine the accuracy and reliability of the imputed data. By using the appropriate model, we believed that reliable imputed records could be obtained from the existing records. Another potential area for future studies is to use multiple imputation to impute continuous variables such as the injection counts. Caporaso et al. used the concept of "smoking pack-year," which is defined as the number of packs smoked in a year. (Caporaso et al., 2009) In this study, we did not deal with the intensity of single injection, but we think that it could be a worthwhile area of investigation.

Figure 3: Trajectories of Injection Drug Use Among 1,716 Injection Drug Users in The AIDS Linked to The Intravenous Experience (ALIVE) Study, Baltimore, Maryland, 1988–2008.[+]

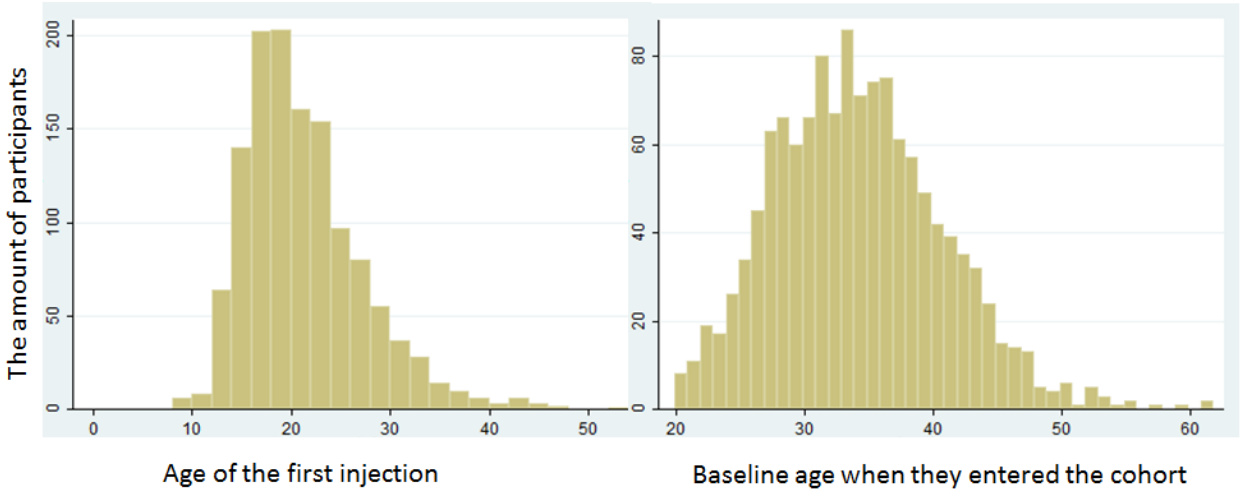


[+] "The dotted lines represent the predicted probabilities of injection drug use conditional on membership in one of the 5 drug-use groups, while the solid lines represent the observed proportion of injection drug use given group membership. The y-axis represents the conditional probability of injection drug use, while the x-axis reflects time since study enrollment. The 5 groups (and prevalence of group within sample) are depicted with the following colors: blue, early cessation (19%); green, delayed cessation (16%); purple, late cessation (18%); red, frequent relapse (16%); orange, persistent injection (32%)."

Table 2: ALIVE GWAS sample N=1197

	First follow up visit	Last follow up visit
Age(median & IQR)	34(30-39)	47(40-53)
Female	287(24%)	287(24%)
African American	1169(98%)	1169(98%)
HIV positive	627(52%)	828(69%)
Current user	1029(86%)	502(42%)
Frequency of iv use		
non	159(16%)	691(58%)
<1/day	415(41%)	265(22%)
>=1/day	449(44%)	231(19%)
Any non-iv use	620(61%)	343(29%)
Cigarette use:		
None	100(10%)	233(20%)
<1/2 pack/day	132(13%)	219(18%)
1/2-<1 packs/day	298(294%)	346(29%)
1-<2 packs/day	368(36%)	328(28%)
>=2 packs/day	128(12%)	63(5%)
Alcohol use		
None	197(19%)	585(49%)
<daily	604(59%)	491(41%)
Daily	222(22%)	114(10%)
N follow up visit(med & IQR)	19(11-30)	

Figure 4: Distribution of first injection age and baseline age

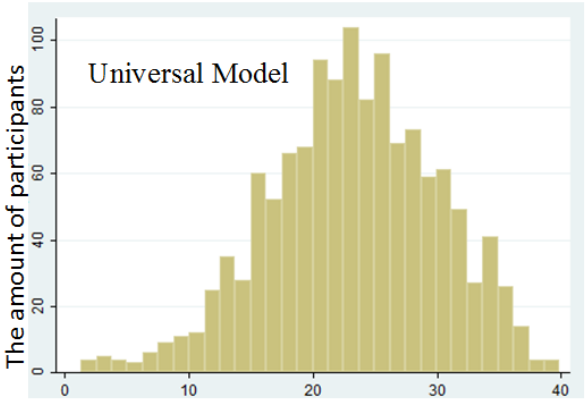


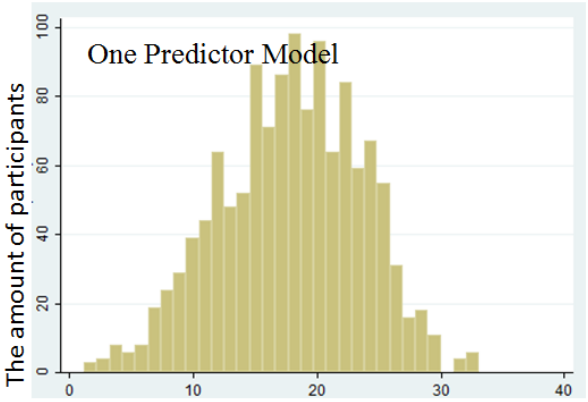
	Participants	Mean (S.D.)	Min	Max
First injection age	1278	20.79 (5.96)	8	47
Baseline age	1278	34.19 (6.54)	19.85	61.25

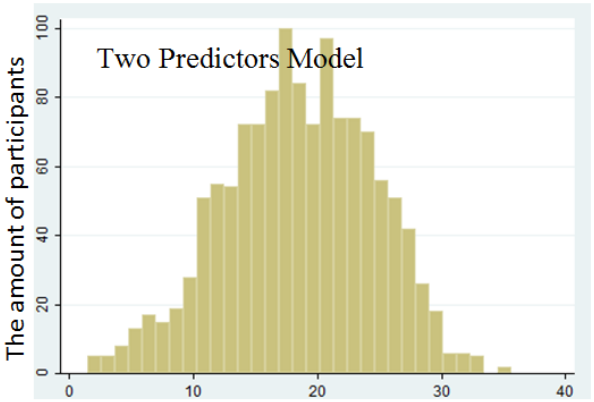
Table 3: Distribution of missing data in the ALIVE GWAS subset

Injection status of last 6 months	Not use (%)	Use (%)	Missing (%)	Total
Age 20 to 60	19,398 (18.72)	15,001 (14.48)	69,200 (66.8)	103,599
Age 20 to 30	4,715 (18.93)	1,506 (6.05)	18,685 (75.02)	24,906
Age 30 to 40	4,129 (15.53)	6,699 (25.50)	15,758 (59.27)	26,586
Age 40 to 50	5,581 (20.80)	5,862 (21.84)	15,395 (57.36)	26,838

Figure 5: Distribution of the imputed injection years of 1,279 individuals from the ALIVE GWAS subset







Sum of Injection year	Mean (S.D.)	Min	Max
Universal	23.21 (6.85)	1.3	39.9
2 predictors	18.53 (6.03)	1.5	35.7
1 predictor	17.90 (5.72)	1.2	33.1

Table 4: Actual value and imputed result agreement in one predictor, two predictor and universal imputation models

	Actual value: 0	Actual value: 1	Agreement, % (Kappa)
One predictor model			
Imputed result: 0	6362	3138	62 (0.22)
Imputed result: 1	3283	4218	Fair
Two predictor model			
Imputed result: 0	7357	2063	74 (0.47)
Imputed result: 1	2288	5293	Moderate
Universal model			
Imputed result: 0	7023	1796	74 (0.48)
Imputed result: 1	2622	5560	Moderate
Total	9645	7356	

Table 5: Distribution of the summary of injection years among different trajectory groups of injection drug user+

Sum of Injection years	Participants	Mean(S.D.)	Min	Max
Injection trajectory groups				
Early	153	16.9 (5.36)	1.3	28.3
Late	146	21.5 (5.30)	7.3	32.3
Delayed	149	22 (4.12)	11.5	30.9
Relapse	152	24.25 (4.92)	12.4	37.5
persistent	299	29 (4.95)	13.9	39.9
Total	899	23.87 (6.63)	1.3	39.9

+ The detail of the trajectories of injection drug use among 1,716 injection drug users in the ALIVE Study were mentioned in Figure 3

Chapter 4: Measuring genetic influence on injection years: GWAS, polygenic risk score and pathway analysis

4.1 Abstract

BACKGROUND: Opioid addiction is heritable. We posit that injecting heroin may have genetic underpinnings. Using the ALIVE cohort, Genberg et al. tracked the trajectories of intravenous drug users and categorized them into five groups. It was possible to quantify injection behavior patterns. We devised the injection years variable to quantify the duration of longitudinal injection behavior. We propose that this variable may have a genetic basis.

METHODS: The AIDS Linked to the Intravenous Experience (ALIVE) GWAS cohort (N=1197) was used. Each subject has twenty-five years of longitudinal data and a single-nucleotide polymorphism (SNP). We created a variable, injection years, the length of injection time. Genome wide association analysis was used to examine the associations between genetic factors and injection years. Data management, genotype quality control, population stratification, and correction for multiple testing were conducted. After a genome wide association analysis, a polygenic risk score analysis and a pathway analysis were performed.

RESULTS: Using PLINK, 1,107 people and 735,081 variants passed filters and quality control. Phenotype data, injection years, is quantitative. Using regression analysis, we tested the SNP associations after adjusting for gender and genome-wide ancestry covariates. No single SNP in the GWAS reached the level of genome-wide significance.

Eleven SNPs exceeded the $p < 5 \times 10^{-5}$ threshold, including one SNP (rs10168062) with $p < 5 \times 10^{-7}$. The closest SNP with a previous report of a genome-wide significant association result is rs7602960 which is associated with memory. The findings from a polygenic risk score analysis support a polygenic basis for injection years that involves many common SNPs. 31 intervals and 5,225 gene sets from Gene Ontology (GO) terms were used in a pathway analysis. None of the gene sets was statistically significant. The top two gene sets were metallocarboxypeptidase activity and DNA dealkylation which are involved in DNA repair.

LIMITATIONS: Genetic and environmental factors are inextricably linked. There is insufficient data for a powerful polygenic risk score analysis.

CONCLUSION: We used longitudinal addiction behavior to perform genome-wide association analyses. The result suggested that injection years might be affected by polygenes. Polygenic risk score analysis further supported this conclusion. The findings from pathway analysis suggest that DNA repair may play a role in our risk phenotype, injection years. Using injection years as a phenotype in genome wide analysis proved to be a successful way to explore genetic influences on specific behavior traits in a longitudinal cohort.

KEY WORDS: injection years, GWAS, polygenic risk score, pathway analysis

4.2 Introduction

As we reviewed in chapter 2, substance use disorders are heritable; many articles provide evidence to support the role of genetics in nicotine dependence and alcohol use disorders. Heroin, one of the notorious illicit drugs, had the greatest overall genetic variance, 0.54. (Tsuang et al., 1998) Kreek et al. demonstrated that developing substance

use disorders may be partly due to heritable factors through underlying behavioral risk dimensions like impulsivity and risk taking. (Kreek et al., 2005) Her conclusion suggested that specific behavior patterns related to addiction are associated with genetics. Borrowing her concept, we believe that injection, a harmful behavior related to heroin addiction, may be associated with genetics. Injection drug use affects personal health and wellbeing through transmission of blood-borne diseases; moreover, it causes huge losses in a nation's economic productivity.

The ALIVE study is a longitudinal prospective community-based study of injection drug users in Baltimore. It provides valuable information of injection behavior patterns and HIV risk behaviors from a group of injection drug users for twenty-five years. By tracking the trajectories of injection drug users, we can categorize the ALIVE cohort into several subgroups: early quit, persistent, and relapse group or quantify the ALIVE participants' injection behavior patterns. Bergen et al. developed a novel variable, smoking pack-year, and their linkage study found that it is associated with a gene region on chromosome 5q. (A. W. Bergen et al., 1999) Using genome wide association analysis, Caporaso et al. associated a similar variable, cigarettes smoked per day, to several genetic markers. (Caporaso et al., 2009) We borrowed this concept to formulate a variable, injection years, to quantify the intensity of longitudinal injection behavior. In this chapter, we examined whether this variable is associated with SNPs, both singly and in composite as polygenic or systems-level scores. Furthermore, in order to reveal the associated brain mechanism, we used the existing dataset to examine whether our findings links to the specific proteins or receptors.

4.3 Materials and Methods

4.3.1 Subjects

The AIDS Linked to the Intravenous Experience (ALIVE) Study was used in our study. (Vlahov et al., 1991) More specifically, we used the subgroup with genome-wide single nuclear polymorphism data (N=1197) from the ALIVE cohort. The details of the ALIVE cohort have been already described in the chapter 3.

4.3.2 Selection of phenotypes for longitudinal injection information

In 1999, Bergen et al. conducted a linkage study and found smoking pack-year is associated with a gene region on chromosome 5q. (A. W. Bergen et al., 1999) In 2009, Caporaso et al. conducted a genome wide association study and linked a similar phenotype, cigarette smoked per day, to several genetic markers. (Caporaso et al., 2009) As well as smoking, we believe that genetics have a high probability to be involved in injection behaviors. The procedure of developing our phenotype, injection years, and the detailed description of the relationship between genetic influences and drug use behavior and more specifically, the previous successes in drug use genetics by studying variable use behavior in drug using populations has been already described in the chapter 3.

4.3.3 Genome wide association analysis

A genome wide association study is a way of measuring and analyzing common DNA sequence variations, in particular single nucleotide polymorphisms, to determine whether a variant is associated with a specific trait from a common disease in the population.(Hirschhorn & Daly, 2005) Single nucleotide polymorphisms are single base-pair changes in the DNA sequence, occurring in more than one percent within a

population; such variants are also regarded as potential genetic risk factors or as proxies for nearby SNPs in linkage disequilibrium.(Genomes Project Consortium, 2010) A goal of human genetics is to identify genetic risk factors for traits or diseases. This goal was partly achieved by identifying the genetic risk factors for rare Mendelian diseases such as cystic fibrosis using linkage analysis. (Kerem et al., 1989) This approach has been also successfully applied to other rare diseases such as Huntington's (MacDonald et al., 1992) , but it failed to identify the genetic risk factors for common diseases like various types of cancer and schizophrenia. This led to the common disease/common variant hypothesis which states that common diseases are likely influenced by genes that are also common in the population. (D. E. Reich & Lander, 2001) Assuming the validity of the common disease/common variants hypothesis, the first successful genome wide association study was published in 2005, identifying the Complement Factor H gene for age related macular degeneration. (Haines et al., 2005) Unlike genome association studies or candidate gene studies, genome wide association studies search the entire human genome instead of focusing on a small number of candidate gene regions. This approach has now been applied to numerous common diseases; the National Human Genome Institute GWAS catalog lists over 3,600 SNPs implicated in common diseases or traits. (Hindorff et al., 2009)

The most common study design among genome wide association studies is the case-control study in which the case group is affected individuals and the control group is healthy individuals. An alternative to the case-control study is the quantitative phenotype, as opposed to the binary phenotype, for example using a variable such as height. All individuals are genotyped for common SNPs, most frequently using one of two primary

platforms: Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA). The data form of GWAS analysis includes two basic parts: phenotype and genotype. The phenotype data contains demographic information such as gender, age, and risk phenotypes. This information will be obtained from the former chapter. The genotype data was obtained from genotyping. Because GWAS data contain vast amounts of information on genotype (750,000 SNPs for individual) I needed to consider a number of issues to deal with large scale that need to be addressed before, during or after genome wide association testing. These issues include data management, genotype quality control, population stratification, genotype imputation, and correction for multiple testing. Quality control procedures discard genotypic information of poor quality—for example, poor calling rates. Population stratification adjusts for the different geographical and ethnical backgrounds of participants. Genotype imputation increases the number of SNPs that can be tested for association and increases the power of the study. Correction of multiple testing prevents type I errors.

In this study, I analyzed larger scale genetic association data with the appropriate computing hardware: the High Performance Computing Center at the JHSPH, and software which will include PLINK and R. In addition to the appropriate devices and software, the advisor to this study, Dr. Brion Maher, has considerable experience managing large datasets and has also been a primary analyst on several GWAS projects. The conductor of this proposed study, Shaocheng Wang, has already taken the required lectures on genetic epidemiology and has performed GWAS analysis under Dr. Brion Maher's supervision.

Genome wide association analysis may bring an increase in the number of false positive associations because of biases in study design and errors in genotype calling. Several quality control steps will be taken to remove individuals or makers with extremely high error rates in order to avoid these false positive associations. First, I identified and eliminated subjects with discordant sex information. Second, because the impact of removing a potential marker is greater than the removal of one individual, I implemented quality control on a "per individual" basis prior to conducting quality control on a "per-marker" basis. (Anderson et al., 2010) I eliminated subjects with high levels (>5%) of missing genotypes and then eliminated markers which exhibited high levels (>5%) of missing data. Third, I tested Hardy-Weinberg equilibrium (HWE) in order to eliminate markers that exhibit large deviations from expected genotype distribution. I eliminated markers exhibiting Hardy-Weinberg disequilibrium ($p < 0.00001$). By using these stringent criteria, I attempted to avoid the inclusion of markers exhibiting systematic genotyping errors. In brief, the purpose of the above quality control measures is to ensure data precision and ultimately to detect true genotype phenotype relationships.

In GWAS, population substructure may cause bias from systematic differences between cases and controls or between individuals at extremes of the phenotypic distribution. Differences in phenotype prevalence by ethnicity and in allele frequencies by ethnic sub-population highlight the notion that specific genetic markers may be more likely associated with the diseases/traits spuriously because of population stratification. I applied multidimensional scaling as implemented in PLINK to generate genome-wide ancestry covariates prior to performing GWAS to examine the relationships between

genotype and phenotype. I used complete linkage agglomerative clustering based on pairwise identity-by-state distance.

In a genome wide association study, hundreds of thousands of tests are conducted, which means the cumulative likelihood of finding one or more false positives is higher than if only one test is done. Several simple techniques are used to correct for multiple testing. These include Bonferroni correction, permutation testing and adjusting for false discovery rate (FDR). In this study, I focused on Bonferroni correction. (Gao, Starmer, & Martin, 2008)

4.3.4 Polygenic risk score and Pathway analysis

After obtaining the results from genome wide association analysis, we planned to conduct a polygenic risk score analysis and a pathway analysis. The polygenic risk score analysis creates a score from the top SNPs from the GWAS on our target phenotype and the pathway analysis tests whether the top SNPs from the GWAS of our target phenotype yield information on the possible biological mechanisms due to an enrichment of SNPs in genes in specific biological pathways. Both approaches are commonly used to examine the common disease-common variant hypothesis.

The polygenic risk score analysis is a statistical approach which is used to summarize genetic effects among a group of SNPs which do not have significant associations with diseases/traits. The approach is based on the assumption that although many SNPs do not reach significance after correcting for genome-wide testing, the tail of the distribution of p-values less than some target threshold will be enriched for true signal. A GWAS is conducted first on a training sample and the p values of SNPs are obtained. A polygenic risk score is then constructed in an independent sample, as a

weighted sum scores trait-associated alleles for each subject, for different subsets of top ranking markers.(Dudbridge, 2013)

The first successful polygenic score analysis in/within a GWAS was applied to schizophrenia and bipolar disorder.(Purcell et al., 2009) This approach has two possible applications. First, the polygenic scores can be used to determine the association between a disease/trait and selected SNPs. Second, the polygenic scores can be used to predict individual disease/trait value. This approach has attracted much attention and has been used with several common and complex disease including multiple sclerosis, cardiovascular risk, and rheumatoid arthritis.(Consortium, 2010; Simonson, Wills, Keller, & McQueen, 2011; Stahl et al., 2012) In this study, I applied polygenic risk score analysis using the injection years phenotype to search for associations between selected genetic risk factors and addiction, and the findings of polygenic risk score analysis have the potential to be used as a prediction tool for injection years. The male group is used as discovery sample and the female group is our target sample.

In order to compensate for the limitations of GWA studies, several alternative approaches to GWA studies have been developed in recent years. In this section, I focus on one of them, the pathway analysis. Unlike the analysis of a single marker as is the method in a GWAS, pathway-based approaches examine whether a group of biologically related genes which are discovered from previous research have significant association with a disease/trait.(K. Wang, Li, & Hakonarson, 2010) Most common and complex diseases are caused by the interaction of multiple genes which involve complex molecular networks and cellular pathways.(Schadt, 2009) In brief, a pathway-based analysis is an alternative to GWAS by combining association results with known

biological pathways; it also offers a biological interpretation of diseases/traits, in particular risk phenotypes used in GWAS.(Cantor, Lange, & Sinsheimer, 2010)

Pathway based analyses have been used with several neuropsychiatric disorders. The associations between the pathway of neuronal cell adhesion molecules and either autism, schizophrenia or bipolar have been examined in two studies.(O'Dushlaine et al., 2010; K. Wang et al., 2009) In addition, a pathway analysis found ion channel activity and synaptic neurotransmission to be associated with bipolar disorder.(Askland, Read, & Moore, 2009; Holmans et al., 2009)

In this study, I used pathway analysis with one of the most common of neuropsychiatric disorders: addiction. I used the GWAS results from our previous work and performed a pathway analysis. I used the software tool, Interval-based Enrichment Analysis Tool for Genome Wide Association Studies (INRICH), to perform our pathway analysis and the reference biological pathways were obtained from the Gene Ontology (GO). (Lee, O'Dushlaine, Thomas, & Purcell, 2012)

4.4 Results

Using PLINK, we applied multidimensional scaling to generate genome-wide ancestry covariates prior to analyzing the genome wide associations. None of the 1,171 subjects was excluded because of population stratification. The calculated ancestry covariates were used to account for genetic ancestry in subsequent association analysis.

Before quality control, there were 1,171 individuals and 906,709 genotypes in the ALIVE GWAS cohort. Because of missing genotype data, 64 subjects had to be excluded. Additionally, 59,813 variants were removed also due to missing genotype data. 13,578 variants were excluded from the analysis due to departure from HWE exceeding $\alpha =$

0.00001. Another 98,237 variants were removed due to minor allele threshold. In the final analysis, 735,081 variants and 1,107 participants passed filters and quality control.

The injection years phenotype data is quantitative. Using regression analysis, we tested the SNP associations after adjusting for risk factors such as gender and eight genome-wide ancestry covariates. We found that no single SNP in the GWAS reached the level of genome-wide significance. ($p < 5 \times 10^{-8}$) (Figure 6) However, eleven SNPs exceeded the $p < 5 \times 10^{-5}$ threshold, including one SNP (rs10168062) with a $p < 5 \times 10^{-7}$. (Table 5) The closest SNP is rs7602960 which is associated with memory. (Seshadri et al., 2007)

Because the ALIVE GWAS subset is unique, it is not possible to find another independent GWAS sample with which to replicate the polygenic risk score analysis. We followed the process described in the supplementary information in Purcell's article (Purcell et al., 2009) and divided the ALIVE GWAS sample into a female group (N=287) and a male group (N=910). We used the ALIVE GWAS male group for the discovery sample, considering four P_T thresholds (arbitrary p-value thresholds below which all SNPs are summed) from the intra-ALIVE GWAS male group analyses. The ALIVE GWAS female group was used as our target sample. The male-derived polygenic risk scores were highly positively correlated with injection years in the entire ALIVE GWAS subset. The value of correlation coefficient is 0.7420077. (Supplementary Figure. 1) Using the male-derived scores in the female group with different P_T thresholds, the correlation coefficient increased and then reached a plateau as the P_T increased. (Supplementary Figure. 1, Figure. 7) The highest correlation coefficient is 0.1041739 when P_T is 0.01. However, due to the relatively small sample size, the correlation was not

statistically significant. In conclusion, the ALIVE GWAS subset suggested that injection years involve many common SNPs.

Using Interval-based Enrichment Analysis Tool for Genome Wide Association Studies (INRICH), we performed Gene Ontology (GO) enrichment analyses. Human genome references, hg19, were obtained from the UCSC Genome Browser. I selected the SNPs which reached the $P_T < 0.0001$ and the linkage disequilibrium is under 0.5 from our genome wide association analysis results with risk phenotype as injection years. 61 clumps (interval) were formed from 78 top SNPs. I excluded 22 intervals which were not on gene regions and 8 intervals which were overlapping; 31 intervals remained in the pathway analysis. After merging and size filtering, we restricted the analysis to terms with at least 3 human genes and considered gene sets with at least 2 overlapping intervals, leaving 5,225 from the total 10,365 GO terms. A permutation procedure was conducted to rule out the probability of observing the number of intersecting intervals by chance alone. We repeated the first-pass permutations and second-pass permutations 1,000 times each to correct for bias from multiple testing. The corrected P values account for the dependence of the GO terms.

These 31 intervals were selected for the pathway analysis, but none of the gene sets proved to be statistically significant. One of the top gene sets is metallocarboxypeptidase activity, including AGBL4 (ATP/GTP binding protein like 4), CPM (Carboxypeptidase M), CPE (Carboxypeptidase E), and AGBL1 (ATP/GTP binding protein like 1). Another is DNA dealkylation involved in DNA repair, including MGMT (O-6-methylguanine DNA methyltransferase), and ALKBH2 (AlkB, alkylation repair

homolog 2). (Table 6) In the global enrichment analysis, 6 and 77 unique genes have empirical p value less than 0.01 and 0.05, respectively. (Table 7)

4.5 Discussion

In this study, we tested the association between longitudinal addiction behavior and genetic risk factors. Longitudinal studies of this population are rare; missing data is a common problem in such studies. Having access to the ALIVE GWAS cohort, a longitudinal community-based cohort, made this study possible. Using multiple imputation, we created a risk phenotype, "injection years". This risk phenotype was used in the genome wide association analysis. However, no single SNP in the GWAS reached the threshold of genome-wide significance. Eleven SNPs exceeded the $p < 5 \times 10^{-5}$ threshold; the top three findings were rs10168062, rs2200572, and rs10187215 clustered on chromosome 2, in an uncharacterized locus. The closest SNP with a previous association was rs7602960 that has been associated with memory. (Seshadri et al., 2007) This result suggests that injection years might be affected by polygenes and/or gene-gene interactions and/or gene-environment interactions, rather than by a single gene.

Traditional genome wide association studies have to potential explain a large proportion of genetic risk factors for diseases but cannot detect genetic subsets in which each gene has a very small size effect without very large sample sizes. Polygenic risk score analyses and pathway analyses are used to detect those genetic subsets, using the results from a genome wide association analysis. Vink et al. examined polygenic risk scores for smoking and found that smoking, alcohol and cannabis use are influenced by many common genetic variants.(Vink et al., 2014) The results of Vink's polygenic risk score analysis support the proposition that injection years are influenced by polygenes.

However, no significant results were found in the pathway analysis. The top two gene sets are metallocarboxypeptidase activity and DNA dealkylation involved in DNA repair. This suggests that DNA repair may play a role in our risk phenotype, injection year.

There were a number of limitations in this study. First, we have proposed that gene-environment interactions may play a role, for example life events. However, in a longitudinal study like the ALIVE GWAS cohort, it is hard to identify and isolate an environmental variable that may have impacted drug use trajectories at a particular time point. Although our target is the trait of injection behavior, which is influenced by genetic factors, we cannot disentangle the interactions between genetic factors and environmental factors; the results of our analysis may be confounded by environmental factors. Second, there were only 1,200 research subjects. Second, we only had 1,200 research subjects. Because there are so few studies of injection drug users, we were not able to find an independent sample to serve as a discovery sample for the polygenic risk score analysis. We attempted to make up for this by creating gender-based subsets. However, the numbers in these subgroups were not high enough for a satisfactory polygenic risk score analysis.

Linking injection years to genetics, this study offers a practical method for exploring genetic influences on specific behavior traits in a longitudinal cohort. We believe that there are two areas that warrant future research. First, there have been few studies of the genetic influences on drug-related behavioral phenotypes; we believe that this study's use of the latest techniques such as polygenic risk score analysis and pathway analysis provides valuable information on how genetic factors influence injection behavior. We have attempted to determine whether there is a large portion of genetic risk

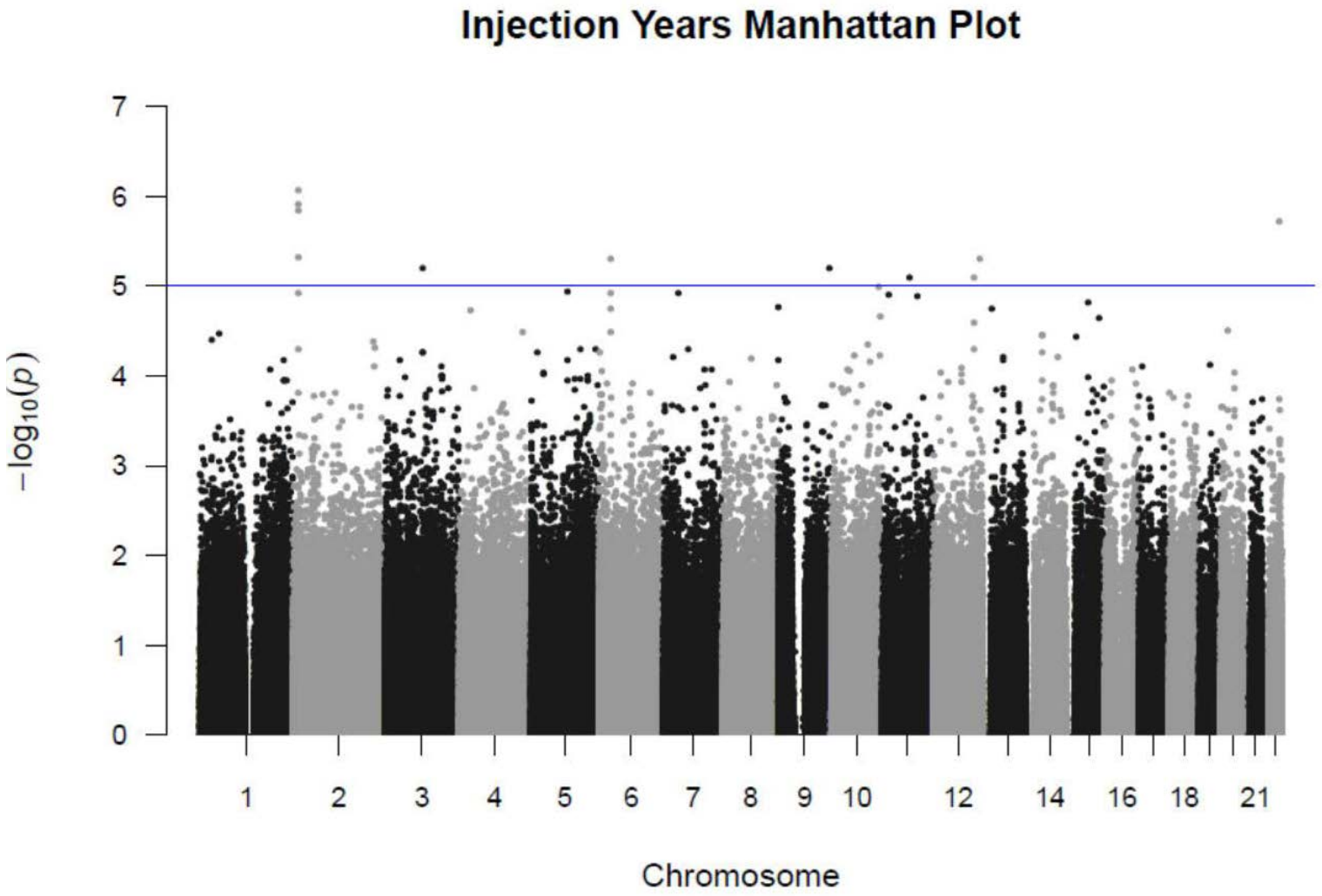
for the phenotype of injection years, whether there are genetic subsets in which each gene has a very small size effect, and whether there are gene sets which influence specific biological pathways.

Second, there are few studies of the genetic influences on longitudinal phenotypes. We used injection years to represent longitudinal trends in injection behavior. Though there are challenges posed by using a longitudinal cohort, we believe that genetic influences persist throughout a person's life; thus, for the purpose of this research, the longitudinal study provides more valuable information than would a cross-sectional study. In conclusion, we recommend using behavioral phenotypes for genome wide analyses to test the association between genetics and observable behaviors; we believe that this method can provide valuable information about how genetic factors influence behavior through specific biological mechanisms.

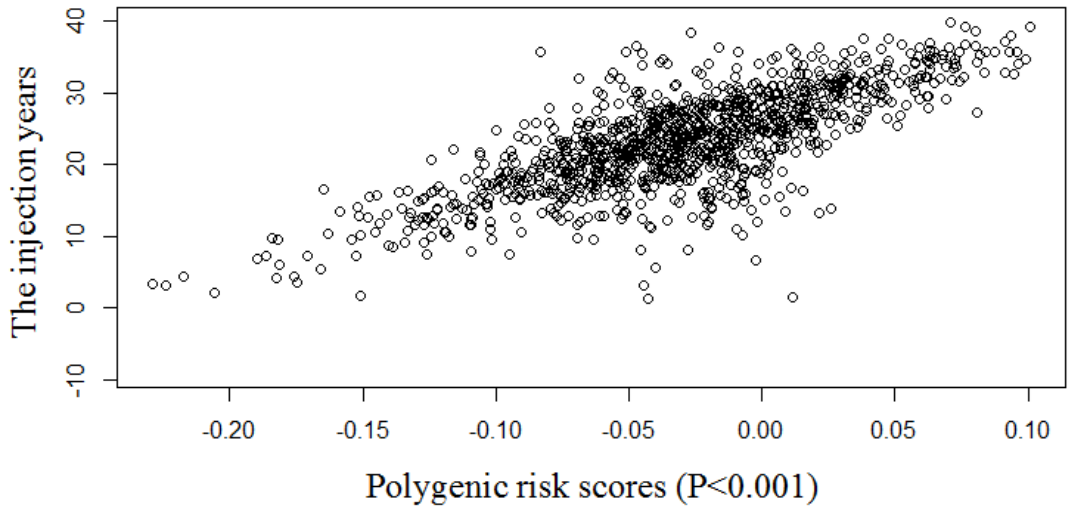
Table 6: Top 20 SNPs, their chromosomal locations, minor alleles, test mode, numbers of non-missing individuals included in the analysis, β for linear regression, coefficient for t-test, and p-value.

Chr	db SNP ID	Physical positon (BP)	Minor/Major allele	TEST	Non-missing individuals (N)	β (Regression coefficient)	T-test coefficient	P value
2	rs10168062	13418345	A/T	ADD	1112	1.386	4.946	8.75E-07
2	rs2200572	13450765	T/A	ADD	1101	-1.412	-4.878	1.23E-06
2	rs10187215	13459528	C/G	ADD	1112	-1.38	-4.844	1.45E-06
22	rs742837	43880111	A/T	ADD	1111	-2.138	-4.79	1.90E-06
2	rs13416262	13458652	G/C	ADD	1113	1.293	4.594	4.84E-06
12	rs12371640	124636328	A/T	ADD	1112	-2.126	-4.587	5.02E-06
6	---	31134599	A/T	ADD	1053	-1.92	-4.587	5.04E-06
3	rs771795	101718657	T/A	ADD	1096	-1.274	-4.54	6.25E-06
9	rs12002290	136747854	C/G	ADD	1093	1.577	4.539	6.29E-06
11	rs666839	73346041	A/T	ADD	1106	-1.434	-4.486	8.04E-06
12	rs11067936	110144932	T/A	ADD	1112	1.424	4.481	8.19E-06
10	rs3781452	126355129	C/G	ADD	1074	1.356	4.431	1.04E-05
5	rs13174024	97930167	T/A	ADD	1109	-1.719	-4.404	1.17E-05
6	---	31130019	C/G	ADD	1112	-1.746	-4.401	1.18E-05
7	rs10486704	41055230	T/A	ADD	1113	-1.303	-4.398	1.20E-05
2	rs6432374	13463725	G/C	ADD	1108	-1.283	-4.393	1.22E-05
11	rs537338	17926525	A/T	ADD	1113	1.495	4.387	1.26E-05
11	rs7940754	93345679	T/A	ADD	1106	-1.783	-4.382	1.29E-05
15	rs12440934	58425804	G/C	ADD	1105	1.797	4.348	1.50E-05
9	rs7869992	462759	C/G	ADD	1100	-1.385	-4.317	1.73E-05

Figure 6: The manhattan plot of the p-values from all 735,081 SNPs. The x-axis shows the chromosome numbers. The y-axis is the-log10 (p-value). The blue line is p-value of 10^{-5}.



Supplementary Figure 1: Correlation between injection years and polygenic risk scores among all participants when P value is less than 0.001.



Supplementary Figure 2: Correlation between injection years and polygenic risk scores among females with P value less than 0.001, 0.01, 0.05, and 0.1.

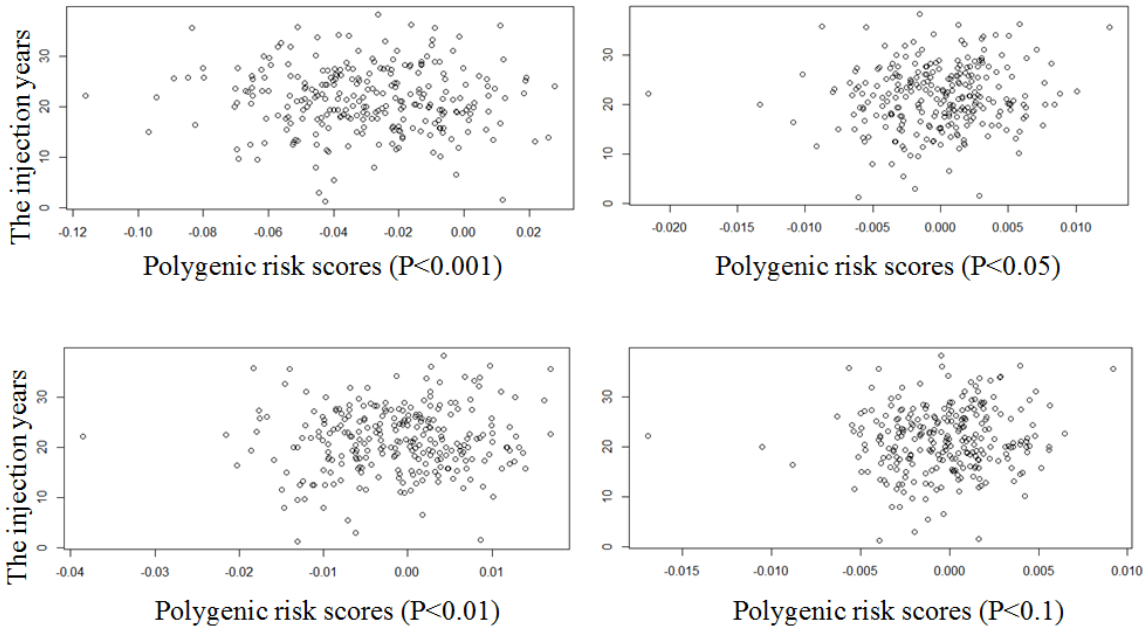


Figure 7: Correlations between injection years and polygenic risk score with P value less than 0.001, 0.01, 0.05, and 0.1.

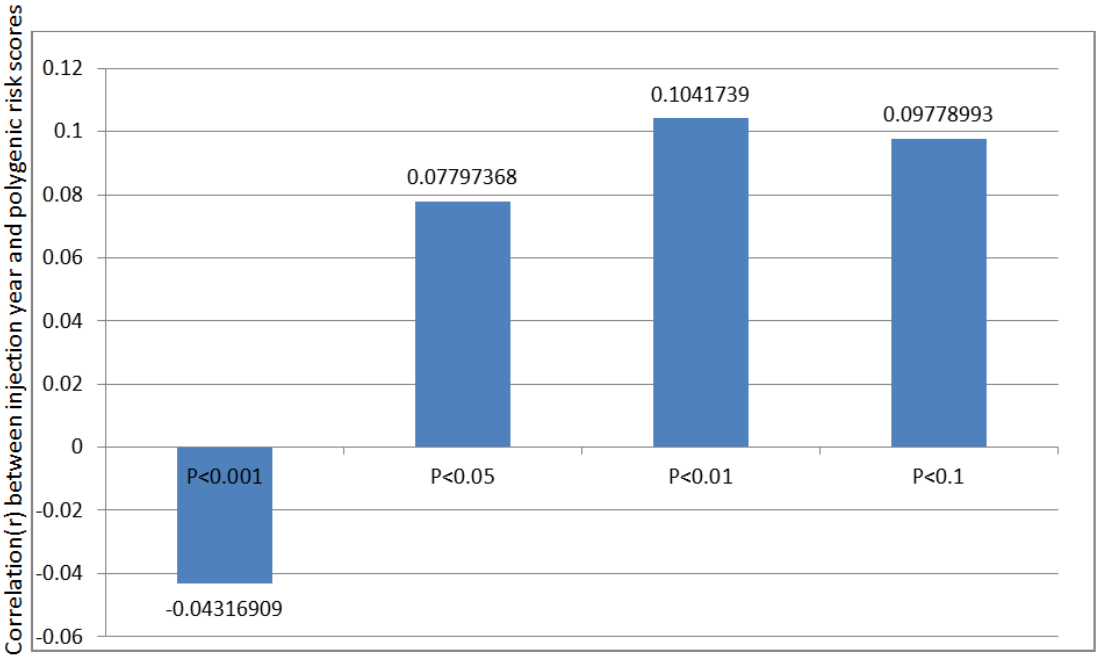


Table 7: Interval-based enrichment analysis results of Gene ontology (GO) biological processes for injection years.

Gene set id	Gene set annotation	Number of gene [a]	Number of interval [b]	Empirical P	Corrected P
GO:0001516	prostaglandin biosynthetic process	12	4	0.030969	0.99001
GO:0002576	platelet degranulation	77	5	0.030969	0.99001
GO:0004181	metallocarboxypeptidase activity	25	4	0.00999001	0.906094
GO:0005085	guanyl-nucleotide exchange factor activity	119	6	0.047952	0.999001
GO:0006307	DNA dealkylation involved in DNA repair	6	2	0.00999001	0.906094
GO:0006476	protein deacetylation	10	3	0.027972	0.987013
GO:0006541	glutamine metabolic process	14	4	0.048951	0.999001
GO:0007405	neuroblast proliferation	14	4	0.025974	0.984016
GO:0008266	poly(U) RNA binding	6	3	0.038961	0.995005
GO:0009881	photoreceptor activity	7	2	0.034965	0.992008
GO:0010460	positive regulation of heart rate	7	4	0.028971	0.988012
GO:0016477	cell migration	66	5	0.037962	0.994006
GO:0018298	protein-chromophore linkage	15	3	0.014985	0.942058
GO:0035609	C-terminal protein deglutamylation	3	2	0.025974	0.984016
GO:0035610	protein side chain deglutamylation	3	2	0.025974	0.984016
GO:0040007	growth	31	4	0.045954	0.999001
GO:0042326	negative regulation of phosphorylation	15	3	0.045954	0.999001
GO:0046887	positive regulation of hormone secretion	13	4	0.015984	0.948052
GO:0048709	oligodendrocyte differentiation	11	3	0.034965	0.992008

[a] Number of genes in gene set; [b] Number of associated intervals overlapping with genes in gene set

Table 8: Global interval-based enrichment analysis results under p value, 0.001, 0.01, and 0.05

P value threshold [a]	Number of genes in the enriched pathways [b]	Significance [c]
<0.001	0	1
<0.01	6	0.758242
<0.05	77	0.703297

[a] P-value threshold to select significantly enriched pathways; [b] number of unique genes in the enriched pathways with empirical p-value ≤ P value threshold; [c] % of bootstrapping samples where number of unique genes in enriched pathways is not less than that of original association data

Chapter 5: Conclusion

5.1 Review of guiding principles

The goals of this dissertation project were to a) review studies that link addiction to injection behavior and to genetics, b) apply a selected multiple imputation model to manage the large amount of missing data in the longitudinal study, c) create a risk phenotype for genome wide analysis.

Substance use disorders are chronic and recurrent. In order to study HIV risk behaviors like the injection of addictive drugs, data from a longitudinal study rather than a cross-sectional study yields more useful information. The reason is that a longitudinal database provides more valuable information about the long-term trajectories of the HIV risk behaviors of injection drug users. Using this longitudinal sample, we borrowed the concept of "smoking pack-years" to create a phenotype, injection years. (Caporaso et al., 2009) However, the lives and relationships of injection drug users tend to be unstable; hence, there exists virtually no perfect attendance record for any single participant in this study. Using the ALIVE cohort, a longitudinal community-based study, I created three imputation models and examined them by sensitivity test. The results of the universal imputation model served as the basis for the genome wide analysis.

The ALIVE cohort is a longitudinal community-based study of IV drug users. This urban minority population is at high risk of illicit drug use and dependence (Reuter et al., 1998) yet there is scant research on the association between their genetic makeup and HIV risk behaviors. I believe that the findings may enhance prevention and treatment strategies and aid in identifying subpopulations at high risk of persistent drug use. In

addition, in studying this target population we avoid a common pitfall of genome wide association analyses: the use of samples from different geographical regions with individuals of different genetic backgrounds. In other words, population stratification, which is a major quality control issue in GWAS, was not a big concern in this study. Still, GWAS have limitations. The first is that current diagnostic categories might not reflect the heritability of a particular disorder. (Psychiatric GWAS Consortium Coordinating Committee, 2009) Borrowing the concept of "smoking pack-years," I created a sub-phenotype, injection years. I believe that this phenotype may be at least partly genetic in origin; moreover, this phenotype can be used to measure the intensity of injection behavior. It is hoped that this finding will elucidate our understanding of modifiable and fixed components of injected drug use, a major HIV risk behavior, in this population, and lead to improved treatment and prevention.

In this thesis, I have attempted to link behavioral phenotypes with genetics. The genetic underpinning of substance use disorders is an area of great interest. Classical genetic approaches have shown that addiction is heritable; molecular genetic approaches suggest that specific addiction-related behaviors are associated with specific genes. Herein, I created a phenotype from the ALIVE dataset: injection years. I believe that this phenotype is genetically influenced. Therefore, I used it in a genome wide association analysis. With the findings from the genome wide association analysis, I conducted a polygenic risk score analysis and a pathway analysis.

5.2 Summary of findings

Review of Chapter 2. The purpose of this critical review was to find articles to provide an answer to the question: Is this phenotype, injection years, at least partly

influenced by genetic factors? To that end, I reviewed articles related to injection of illegal drugs, addiction, and the genetic evidence for opioid addiction. First, I reviewed the literature on substance use disorders and HIV. Injection drug use is a common form of administration among opiate dependent individuals, and the use of injectable drugs increases the risk of HIV. (Friedland & Klein, 1987) In other words, injection drug users are at high risk for HIV, with heroin the most commonly injected drug. Second, I reviewed the literature on symptoms and signs of substance use disorder. DSM-IV states that the persistent use of a substance can induce changes in brain circuitry and cause specific behavior patterns such as craving. (American Psychiatric Association, 2000) These findings suggested that substance use disorders are chronic and recurrent, and a substance dependent individual's behavior will change over time. Third, I reviewed the literature on the genetic epidemiology of substance use disorders. Tsuang's twin study, Gelernter's genome wide linkage study and genome wide associations study provided ample evidence that genetic factors influence opioid addiction. (Gelernter et al., 2014; Gelernter et al., 2006; Tsuang et al., 1998) I created a risk phenotype, injection years to evaluate the intensity of injection behavior. Because there is a genetic basis for opioid addiction, I am confident that injection years which is highly associated with addiction, is also associated with genetics.

Study in Chapter 3. The purpose of this study was to deal with the missing data in a longitudinal cohort and create a risk phenotype, injection year, for the genome wide analysis. The ALIVE GWAS subset included 1,197 participants drawn from the ALIVE cohort in 1998. They visited semiannually with an average of 19 visits per person. The population was predominantly African-American (98%). The mean age of first injection

was 21. The mean age at entering was 34. This age difference presented a problem since so much data was not available. Sixty-six percent of the recordable visits were missing. I imputed the ALIVE GWAS subset with three models: universal, two-predictors, and one-predictor. Respectively, the mean of the summary of injection years was 23, 19, and 18 years in universal, two-predictors, and one-predictor models. The mean of the summary of injection years in the universal model was much greater than that of the other two models. After conducting a sensitivity exam, I chose the universal model for the subsequent analysis. I compared the findings in this study with Genberg's, and the most important finding, that the mean injection years were 24, is consistent with hers. (Genberg et al., 2011)

Study in chapter 4. The purpose of this study was to examine the associations between injection years and genetic factors by using genome wide association analysis, polygenic risk score analysis and pathway analysis. Using PLINK, 735,081 variants and 1107 people passed filters and quality control. Phenotype data is quantitative. Using regression analysis, I tested the SNP associations after adjusting for ancestry components and found no single SNP in the GWAS reached the level of genome-wide significance. ($p < 5 \times 10^{-8}$) Eleven SNPs exceeded the $p < 5 \times 10^{-5}$ threshold, including one SNP (rs10168062) with a $p < 5 \times 10^{-7}$. Using polygenic risk score analysis, I used the ALIVE GWAS male group for the discovery sample, considering the four most informative P_T thresholds from the intra- ALIVE GWAS male group analyses. The highest correlation coefficient is around 0.10 when P_T is 0.01. The ALIVE GWAS subset supports a polygenic basis to injection years that involves many common SNPs. I used INRICH for pathway analysis. The 31 intervals were selected for the pathway analysis but none of the

gene sets was statistically significant. Two of the top gene sets are metallocarboxypeptidase activity and DNA dealkylation involved in DNA repair.

5.3 Discussion of findings

In this dissertation, I created a risk phenotype which can be used to represent the intensity of injection behavior; and examined the associations between genetics and this risk phenotype. The literature supports the proposition that a) injection drug users are at high risk for HIV, with heroin the most commonly injected drug, b) substance use disorders are chronic and recurrent, and an addict's behaviors changes with time, c) genetic factors influence opioid addiction. Thus, there is a high possibility that injection year is associated with genetics. Assuming that multiple genes with small effects contribute to injection years, I performed a polygenic risk score analysis and pathway analysis in an attempt to ascertain which biological pathways are related to this variable.

Substance use disorders are chronic and recurrent. Changes in behavior patterns occur in individuals with substance abuse disorders; observing these changes over the long term can perhaps help us to understand more about the outcomes of addiction. In the ALIVE cohort, Genberg et al. found that injection drug users can be categorized into five groups with unique trajectories. (Genberg et al., 2011) From Genberg's findings, I believe that measuring the length of time someone has injected drugs can be used to evaluate the intensity of injection behavior. However, it is virtually impossible to track and collect data from this highly unstable population for any length of time, much less the amounts of time required for this study. The longer the study, the higher the amounts of missing information. In the ALIVE GWAS subset, subjects had been tracked for a mean of approximately ten years; thus, there are vast amounts of missing data. In studying

injection drug users, we borrowed the concept of "smoking pack-year" and created a phenotype, injection years, to track changes in the status of injected drug use in our longitudinal dataset.

It was necessary to account for the massive amount of missing data before starting the analysis. To manage this problem, I imputed injection data in the ALIVE GWAS subset with three models: universal, two predictors, and one predictor. In the universal model, I supposed that several causal factors continuously influence injection behavior in all phases. In the two-predictor model, I supposed that causal factors which influence injection behavior change in different phases but may remain the same in the identical phase. In the one-predictor model, I used only the existing report, six months prior, as a predictor, so the imputed reports were only influenced by the existing reports at the identical time point. The results of the sensitivity test supported the use of the universal model. With the universal model, the mean of the sum of imputed injection years is consistent with the results of Genberg's study. (Genberg et al., 2011) The findings of this study support the use of multiple imputation in a longitudinal cohort such as the ALIVE GWAS subset. Clearly, the missing data presents a major obstacle for conducting longitudinal studies; multiple imputation using a specific model may provide a solution to this problem.

In this study, I tested the association between longitudinal addiction behavior and genetics. Longitudinal studies are rare in the field of addiction; missing data is a common problem in such studies. Using multiple imputation, I created a risk phenotype, injection year, which was used in the genome wide association analysis. However, no single SNP in the GWAS reached the level of genome-wide significance. This result suggested that

injection years might be affected by polygenes and/or gene-gene interactions and/or gene-environment interactions, instead of a single gene. Traditional genome wide association studies can identify a large portion of genetic risk for diseases but cannot detect genetic subsets in which each gene has a very small size effect. Polygenic risk score analysis and pathway analysis are used to detect such genetic subsets, using the results from genome wide association analysis. In this study, the findings of polygenic risk scores analysis support the idea that injection year is influenced by polygenes. No significant result was found in pathway analysis. The top two gene sets, metallocarboxypeptidase activity and DNA dealkylation, are both involved in DNA repair. This result suggests that DNA repair may play a role in our risk phenotype, injection year. Linking injection years to genetics, this study offers a practical method for exploring genetic influences on specific behavior traits in a longitudinal cohort. I believe that two types of studies would be beneficial follow up studies: studies of genetic influences on behavioral phenotypes, and studies of genetic influences on longitudinal phenotypes.

5.4 Implications for public health

Substance use disorders are a major public health issue. Illicit psychoactive drugs are associated with a variety of psychological and physical problems. From a public health perspective, in the U.S. there are enormous societal consequences to the use of illicit drugs. (Birnbaum et al., 2011) These include costs to the health care, law enforcement, and criminal justice systems as well as losses workplace productivity of up to $50 billion *annually*. (Birnbaum et al., 2011; Hansen et al., 2011) From a personal health perspective, injection drug use leads to a host of physical problems, one of the

most serious being the high risk of acquiring HIV. (Friedland & Klein, 1987) There have been few studies of injection drug users; it is hoped that this study can be of benefit to this group and to those working on all fronts to help them.

Since treatment for injection drug use comes with no guarantee of success, prevention is the preferred strategy. From a public health perspective, disease prevention can be categorized as primary, secondary, and tertiary prevention. Secondary prevention aims to reduce the impact of a disease or injury that has already occurred. In this dissertation, I created a risk phenotype which can be used to measure the intensity of injection behavior. I believe injection years can be used to identify those who have injected drugs persistently, and who are therefore at highest risk of HIV infection. In order to find biological evidence, I tested association between injection and genetics. The findings suggest that injection year is influenced by polygenes, and probably by gene-gene interactions and gene-environment interaction.

5.5 Limitations of the studies

The analyses that are the basis for this dissertation had a number of limitations. There are two limitations resulting from multiple imputation analysis. The first was the amount of missing data in the ALIVE GWAS subset. Overall, 66 percent of the data is missing. The accuracy of imputation is associated with the amount of existing data. These amounts of missing data definitely decrease the accuracy of imputation. Another limitation was the definition of "injection years." The definition only describes the number of years the person has been injecting; the definition does not take into account other aspects like dosage or injection counts. I maintain that the study target, the trajectory of injection behavior, is mostly influenced by injection years; however, I

cannot exclude the possibility that the target is also influenced by these other factors. There were two limitations in the genome wide analyses. First, I did not include the variable related to gene-environment interactions because, in a longitudinal study such as the ALIVE GWAS cohort, it is hard to define an environmental variable. Although our target is the trait of injection behavior which is affected by genetic factors, we cannot disentangle the interactions between genetic factors and environmental factors. Second, there were fewer than 1,200 subjects. Because there are so few studies of injection drug users, we could not find another independent sample to serves as a discovery sample in the polygenic risk score analysis. We divided the ALIVE GWAS subset into male and female groups, but the number of people in each group was insufficient for polygenic risk score analysis.

5.6 Conclusions and future directions for research

In this dissertation, I reviewed literature on addiction, HIV risk behaviors and genetics to build a solid foundation for my hypothesis. I created a risk phenotype, injection years, to examine the sensitivity of different imputation models. The mean of imputed injection years were consistent with Genberg's findings. (Genberg et al., 2011) Thus, it could potentially be used as a simple and clear tool to measure the intensity of injection behavior. Using the imputed results, I conducted genome wide analyses: genome wide association analysis, polygenic risk score analysis, and pathway analysis. The findings link genetics to injection years. In brief, I recommend using behavioral phenotypes for genome wide analysis to link genetic studies to observable behaviors; I believe that this method can provide valuable information about how genetic factors influence behavior through specific biological mechanisms. Another potential area for

future studies is to analyze longitudinal information to serve as risk phenotype for genome wide analysis.

Longitudinal data can provide information about changes in people's lives and behavior patterns; however, missing data is a major problem. I recommend using multiple imputation in other longitudinal studies and selecting an appropriate model by examining the accuracy and reliability of the imputed data. By using the appropriate model, I believe that reliable imputed records can be obtained from the existing data. Another potential area for future studies is to impute a continuous variable such as the injection counts. Caporaso et al. used the concept of "smoking pack-year," defined as the number of packs smoked in a year.(Caporaso et al., 2009) In this study, we did not deal with the issue of the intensity of single injection, but we think that it could be a worthwhile area of investigation.

Bibliography

American Psychiatric Association. (2000). Diagnostic and Statistical Manual of Mental Disorders. 4th text revision ed. *Washington, DC: American Psychiatric Association*.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*: American Psychiatric Pub.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols, 5*(9), 1564-1573.

Armstrong, G. L. (2007). Injection drug users in the United States, 1979-2002: an aging population. *Archives of internal medicine, 167*(2), 166-173.

Askland, K., Read, C., & Moore, J. (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human genetics, 125*(1), 63-79.

Bergen, A., Kokoszka, J., Peterson, R., Long, J., Virkkunen, M., Linnoila, M., & Goldman, D. (1996). Mu opioid receptor gene variants: lack of association with alcohol dependence. *Molecular psychiatry, 2*(6), 490-494.

Bergen, A. W., Korczak, J. F., Weissbecker, K. A., & Goldstein, A. M. (1999). A genome-wide search for loci contributing to smoking and alcoholism. *Genetic Epidemiology, 17*(S1), S55-S60.

Birnbaum, H. G., White, A. G., Schiller, M., Waldman, T., Cleveland, J. M., & Roland, C. L. (2011). Societal costs of prescription opioid abuse, dependence, and misuse in the United States. *Pain Medicine, 12*(4), 657-667.

Bosron, W., Ehrig, T., & Li, T. (1993). *Genetic factors in alcohol metabolism and alcoholism.* Paper presented at the Seminars in Liver Disease.

Brennan, C. (2012). Smoking can have negative impact on immune response to HAART. *HIV clinician/Delta Region AIDS Education & Training Center, 25*(1), 6-9.

Burnam, M. A., Bing, E. G., Morton, S. C., Sherbourne, C., Fleishman, J. A., London, A. S., . . . Shapiro, M. F. (2001). Use of mental health and substance abuse treatment services among adults with HIV in the United States. *Archives of general psychiatry, 58*(8), 729-736.

Burton, P. R., Tobin, M. D., & Hopper, J. L. (2005). Key concepts in genetic epidemiology. *The Lancet, 366*(9489), 941-951.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software, 45*(3).

Cadoret, R. J., Troughton, E., O'Gorman, T. W., & Heywood, E. (1986). An adoption study of genetic and environmental factors in drug abuse. *Archives of general psychiatry, 43*(12), 1131-1136.

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics, 86*(1), 6-22.

Caporaso, N., Gu, F., Chatterjee, N., Sheng-Chih, J., Yu, K., Yeager, M., . . . Landi, M. T. (2009). Genome-wide and candidate gene association study of cigarette smoking behaviors. *PloS one, 4*(2), e4653.

Carmelli, D., Swan, G. E., Robinette, D., & Fabsitz, R. (1992). Genetic influence on smoking—a study of male twins. *New England Journal of Medicine, 327*(12), 829-833.

Center for Disease Control. (1981). Pneumocystis pneumonia--Los Angeles. *MMWR. Morbidity and Mortality Weekly Report, 30*(21), 250.

Cheng, L. S. C., Swan, G. E., & Carmelli, D. (2000). A genetic analysis of smoking behavior in family members of older adult males. *Addiction, 95*(3), 427-435.

Cloninger, C. R., Bohman, M., & Sigvardsson, S. (1981). Inheritance of alcohol abuse: Cross-fostering analysis of adopted men. *Archives of general psychiatry, 38*(8), 861-868.
Comings, D., Gade, R., Wu, S., Chiu, C., Dietz, G., Muhleman, D., . . . Lesieur, H. (1997). Studies of the potential role of the dopamine D 1 receptor gene in addictive behaviors. *Molecular psychiatry, 2*(1), 44-56.
Consortium, I. M. S. G. (2010). Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. *The American Journal of Human Genetics, 86*(4), 621-625.
Cornelis, M. C., Monda, K. L., Yu, K., Paynter, N., Azzato, E. M., Bennett, S. N., . . . Chatterjee, N. (2011). Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS Genet, 7*(4), e1002033.
Crabb, D. W., Dipple, K. M., & Thomasson, H. R. (1993). Alcohol sensitivity, alcohol metabolism, risk of alcoholism, and the role of alcohol and aldehyde dehydrogenase genotypes. *The Journal of laboratory and clinical medicine, 122*(3), 234-240.
Crabb, D. W., Edenberg, H. J., Bosron, W. F., & Li, T.-K. (1989). Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2 (2) allele is dominant. *Journal of Clinical Investigation, 83*(1), 314.
de Jong, B. C., Prentiss, D., McFarland, W., Machekano, R., & Israelski, D. M. (2005). Marijuana use and its association with adherence to antiretroviral therapy among HIV-infected persons with moderate to severe nausea. *JAIDS Journal of Acquired Immune Deficiency Syndromes, 38*(1), 43-46.
Degenhardt, L., & Hall, W. (2012). Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *The Lancet, 379*(9810), 55-70.
Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. [Research Support, Non-U.S. Gov't]. *PLoS Genet, 9*(3), e1003348. doi: 10.1371/journal.pgen.1003348
Duggirala, R., Almasy, L., & Blangero, J. (1999). Smoking behavior is under the influence of a major quantitative trait locus on human chromosome 5q. *Genetic Epidemiology, 17*(S1), S139-S144.
Edelman, E. J., Tetrault, J. M., & Fiellin, D. A. (2014). Substance use in older HIV-infected patients. *Current Opinion in HIV and AIDS, 9*(4), 317-324.
Edenberg, H. J. (2002). The collaborative study on the genetics of alcoholism: an update. *Alcohol Research and Health, 26*(3), 214-218.
Ersche, K. D., Clark, L., London, M., Robbins, T. W., & Sahakian, B. J. (2006). Profile of executive and memory function associated with amphetamine and opiate dependence. *Neuropsychopharmacology, 31*(5), 1036-1047.
Foroud, T., Edenberg, H. J., Goate, A., Rice, J., Flury, L., Koller, D. L., . . . Bucholz, K. K. (2000). Alcoholism susceptibility loci: confirmation studies in a replicate sample and further mapping. *Alcoholism: Clinical and experimental research, 24*(7), 933-945.
Friedland, G. H., & Klein, R. S. (1987). Transmission of the human immunodeficiency virus. *New England Journal of Medicine, 317*(18), 1125-1135.
Friedman-Kien, A., Laubenstein, L., Marmor, M., Hymes, K., Green, J., Ragaz, A., . . . Weintraub, M. (1981). Kaposis sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR. Morbidity and Mortality Weekly Report, 30*(25), 305-308.
Galai, N., Safaeian, M., Vlahov, D., Bolotin, A., & Celentano, D. (2003). Longitudinal patterns of drug injection behavior in the ALIVE Study cohort, 1988–2000: description and determinants. *American Journal of Epidemiology, 158*(7), 695-704.
Galvan, F. H., Bing, E. G., Fleishman, J. A., London, A. S., Caetano, R., Burnam, M. A., . . . Shapiro, M. (2002). The prevalence of alcohol consumption and heavy drinking among people with HIV in the United States: results from the HIV Cost and Services Utilization Study. *Journal of studies on alcohol, 63*(2), 179-186.

Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology, 32*(4), 361.

Garriock, H. A., Tanowitz, M., Kraft, J. B., Dang, V. C., Peters, E. J., Jenkins, G. D., . . . Slager, S. L. (2010). Association of mu-opioid receptor variants and response to citalopram treatment in major depressive disorder. *American Journal of Psychiatry, 167*(5), 565-573.

Gelernter, J., Kranzler, H. R., Sherva, R., Koesterer, R., Almasy, L., Zhao, H., & Farrer, L. A. (2014). Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. *Biological psychiatry, 76*(1), 66-74.

Gelernter, J., Panhuysen, C., Wilcox, M., Hesselbrock, V., Rounsaville, B., Poling, J., . . . Farrer, L. (2006). Genomewide linkage scan for opioid dependence and related traits. *The American Journal of Human Genetics, 78*(5), 759-769.

Genberg, B. L., Gange, S. J., Go, V. F., Celentano, D. D., Kirk, G. D., & Mehta, S. H. (2011). Trajectories of injection drug use over 20 years (1988–2008) in Baltimore, Maryland. *American Journal of Epidemiology*, kwq441.

Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073.

Goldman, D., Oroszi, G., & Ducci, F. (2005). The genetics of addictions: uncovering the genes. *Nature Reviews Genetics, 6*(7), 521-532.

Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., . . . Kaplan, K. (2004). Prevalence and co-occurrence of substance use disorders and independentmood and anxiety disorders: Results from the national epidemiologic survey on alcohol and relatedconditions. *Archives of general psychiatry, 61*(8), 807-816.

Haerian, B. S., & Haerian, M. S. (2013). OPRM1 rs1799971 polymorphism and opioid dependence: evidence from a meta-analysis. *Pharmacogenomics, 14*(7), 813-824.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., . . . Gilbert, J. R. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science, 308*(5720), 419-421.

Hansen, R. N., Oster, G., Edelsberg, J., Woody, G. E., & Sullivan, S. D. (2011). Economic costs of nonmedical use of prescription opioids. *The Clinical journal of pain, 27*(3), 194-202.

Harada, S., Agarwal, D., Goedde, H., Tagaki, S., & Ishikawa, B. (1982). Possible protective role against alcoholism for aldehyde dehydrogenase isozyme deficiency in Japan. *The Lancet, 320*(8302), 827.

Heath, A., Kirk, K., Meyer, J., & Martin, N. (1999). Genetic and social determinants of initiation and age at onset of smoking in Australian twins. *Behavior Genetics, 29*(6), 395-407.

Heath, A. C., Madden, P. A., Slutske, W. S., & Martin, N. G. (1995). Personality and the inheritance of smoking behavior: a genetic perspective. *Behavior Genetics, 25*(2), 103-117.

Hettema, J. M., Prescott, C. A., Myers, J. M., Neale, M. C., & Kendler, K. S. (2005). The structure of genetic and environmental risk factors for anxiety disorders in men and women. *Archives of general psychiatry, 62*(2), 182-189.

Hill, K. G., White, H. R., Chung, I.-J., Hawkins, J. D., & Catalano, R. F. (2000). Early adult outcomes of adolescent binge drinking: Person-and variable-centered analyses of binge drinking trajectories. *Alcoholism, clinical and experimental research, 24*(6), 892.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362-9367.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., . . . Craddock, N. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *The American Journal of Human Genetics, 85*(1), 13-24.

Hopper, J. L., Bishop, D. T., & Easton, D. F. (2005). Population-based family studies in genetic epidemiology. *The Lancet, 366*(9494), 1397-1406.

Hou, Q.-F., & Li, S.-B. (2009). Potential association of DRD2 and DAT1 genetic variation with heroin dependence. *Neuroscience letters, 464*(2), 127-130.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: a methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19*(4), 339.

Johnston, L., O'Malley, P. M., Bachman, J. G., & Schulenberg, J. E. (2013). Monitoring the Future National Survey Results on Drug Use, 1975–2011. *Institue for Social Research, the University of Michigan: Rockville, Michigan, USA, 1*.

Jung, T., & Wickrama, K. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302-317.

Kaij, L., & Rosenthal, D. (1961). Alcoholism in Twins. Studies on the Etiology and Sequels of Abuse of Alcohol. *The Journal of Nervous and Mental Disease, 133*(3), 272.

Kendler, K. S., Jacobson, K. C., Prescott, C. A., & Neale, M. C. (2003). Specificity of genetic and environmental risk factors for use and abuse/dependence of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates in male twins. *American Journal of Psychiatry*.

Kendler, K. S., Neale, M. C., Heath, A. C., Kessler, R. C., & Eaves, d. L. J. (1994). A twin-family study of alcoholism in women. *American Journal of Psychiatry, 151*(5), 707-715.

Kendler, K. S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Archives of general psychiatry, 60*(9), 929-937.

Kendler, K. S., Thornton, L. M., & Pedersen, N. L. (2000). Tobacco consumption in Swedish twins reared apart and reared together. *Archives of general psychiatry, 57*(9), 886-892.

Kerem, B.-s., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., . . . Tsui, L.-C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science, 245*(4922), 1073-1080.

Koopmans, J. R., Slutske, W. S., Heath, A. C., Neale, M. C., & Boomsma, D. I. (1999). The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. *Behavior Genetics, 29*(6), 383-393.

Kreek, M. J., Nielsen, D. A., Butelman, E. R., & LaForge, K. S. (2005). Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction. *Nature neuroscience, 8*(11), 1450-1457.

Kreek, M. J., Nielsen, D. A., & LaForge, K. S. (2004). Genes associated with addiction. *Neuromolecular medicine, 5*(1), 85-108.

Kreuter, F., & Muthén, B. (2007). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically derived criminal trajectory profiles. *Advances in latent variable mixture models*, 53-75.

Lachman, H. M., Fann, C. S., Bartzis, M., Evgrafov, O. V., Rosenthal, R. N., Nunes, E. V., . . . Riddick, A. (2007). Genomewide suggestive linkage of opioid dependence to chromosome 14q. *Human Molecular Genetics, 16*(11), 1327-1334.

Lee, P. H., O'Dushlaine, C., Thomas, B., & Purcell, S. M. (2012). INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics, 28*(13), 1797-1799.

Lofwall, M. R., Brooner, R. K., Bigelow, G. E., Kindbom, K., & Strain, E. C. (2005). Characteristics of older opioid maintenance patients. *Journal of Substance Abuse Treatment, 28*(3), 265-272.

Long, J. C., Knowler, W. C., Hanson, R. L., Robin, R. W., Urbanek, M., Moore, E., . . . Goldman, D. (1998). Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. *American journal of medical genetics, 81*(3), 216-221.

Lusher, J., Chandler, C., & Ball, D. (2001). Dopamine D4 receptor gene (DRD4) is associated with Novelty Seeking (NS) and substance abuse: the saga continues. *Molecular psychiatry*.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual review of psychology, 51*(1), 201-226.

MacDonald, M. E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., . . . Myers, R. (1992). The Huntington's disease candidate region exhibits many different haplotypes. *Nature genetics, 1*(2), 99-103.

McGue, M., Elkins, I., & Iacono, W. G. (2000). Genetic and environmental influences on adolescent substance use and abuse. *American journal of medical genetics, 96*(5), 671-677.

Merikangas, K. R., Stolar, M., Stevens, D. E., Goulet, J., Preisig, M. A., Fenton, B., . . . Rounsaville, B. J. (1998). Familial transmission of substance use disorders. *Archives of general psychiatry, 55*(11), 973-979.

Mimiaga, M. J., Reisner, S. L., Grasso, C., Crane, H. M., Safren, S. A., Kitahata, M. M., . . . Mayer, K. H. (2013). Substance use among HIV-infected patients engaged in primary care in the United States: findings from the Centers for AIDS Research Network of Integrated Clinical Systems cohort. *American journal of public health, 103*(8), 1457-1467.

Mizoi, Y., Tatsuno, Y., Adachi, J., Kogame, M., Fukunaga, T., Fujiwara, S., . . . Ijiri, I. (1983). Alcohol sensitivity related to polymorphism of alcohol-metabolizing enzymes in Japanese. *Pharmacology Biochemistry and Behavior, 18*, 127-133.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American journal of human genetics, 7*(3), 277.

Morton, N. E. (1982). *Outline of genetic epidemiology*: Karger Basel.

Muramatsu, T., Zu-Cheng, W., Yi-Ru, F., Kou-Bao, H., Heqin, Y., Yamada, K., . . . Kono, H. (1995). Alcohol and aldehyde dehydrogenase genotypes and drinking behavior of Chinese living in Shanghai. *Human genetics, 96*(2), 151-154.

Muthén, B., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. *Longitudinal data analysis*, 143-165.

Nagin, D. (2005). *Group-based modeling of development*: Harvard University Press.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change.

O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D., & Corvin, A. (2010). Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Molecular psychiatry, 16*(3), 286-292.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *American journal of human genetics, 26*(5), 588.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., . . . Nakamura, Y. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics, 32*(4), 650-654.

Panel on Antiretroviral Guidelines for Adults Adolescents. (2009). Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. *Department of Health and Human Services*, 1-161.

Partanen, J., Bruun, K., & Markkanen, T. (1966). *Inheritance of drinking behavior: A study on intelligence, personality, and use of alcohol of adult twins* (Vol. 14): Finnish Foundation for Alcohol Studies.

Petry, N. M. (1999). Alcohol use in HIV patients: what we don't know may hurt us. *International journal of STD & AIDS, 10*(9), 561-570.
Pilowsky, D. J., & Wu, L.-T. (2015). Sexual risk behaviors and Hiv risk among Americans aged 50 years or older: a review. *Substance abuse and rehabilitation, 6*, 51.
Pouletty, P. (2002). Drug addictions: towards socially accepted and medically treatable diseases. *Nature Reviews Drug Discovery, 1*(9), 731-736.
Prentiss, D., Power, R., Balmas, G., Tzuang, G., & Israelski, D. M. (2004). Patterns of marijuana use among patients with HIV/AIDS followed in a public health care setting. *JAIDS Journal of Acquired Immune Deficiency Syndromes, 35*(1), 38-45.
Prescott, C. A., & Kendler, K. S. (1999). Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *American Journal of Psychiatry, 156*(1), 34-40.
Psychiatric GWAS Consortium Coordinating Committee. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American journal of psychiatry, 166*(5), 540-556.
Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460*(7256), 748-752. doi: 10.1038/nature08185
Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *TRENDS in Genetics, 17*(9), 502-510.
Reich, T., Edenberg, H. J., Goate, A., Williams, J. T., Rice, J. P., Van Eerdewegh, P., . . . Begleiter, H. (1998). Genome-wide search for genes affecting the risk for alcohol dependence. *American journal of medical genetics, 81*(3), 207-215. doi: 10.1002/(sici)1096-8628(19980508)81:3<207::aid-ajmg1>3.0.co;2-t
Reuter, P., Hsu, M., Petronis, K., & Wish, E. (1998). *Estimating the need for substance abuse treatment in Maryland*: Center for Substance Abuse Research (CESAR).
Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517.
Rosen, D., Smith, M. L., & Reynolds, C. F. (2008). The prevalence of mental and physical health disorders among older methadone patients. *The American Journal of Geriatric Psychiatry, 16*(6), 488-497.
Rounsaville, B. J., & Kleber, H. D. (1985). Untreated opiate addicts: how do they differ from those seeking treatment? *Archives of general psychiatry, 42*(11), 1072-1077.
Rounsaville, B. J., Tierney, T., Crits-Christoph, K., Weissman, M. M., & Kleber, H. D. (1982). Predictors of outcome in treatment of opiate addicts: evidence for the multidimensional nature of addicts' problems. *Comprehensive Psychiatry, 23*(5), 462-478.
Royston, P. (2011). ICE: Stata module for multiple imputation of missing values. *Statistical Software Components*.
Saccone, N. L., Kwon, J. M., Corbett, J., Goate, A., Rochberg, N., Edenberg, H. J., . . . Reich, T. (2000). A genome screen of maximum number of drinks as an alcoholism phenotype. *American journal of medical genetics, 96*(5), 632-637.
Saitz, R. (2005). Unhealthy alcohol use. *New England Journal of Medicine, 352*(6), 596-607.
Saunders, J. B., Schuckit, M. A., Sirovatka, P. J., & Regier, D. A. (2008). *Diagnostic Issues in Substance Use Disorders: Refining the Research Agenda for DSM-V*: American Psychiatric Publishing.
Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature, 461*(7261), 218-223.
Schinka, J., Letsch, E., & Crawford, F. (2002). DRD4 and novelty seeking: Results of meta-analyses. *American journal of medical genetics, 114*(6), 643-648.
Seshadri, S., DeStefano, A. L., Au, R., Massaro, J. M., Beiser, A. S., Kelly-Hayes, M., . . . Atwood, L. D. (2007). Genetic correlates of brain aging on MRI and cognitive test

measures: a genome-wide association and linkage analysis in the Framingham Study. *BMC medical genetics, 8*(Suppl 1), S15.
Shih, R. A., Belmonte, P. L., & Zandi, P. P. (2004). A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *International Review of Psychiatry, 16*(4), 260-283.
Shuter, J., Bernstein, S. L., & Moadel, A. B. (2012). Cigarette smoking behaviors and beliefs in persons living with HIV/AIDS. *American journal of health behavior, 36*(1), 75.
Simonson, M. A., Wills, A. G., Keller, M. C., & McQueen, M. B. (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC medical genetics, 12*(1), 146.
Skeer, M. R., Mimiaga, M. J., Mayer, K. H., O'Cleirigh, C., Covahey, C., & Safren, S. A. (2012). Patterns of substance use among a large urban cohort of HIV-infected men who have sex with men in primary care. *AIDS and Behavior, 16*(3), 676-689.
Smith, D. K., Pan, Y., Rose, C. E., Pals, S. L., Mehta, S. H., Kirk, G. D., & Herbst, J. H. (2015). A Brief Screening Tool to Assess the Risk of Contracting HIV Infection Among Active Injection Drug Users. *Journal of addiction medicine*.
Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., . . . Kurreeman, F. A. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics, 44*(5), 483-489.
Stallings, M. C., Corley, R. P., Hewitt, J. K., Krauter, K. S., Lessem, J. M., Mikulich, S. K., . . . Crowley, T. J. (2003). A genome-wide search for quantitative trait loci influencing substance dependence vulnerability in adolescence. *Drug and alcohol dependence, 70*(3), 295-307.
StataCorp. (2011). Stata Statistical Software (Version Release 12): College Station, TX: StataCorp LP.
Sulem, P., Gudbjartsson, D. F., Geller, F., Prokopenko, I., Feenstra, B., Aben, K. K., . . . Stumvoll, M. (2011). Sequence variants at CYP1A1–CYP1A2 and AHR associate with coffee consumption. *Human Molecular Genetics, 20*(10), 2071-2077.
Sullivan, P. F. (2010). The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron, 68*(2), 182-186.
Teare, M. D., & Barrett, J. H. (2005). Genetic linkage studies. *The Lancet, 366*(9490), 1036-1044.
Thompson, W. D., Orvaschel, H., Prusoff, B. A., & Kidd, K. K. (1982). An evaluation of the family history method for ascertaining psychiatric disorders. *Archives of general psychiatry, 39*(1), 53-58.
Thomson, A. D. (2000). Mechanisms of vitamin deficiency in chronic alcohol misusers and the development of the Wernicke-Korsakoff syndrome. *Alcohol and Alcoholism, 35*(Supplement 1), 2-1.
Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., . . . Ingason, A. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature, 452*(7187), 638-642.
Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics, 42*(5), 441-447.
Tsuang, M. T., Lyons, M. J., Harley, R. M., Xian, H., Eisen, S., Goldberg, J., . . . Faraone, S. V. (1999). Genetic and environmental influences on transitions in drug use. *Behavior Genetics, 29*(6), 473-479.
Tsuang, M. T., Lyons, M. J., Meyer, J. M., Doyle, T., Eisen, S. A., Goldberg, J., . . . Eaves, L. (1998). Co-occurrence of abuse of different drugs in men: the role of drug-specific and shared vulnerabilities. *Archives of general psychiatry, 55*(11), 967-972.
Van Eerdewegh, P., Foroud, T., Hesselbrock, V., Schuckit, M. A., Bucholz, K., Porjesz, B., . . . Tischfield, J. A. (1998). Genome-wide search for genes affecting the risk for alcohol

dependence. *American Journal of Medical Genetics (Neuropsychiatric Genetics), 81*, 207-215.
Vereczkei, A., Demetrovics, Z., Szekely, A., Sarkozy, P., Antal, P., Szilagyi, A., . . . Barta, C. (2013). Multivariate analysis of dopaminergic gene variants as risk factors of heroin dependence. *PloS one, 8*(6), e66592.
Victor, M., Adams, R. D., & Collins, G. H. (1971). The Wernicke-Korsakoff syndrome. A clinical and pathological study of 245 patients, 82 with post-mortem examinations. *Contemporary neurology series, 7*, 1.
Vink, J. M., Hottenga, J. J., Geus, E. J., Willemsen, G., Neale, M. C., Furberg, H., & Boomsma, D. I. (2014). Polygenic risk scores for smoking: predictors for alcohol and cannabis use? *Addiction, 109*(7), 1141-1151.
Vlahov, D., Anthony, J. C., Muñoz, A., & Margolick, J. (1991). The ALIVE Study: a longitudinal study of HIV-1 infection in intravenous drug users: description of methods. *Journal of Drug Issues*.
Vlahov, D., Graham, N., Hoover, D., Flynn, C., Bartlett, J. G., Margolick, J. B., . . . Holmberg, S. (1998). Prognostic indicators for AIDS and infectious disease death in HIV-infected injection drug users: plasma viral load and CD4+ cell count. *Jama, 279*(1), 35-40.
Wang, C. N., Little, R., Nan, B., & Harlow, S. D. (2011). A Hot-Deck Multiple Imputation Procedure for Gaps in Longitudinal Recurrent Event Histories. *Biometrics, 67*(4), 1573-1582.
Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics, 11*(12), 843-854.
Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., . . . Sleiman, P. M. (2009). Common genetic variants on 5p14. 1 associate with autism spectrum disorders. *Nature, 459*(7246), 528-533.
Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics, 6*(2), 109-118.
Wetherill, L., Agrawal, A., Kapoor, M., Bertelsen, S., Bierut, L. J., Brooks, A., . . . Koller, D. L. (2015). Association of substance dependence phenotypes in the COGA sample. *Addiction biology, 20*(3), 617-627.
Wilens, T. E., Biederman, J., Bredin, E., Hahesy, A. L., Abrantes, A., Neft, D., . . . Spencer, T. J. (2002). A family study of the high-risk children of opioid-and alcohol-dependent parents. *The American journal on addictions, 11*(1), 41-51.
Williams, M. A., Carson, R., Passmore, P., Silvestri, G., & Craig, D. (2011). Introduction to genetic epidemiology. *Optometry-Journal of the American Optometric Association, 82*(2), 83-91.
Wise, R., & Bozarth, M. (1984). Brain mechanisms of drug reward and euphoria. *Psychiatric medicine, 3*(4), 445-460.
Wray, N. R. (2007). Genes, Environment and Psychopathology: Understanding the Causes of Psychiatric and Substance Use Disorders Kenneth S. Kendler and Carol A. Prescott (2006). New York: The Guildford Press, 412 pp, US $45.00, ISBN-10: 1-59385-316-5. *Twin Research and Human Genetics, 10*(01), 231-233.
Wu, L.-T., & Blazer, D. G. (2014). Substance use disorders and psychiatric comorbidity in mid and later life: a review. *International journal of epidemiology, 43*(2), 304-317.
Xie, P., Kranzler, H. R., Krauthammer, M., Cosgrove, K. P., Oslin, D., Anton, R. F., . . . Zhao, H. (2011). Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biological psychiatry, 70*(6), 528-536.
Xie, X., Xu, L., Liu, H., Chen, W., Zhuang, D., Zhang, J., . . . Zhou, W. (2013). Positive association between− 1021TT genotype of dopamine beta hydroxylase gene and progressive behavior of injection heroin users. *Neuroscience letters, 541*, 258-262.

Curriculum Vitae

Shao-Cheng Wang, M.D., Ph.D.
Email: Shaowang@jhsph.edu
Phone: 443-956-8459

SUMMARY

- M.D. in National Yang-Ming University, Ph.D. in Johns Hopkins Bloomberg School of Public Health
- Visiting Fellow at the Intramural Research Program (IRP), National Institute of Drug Abuse (NIDA), National Institutes of Health
- Recipient of Taiwanese government scholarship for pursuit of PhD in U.S.A. (Government funding, full tuition support for 3 years)
- Capability for research in genetic epidemiology with peer-reviewed poster and paper by preparing and analyzing files with PLINK, R, STATA, haploview, computer cluster
- Trained in psychopathology, psychopharmacology, psychotherapy, substance abuse clinical treatment and methadone maintenance treatment at Taoyuan Psychiatry Center, Taiwan
- Board certification: Medicine in Taiwan, Psychiatric specialist doctor in Taiwan

EDUCATION

- **Johns Hopkins Bloomberg school of Public Health:** Baltimore, Maryland

  Doctor of Philosophy in Genetic Epidemiology, Department of Mental Health; 2011 to 2015

  Dissertation topic: Genetic Influences on Injection Drug Use

  Advisor: Brion Maher, Ph.D.

- **National Yang-Ming University**: Taipei, Taiwan

  Doctor of Medicine; 1999-2006

RESEARCH INTERESTS

Genetic Epidemiology, Epigenetics, Big Data Analysis, Substance Use Disorders

POSTGRADUATE TRAINING

2006-2007 Post Graduate Year Training Program, Department of Internal Medicine, National Taiwan University Hospital

2007-2009 Psychiatry Residency, Taoyuan Psychiatric Center, Department of Health, Executive Yuan

2009-2011 Chief Resident and Fellowship, Taoyuan Psychiatric Center, Department of Health, Executive Yuan

2012 Visiting Fellow at the Intramural Research Program (IRP), National Institute of Drug Abuse (NIDA), National Institutes of Health

RESEARCH EXPERIENCE

- **National Yang-Ming University and Taipei Veterans General Hospital Psychiatry Department: Taipei, Taiwan**

  Summer research program, topic: "The relationship between emotional difficulty and post operation treatment in epilepsy patients". Supervisor: Dr. Ton-Ping Su, Chief, Department of Psychiatry, 2003

- **The National Institute on Drug Abuse (NIDA), National Institutes of Health: Bethesda, MD**

  Summer research visitor in Intramural Research Program (IRP) August 2010

- **Johns Hopkins Bloomberg school of Public Health, Department of Mental Health: Baltimore, MD**

  Research assistant, Data Manager of the ALIVE study, 2014 present

TEACHING EXPERIENCE

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Teaching assistant: "Introduction to Behavioral and Psychiatric Genetics" (330.612), 2014

BOARD CERTIFICATION

2006 Ministry of Health, Taiwan - Medical Doctor

2012 Ministry of Health, Taiwan - Psychiatric Specialist

AWARD

2008 *Taiwanese Government Scholarship* by the Taiwanese Ministry of Education (full tuition support for 3 years)

2012-2014 Annual travel fund by Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

2015 Student Conference Fund Award of Johns Hopkins Bloomberg School of Public Health

PUBLICATIONS

Shao-Cheng Wang, Brion Maher, The Risk Difference in Borderline Personality Symptoms Between With and Without Alcohol Use Disorder in the NESARC Wave 2 (2004-2005), Poster in The 167th Annual Meeting of American Psychiatric Association, New York, NY, May 3rd-7th, 2014

Shao-Cheng Wang, Ramin Mojtabai, Deborah Hasin, The Prevalence Difference Of Bipolar I Disorder Criteria B Symptoms Between With Alcohol Use Disorder And Without Alcohol Use Disorder, Poster in The 166th Annual Meeting of American Psychiatric Association, San Francisco, CA, May 18th-22nd, 2013

Shao-Cheng Wang, "Review of Genomics and Psychiatry Medicine", Psychiatry Newsletter P.16-P.18, Vol 29, No.6, June 2010

Shao-Cheng Wang, WC Huang, Case Report：From Emergency Management of A Case of Wernicke-Korsakoff Syndrome, and Review of Related Neurobiological Mechanism literature, Poster in The 49th Congress of Taiwanese Society of Psychiatry, Taipei, Taiwan, November 11th, 2008

PRESENTATIONS

Shao-Cheng Wang, Brion Maher, Genome-Wide Association Study Implicates Protein Kinase N2 as a Risk Factor in Persistent Drug Abuse in Intravenous Drug Users, (in press) oral presentation in The 168th Annual Meeting of American Psychiatric Association, Toronto, Canada, May 16th-20th, 2015

Shao-Cheng Wang, Brion Maher, A Longitudinal Analysis: The Odds Ratio of Manic Episode Symptoms between Individuals with Alcohol Use Disorders and Without Alcohol Use Disorders, oral presentation in The 167th Annual Meeting of American Psychiatric Association, New York, NY, May 3rd-7th, 2014

Shao-Cheng Wang, KH Wu, CM Huang, The depressive symptoms of a heroin addiction patient, from health, financial and social support difficulty. The seasonal seminar in Taoyuan Psychiatric Center, Department of Health, Executive Yuan, Taoyuan, Taiwan, October 10th, 2010

Shao-Cheng Wang, EL Wu, SF Deng, M Lin, LF Guo, LH Liu. A difficult patient, old with repeated admissions. The seasonal seminar in Taoyuan Psychiatric Center, Department of Health, Executive Yuan, Taoyuan, Taiwan, April 10th, 2009

Shao-Cheng Wang, HC Chang, L Yun, LH Chia. Mandatory community treatment: the experience of two home care cases visiting. The seasonal seminar in Taoyuan Psychiatric Center, Department of Health, Executive Yuan, Taoyuan, Taiwan, February 13rd, 2009

INVITED SPEAKER

Shao-Cheng Wang, Psychiatry Department of Taichung Veterans General Hospital, topic "Basic concept of evaluation and treatment of psychotic and mood symptoms of psychiatry medicine", Taichung, Taiwan, May 5th, 2010

EXTRACURRICULAR ACTIVITIES

1999-2000	Medical service teams, Volunteer in promoting education of aboriginal elementary high school students in Taitung Changpin and Taoyuan Fusing, Taiwan
1999-2003	Tenor, Chinyun Chorus, performed in the National Concert Hall
2001-2002	Chief, Standing Committee on Professional Exchange in Yang-Ming University
2001	Attendee, annual meeting of the International Federation of Medical Students' Associations in Yugoslavia
2002	Attendee, annual meeting of the International Federation of Medical Students' Associations in Taiwan
2012-2013	Secretary, student assembly of Mental Health department of Johns Hopkins Bloomberg school of Public Health